Nucleic Acids Research, 2005, Vol. 33, Database issue ***D495–D497***
doi:10.1093/nar/gki090

The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics

Christian Roth, Matthew J. Betts, Pär Steffansson, Gisle Sælensminde and David A. Liberles*

Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway

Received August 14, 2004; Revised and Accepted October 12, 2004

ABSTRACT

From 138 662 embryophyte (higher plant) and 348 142 chordate genes, 4216 embryophyte and 15 452 chordate gene families were generated. For each of these gene families, multiple sequence alignments, phylogenetic trees, ratios of non-synonymous to synonymous nucleotide substitution rates (K_a/K_s), mappings from gene trees to the NCBI taxonomy and structural links to solved three-dimensional protein structures in the Protein Data Bank (PDB) with Grantham-weighted mutational factors were all calculated. Of the 'gene family trees', 173 embryophyte and 505 chordate branches show $K_a/K_s \gg 1$ and are candidates for functional adaptation. The calculated information is available both as a gene family database and as a phylogenetically indexed resource, called 'The Adaptive Evolution Database' (TAED), available at http://www.bioinfo.no/tools/TAED.

INTRODUCTION

The Adaptive Evolution Database (TAED) was first presented as a collection of branches from chordate and embryophyte gene families with fast evolutionary rates mapped onto the NCBI taxonomy (1,2). The original gene families were from the Master Catalog and are proprietary (3). A new version of TAED is now presented as a taxonomic shell together with a gene family database. In addition to multiple sequence alignments and phylogenetic trees for all families of chordate and embryophyte sequences, the ratio of non-synonymous to synonymous nucleotide substitution rates (K_a/K_s) is provided for each branch of every phylogenetic tree. This ratio, when significantly greater than 1, is an indicator of positive selection and potentially a change of function of the encoded protein. With a gene tree to species tree mapping, the branches significantly greater than 1 are collated together in a phylogenetic context. The framework is expandable to incorporate other genomic-scale information in a phylogenetic context.

Ultimately, the database is designed both to provide high-quality gene families with multiple sequence alignments and phylogenetic trees for chordates and embryophytes, and to enable asking the question, 'What makes each species unique at the molecular genomic level?'.

METHODS

A total of 138 662 embryophyte and 348 142 chordate protein-encoding gene sequences were extracted from GenBank 136 for embryophytes and GenBank 138 for chordates. Pseudogenes and genes with a protein length of less than 10 amino acids were ignored. Independently, all-against-all BLAST searches were calculated for the embryophyte and chordate genes with an *E*-value cutoff of 1.0. For each hit, global PAM distances were calculated using Darwin (4).

Gene families were prepared from this dataset using single linkage clustering of genes annotated as complete, with pairwise PAM distances of 100 PAM units or less and where the length of the shorter fragment divided by the longer fragment was at least 0.9. After formation of these families, partial sequences with matches to a family of not more than 100 PAM units were added back to existing families. Families containing sequences from only one species were not considered further.

For each gene family, multiple sequence alignments were calculated using POA (5) with its default Blosum 80 substitution matrix. For embryophytes, large families with poor quality alignment were subdivided until every sequence in a family aligned with at least 85% of every other sequence in the family. For chordates, such families were refined with complete linkage clustering at PAM $\leqslant 70$. In the future, resulting families where the most ancient node is a gene duplication event will be subdivided until the most ancient node represents a speciation event.

Phylogenetic trees were estimated by Bayesian inference using MrBayes (6). Using the Jones amino acid matrix, 4 chains were calculated for 500 000 generations after a burnin of 250 000 generations. A majority rule consensus tree was calculated from trees sampled every 100 generations after the

*To whom correspondence should be addressed. Tel: +47 55584043; Fax: +47 55584295; Email: liberles@cbu.uib.no initial burnin. In some cases, additional generations were run to reach convergence and trees were only sampled after convergence was reached.

A novel soft parsimony approach (7) was used to simultaneously root trees and map them onto the NCBI taxonomy. Nodes with low posterior probabilities (<0.7) that conflicted with the NCBI taxonomy were corrected according to the NCBI taxonomy. Nodes that remained non-binary were ultimately resolved using UPGMA as a last resort.

For each branch of every phylogenetic tree, K_a/K_s ratios were calculated using a previously published method combining parsimony ancestral sequence reconstruction and a standard K_a/K_s estimation method (8). Branches were considered significant if $K_a \gg K_s$ and at least two non-synonymous substitutions occurred along the branch.

