

This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

## The Adaptive Evolution Database (TAED)

David A Liberles<sup>1,4</sup>, David R Schreiber<sup>1</sup>, Sridhar Govindarajan<sup>3,5</sup>, Stephen G Chamberlin<sup>3</sup>, and Steven A Benner<sup>1,2</sup>

Addresses: <sup>1</sup>Departments of Chemistry and of <sup>2</sup>Anatomy and Cell Biology, University of Florida, Gainesville, FL 32611 USA. <sup>3</sup>Bioinformatics Division, EraGen Biosciences, 12085 Research Drive, Alachua, FL 32615 USA. <sup>4</sup>Current Address: Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden. <sup>5</sup>Current Address: Maxygen, 515 Galveston Drive, Redwood City, CA 94063, USA.

Correspondence: David A Liberles. E-mail: liberles@sbc.su.se

Posted: 9 March 2001

Received: 7 March 2001

*Genome Biology* 2001, **2(4)**:preprint0003.1-0003.18

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2001/2/4/preprint/0003>

This is the first version of this article to be made available publicly. This article has been submitted to *Genome Biology* for peer review.

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)



deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY PRIMARY RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.



The Adaptive Evolution Database (TAED)

David A. Liberles<sup>1,4\*</sup>, David R. Schreiber<sup>1</sup>, Sridhar Govindarajan<sup>3,5</sup>, Stephen G. Chamberlin<sup>3</sup>, and  
Steven A. Benner<sup>1,2</sup>

<sup>1</sup>Departments of Chemistry and of <sup>2</sup>Anatomy and Cell Biology, University of Florida, Gainesville,  
FL 32611 USA

<sup>3</sup>Bioinformatics Division, EraGen Biosciences, 12085 Research Drive, Alachua, FL 32615 USA

<sup>4</sup>Current Address: Department of Biochemistry and Biophysics and Stockholm Bioinformatics  
Center, Stockholm University, 10691 Stockholm, Sweden

<sup>5</sup>Current Address: Maxygen, 515 Galveston Drive, Redwood City, CA 94063, USA

\*To whom correspondence should be addressed at Department of Biochemistry and Biophysics and  
Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden.

liberles@sbc.su.se

running head: The Adaptive Evolution Database (TAED)

keywords: species diversification, protein, DNA, phylogeny

March 6, 2001

## ABSTRACT

### BACKGROUND

Developing an understanding of the molecular basis for the divergence of species lies at the heart of biology. The Adaptive Evolution Database (TAED) serves as a starting point to link events that occur at the same time in the evolutionary history (tree of life) of species, based upon coding sequence evolution analyzed with the Master Catalog. The Master Catalog is a collection of evolutionary models, including multiple sequence alignments, phylogenetic trees, and reconstructed ancestral sequences, for all independently evolving protein sequence modules encoded by genes in GenBank [1].

### RESULTS

We have estimated from these models the ratio of nonsynonymous to synonymous nucleotide substitution ( $K_a/K_s$ ), for each branch in their respective evolutionary trees of every subtree containing only chordata or only embryophyta proteins. Branches with high  $K_a/K_s$  values represent candidate episodes in the history of the family where the protein may have undergone positive selection, a phenomenon in molecular evolution where the mutant form of a gene must have conferred more fitness than the ancestral form. Such episodes are frequently associated with change in function. We have found that an unexpectedly large number of families (between 10 and 20% of those families examined) have at least one branch with a notably high  $K_a/K_s$  value (putative adaptive evolution). As a resource for biologists wishing to understand the interaction between protein sequences and the Darwinian processes that shape these sequences, we have collected these into The Adaptive Evolution Database (TAED).

## CONCLUSIONS

Placed in a phylogenetic perspective, candidate genes that are undergoing evolution at the same time in the same lineage can be viewed together. This framework based upon coding sequence evolution can be readily expanded to include other types of evolution. In its present form, TAED provides a resource for bioinformaticists interested in data mining and for experimental evolutionists seeking candidate examples of adaptive evolution for further experimental study.

