D442–D446 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki094

DED: Database of Evolutionary Distances

Vamsi Veeramachaneni and Wojciech Makałowski*

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University,
University Park, PA 16802, USA

Received August 15, 2004; Revised and Accepted October 12, 2004

ABSTRACT

A large database of homologous sequence alignments with good estimates of evolutionary distances can be a valuable resource for molecular evolutionary studies and phylogenetic research in particular. We recently created a database containing 159 921 transcripts from human, mouse, rat, zebrafish and fugu species. Approximately 16 000 homology groups were identified with the help of Ensembl homology evidence. At the macro-level, the database allows us to answer queries of the form:

(1) What is the average *k*-distance between 5′ untranslated regions of human and mouse?
(2) List the 10 groups with the highest K_a/K_s ratio between mouse and rat.
(3) List all identical proteins between human and rat.

Researchers interested in specific proteins can use a simple web interface to retrieve the homology groups of interest, examine all pairwise distances between members of the group and study the conservation of exon–intron gene structures using a graphical interface. The database is available at http://warta.bio.psu.edu/DED/.

INTRODUCTION

The previous decade in biology witnessed unprecedented accumulation of molecular sequence data. However, as Sydney Brenner remarked 'The great challenge in biological research today is how to turn data into knowledge' (1). Evolution, inspite of being recognized for decades as crucially important for understanding life, was until recently the most speculative area of biology. This situation has been radically changed with the molecular approaches that are now possible, thanks to the availability of large amounts of molecular sequences. However, in order to be useful for evolutionary studies, sequences have to be carefully selected and grouped into homology clusters. This is the most important preparatory step and the most tedious one in any evolutionary analysis. For many analyses, homologous sequences have to be further classified as orthologous (i.e. sequences that shared their last common ancestor during speciation time) or paralogous (i.e. sequences that were created by ancestral gene duplication). This distinction is especially important for molecular phylogeny as it is necessary to work with orthologous genes to infer species phylogeny based on gene phylogeny. Interestingly, despite vast amount of sequence data from different organisms, there have been surprisingly few large scale gene comparison studies between different species or groups of organisms (2–6). Information on expected evolutionary distances or protein/gene identity between different organisms (e.g. human and zebrafish) or taxonomy groups (e.g. mammals and reptiles) is difficult to obtain. To fill this gap, we have created the Database of Evolutionary Distances (DED) which contains sequence information from several vertebrate species clustered into homology groups. It also includes multiple sequence alignments for both protein and nucleotide sequences along with the phylogenetic trees and graphical representation of sequence relationships within a homology group. Large number of links to external databases makes further data exploration 'as easy as a click of a mouse'.

Our DED should be useful for gene function assignment, molecular phylogenetic studies, search for lateral gene transfer, reconstruction of identification of biochemical pathways in poorly characterized organisms and sequence evolution patterns. Simple, yet powerful, web interfaces provide a convenient way to access the data. The results are displayed in easy-to-understand tabulated and/or graphical forms.

SEQUENCE DATA

The basic objects stored in our database are genes and their associated transcripts. For each gene, we maintain all its known transcript variants and for each transcript we store its sequence, coding region annotation and exon–intron structure. Currently, our database is based on Ensembl release 20 (7) of human, mouse, rat, zebrafish and fugu data (see Table 1). A total of 159 921 vertebrate transcripts stored in the database

*To whom correspondence should be addressed at Department of Biology, Pennsylvania State University, 514 Mueller Lab, University Park, PA 16802, USA. Tel: +1 814 865 5025; Fax: +1 814 865 9366; Email: wojtek@psu.edu

Table 1. Number of genes and transcripts stored in the DED (August 2004)

Species	Number of genes	Number of transcripts
Human	21 787	29 802
Mouse	25 307	32 281
Rat	22 159	28 545
Fugu	35 180	38 510
Zebrafish	22 409	30 783
Total	126 842	159 921

