Published online September 24, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 17 ***5029–5035***
doi:10.1093/nar/gkh833

Comparative analysis of complete genomes reveals gene loss, acquisition and acceleration of evolutionary rates in Metazoa, suggests a prevalence of evolution via gene acquisition and indicates that the evolutionary rates in animals tend to be conserved

Vladimir N. Babenko and Dmitri M. Krylov*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received July 2, 2004; Revised August 4, 2004; Accepted August 30, 2004

ABSTRACT

In this study we systematically examined the differences between the proteomes of Metazoa and other eukaryotes. Metazoans (*Homo sapiens*, *Ceanorhabditis elegans* and *Drosophila melanogaster*) were compared with a plant (*Arabidopsis thaliana*), fungi (*Saccharomyces cerevisiae* and *Schizosaccaromyces pombe*) and *Encephalitozoan cuniculi*. We identified 159 gene families that were probably lost in the Metazoan branch and 1263 orthologous families that were specific to Metazoa and were likely to have originated in their last common ancestor (LCA). We analyzed the evolutionary rates of pan-eukaryotic protein families and identified those with higher rates in animals. The acceleration was shown to occur in: (i) the LCA of Metazoa or (ii) independently in the Metazoan phyla. A high proportion of the accelerated Metazoan protein families was found to participate in translation and ribosome biogenesis, particularly mitochondrial. By functional analysis we show that no metabolic pathway in animals evolved faster than in other organisms. We conclude that evolution in the LCA of Metazoa was extensive and proceeded largely by gene duplication and/or invention rather than by modification of extant proteins. Finally, we show that the rate of evolution of a gene family in animals has a clear, but not absolute, tendency to be conserved.

INTRODUCTION

Paleontology data suggest that Metazoa radiated from the rest of eukaryotes ~600 million years ago; they are also known to have undergone an explosive diversification in the Cambrian era (1,2). Molecular phylogeny data provide evidence for an earlier emergence of Metazoa, up to 1400 million years ago (3–5), with the apparent discrepancy originating from the incompleteness of paleontology records or, alternatively, from the inapplicability of the molecular clock hypothesis to animal evolution (6–9). The distinguishing traits of the Metazoa kingdom, at least at a present evolutionary stage, include heterotrophy, multicellularity, presence of different tissues, a nervous system and a locomotion apparatus. All of these characteristic traits are shared by the vast majority of Metazoa including the phyla of Chordata, Nematoda and Arthropoda. Biochemically, Metazoa are typified by a vast array of molecules involved in reception and signal transduction; many of these molecules play a role in the communication between different cell types. *Hox* genes are believed to be responsible for the body plans that are characteristic of animals.

While the phenotypical differences between eukaryotes have been well studied, historically, and form the foundation of biological classification (10), genomic differences have, to date, been characterized in a less systematic manner, although considerable advances have been made for Metazoa (11–14). The limited scope of these previous genomic studies stems from the fact that complete genome sequences were not available. With the completion of the sequencing and annotation of a number of genomes, such comparisons have become feasible. Furthermore, comparisons of all protein sequences from one genome to another is technically possible within comprehensive databases such as KOG (eukaryotic orthologous groups) (15,16). The KOG database organizes proteins into families based on the principle of orthology (17–21). In short, orthologs are genes which are inherited from the last common ancestor (LCA) and which retain the original function. This underlying principle of orthology makes the KOG database a tool of choice for a direct comparison of proteins of the same family across different species.

Before the relationship between the genotype and the phenotype can be understood in precise terms, a simple enumeration of the differences between genotypes seems to be necessary. We divided all such differences pertaining to Metazoa into three major categories: (i) genes that had been present before the origin of Metazoa, but were lost in

*To whom correspondence should be addressed. Tel: +1 301 594 6993; Fax: +1 301 435 7794; Email: krylov@ncbi.nlm.nih.gov

them, (ii) genes that are specific to Metazoa and thus appear to have been invented in them and (iii) genes that have an accelerated rate of evolution in Metazoa.

MATERIALS AND METHODS

Eukaryotic genomes

All analysis was performed on seven completely sequenced genomes: *Arabidopsis thaliana*, *Ceanorhabditis elegans*, *Drosophila melanogaster*, *Encephalitozoan cuniculi*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Schizosaccaromyces pombe*. The division of proteins into families according to the principle of expanded orthology was adopted from the KOG database (16). Out of 4852 KOG families, 1817 were selected for analysis based on the following criteria: (i) presence of one or more proteins from each of *C.elegans*, *D.melanogaster*, *H.sapiens*, *S.cerevisiae* and *S.pombe* and (ii) under 30 proteins total per KOG family.