## **BACKGROUND**

The growth of gene and genomic databases motivates efforts to develop tools to extract information about the function of a protein from sequence data with the ultimate goal of understanding the collection of functions represented in an organismal genome. A long history of work in molecular evolution extending over thirty years has shown that such questions must be phrased carefully, and always with cognizance of the Darwinian paradigm that insists that the only way to obtain functional behavior in living systems is through natural selection superimposed upon random variation in structure [2]. A behavior is functional if the host would be less able to survive and reproduce if that behavior were different. An amino acid residue is functional if, upon mutation, the host is less able to survive and reproduce.

A long literature has sought to interpret the evolutionary behavior of protein sequences, in the hope of drawing inferences about the relationship between fitness and sequence [3]. What has emerged is the recognition that a family of orthologous proteins displays a continuum of structure and a corresponding continuum in behavior, where some of the behavioral differences have a strong impact on fitness (are functional), while others are neutral (or nearly so). Without resolving, in a general way, questions regarding the relationship (neutrality vs. selection) between fitness and protein sequence, we can build interpretive tools that capture information from patterns of evolution of genomic sequences that is informative about function, in particular, events that are characterized by the biological scientist as a change in function.

For a protein to change its function, it must change its behavior; this in turn requires that it change its amino acid sequence. A protein being recruited for a very different function over a very short time (geologically speaking) frequently experiences an episode of rapid sequence evolution, an episode where the number of amino acid substitutions per unit time is large. Therefore, molecular evolutionists have long been interested in the rates with which substitutions accumulate in protein sequences. These rates are known to vary widely in different protein families.

Calculating rates in the units substitutions/time requires knowledge of the geological dates of divergence of protein sequences. Because geological times are frequently not known (and almost never known precisely), alternative approaches for identifying episodes of rapid sequence evolution have been sought. One of these examines nucleotide substitutions, and divides the number of nucleotide substitutions that change the sequence of the encoded protein (nonsynonymous substitution) by the number of nucleotide substitutions that do not change the sequence of the encoded protein (synonymous substitution), and then normalizes these for the number of nonsynonymous and synonymous sites. This is the Ka/Ks ratio [4-5]. High Ka/Ks ratios for reconstructed ancestral episodes of sequence evolution are known to be signatures of positive adaptation, which in turn indicate significant change in function [6-7].

In general, Ka/Ks values are low. For example, the average Ka/Ks value in proteins between rodents and primates is 0.2 [8]. This is taken to indicate that most of these proteins, selected for millions of years, attained an optimum function prior to the divergence of rodents and primates. This implies that subsequent evolution was conservative; most nonsynonymous mutations were detrimental to the fitness of the organism.

Functional change can be defined as mutation that alters organismal fitness and is subject to selective pressure. For an example of intraspecific variation, phosphoglucose isomerase in montane beetles shows adaptation to local temperature variations [9]. Orthologous proteins also suffer positive selection. For example, the hemoglobin in the bar-headed goose has suffered adaptive change relative to the hemoglobin from the closely related greylag goose in response to a reduced partial pressure of oxygen at high altitudes [10]. Adaptive evolution is also believed to be displayed in paralogous mammalian MHC class I genes and relate to a birth and death model of gene duplication [11].

Traditionally, positive selection is defined by a Ka/Ks ratio significantly greater than unity. However, the theoretical cutoff of 1 is well known to miss significant functional changes in proteins for several reasons [12]. Long branches can dilute an episode of positive adaptation with episodes

of conservative evolution.  $Ka/Ks$  values can miss positive selective pressures on individual amino acids because they average events over the entire protein sequence. Behavior in a protein can change significantly if only a few amino acids change while the remainder of the sequence is conserved in order to retain core behaviors of the old and new functions (e.g. the protein fold). These adaptive events will only be detected on sufficiently short branches which pinpoint the adaptive change.