Table 2. Number of external links present in the DED

Database	Human	Mouse	RAT	Fugu	Zebrafish	Total
GKB	526	0	0	0	0	526
ZFIN_ID	0	0	0	0	1397	1397
PDB	1174	351	228	2033	0	3786
Sanger_Hver1_3_1	4 976	0	0	0	0	4976
UMCU_Hsapiens_19Kv1	12 311	0	0	0	0	12 311
RefSeq	5255	1219	1592	6197	29	14 292
HUGO	11 075	0	0	3510	0	14 585
MIM	8738	205	148	5615	0	14 706
MarkerSymbol	0	17 177	0	0	0	17 177
SPTREMBL	5500	2174	789	6758	2123	17 344
GO	13 247	0	0	4171	0	17 418
SWISS-PROT	962	211	3250	24 711	5	29 139
LocusLink	15 903	15 443	4252	4833	1079	41 510
Protein_id (at EMBL)	19 414	19 989	5223	7862	3483	55 971
EMBL (nucleotide records)	19 434	20 014	5255	7874	3483	56 060
Ensembl	21 787	25 307	22 159	35 180	22 409	126 842
Total	140 302	102 090	42 896	108 744	34 008	428 040

represent 126 842 unique genes clustered in homology groups (see later).

Based on the information retrieved from Ensembl, the gene and transcript objects in our database were cross-referenced with objects in external databases such as RefSeq, Pfam, GO, etc. As expected, the human genes and transcripts have the most external links associated with them (140 302), while those of zebrafish have the least (34 008). Surprisingly, rat records have relatively few external links (42 896) possibly reflecting the transient status of the rat genome annotation. Obviously, Ensembl is the most frequently linked external database, followed by EMBL database, and LocusLink (for details see Table 2).

HOMOLOGY GROUPS

Single linkage clustering was used to create homology groups from pairwise homology information obtained through Ensmart (8). Overall, 16 127 groups are formed from 150 158 pairwise homology relations. Although not all species are present in each group, there are 8402 groups that contain transcripts from all five species. There are several one-to-many homology relationships annotated in Ensembl. In such cases, our use of single-linkage clustering results in homology groups that contain multiple genes from the same species. Figure 1 shows the distribution of group sizes. For each homology group, CLUSTAL W (9) is used to compute two multiple sequence alignments—one from the mRNA sequences and one from the amino acid sequences.

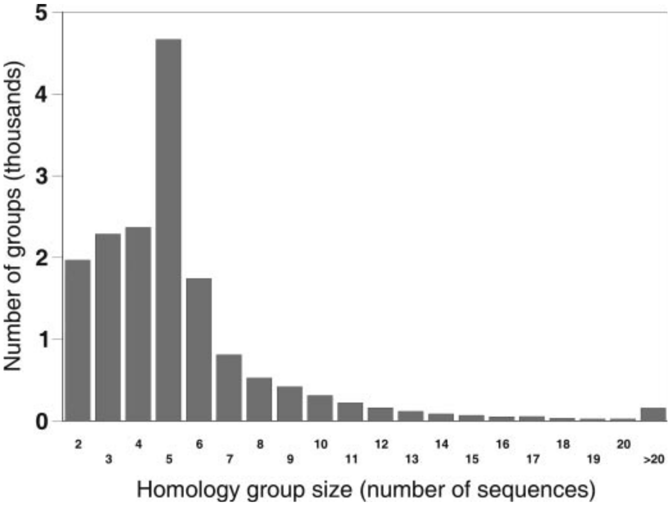


Figure 1. The distribution of group sizes.

The multiple sequence alignments are then stored in a compressed format within the Mysql database. Compression is achieved by noting that a gapped sequence that belongs to an alignment can be obtained from the ungapped transcript (or protein) sequence already stored in the database if one knows the location of the gaps. Instead of storing a whole alignment, we store only information about location and length of gaps in the alignment. This procedure results in a 100-fold reduction of the required storage.