Functional category analysis

The functional categories were adopted from the KOG database (16). They are: Nucleotide transport and metabolism, Signal transduction mechanisms, Cell motility, Transcription, Nuclear structure, Amino acid transport and metabolism, Defense mechanisms, Cytoskeleton, Secondary metabolites biosynthesis, transport and catabolism, Cell wall/membrane/envelope biogenesis, Energy production and conversion, Replication, recombination and repair, RNA processing and modification, Posttranslational modification, protein turnover, chaperones, Translation, ribosomal structure and biogenesis, Extracellular structures, Inorganic ion transport and metabolism, Chromatin structure and dynamics, Coenzyme transport and metabolism, Cell cycle control, Cell division, Chromosome partitioning, Lipid transport and metabolism, Carbohydrate transport and metabolism, Intracellular trafficking, secretion and vesicular transport, Function unknown and General function prediction. The fraction of gene families belonging to each category was calculated as a ratio between those in the functional category to all gene families in the data pool. The statistical significance of differences between gene pools (e.g. 'slow' versus 'rapid') was established by Fisher's Exact test.

Identifying homologues of Metazoa-specific proteins

The longest protein sequence out of each of the 1147 Metazoa-specific KOGs was selected for BLAST searches against the non-redundant protein database (NR BLAST) at NCBI (22). Two hundred and fifty best hits with an *E*-value of 0.001 or lower were selected from each search, and their phyletic pattern was analyzed. Furthermore, the Metazoan KOGs were searched using RPS-BLAST in the Pfam database (23) and the phyletic pattern of hits was analyzed. The NR BLAST and Pfam RPS-BLAST results were combined together in a union, i.e. a protein was considered as having homologues if at least one of the BLAST searches produced hits.

Alignments

The 1817 KOGs were aligned using MAP (24). All columns with insertions or deletions were excised by an in-house script. This resulted in 821 KOGs with <20 aligned positions after excision of indels. These poor alignments typically resulted from a complex domain structure or incomplete protein sequences in the genomes. For these poorly aligned KOGs, the most poorly aligned sequence was removed and indels were re-excised. As a result, another 290 aligned KOGs were obtained. All together, 1308 good alignments were obtained.

Phylogenetic tree construction

Neighbor-joining (NJ) trees were constructed automatically using a pipeline based on PHYLIP (25). It should be noted that NJ in PHYLIP does not restrict for negative branch length, and trees with zero or negative lengths for the branches are routinely produced in mass scale topology reconstruction. They were discarded and the 1308 trees with only positive branch lengths were selected for analysis.

Tree analysis

The trees were analyzed using an in-house program based on the Tree Bioperl Module (26). It utilizes the standard NJ tree reconstruction data to calculate the evolutionary rates with respect to the topology of the tree and it is able to process large volumes of data. For each KOG, the program calculated: (i) the length of the internal branches leading to the monophyletic Metazoa, fungal and *A.thaliana* clades, to the exclusion of the leaves (i.e. outside branches) and (ii) the lengths of the leaf branches for Metazoa and fungi to the exclusion of the internal branches. The average value per branch in both cases was calculated and the evolutionary rates were compared with Metazoa and fungi based on this calculated branch length average. The evolutionary distances were normalized for the time of divergence: Metazoa were assumed to have diverged from the rest of the eukaryotic taxa 1415 million years ago, while fungi were assumed to have diverged 837 million years ago (4,27) (D. M. Krylov, unpublished data). The evolutionary rate was assumed to have been greater in Metazoa if the corrected branch length in Metazoa was over 25% greater in Metazoa than in fungi over the average corrected rate for this dataset. Genes with a rate above the threshold were designated 'rapid' and the rest were designated 'slow'. Thresholds of 10, 15 and 25% were also sampled. They produce the same qualitative picture for functional group distribution: the same functional groups stand out. Larger thresholds produced smaller sample size for individual functional groups. The 25% threshold was chosen as the maximal threshold before the sample size for each functional group became too small for statistical analysis. The rates in 'slow' and 'rapid' pools were further compared under the cutoff of 25% and it was found that they differ with probabilities of 8.1×10^{-19} and 1.6×10^{-17} for the leaves and internal branches, respectively.

Measuring K_n and K_s

The following pairs of orthologs were identified: *C.elegans–Caenorhabditis briggsae*, *D.melanogaster–Anopheles gambiae* and *H.sapiens–Mus musculus*. The pairs of proteins were aligned and these protein alignments were used to construct DNA alignments of the corresponding CDS.

The K_n and K_s values were measured using ML method with default parameters in pairwise alignment in PAML (28).