Alternative ways to identify  $Ka/Ks$  values below unity that are suggestive of adaptive evolution involve comparison of these values for an individual branch of a tree with those values for branches in the tree generally. If one branch has a  $Ka/Ks$  value far outside of the norm for the family (but still below 1), we can guess that this branch represents an episode of positive selection. This will work for gene families that generally display conservative evolution (such as the SH2 domains) [13], but not for others. For example, many immune system genes show a much more continuous distribution of values, which may indicate that they are perpetually under different amounts of positive selective pressure [11]. In this case, the designation of a cut-off value of  $Ka/Ks$ , below which two homologous genes have the same function, and above which they have different functions, is arbitrary. Ultimately this level should be determined by benchmarking adaptivity with specific functions and specific protein folds.

$Ka/Ks$  ratios are well known to be useful starting points for generating stories about the interaction between protein sequences and the Darwinian processes that shape these sequences. These stories help us understand how these sequences contribute to the fitness of the host. This means that biologists would find useful a comprehensive database of examples where  $Ka/Ks$  values are high. Most useful would be a database that presents families where  $Ka/Ks$  is greater than 1, and a separate family where  $Ka/Ks$  is greater than some arbitrary cut-off less than 1, but still relatively high compared to the average value in the average protein.

We report here such a database, The Adaptive Evolution Database (TAED). TAED is designed as a database to collect molecular events that are candidates for driving divergent

evolution along each branch of the chordate and embryophyta trees of life. TAED contains a collection of protein families where at least one branch in the reconstructed molecular record has a Ka/Ks value greater than 1, or greater than 0.6. The second is an arbitrary cut-off that is high relative to the average Ka/Ks value for the average protein and seems to include many additional examples of genes that are likely to be true positives. Thus, it collects events in the molecular history recorded in the contemporary genomic database that are candidates for adaptive evolution. TAED can be utilized for benchmarking purposes and as a list of potentially adaptively evolving genes for experimentalists interested in studying these candidate genes in further detail and for bioinformaticists interested in studying large datasets of examples of genes with high Ka/Ks ratios.

## **METHODS**

Starting with the Master Catalog (version 1.1 derived from Genbank release 113) [1], Ka/Ks ratios were reconstructed database-wide for each ancestral branch in every evolutionary tree containing genes from chordata and embryophyta. Analysis here was restricted to these organisms because there is less evidence for codon and GC-content biases which complicate the accurate calculation of Ks. Ka/Ks calculations used a modification of the method of Li and Pamilo and Bianchi [4-5] to incorporate reconstructed ancestral sequences and thus allow specific branches undergoing putative adaptive evolution to be identified. The Master Catalog uses multiple sequence alignments generated from Clustal W and neighbor joining trees. Reconstruction of ancestral sequences was done using the Fitch maximum parsimony methodology [14]. While reconstructed ancestral sequences contain ambiguities, using probabilistic ancestral sequences embraces this and allows us to construct a model of evolutionary history that is robust.

While maximum likelihood methodologies perform better in some situations, they are too computationally intensive to apply exhaustively. Further, they are based upon an explicit model of evolution that may not be appropriate along all branches analyzed, a situation where maximum parsimony may outperform maximum likelihood on some branches [15]. Therefore, to generate the

initial version of this database, more computationally simple methods were used. As new methodology is developed, this database will be recalculated using different methodologies. Since ancestral sequence reconstruction, is approximate, these branches should be viewed as candidates rather than absolutely definitive statements of adaptivity.