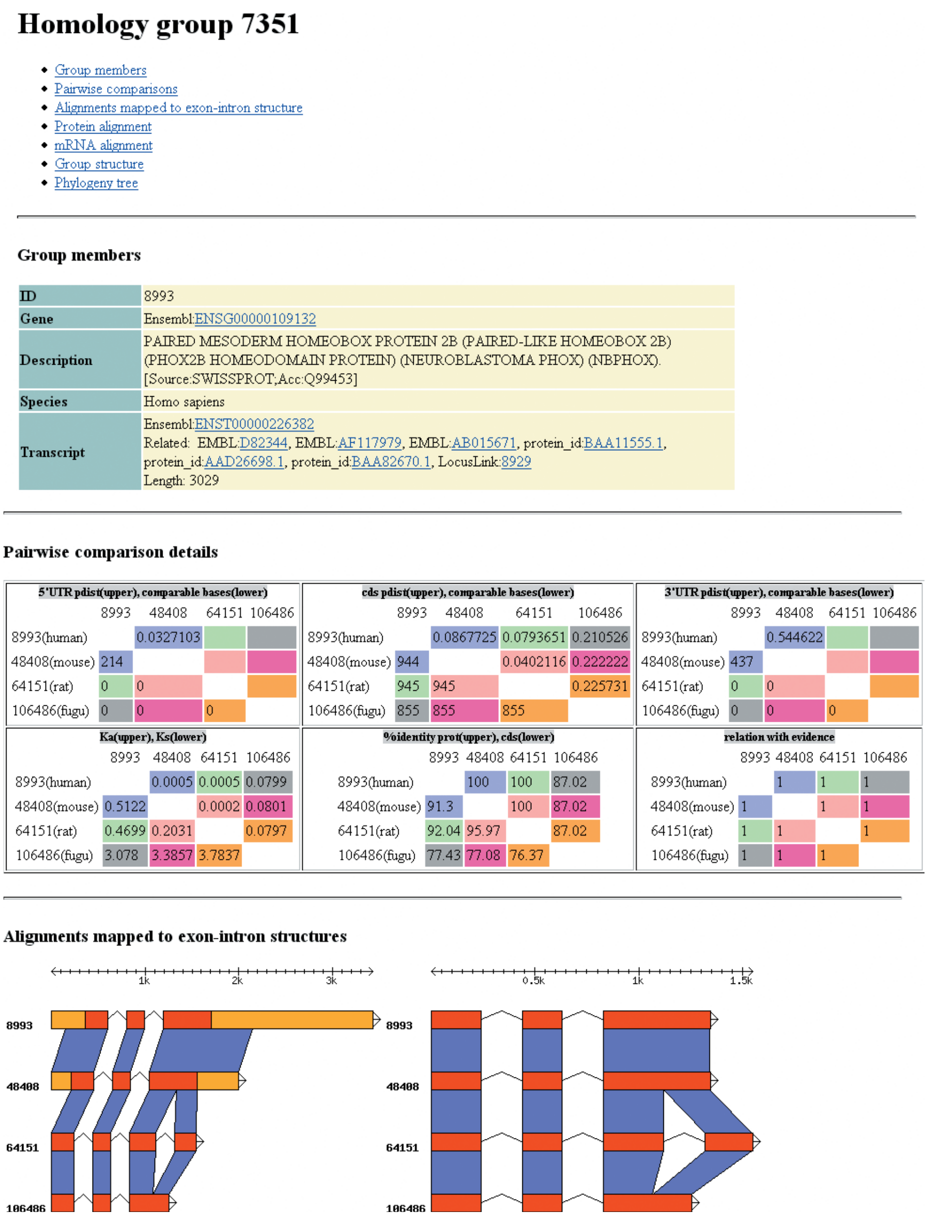
DISTANCE COMPUTATION

In calculating distances, only the transcript with the longest coding region is taken into consideration. mRNA alignments are used for calculation of *p* and *k* distances of coding sequences and untranslated regions. We use Kimura's two-parameters model to compute *k* distances. In case the coding regions do not align perfectly with each other, only the common part of each distinct mRNA region is considered for calculation.

Protein sequence alignments are used for protein identity calculations and serve as a template for the coding sequence alignment that is used in synonymous (K_s) and non-synonymous (K_n) distance calculations. Currently K_s, K_n are obtained using the Nei–Gojobori method (10) as implemented in the PAML package (11). All other pairwise comparison analyses were carried out using Bioperl 1.4 modules (12).