RESULTS

Genes lost in Metazoa

Out of the 5387 total protein families in KOG, there are 159 with no orthologous representatives in Metazoa. They constitute 2.95% of all eukaryotic gene families in seven fully sequenced genomes. We examined the distribution of these genes over different functional categories (16). In short, these functional categories attribute proteins to one of the major cellular metabolic processes and structural roles (see Materials and Methods). We found that the largest gene loss occurred in the 'Coenzyme transport and metabolism', 'Amino acid transport and metabolism' and 'Cytoskeleton' categories (Figure 1). We also found that signal transduction genes have been lost more seldom. We measured the probability of this pattern of loss across all functional categories and found that in every case, it was statistically significant ($P < 0.05$) by Fisher's Exact test.

Genes acquired in Metazoa

Within the present set of KOG organisms, there are 1147 gene families that are found in all three Metazoan phyla but have no detectable orthologs in other eukaryotes in KOG. They constitute 21.45% of all eukaryotic families in seven genomes. The most likely evolutionary scenario for them is the acquisition of these genes in the LCA of the three Metazoan phyla. It is noteworthy, however, that these genes may, in some cases, have homologues (but not orthologs) in non-metazoan eukaryotes and other taxa. The largest proportion of gene acquisition took place in the functional categories of 'Signal transduction' (T) and 'Extracellular structures' (W), while a relatively small proportion was acquired in 'Posttranslational modification, protein turnover, chaperones' (O) and in 'Translation, ribosomal structure and biogenesis' (J) (Figure 2A).

Figure 2B provides an insight into the phylogenetic origin of these Metazoa-specific orthologous sets of genes. Strikingly, 50% of them have no detectable homologues outside Metazoa. Twenty-seven percent have detectable homologues in other eukaryotes, but nowhere else; 11% are found in various eukaryotic lineages and in bacteria; 7% in various eukaryotes, bacteria and Archaea; 3% in Metazoa and bacteria; while 1% or fewer are in Metazoa, bacteria and Acrhaea, or in Metazoa and Archaea or in eukaryotes and Archaea.

Accelerated evolutionary rates in Metazoan proteins

A large number of Metazoan proteins have orthologs in fungi and plants. For these proteins, it is possible to make a direct comparison of the rate of evolution in Metazoa and two other kingdoms: fungi and plants.

First, we compared the evolutionary rates in the LCAs of Metazoa and fungi, disregarding the rates of evolution after the split of Metazoa into different phyla (Figure 3A). We found 211 gene families (Figure 3B), whose rate of evolution in Metazoa is higher than in fungi. We further analyzed them for their function and found that such genes are distributed evenly among different functional categories.

Next, we studied the evolutionary rates in individual Metazoan phyla. These rates describe how rapidly proteins have accumulated mutations after Metazoa had split into Chordata, Nematoda and Arthropoda (Figure 4A). For this analysis, we used the 1817 protein families represented in Metazoa and fungi as well as, possibly, in other eukaryotes. Of these 1817 protein families, 1245 produced good phylogenetic trees, while 572 resulted in phylogenetic trees with negative branch lengths. This reflects the limitations of the phylogenetic tree reconstruction method. Out of the 1245 protein families we analyzed, 421 showed an acceleration in Metazoa as compared with fungi. This constitutes 33.81% of all selected families. Figure 4B provides a distribution of the accelerated gene families across different functional categories. It includes only those categories wherein accelerated families were overrepresented or underrepresented in a statistically significant manner. Notably, a considerably high

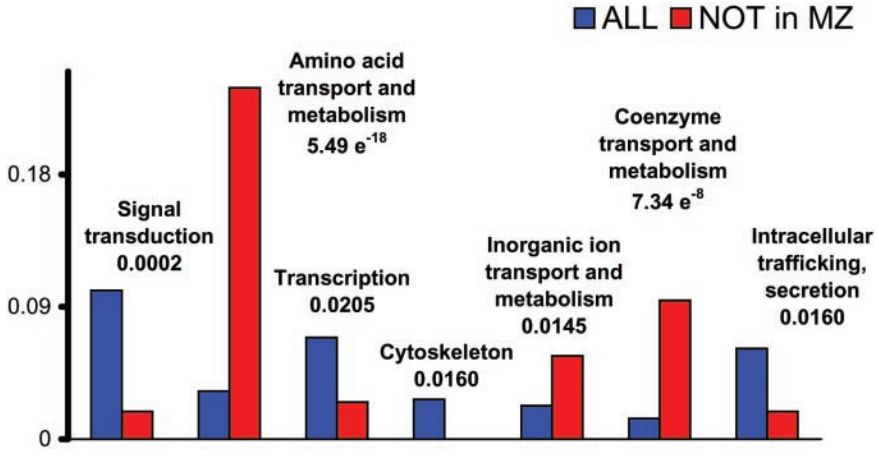

Figure 1. Functional categories of genes lost in Metazoa. Gene loss specific to Metazoa is analyzed for distribution among functional categories. The fraction of lost genes belonging to each category (red bars) is compared with the fraction of genes among all studied gene families belonging to the same functional category (blue bars). Categories for which there is a statistically significant ($P < 0.05$) difference within a functional category are shown. The significance value (P-value) is shown above the bars.