Two cutoffs of Ka/Ks ratio were utilized, branches with values >1 and with values >0.6. While reconstruction back to the last common ancestor of chordates or embryophyta with no intermediates frequently bears the signature of synonymous position equilibration, synonymous position saturation can be avoided if individual branches are shorter than the period required for saturation to occur ( $t_{1/2}$  to saturation ~120 million years). Saturation was measured through the calculation of neutral evolutionary distances (NED) along branches. NED is defined as the synonymous substitution rate in two-fold redundant codons interchanged by a pyrimidine-pyrimidine transition [16]. These are the fastest equilibrating sites. Branches which showed NED values greater than 5 half lives towards saturation were excluded from TAED based upon differences between reconstructed ancestral sequences at the beginning of branches and sequences at the end.

A second problem of significance is that of short branches bearing fractional mutations. In order to exclude these, a new test was implemented. The modified Ka/Ks calculation is simple and is described below:

$$\text{modified Ka/Ks} = (\text{Ka}_{\text{mod}}) / (\text{Ks}_{\text{mod}})$$

where

$$\text{Ka}_{\text{mod}} = (\text{number of nonsyn} - 1) / \text{total nonsyn. sites}$$

$$\text{Ks}_{\text{mod}} = (\text{number of syn} + 1) / \text{total syn. sites}$$

In general, the smaller the difference between  $Ka/Ks$  and  $Ka_{mod}/Ks_{mod}$ , the more significant or robust the branch. To exclude short branches with fractional mutations without excluding other short branches, branches with  $Ka_{mod}/Ks_{mod}$  values below 0.5 were excluded from TAED.

The resulting dataset is available for further analysis at <http://www.sbc.su.se/~liberles/TAED.html>.

## RESULTS

The Master Catalog is a database of 26,843 families of protein modules [1]. This database was generated from an all-against-all search of Genbank release 113. A protein is broken into independently evolving modules by the presence of a subsection of a gene as a complete open reading frame in another species. Pairs that were within 180 PAM units with a minimum length requirement were grouped into the same family. Each family contains an evolutionary tree and a multiple sequence alignment. This database was the starting point for the exhaustive calculation of  $Ka/Ks$  ratios.

The Master Catalog is different, both in concept and execution, from other resources (e.g. Hovergen [17] Pfam [18], and COGs) that offer databases of protein families. The Master Catalog incorporates reconstructed ancestral states within its data structure, in addition to multiple sequence alignments and evolutionary trees. Having these reconstructed ancestral states provides a dimension of value to the database, especially for functional interpretation, that is not offered by databases that contain only trees, or only multiple sequence alignments, or only trees and multiple sequence alignments. Further, because the Master Catalog is explicitly developed as a tool for doing functional genomics relying reconstructed intermediates, and as the information about function is extracted from analysis of patterns of variation and conservation in genes and proteins within a family, it emphasizes obtaining high quality trees, MSAs, and reconstructed ancestral states. For this reason, the Master Catalog does not attempt to build superfamilies (like Pfam does, for

example). Instead, the Master Catalog constructs nuclear families, where the trees, MSAs, and ancestral states are quite reliable.

Of 5305 families of modules containing chordate proteins, 280 contained at least one branch with a  $Ka/Ks$  value greater than 1, representing 643 branches emanating from 63 different nodes of the tree of life. Some 778 families had at least one branch with a  $Ka/Ks$  value greater than 0.6, totaling 2232 branches emanating from 92 nodes of the tree of life. Thus 15% of all families of chordate modules are likely to have modified their function at least once during the course of evolution.

Of 3385 families of modules representing embryophyta proteins, 123 have at least one branch with a  $Ka/Ks$  value greater than 1, representing 228 families emanating from 25 nodes. Some 407 families had at least 1 branch with a  $Ka/Ks$  value greater than 0.6, totaling 1105 branches from 43 nodes. Here, perhaps 12% of all embryophyta families have modified their function along at least one branch.

This result based upon ancestral sequence reconstruction contrasts greatly with the result of Endo, Ikeya, and Gojobori, where the search for gene families undergoing adaptive evolution yielded only 2 families [19]. These scientists compared extant sequences rather than reconstructed evolutionary intermediates, counted families only where a majority of the pairs at high  $Ka/Ks$  values, and used a smaller database.

