

Research

The Adaptive Evolution Database (TAED)

David A Liberles^{*†}, David R Schreiber^{*}, Sridhar Govindarajan^{‡§},
Stephen G Chamberlin[‡] and Steven A Benner^{*¶}

Addresses: ^{*}Departments of Chemistry and [¶]Anatomy and Cell Biology, University of Florida, Gainesville, FL 32611, USA. [‡]Bioinformatics Division, EraGen Biosciences, 12085 Research Drive, Alachua, FL 32615, USA. Current addresses: [†]Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden. [§]Maxygen, 515 Galveston Drive, Redwood City, CA 94063, USA.

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

Published: 24 July 2001

Genome Biology 2001, **2**(8):research0028.1-0028.6

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2001/2/8/research/0028>

© 2001 Liberles et al., licensee BioMed Central Ltd
(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 7 March 2001

Revised: 21 May 2001

Accepted: 6 June 2001

A previous version of this manuscript was made available before peer review at <http://genomebiology.com/2001/2/4/preprint/0003/> (*Genome Biology* 2001, **2**(4):preprint0003.1-0003.18)

Abstract

Background: The Master Catalog is a collection of evolutionary families, including multiple sequence alignments, phylogenetic trees and reconstructed ancestral sequences, for all protein-sequence modules encoded by genes in GenBank. It can therefore support large-scale genomic surveys, of which we present here The Adaptive Evolution Database (TAED). In TAED, potential examples of positive adaptation are identified by high values for the normalized ratio of nonsynonymous to synonymous nucleotide substitution rates (K_A/K_S values) on branches of an evolutionary tree between nodes representing reconstructed ancestral sequences.

Results: Evolutionary trees and reconstructed ancestral sequences were extracted from the Master Catalog for every subtree containing proteins from the Chordata only or the Embryophyta only. Branches with high K_A/K_S values were identified. These represent candidate episodes in the history of the protein family when the protein may have undergone positive selection, where the mutant form conferred more fitness than the ancestral form. Such episodes are frequently associated with change in function. An unexpectedly large number of families (between 10% and 20% of those families examined) were found to have at least one branch with high K_A/K_S values above arbitrarily chosen cut-offs (1 and 0.6). Most of these survived a robustness test and were collected into TAED.

Conclusions: TAED is a raw resource for bioinformaticists interested in data mining and for experimental evolutionists seeking candidate examples of adaptive evolution for further experimental study. It can be expanded to include other evolutionary information (for example changes in gene regulation or splicing) placed in a phylogenetic perspective.

Background

The growth of gene and genomic databases provides motivation for developing 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 organism's genome. Work on molecular evolution over

30 years has shown that such questions must be phrased carefully, and always with cognizance of the Darwinian paradigm that insists that the only way of obtaining functional behavior in living systems is through natural selection superimposed on random variation in structure [1]. A behavior is functional if the organism would be less able to survive

and reproduce if that behavior were different. An amino-acid residue is functional if, upon mutation, the organism 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 [2]. What has emerged is the recognition that a family of orthologous proteins displays a diversity of structure and a corresponding diversity in behavior, where some of the behavioral differences have a strong impact on fitness (are functional), and others are neutral (or nearly so). Without resolving, in a general way, questions regarding the relationship (neutrality versus 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 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 at which substitutions accumulate in protein sequences. These rates are known to vary widely in different protein families.

Calculating rates in the units of 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. It 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 K_A/K_S rate ratio [3-5]. High K_A/K_S rate ratios for reconstructed ancestral episodes of sequence evolution are hypothesized to be signatures of positive adaptation, and have been associated with significant change in function [6,7].

In general, K_A/K_S values are low. For example, the average K_A/K_S value in proteins between rodents and primates is 0.2 [8]. This is taken to indicate that most of these proteins, selected over 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 undergone 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 by 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 K_A/K_S rate ratio significantly greater than unity. While $0.6 < K_A/K_S < 1$ can occur by relaxation of functional constraint, the theoretical cut-off 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 $K_A/K_S > 1$) with episodes of conservative evolution. K_A/K_S 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 (for example, the protein fold). These adaptive events will only be detected on sufficiently short branches which pinpoint the adaptive change.

Alternative ways of identifying K_A/K_S 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 K_A/K_S 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 (Src homology 2) 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 K_A/K_S , 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.