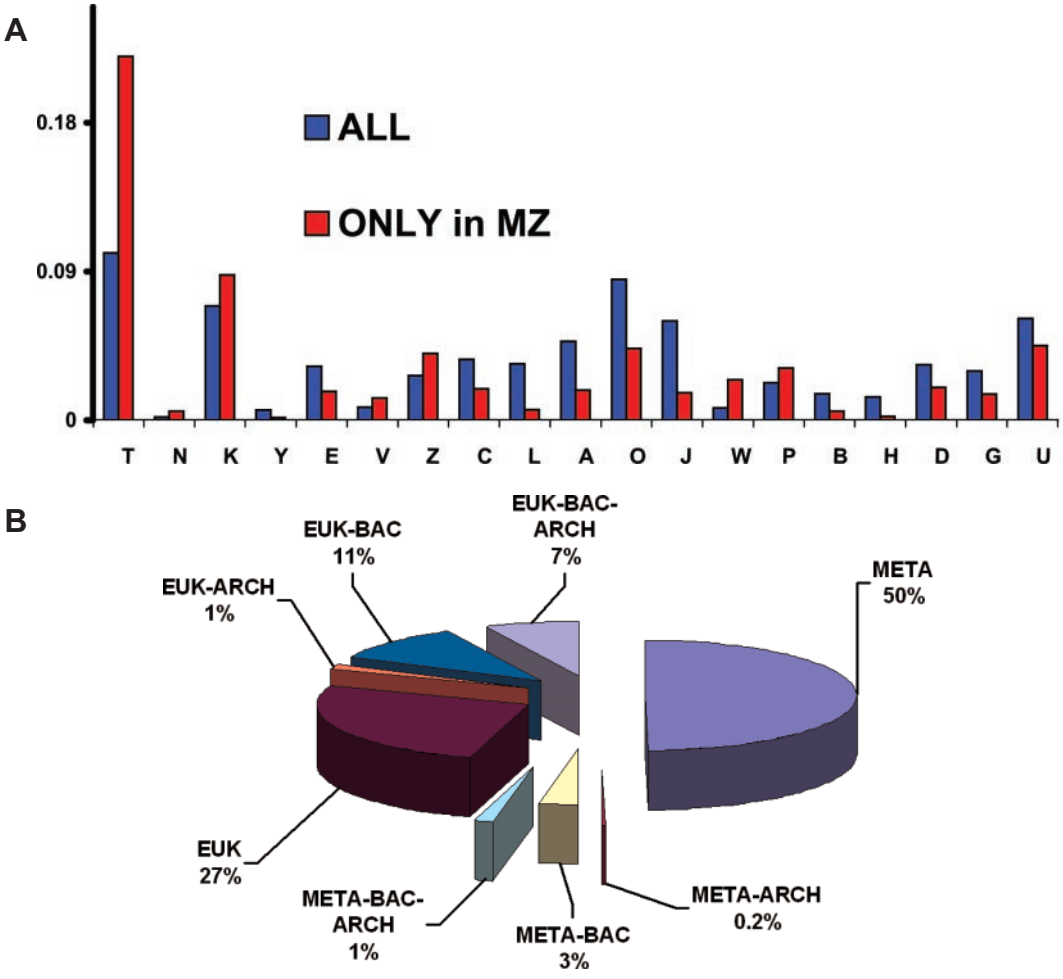

Figure 2. (**A**) Functional categories of Metazoa-specific genes. The fraction of Metazoa-specific genes belonging to each category (red bars) is compared with the fraction of genes among all studied gene families belonging to the same functional category (blue bars). Categories for which there is a statistically significant difference within a functional category are shown. The functional categories are abbreviated as follows, the significance value (*P*-value) is in parenthesis: T, Signal transduction mechanisms (1.64×10^{-21}); N, Cell motility (0.0293); K,Transcription (0.0160); Y, Nuclear structure (0.0266); E, Amino acid transport and metabolism (0.0026); V, Defense mechanisms (0.0454); Z, Cytoskeleton (0.0097); C, Energy production and conversion (0.0009); L, Replication, recombination and repair (2.47×10^{-9}); A, RNA processing and modification (8.27×10^{-7}); O, Posttranslational modification, protein turnover, chaperones (8.02×10^{-7}); J, Translation, ribosomal structure and biogenesis (2.08×10^{-11}); W, Extracellular structures (0.000002); P, Inorganic ion transport and metabolism (0.0420); B, Chromatin structure and dynamics (0.0020); H, Coenzyme transport and metabolism (0.00009); D, Cell cycle control, cell division, chromosome partitioning (0.0075); G, Carbohydrate transport and metabolism (0.0037); U, Intracellular trafficking, secretion and vesicular transport (0.0217). (**B**) The origin of Metazoa-specific proteins. Metazoa-specific proteins without direct orthologs in other taxa are analyzed for detectable homologues in the following groups: META, only in Metazoa; METBAC, in Metazoa and bacteria; META-BAC-ARCH, in Metazoa, bacteria and Archaea; EUK-ARCH, in different eukaryotes and Archaea; EUK, in different eukaryotes only; EUK-BAC, in different eukaryotes and bacteria; EUK-BAC-ARCH, in different eukaryotes, bacteria and Archaea; EUK-ARCH, in different eukaryotes and Archaea.