USER INTERFACES

A simple search interface allows users to search the database by keyword or accession number from Ensembl (or other databases linked to Ensembl records such as Swiss-Prot, RefSeq, Gene Ontology, etc.). Genes matching the search criteria and the homology groups that they belong to are displayed. Clicking on the hyperlink for a homology group listed in the search results leads to a page with the full description of the group consisting of seven sections (see Figure 2): (i) description of group members with links to external databases; (ii) pairwise comparison analysis results in a tabular format; (iii) pictorial representation of alignments mapped to



Homology group 7351

- Group members
- Pairwise comparisons
- Alignments mapped to exon-intron structure
- Protein alignment
- mRNA alignment
- Group structure
- Phylogeny tree

Group members

ID	8993
Gene	Ensembl:ENSG00000109132
Description	PAIRED MESODERM HOMEOBOX PROTEIN 2B (PAIRED-LIKE HOMEOBOX 2B) (PHOX2B HOMEODOMAIN PROTEIN) (NEUROBLASTOMA PHOX) (NBPHOX). [Source:SWISSPROT;Acc:Q99453]
Species	Homo sapiens
Transcript	Ensembl:ENST00000226382 Related: EMBL:D82344, EMBL:AF117979, EMBL:AB015671, protein_id:BAA11555.1, protein_id:AAD26698.1, protein_id:BAA82670.1, LocusLink:8929 Length: 3029

Pairwise comparison details

5'UTR pdist(upper), comparable bases(lower)

	8993	48408	64151	106486
8993(human)		0.0327103		
48408(mouse)	214			
64151(rat)	0	0		
106486(fugu)	0	0	0	

cds pdist(upper), comparable bases(lower)

	8993	48408	64151	106486
8993(human)		0.0867725	0.0793651	0.210526
48408(mouse)	944		0.0402116	0.222222
64151(rat)	945	945		0.225731
106486(fugu)	855	855	855	

3'UTR pdist(upper), comparable bases(lower)

	8993	48408	64151	106486
8993(human)		0.544622		
48408(mouse)	437			
64151(rat)	0	0		
106486(fugu)	0	0	0	

Ka(upper), Ks(lower)

	8993	48408	64151	106486
8993(human)		0.0005	0.0005	0.0799
48408(mouse)	0.5122		0.0002	0.0801
64151(rat)	0.4699	0.2031		0.0797
106486(fugu)	3.078	3.3857	3.7837	

%identity prot(upper), cds(lower)

	8993	48408	64151	106486
8993(human)		100	100	87.02
48408(mouse)	91.3		100	87.02
64151(rat)	92.04	95.97		87.02
106486(fugu)	77.43	77.08	76.37	

relation with evidence

	8993	48408	64151	106486
8993(human)		1	1	1
48408(mouse)	1		1	1
64151(rat)	1	1		1
106486(fugu)	1	1	1	

Alignments mapped to exon-intron structures

Figure 2. Sample homology group details. The member section has been truncated. Note that while the proteins are 100% identical, the alignment picture shows that the gene structure is not—there appears to be an intron gain in the rat lineage.

exon–intron structures which help visualize conservation of gene structure; (iv) protein alignment; (v) mRNA alignment; (vi) phylogenetic tree; (vii) group structure picture which shows the pairwise homology relationships that resulted in the construction of the group (Figure 3 shows a case where one possibly false homology relationship resulted in the merging of two distinct homology groups).

By default, only the description of group members and pairwise comparison results are shown. User preferences stored in a cookie are used to determine the set of sections to be shown.

A more elaborate accession search interface can be used for larger scale analyses. It enables calculation of some evolutionary parameters at a global scale (i.e. it summarizes results for a selected group of genes or if there is no limit specified, for all genes present in the database). Extensive filtering options allow a user to restrict analysis to alignments which satisfy certain length and similarity constraints. This helps avoid some statistical biases due to data sampling artefacts or erroneous comparison of paralogous genes. The summary of overall evolutionary statistics, shown in Figure 4, is in agreement with published literature (2,3,13–15).

This interface makes it convenient to verify published results regarding evolutionary rates of groups of proteins. For instance, it was shown in a recent study that sperm-specific proteins evolve at a faster rate than other proteins (16). The paper listed either the RefSeq id or the EMBL accession number for each of the analyzed proteins. By entering the RefSeq ids in one entry box and the EMBL accession numbers in another entry box, one can confirm these results in seconds in the accession search page.