K_A/K_S rate 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 K_A/K_S values are high. Most useful would be a database that presents families where K_A/K_S is greater

than 1, and a separate family where K_A/K_S 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 to provide, in raw form, evolutionary episodes in specific chordate and embryophyte (flowering plants, conifers, ferns, mosses and liverworts) protein families that might be candidates for adaptive evolution. TAED contains a collection of protein families where at least one branch in the reconstructed molecular record has a K_A/K_S value greater than unity, or greater than 0.6. The second cut-off is arbitrary, chosen to be high relative to the average K_A/K_S value for the average episode of evolution in a protein family. Empirically, the lower cut-off seems to admit additional examples of gene families that might have undergone adaptive evolution.

TAED should be used as a raw list of potentially adaptively evolving genes for experimentalists seeking gene families to study in further detail, and for bioinformaticists interested in studying large datasets of examples of genes with high K_A/K_S rate ratios.

Results

The Master Catalog [14] is a database of 26,843 families of protein modules generated from an all-against-all search of GenBank release 113. A protein is broken into independently evolving modules on the criterion of the presence of a subsection of a gene as a complete open reading frame in another species. Pairs that were within 180 PAM (point accepted mutation) 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 K_A/K_S rate ratios.

The Master Catalog is different, both in concept and execution, from other resources (for example Hovergen [15] Pfam [16], 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 (MSAs) and evolutionary trees. Having these reconstructed ancestral states provides a 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 on 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 the generation of high-quality trees, MSAs, and reconstructed ancestral states. For this reason, the Master Catalog does not attempt to build superfamilies (like Pfam

does). Instead, it constructs nuclear families, where the trees, MSAs, and ancestral states are not compromised by poor gap placement, a common problem in Clustal-based multiple sequence alignments of sets of highly divergent protein sequences.

Alternative methods were considered for reconstructing ancestral sequences. Whereas 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 [17]. Therefore, to generate the initial version of this database, more computationally simple methods were used. As improved methodologies are developed, these will undoubtedly be applied to recalculate this database.

Two issues concerned the scope of the K_A/K_S analysis. First, we were concerned that silent positions would be 'saturated' with substitutions, rendering the K_S measurements meaningless. Whereas reconstruction back to the last common ancestor of chordates or embryophytes 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 of approximately 120 million years). Saturation was measured through the examination of the extent to which twofold redundant codon systems had reached equilibration [18]. Branches that showed equilibration greater than five half-lives towards saturation were excluded from TAED on the basis of differences between reconstructed ancestral sequences at the beginning of branches and sequences at the end.

A second significant problem is that of short branches bearing fractional mutations. These are known to generate K_A/K_S values with large errors. To prevent these errors from biasing the database, a new simple robustness test was implemented to ensure that an 'interesting' K_A/K_S value (one above the cut-off) was not recorded in the database if it became 'uninteresting' (below the cut-off) through the shift of a single mutation reconstructed in the branch. The test modified the K_A/K_S calculation in a simple way, as described below:

$$\text{modified } K_A/K_S = K_{A\text{mod}}/K_{S\text{mod}}$$

where

$$K_{A\text{mod}} = (\text{number of nonsynonymous} - 1)/\text{total nonsynonymous sites}$$

$$K_{S\text{mod}} = (\text{number of synonymous} + 1)/\text{total synonymous sites}$$

In general, the smaller the difference between K_A/K_S and $K_{A\text{mod}}/K_{S\text{mod}}$, the more significant or robust the branch.

To exclude short branches with fractional mutations (arising through ambiguous ancestral sequence reconstruction) without excluding other short branches, branches with K_A/K_S values below 0.5 were excluded from TAED.

Of 5,305 families of modules containing chordate proteins, 280 contained at least one branch with a K_A/K_S 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 K_A/K_S value greater than 0.6, totaling 2,232 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 3,385 families of modules representing embryophyte proteins, 123 have at least one branch with a K_A/K_S value greater than 1, representing 228 families emanating from 25 nodes. Some 407 families had at least one branch with a K_A/K_S value greater than 0.6, totaling 1,105 branches from 43 nodes. Here, perhaps 12% of all embryophyte families have modified their function along at least one branch.

This result based on ancestral sequence reconstruction contrasts greatly with the result of Endo, Ikeo and Gojobori, where the search for gene families undergoing adaptive evolution yielded only two families [19]. They compared extant sequences rather than reconstructed evolutionary intermediates, counted families only where a majority of the pairs were at high K_A/K_S values, and used a smaller database.

A list of candidate protein module families that have undergone modification of function is available at [20]. 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 embryophytes), 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 [21].

Discussion

This study represents the first comprehensive analysis of K_A/K_S rate ratios throughout the Chordata and the Embryophyta. Although the methods utilized were rough and designed to give a quick snapshot into a global picture of evolution, the TAED, as a raw resource, should be valuable in the analysis of much of chordate evolution. Functional genomics analyses of many of the protein 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.