proportion of such accelerated genes fall into the category of 'Translation, ribosomal structure and biogenesis' (J). These translation machinery families were examined individually and it was found that most of them function in the mitochondrion, while their genes are encoded in the nuclear genome.

Comparing K_n and K_s with the rate of protein evolution

We measured the rate of non-synonymous (K_n) and synonymous (K_s) substitution in the DNA sequences of the genes that have been described in the Materials and Methods section. It is noteworthy that the protein substitution rate was measured over a large evolutionary distance by comparing sequences from three distant Metazoan lineages: *C.elegans*, *D.melanogaster* and *H.sapiens*. The DNA substitution rate, on the contrary, is measured here by comparing sequences of closely related species and therefore reflects the rate of change on a much smaller scale; it is also worth pointing out that the rate is assessed by an independent method. Table 1 shows a significant correlation between the rate of protein substitution and K_n, while a weaker correlation is observed between protein substitution and K_s.

We further compared the K_n values in genes that were designated as 'slow' and 'rapid' based on the rate of amino acid substitution. Those that have a higher substitution rate in Metazoa than in fungi and, in some cases, in plants are designated 'rapid', while those with a comparatively lower substitution rate are designated 'slow'. An average value for K_n was

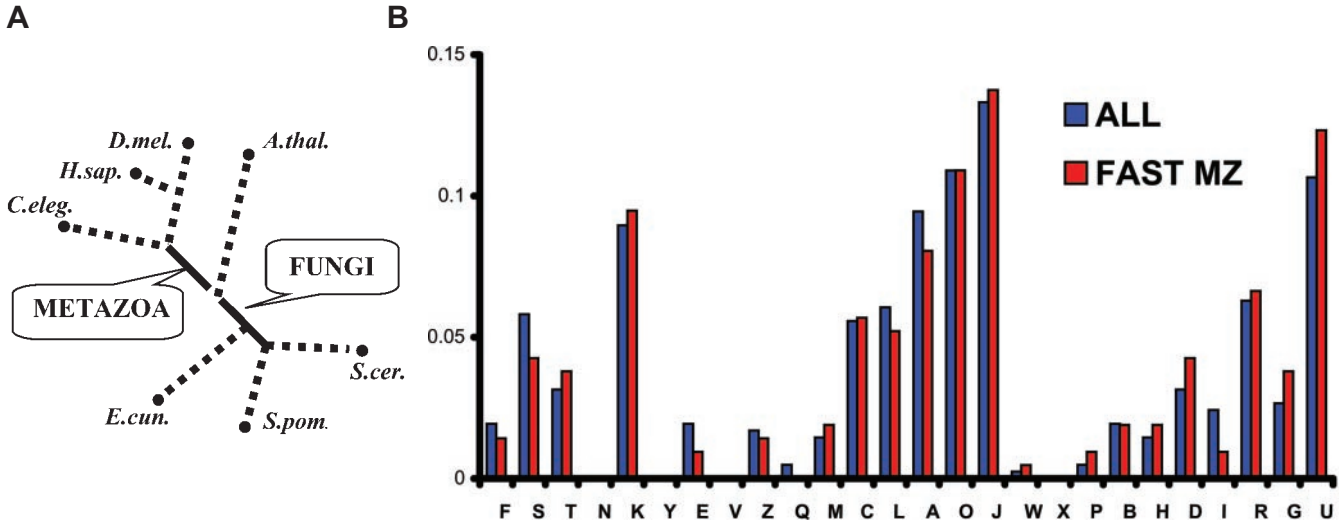

Figure 3. (**A**) Evolutionary rates in the LCAs of Metazoa and fungi. The protein substitution rates in the LCAs are reflected by the length of branches preceding the split into individual clades. (**B**) Function of genes with rapid evolutionary rates in the LCA of Metazoa. The fraction of Metazoa-specific genes belonging to each category (red bars) is compared with the fraction of genes among all studied gene families belonging to the same functional category (blue bars). Functional categories are abbreviated as follows: F, Nucleotide transport and metabolism; S, Function unknown; T, Signal transduction mechanisms; N, Cell motility; K, Transcription; Y, Nuclear