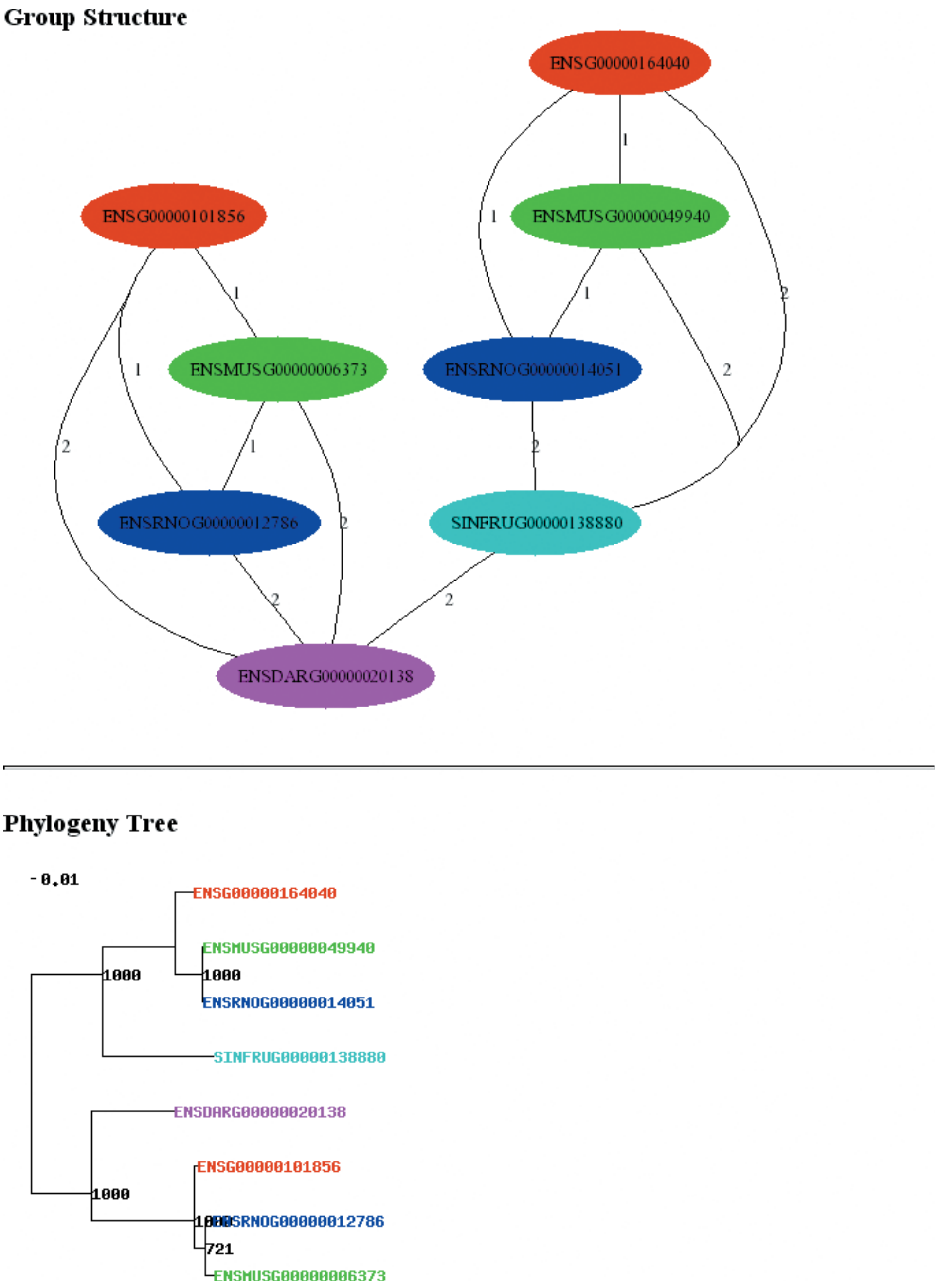