A list of protein module family candidates for having undergone modification of function is available on the web at <http://www.sbc.su.se/~liberles/TAED.html>. The version described here is designated TAED 2.1 and will remain available at this site. As more sophisticated methods are developed and applied, as correlations with functional and structural databases are pursued, and as data from other types of evolution beyond coding sequence evolution is added, links to these datasets will be provided. TAED 2.1 contains two image mapped trees (for chordates and embryophyta), where the node that an adaptive branch emanates from can be clicked on to obtain a

list and Master Catalog reference number. Multiple sequence alignments and phylogenetic trees corresponding to these entries can be obtained from EraGen Biosciences (<http://www.era-gen.com>).

Genes that appear on this list appear for several possible reasons. Branches resulting from changes during speciation events to orthologues or following gene duplication events in paralogues will appear. Because this search was done without knowledge of genomic location of genes, paralogues will be indistinguishable from genes with alternative splice patterns or from intraspecific variation. However, for the purpose of this analysis, all four sets of information (orthologues, paralogues, changes in alternative splicing detected from cDNA analysis, and intraspecific variation) reflect organismal mechanisms of adaptation and are relevant for our purposes.

Because there is no reliable truth set for functional adaptation, it is not possible to score the results of this tool. It is important to remember that a Darwinian definition of function differs from the functional annotation of genomes and it is possible for a protein to alter or change its function while retaining the same annotation. To examine this dataset, specific proteins must be examined individually.

In viewing the list of proteins, many of these are already believed to be candidates for functional recruitment. These include plasminogen activator in vampire bats which is expressed in saliva and involved in blood clotting [20], phospholipase A2 in snakes which is expressed in venom and involved in tissue damage [21], and MHC genes in mammals which are involved in the immune system as part of the host-parasite arms race [22], all having obvious stories to explain why they may have suffered functional change. Several families are newly identified as being candidates for functional change, such as the obesity gene protein leptin in primates. A third category of discovery in TAED is in the detection of episodes of adaptive change at new points in the divergent evolution of proteins, for example myostatin in bovidae [23]. A sample table from TAED representing bovidae is presented as Table 1. These are the candidate genes that were identified as showing rapid sequence evolution emanating from this node in the tree of life. They potentially include orthologues between two species of bovidae, paralogues, alternatively spliced transcripts,

and intraspecific evolution. The genes on the list play roles in the immune system, body musclation, and reproduction, traits frequently under selective pressure. These examples and many others are candidates for further experimental study through cloning from additional species and from functional study for labs expert in those specific proteins.

## CONCLUSION

This study represents the first comprehensive analysis of Ka/Ks ratios throughout chordata and embryophyta. While the methods utilized were rough and designed to give a quick snapshot into a global picture of evolution, this resource should be valuable in the analysis of much of chordate evolution. Functional genomics analyses of many of the families that have suffered recruitment and functional change within the past 500 million years will soon emerge. Many of the episodes of functional change recorded in TAED can be correlated with events in the geological or paleontological record, in response to changing environments, evolving paleoecology, or the development of new physiology.

From a phylogenetic perspective, the knowledge of candidate genes evolving at the same time in the same organism can allow one to begin to ask if entire pathways or phenotypic functions are under selective pressure at specific points in evolutionary history. Where tertiary structures exist, mutations along branches can be mapped onto three dimensional structures first to evaluate the validity of specific examples, and second, to understand the nature of adaptive evolution at a structural level.

One statistical analysis of this database indicates that among branches with Ka/Ks ratios >1, only 3% of synonymous sites had mutated compared with 10% on the average branch in the database. This is consistent with the notion that episodes of adaptive evolution can be lost in long branches, as these are combined with prior and/or subsequent episodes characterized by lower Ka/Ks ratios characteristic of functional constancy. As more genes are sequenced from more species, the greater articulation of trees will not only increase the accuracy of sequence

reconstructions, but will also allow us to detect new examples of functional change that are buried in long branches.