Sequences at nodes immediately preceding branches with $K_a/K_s \gg 1$ were blasted against the Protein Data Bank (PDB) (9), and BLAST hits with E-values < 1 were used to calculate pairwise PAM distances. When a hit within 70 PAM units was found, the structure was linked to the family. For mutations occurring along the high K_a/K_s branch, all transitions along the most parsimonious pathway were multiplied probabilistically with the Grantham matrix (10) score of the physicochemical severity of the transition. The temperature factor column of the PDB structure was annotated with this number for each position, normalized by the highest possible value (215, representing a Trp to Cys substitution with a probability 1).

The database is set up for periodic updating.

RESULTS

A total of 4216 embryophyte and 15 452 chordate gene families were generated and are available online at http://www.bioinfo.no/tools/TAED. Of these, 4211 embryophyte and 14 772 chordate families have been fully processed at the time of submission. These families contain multiple sequence alignments, phylogenetic trees, ratios of non-synonymous to synonymous nucleotide substitution rates (K_a/K_s) calculated for each branch of every phylogenetic tree and, where available from the PDB, structural information including calculated Grantham-weighted mutational factors for each amino acid along branches where $K_a/K_s \gg 1$. Of the missing families, several are large families from the chordate immune system. The embryophyte families vary in size from 2 to 1701 members, while the chordate families vary in size from 2 to 14 001 members (the largest chordate gene family currently in the database has 256 members). Both taxonomic groupings show a power-law distribution of sizes.

Embryophyte gene families had significant K_a/K_s ratios ranging from 0 to 7.79. The most positively selected branches are drought induced aldehyde dehydrogenase in the common ice plant and fructose-1,6-bisphosphatase in the lineage leading to tomato. Chordate gene families had significant K_a/K_s ratios ranging from 0 to 13.18. The most positively selected branch in chordates was neuronal apoptosis inducing protein in rodents. Overall, 173 branches of 116 gene families had $K_a/K_s \gg 1$ in embryophytes and 505 branches of 381 gene families had $K_a/K_s > 1$ in chordates. Of these, 79 embryophyte protein branches and 139 chordate protein branches had solved three-dimensional structures for close homologs.

Intriguingly, the distributions of K_a/K_s ratios across branches were somewhat multiphasic for both embryophytes and chordates and future modeling will determine if this is significant. Also interesting is that the percentage of positively selected branches of gene family trees has gone down as the database has grown [compared with (1)]. Possible explanations include the trivial explanation of missing chordate immune system gene families, but also less random sampling of genes through genomic sequencing, more conservative family building, and better articulation of trees reducing the number of substitutions along any branch. This last point is coupled with the fact that small numbers of mutations can significantly alter protein function (11) and that $K_a/K_s \gg 1$ can be too conservative a test to detect this, especially along short branches (12,13). For this reason, other tests will also be applied in the future.

Among the positively selected genes are some cases that have been previously characterized in the literature. These include chitinases (14) and self-incompatibility proteins (15) from embryophytes and olfactory receptors (16), primate leptin (13), snake venom phospholipase A2 (17) and mammalian defensins (18) from chordates.

DISCUSSION AND FUTURE DIRECTIONS

TAED is an expandable resource for functional genomics and molecular evolution research. At the first level, it provides high-quality gene families with multiple sequence alignments and Bayesian phylogenetic trees annotated with K_a/K_s values and searchable by text. On top of this, a phylogenetic context enables one to examine molecular events occurring along the same lineage of species divergence in evolutionary history. This should enable a fuller understanding of the molecular basis for phenotypic divergence along such lineages.

In the future, the database is expandable to include many additional features. Other types of analysis on the gene family dataset, including maximum-likelihood K_a/K_s estimation (19), various likelihood ratio tests (20) and other methods that examine amino-acid-based properties of evolution (21) can be included.

The sequence data can be overlaid on metabolic and protein interaction datasets [e.g. the KEGG database (22)] to understand how proteins with rapidly evolving sequences relate to each other in a biological context. This can even be extrapolated to the protein structural level, given the mapping onto known three-dimensional structures.