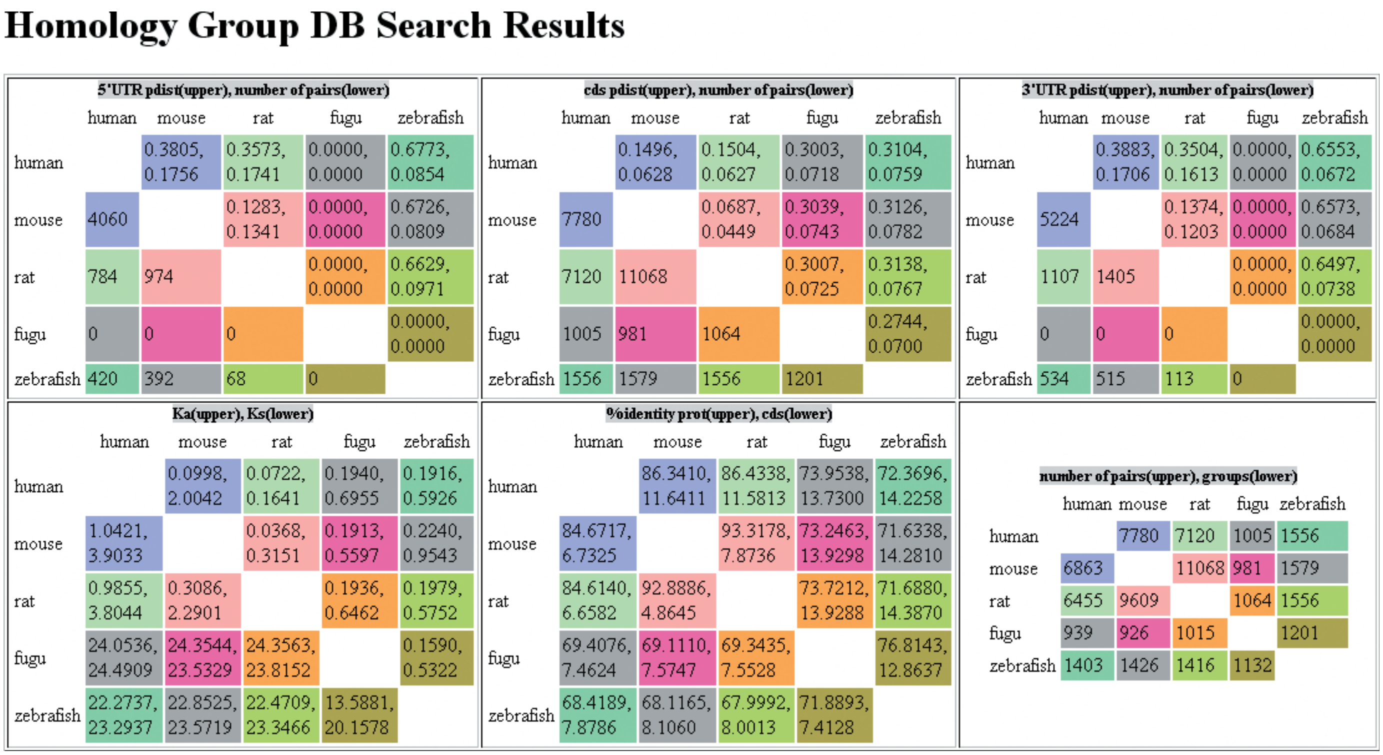