structure; E, Amino acid transport and metabolism; V, Defense mechanisms; Z, Cytoskeleton; Q, Secondary metabolites biosynthesis, transport and catabolism; M, Cell wall/membrane/envelope biogenesis; C, Energy production and conversion; L, Replication, recombination and repair; A, RNA processing and modification; O, Posttranslational modification, protein turnover, chaperones; J, Translation, ribosomal structure and biogenesis; W, Extracellular structures; P, Inorganic ion transport and metabolism; B, Chromatin structure and dynamics; H, Coenzyme transport and metabolism; D, Cell cycle control, cell division, chromosome partitioning; I, Lipid transport and metabolism; R, General function prediction only; G, Carbohydrate transport and metabolism; U, Intracellular trafficking, secretion and vesicular transport.

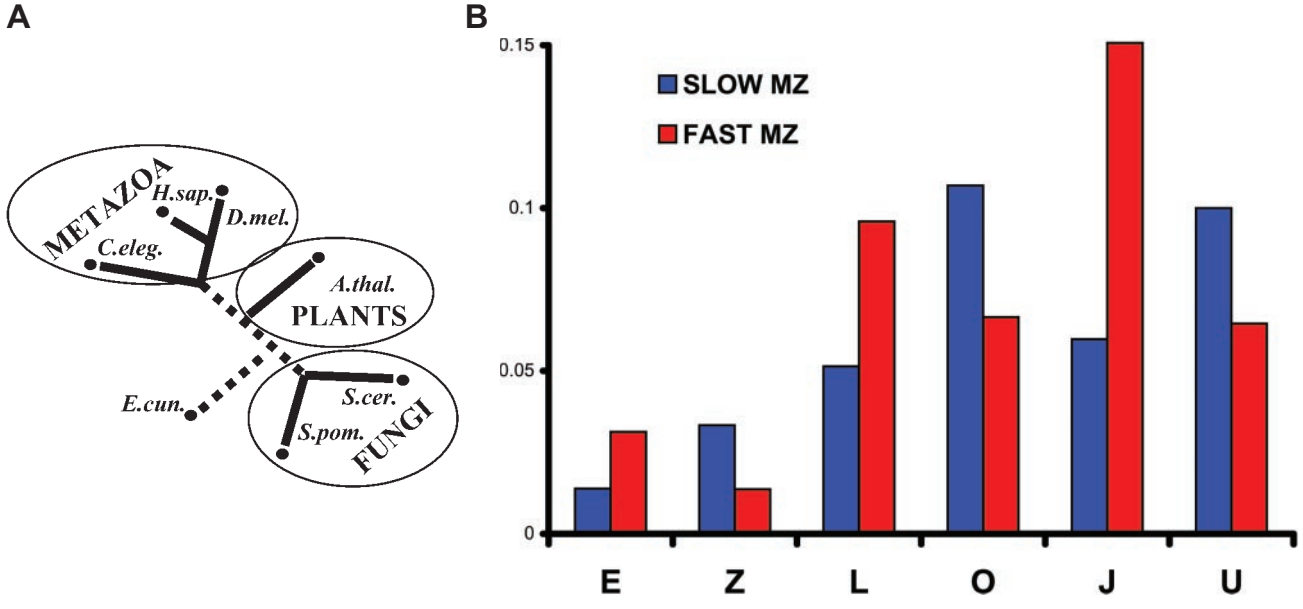

Figure 4. (**A**) Evolutionary rates in the separate lineages of Metazoa, fungi and plants. The protein substitution rates in the lineages are reflected by the length of branches after the split into individual clades. (**B**) Function of genes with rapid evolutionary rates in separate Metazoan phyla. Categories for which there is a statistically significant difference within a functional category are shown. The functional categories are abbreviated as follows, the significance value (*P*-value) is listed here in parenthesis: E, Amino acid transport and metabolism (0.033); Z, Cytoskeleton (0.023); L, Replication, recombination and repair (0.0037); O, Posttranslational modification, protein turnover, chaperones (0.0154); J, Translation, ribosomal structure and biogenesis (0.0000015); U, Intracellular trafficking, secretion and vesicular transport (0.0266).

calculated separately for the 'rapid' and the 'slow' pools and these values were assessed statistically for the probability of belonging to the same general pool (Table 2). As a result, we found that in each separate case, average K_n values for 'rapid' genes were higher than average K_n values for 'slow' genes. This difference is statistically significant (column 4 in Table 2), i.e. it is rather unlikely that 'rapid' and 'slow' genes belong to the same general pool in terms of K_n.