At a biological level, the dataset generated here can be data-mined to provide global pictures of how evolution has occurred. Correlation of data in this database with that in other functional databases will enable a leap from genotype to organismal phenotype. Further, the dataset provides a resource for experimentalists interested in specific genes. The high Ka/Ks ratio in leptin in a branch connecting primates with rodents may have been a useful predictor of changes of function for pharmaceutical companies interested in the mouse model of leptin for human obesity. For the experimentalist, mutations occurring along putatively adaptive branches can be assayed for functional importance in systems of interest.

Finally, this database represents a growing framework for the study of adaptive evolution. As datasets become available, changes in gene expression, alternative splicing patterns, imprinting patterns, recombination events, and other molecular mechanisms of adaptation will be added to this database in a phylogenetic perspective. The ultimate goal is a dynamic resource depicting candidate molecular events that are responsible for phenotypic differences between closely related species.

### **Acknowledgments**

We thank Eric Gaucher for critical reading of this manuscript. We are indebted to the National Institutes of Health (Grants HG 01729 and MH 55479) for partial support of this work. TAED is freely available at <http://www.sbc.su.se/~liberles/TAED.html>. The Master Catalog can be obtained free of charge for academic users through [info@eragen.com](mailto:info@eragen.com).

## References

1. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, and Knecht L: **Functional inferences from reconstructed evolutionary biology involving rectified databases- an evolutionarily grounded approach to functional genomics.** *Res. Microbiol.* 2000, **151**:97-106.
2. Benner SA and Ellington AD: **Interpreting the behavior of enzymes Purpose or pedigree?** *CRC Crit. Rev. Biochem.* 1988, **23**:369-426.
3. Kimura M: *Molecular Evolution Protein Polymorphism and the Neutral Theory.* Berlin: Springer-Verlag, 1982.
4. Li WH, Wu CI, and Luo CC: **A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.** *Mol. Biol. Evol.* 1985, **2**:150-174.
5. Pamilo P and Bianchi NO: **Evolution of the Zfx and Zfy genes: rates and interdependence between the genes.** *Mol. Biol. Evol.* 1993, **10**:271-281.
6. Trabesinger-Ruef N, Jermann TM, Zankel TR, Durrant B, Frank G, and Benner SA: **Pseudogenes in ribonuclease evolution: A source of new biomacromolecular function?** *FEBS Lett.* 1996, **382**:319-322.
7. Messier W and Stewart CB: **Episodic adaptive evolution of primate lysozymes.** *Nature* 1997, **385**:151-154.
8. Makalowski W and Boguski MS: **Evolutionary parameters of the transcribed mammalian genome: An analysis of 2820 orthologous rodent and human sequences.** *Proc. Natl. Acad. Sci., USA* 1998, **95**:9407-9412.
9. Dahlhoff EP and Rank NE: **Functional and physiological consequences of genetic variation at phosphoglucose isomerase: Heat shock protein expression is related to enzyme genotype in a montane beetle.** *Proc. Natl. Acad. Sci., USA* 2000, **97**:10056-10061.