Figure 3. Group structure and phylogenetic tree for a homology group. Pairwise comparison analysis suggests that the homology relationship between fugu and zebrafish genes can be ignored and the group split into two smaller groups.

CONCLUSIONS AND PERSPECTIVES

Evolutionary analysis is a key step in many biological investigations from classical systematics to comparative genomics and bioinformatics. Very often, researchers are interested in knowing how the results of a comparison of a single gene or set of genes fit a 'global picture'. However, such global information is hard to obtain or does not exist. To fill this gap, we have created the DED, which contains sequence information from several vertebrate species clustered into homology groups. This database should be useful in a wide range of biological investigations including gene function assignment, molecular phylogenetic studies and sequence evolution patterns.

Our database depends on other primary databases for sequence, structure and homology information. However, because of the extensive post-processing involved, it is not possible to update our database and keep it synchronized with the primary (source) databases at all times. At present, we plan to update the DED at least twice a year and add new genomes at the time of scheduled updates. In addition, we also plan to add sequence information from organisms whose genomes are not yet fully sequenced.



Homology Group DB Search Results

5'UTR pdist(upper), number of pairs(lower)

	human	mouse	rat	fugu	zebrafish
human		0.3805, 0.1756	0.3573, 0.1741	0.0000, 0.0000	0.6773, 0.0854
mouse	4060		0.1283, 0.1341	0.0000, 0.0000	0.6726, 0.0809
rat	784	974		0.0000, 0.0000	0.6629, 0.0971
fugu	0	0	0		0.0000, 0.0000
zebrafish	420	392	68	0	

cds pdist(upper), number of pairs(lower)

	human	mouse	rat	fugu	zebrafish
human		0.1496, 0.0628	0.1504, 0.0627	0.3003, 0.0718	0.3104, 0.0759
mouse	7780		0.0687, 0.0449	0.3039, 0.0743	0.3126, 0.0782
rat	7120	11068		0.3007, 0.0725	0.3138, 0.0767
fugu	1005	981	1064		0.2744, 0.0700
zebrafish	1556	1579	1556	1201	

3'UTR pdist(upper), number of pairs(lower)

	human	mouse	rat	fugu	zebrafish
human		0.3883, 0.1706	0.3504, 0.1613	0.0000, 0.0000	0.6553, 0.0672
mouse	5224		0.1374, 0.1203	0.0000, 0.0000	0.6573, 0.0684
rat	1107	1405		0.0000, 0.0000	0.6497, 0.0738
fugu	0	0	0		0.0000, 0.0000
zebrafish	534	515	113	0	

Ka(upper), Ks(lower)

	human	mouse	rat	fugu	zebrafish
human		0.0998, 2.0042	0.0722, 0.1641	0.1940, 0.6955	0.1916, 0.5926
mouse	1.0421, 3.9033		0.0368, 0.3151	0.1913, 0.5597	0.2240, 0.9543
rat	0.9855, 3.8044	0.3086, 2.2901		0.1936, 0.6462	0.1979, 0.5752
fugu	24.0536, 24.4909	24.3544, 23.5329	24.3563, 23.8152		0.1590, 0.5322
zebrafish	22.2737, 23.2937	22.8525, 23.5719	22.4709, 23.3466	13.5881, 20.1578	

%identity prot(upper), cds(lower)

	human	mouse	rat	fugu	zebrafish
human		86.3410, 11.6411	86.4338, 11.5813	73.9538, 13.7300	72.3696, 14.2258
mouse	84.6717, 6.7325		93.3178, 7.8736	73.2463, 13.9298	71.6338, 14.2810
rat	84.6140, 6.6582	92.8886, 4.8645		73.7212, 13.9288	71.6880, 14.3870
fugu	69.4076, 7.4624	69.1110, 7.5747	69.3435, 7.5528		76.8143, 12.8637
zebrafish	68.4189, 7.8786	68.1165, 8.1060	67.9992, 8.0013	71.8893, 7.4128	

number of pairs(upper), groups(lower)

	human	mouse	rat	fugu	zebrafish
human		7780	7120	1005	1556
mouse	6863		11068	981	1579
rat	6455	9609		1064	1556
fugu	939	926	1015		1201
zebrafish	1403	1426	1416	1132	

Figure 4. Overall evolutionary statistics with mean and standard deviation shown for all distances. Analysis was restricted to pairs with direct homology evidence. UTR comparisons were made only when UTR size was at least 30 bp. Alignments in which start (or stop) codons were separated by more than 20 columns were ignored.

REFERENCES

1. Brenner,S. (2002) Ontology recapitulates philology. *Scientist*, **16**, 12.
2. Makalowski,W., Zhang,J. and Boguski,M.S. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.*, **6**, 846–857.
3. Makalowski,W. and Boguski,M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.
4. Mushegian,A.R., Garey,J.R., Martin,J. and Liu,L.X. (1998) Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.*, **8**, 590–598.
5. Wheelan,S.J., Boguski,M.S., Duret,L. and Makalowski,W. (1999) Human and nematode orthologs—lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*. *Gene*, **238**, 163–170.
6. Glazko,G.V. and Nei,M. (2003) Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.*, **20**, 424–434.
7. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
8. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
9. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
10. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
11. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
12. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
13. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genomeI. *Nature*, **420**, 520–562.
14. Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.M., Dehal,P., Christoffels,A., Rash,S., Hoon,S., Smit,A. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.
15. Gibbs,R.A., Weinstock,G.M., Metzker,M.L., Muzny,D.M., Sodergren,E.J., Scherer,S., Scott,G., Steffen,D., Worley,K.C., Burch,P.E. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
16. Torgerson,D.G., Kulathinal,R.J. and Singh,R.S. (2002) Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol. Evol.*, **19**, 1973–1980.