Table 1. Correlation coefficients between protein and DNA substitutions

Protein–DNA	r with K_n	r with K_s	d.f.
Homo sapiens/ *Mus musculus*	0.284 ($<10^{-4}$)	0.068 (0.0859)	637
Homo sapiens/ *Rattus norvegicus.*	0.261 ($<10^{-4}$)	0.095 (0.0202)	597
Drosophila melanogaster/ *Anopheles gambiae*	0.339 ($<10^{-4}$)	0.094 (0.0134)	688
Ceanorhabditis elegans/ *Caenorhabditis briggsae*	0.285 ($<10^{-4}$)	0.099 (0.0109)	659

The Pearson correlation coefficients (r) between the rate of amino acid substitution averaged over three major Metazoan phyla (protein) and the rate of DNA mutations (DNA), non-synonymous (K_n) and synonymous (K_s), measured in closely related Metazoan lineages. The significance values of each (r) are provided in parentheses, while the degrees of freedom (d.f.) for the t-test are listed in the far right column.

Table 2. K_n in slow and rapidly evolving Metazoan genes

	Slow	Rapid	Probability
Homo sapiens/ *Mus musculus*	0.049	0.063	0.005
Homo sapiens/ *Rattus norvegicus*	0.057	0.068	0.046
Drosophila melanogaster/ *Anopheles gambiae*	1.333	1.381	0.006
Ceanorhabditis elegans/ *Caenorhabditis briggsae*	0.097	0.116	0.026

The average K_n value was calculated for three pairs of species, separately for 'slow' and 'rapid' genes, designated by protein sequence substitution rates. Genes with a high rate of protein substitution had a higher K_n in all three cases. This difference in K_n is shown to be statistically significant by Student's t-test (column 4 'Probability'). The null hypothesis was that both 'slow' and 'rapid' groups have the same mean K_n.

DISCUSSION

The most dramatic loss in Metazoa occurred in gene families that are involved in: (i) Coenzyme transport and metabolism, (ii) Amino acid transport and metabolism and (iii) Cytoskelton (Figure 1). While the LCA of crown group eukaryotes is likely to have sustained gene loss (16), the loss reported in this study is assumed to have occurred later on in evolution, in the LCA of Metazoa based on the most likely evolutionary scenarios. It is noteworthy that many amino acid and coenzyme metabolism pathways are dispensable, at least in modern day Metazoa, due to heterotrophy, and, as shown previously, gene loss is dependant on dispensability (27,29). The complete picture of Metazoa-specific loss in the three Metazoan phyla presented here reiterates the notion that dispensable pathways are subject to substantial gene loss (30). The loss in cytoskeleton genes may be explained by new requirements for these genes in multicellular, locomoting Metazoa.

We have identified 1147 gene families that have no bona fide orthologs in other eukaryotes under study (Figure 2A). In contrast, there are only 211 orthologous gene families where Metazoa genes have higher rates of evolution in the LCA (Figure 3A and B). It is also conspicuous that these relatively rapidly evolving Metazoa genes are distributed over functional categories in an unbiased manner (Figure 3B), i.e. the distribution is not different from that in all genes. This latter observation suggests that no considerable acceleration in rates of genes belonging to any particular functional pathway in Metazoa has taken place. Considering the large number of Metazoa-specific gene expansions and inventions (1147) and the relatively few gene families with accelerated evolutionary rate with an unbiased distribution over functional categories, we conclude that evolution in the LCA of Chordata, Nematoda and Arthropoda was *largely extensive*. It proceeded mostly by gene duplication and/or acquisition through different mechanisms rather than acceleration of the evolutionary rates of existing genes.

Half of the Metazoa-specific sets of orthologs are shown to have homologues in other taxa, thus suggesting a pre-Metazoan origin of these genes (Figure 2B). Strikingly, we discovered that 50% of Metazoa-specific genes have no detectable homologues outside of Metazoa (Figure 2B). This can be due to two reasons: (i) the substitution rate in these proteins has been so high that at present no similarity can be detected even by rather sensitive methods, including PSI-BLAST and RPS-BLAST with Pfam profiles and (ii) these are true gene inventions in Metazoa. Undoubtedly, this subject merits a more detailed study.

We identified (Figure 4B) a higher than expected number of mitochondrial genes which rapidly evolved in the three separate phyla of Metazoa after their divergence from the LCA. In conjunction with previous observations (31) on the rapid rates of evolution in the mitochondrion and in translation genes in particular, this suggests that genes encoded in the nuclear genome but functioning in the mitochondrion translation machinery appear to coevolve, at least in terms of evolutionary rates, with their rapidly evolving counterparts encoded in the mitochondrion.