Finally, coding sequence evolution is only one part of the molecular evolution of genomes driving phenotypic divergence. Changes in gene content (23), and phylogenetic reconstructions of changes in gene expression and alternative splicing data (24) can indicate where other significant lineage-specific changes have occurred. Altogether, phylogenetic indexing of genomic data presents a powerful approach to understanding the evolution of function in genomes.

VISUALIZATION OPTIONS AND COMPUTATIONAL REQUIREMENTS

Two options are available for visualization. For the first, a simple web interface is all that is required. The gene families are accessible through several search functions. The gene families with branches significantly greater than one can be obtained through their taxonomic index with a mapped tree in a Java applet. Gene families then display structural, phylogenetic and multiple sequence information.

A second viewer is also available. The tree viewer requires Java 1.3 (or higher) and the protein structural viewer requires Java 1.3 (or higher) and a Java 3D plugin. The download and installation instructions for this viewing option are available on the TAED website.

ACKNOWLEDGEMENTS

We are grateful to Malin Almgren, Alexander Roth, Jesper Lundström, Tim Hughes and Saywan Pasha, who all made contributions to improving TAED since its initial publication based on the Master Catalog. Funding for this effort proceeded initially through the Swedish Foundation for Strategic Research and subsequently through FUGE, the functional genomics platform of the Norwegian research council. C.R. received funding from Schweizerischer Nationalfonds—the Swiss national science foundation.

REFERENCES

1. Liberles,D.A., Schreiber,D.R., Govindarajan,S., Chamberlin,S.G. and Benner,S.A. (2001) The Adaptive Evolution Database (TAED). *Genome Biol.*, **2**, research0028.1–research0028.6.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
3. Benner,S.A., Chamberlin,S.G., Liberles,D.A., Govindarajan,S. and Knecht,L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded approach to functional genomics. *Res. Microbiol.*, **151**, 97–106.
4. Gonnet,G.H., Hallett,M.T., Korostensky,C. and Bernardin,L. (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, **16**, 101–103.
5. Grasso,C. and Lee,C. (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, **20**, 1546–1556.
6. Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
7. Steffansson,P. (2004) Building consensus trees using gene sequences—a phylogenetic approach. MSc Thesis, Stockholm University, Sweden.
8. Liberles,D.A. (2001) Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol. Biol. Evol.*, **18**, 2040–2047.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
11. Golding,G.B. and Dean,A.M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.*, **15**, 355–369.
12. Crandall,K.A., Kelsey,C.R., Imamichi,H., Lane,H.C. and Salzman,N.P. (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.*, **16**, 372–382.
13. Siltberg,J. and Liberles,D.A. (2002) A simple covarion-based approach to analyse nucleotide substitution rates. *J. Evol. Biol.*, **15**, 588–594.
14. Tiffin,P. (2004) Comparative evolutionary histories of chitinase genes in the genus zea and family poaceae. *Genetics*, **167**, 1331–1340.
15. Lu,Y. (2002) Molecular evolution at the self-incompatibility locus of Physalis longifolia (Solanaceae). *J. Mol. Evol.*, **54**, 784–793.
16. Gilad,Y., Segre,D., Skorecki,K., Nachman,M.W., Lancet,D. and Sharon,D. (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nature Genet.*, **26**, 221–224.
17. Valentin,E. and Lambeau,G. (2000) What can venom phospholipases A(2) tell us about the functional diversity of mammalian secreted phospholipases A(2)? *Biochimie*, **82**, 815–831.
18. Maxwell,A.I., Morrison,G.M. and Dorin,J.R. (2003) Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. *Mol. Immunol.*, **40**, 413–421.
19. Yang,Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.*, **15**, 568–573.
20. Gaucher,E.A., Gu,X., Miyamoto,M.M. and Benner,S.A. (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.*, **27**, 315–321.
21. Soyer,O.S., Dimmic,M.W., Neubig,R.R. and Goldstein,R.A. (2003) Dimerization in aminergic G-protein-coupled receptors: application of a hidden-site class model of evolution. *Biochemistry*, **42**, 14522–14531.
22. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
23. 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., Rogozin,I.B., Smirnov,S., Sorokin,A.V., Sverdlov,A.V., Vasudevan,S., Wolf,Y.I., Yin,J.J. and Natale,D.A. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
24. Rossnes,R. (2004) Phylogenetic reconstruction of ancestral character states for gene expression and mRNA splicing data. MSc Thesis, Universtiy of Bergen, Norway.