10. Zhang J, Ziqian H, Tame JRH, Lu G, Zhang R, and Gu X: **The crystal structure of a high oxygen affinity species of hemoglobin. bar-headed goose haemoglobin in the oxy form.** *J. Mol. Biol.* 1996, **255**:484-493.
11. Nei M, Gu X and Sitnikova T: **Evolution by the birth-and-death process in multigene families of the vertebrate immune system.** *Proc. Natl. Acad. Sci., USA* 1997, **94**: 7799-7806.
12. Crandall KA, Kelsey CR, Imamichi H, Lane HC, and Salzman NP: **Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection.** *Mol. Biol. Evol.* 1999, **16**:372-382.
13. Wigger M: *Receptor-Assisted Combinatorial Synthesis. RACS.- A New Approach for Combinatorial Chemistry.* Zuerich: PhD Thesis #12929, Swiss Federal Institute of Technology, 1998.
14. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst. Zool.* 1971, **20**:406-416.
15. Page RDM and Holmes EC: *Molecular Evolution, A Phylogenetic Approach.* Oxford: Blackwell Sciences, 1998.
16. Peltier MR, Raley LC, Liberles DA, Benner SA, and Hansen PJ: **Evolutionary history of the uterine serpins.** *J. Exp. Zoo.* 2000, **288**:165-174.
17. Duret L, Mouchiroud D, and Guoy M: **HOVERGEN: a database of homologous vertebrate genes.** *Nucleic Acids Research* 1994, **22**:2360-2365.
18. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, and Sonnhammer ELL: **The Pfam protein families database.** *Nucleic Acids Research* 2000, **28**:263-266.
19. Endo T, Ikeo K, and Gojobori T: **Large-scale search for genes on which positive selection may operate.** *Mol. Biol. Evol.* 1996, **13**:685-690.
20. Bode W and Renatus M: **Tissue-type plasminogen activator: variants and crystal/solution structures demarcate structural determinants of function.** *Curr. Opin. Struc. Biol.* 1997, **7**:865-872.

21. Nakashima KI, Nobuhisa I, Deshimaru M, Nakai M, Ogawa T, Shimohigashi Y, Fukumaki Y, Hattori M, Sakaki Y, Hattori S, and Ohno M: **Accelerated evolution in the protein-coding regions is universal in Crotalinae snake venom gland phospholipase A(2) isozyme genes.** *Proc. Natl. Acad. Sci., USA* 1995, **92**:5605-5609.
22. Hughes AL and Nei M: **Patterns of nucleotide substitution at Major Histocompatibility Complex class-I loci reveals overdominant selection.** *Nature* 1988, **335**: 167-170.
23. Lee SJ and McPherron AC: **Myostatin and the control of skeletal muscle mass.** *Curr. Opin. Gen. Dev.* 1999, **9**:604-607.

Table 1. A sample listing from TAED indicating candidate adaptively evolving genes detected that emanated from the bovidae node. These examples potentially include orthologues between different species of bovidae, paralogues, alternatively spliced cDNAs with potentially different functional effects, and intraspecific modifications.

---

---

The genes with  $Ka/Ks > 1.0$  are:

1. T-cell receptor CD3 epsilon chain from Master Catalog family 9668
  2. AF092740 cytotoxic T-lymphocyte-associated protein 4 precursor from Master Catalog family 9698
  3. CD5 from Master Catalog family 9700
  4. AF110984 intercellular adhesion molecule-1 precursor from Master Catalog family 9802
  5. interferon alpha/beta receptor-2 from Master Catalog family 9817
  6. AF020508 pregnancy-associated glycoprotein 6 from Master Catalog family 15612
  7. MCH OVAR-DQ-ALPHA1 from Master Catalog family 15669
  8. major histocompatibility complex class II from Master Catalog family 21739
  9. TCR gamma from Master Catalog family 21940
- 
- 

Additional genes with  $Ka/Ks > 0.6$  are:

10. interleukin 2 receptor from Master Catalog family 9745
  11. interleukin-3 from Master Catalog family 9775
  12. AF019622 myostatin; growth/differentiation factor-8; GDF-8 from Master Catalog family 20325
  13. Fas gene product from Master Catalog family 21743
  14. calpastatin from Master Catalog family 21751
  15. prolactin receptor from Master Catalog family 21853
  16. pre-pro serum albumin from Master Catalog family 21864
  17. immunoglobulin gamma-1 chain from Master Catalog family 21881
  18. AF110984 intercellular adhesion molecule-1 precursor from Master Catalog family 21997
- 
-