We confirm that our method of measuring protein evolutionary rates is consistent with an independent measurement of K_n in three species representing the three Metazoan phyla (Table 1). This is evident from the considerable correlation of K_n (non-synonymous DNA substitution rate) with the rates we measured and the lack of such correlation between the protein rates and K_s (synonymous DNA substitution rates). Furthermore, we show that the K_n is, on the average, consistently higher among genes that have evolved rapidly on a large evolutionary scale ('rapid') than in those that have been relatively slow (Table 2). Noteworthy, the K_n measurement comes from pairs of closely related species and, therefore, describes the mutational rate on a relatively small scale, while the protein substitution rates are measured over a large evolutionary distance spanning the time since the Vendian divergence of Nematoda, Chordata and Arthropoda. The fact that these measurements correlate suggests that genes which once assumed a rapid evolutionary rate tend to preserve this rapid rate throughout long evolutionary distances. It should be pointed out that this tendency is not absolute, and a rate deviation can occur in a number of protein families in different lineages. The correlation coefficients we calculate can serve as a numerical measure for this tendency.

ACKNOWLEDGEMENTS

The authors wish to express gratitude to Eugene V. Koonin and Igor Rogozin for helpful scientific discussions and to all NCBI staff involved in the construction and maintenance of the COG database.

REFERENCES

1. Knoll,A.H. and Carroll,S.B. (1999) Early animal evolution: emerging views from comparative biology and geology. *Science*, **284**, 2129–2137.
2. Conway Morris,S. (2000) The Cambrian ''explosion'': slow-fuse or megatonnage? *Proc. Natl Acad. Sci. USA*, **97**, 4426–4429.
3. Doolittle,R.F., Cho,G. and Feng,D.F. (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl Acad. Sci. USA*, **94**, 13028–13033.
4. Wang,D.Y., Kumar,S. and Hedges,S.B. (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc R Soc Lond., B, Biol Sci*, **266**, 163–171.
5. Hendy,M.D. and Bromham,L.D. (2000) Can fast early rates reconcile molecular dates with the Cambrian explosion? *Proc. R. Soc. Lond., B, Biol. Sci*., **267**, 1041–1047.
6. Aris-Brosou,S. and Yang,Z. (2003) Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol*., **20**, 1947–1954.
7. Ingram,V.M. (1961) *Nature*, **189**, 704–708.
8. Jukes,T. (1963) *Advan. Biol. Med. Phys*, **9**, 1–41.
9. Wilson,A.C., Carlson,S.S. and White,T.J. (1977) Biochemical evolution. *Annu. Rev. Biochem*., **46**, 573–639.
10. Linnaeus,C. (1736–1740) *Systema Naturae*. Stockholm.
11. Perovic,S., Krasko,A., Prokic,I., Muller,I.M. and Muller,W.E. (1999) Origin of neuronal-like receptors in Metazoa: cloning of a metabotropic glutamate/GABA-like receptor from the marine sponge *Geodia cydonium*. *Cell Tissue Res*., **296**, 395–404.
12. Williams,N.A. and Holland,P.W. (1998) Gene and domain duplication in the chordate Otx gene family: insights from amphioxus Otx. *Mol. Biol. Evol*., **15**, 600–607.
13. Ono,K., Suga,H., Iwabe,N., Kuma,K. and Miyata,T. (1999) Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan-eumetazoan split. *J. Mol. Evol*., **48**, 654–662.
14. Nagy,L. (1998) Changing patterns of gene regulation in the evolution of arthropod morphology. *Am. Zool*., **38**, 818–828.
15. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
16. Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*., **5**, R7.
17. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool*., **19**, 99–113.
18. Fitch,W.M.E. (1970) *Evol. Biol*., **4**, 67–109.
19. Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.
20. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
21. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*., **18**, 619–620.
22. 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.
23. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32** (Database issue), D138–141.
24. Huang,X. (1994) On global sequence alignment. *Comput. Appl. Biosci*., **10**, 227–235.
25. Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. Enzymol*., **266**, 418–427.
26. 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.
27. Krylov,D.M., Wolf,Y.I., Rogozin,I.B. and Koonin,E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*., **13**, 2229–2235.
28. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.
29. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*., **12**, 962–968.
30. Aravind,L., Watanabe,H., Lipman,D.J. and Koonin,E.V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA*, **97**, 11319–11324.
31. Lynch,M. and Jarrell,P.E. (1993) A method for calibrating molecular clocks and its application to animal mitochondrial DNA. *Genetics*, **135**, 1197–1208.