Gene families may display evolutionary episodes with high K_A/K_S values, and therefore appear within TAED, for several possible reasons. For example, branches resulting from gene duplication events that give rise to paralogs with very different behaviors will presumably have high K_A/K_S values, as will orthologous pairs from species that place very different demands on their function. This search was done without distinguishing paralogs from orthologs, and the user of TAED should be careful in the analysis of specific families in recognition of this fact.

Table 1

A sample listing from TAED indicating candidate adaptively evolving genes detected that emanated from the Bovidae node

Genes with $K_A/K_S > 1.0$

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. T-cell receptor gamma from Master Catalog family 21940

Additional genes with $K_A/K_S > 0.6$

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
-

These examples potentially include orthologs between different species of Bovidae, paralogs, alternatively spliced cDNAs with potentially different functional effects and intraspecific modifications.

Because there is no reliable true set of protein families 'known' to have suffered 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.

Individual examination is likely to be productive, however. Many protein families already believed to be candidates for functional recruitment appear on the list. These include plasminogen activator in vampire bats which is expressed in saliva and involved in blood clotting [22], phospholipase A2 in snakes which is expressed in venom and involved in tissue damage [23] and MHC proteins in mammals, which are involved in the immune system as part of the host-parasite arms race [24], all having obvious explanations of why they may have suffered functional change. Several families are newly identified as being candidates for functional change, such as the previously proposed obesity 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 the Bovidae [25]. Table 1 is a sample table from TAED representing bovinds. 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 orthologs between two species of bovinds, paralogs, alternatively spliced transcripts and intraspecific evolution. The genes on the list have roles in the immune system, body musculature and reproduction, traits frequently under selective pressure. These examples and many others are candidates for further experimental study through cloning from additional species and through functional study in laboratories expert in the particular protein.

Conclusions

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 particular points in evolutionary history. Where tertiary structures for the proteins 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 analysis of TAED indicates that among branches with K_A/K_S rate 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 K_A/K_S rate 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 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 K_A/K_S rate ratio in leptin in a branch connecting primates with rodents may have been a useful predictor of change 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.

Materials and methods

Starting with the Master Catalog [14] (version 1.1 derived from GenBank release 113; free of charge for academic users through [21]), K_A/K_S rate ratios were reconstructed database-wide for each ancestral branch in every evolutionary tree containing genes from the Chordata and the Embryophyta. This analysis was restricted to these organisms because there is less evidence for codon and GC-content biases which complicate the accurate calculation of K_S . The Master Catalog uses multiple sequence alignments generated from Clustal W and neighbor-joining trees, both derived from protein sequences. Because the Master Catalog is based on an analysis of nuclear families, rather than extended families, these inexpensive tools generate acceptable multiple sequence alignments.

K_A/K_S values were calculated for branches on an evolutionary tree between nodes using the method of Li and Pamilo and Bianchi [3-5] modified to allow full treatment of probabilistic ancestral sequences [26]. Reconstruction of ancestral sequences was done using the Fitch maximum parsimony methodology [27]. While reconstructed ancestral sequences contain ambiguities, using probabilistic ancestral sequences takes this into account (by weighting ambiguous positions

according to their probabilities) and allows us to construct a model of evolutionary history that is robust. Two cut-offs were used to identify 'interesting' values for the K_A/K_S rate ratio, 1 and 0.6. Separate databases were constructed for each cut-off. The resulting dataset is freely available for further analysis at [20].

Acknowledgements

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.

References

- Benner SA, Ellington AD: **Interpreting the behavior of enzymes. Purpose or pedigree?** *CRC Crit Rev Biochem* 1988, **23**:369-426.
- Kimura M: *Molecular Evolution Protein Polymorphism and the Neutral Theory*. Berlin: Springer-Verlag, 1982.
- Li WH, Wu CI, 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.
- Pamilo P, Bianchi NO: **Evolution of the Zfx and Zfy genes: rates and interdependence between the genes.** *Mol Biol Evol* 1993, **10**:271-281.
- Li WH: **Unbiased estimation of the rates of synonymous and nonsynonymous substitution.** *J Mol Evol* 1993, **36**:96-99.
- Trabesinger-Ruef N, Jermann TM, Zankel TR, Durrant B, Frank G, Benner SA: **Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function?** *FEBS Lett* 1996, **382**:319-322.
- Messier W, Stewart CB: **Episodic adaptive evolution of primate lysozymes.** *Nature* 1997, **385**:151-154.
- Makalowski W, 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.
- Dahlhoff EP, 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.
- Zhang J, Ziqian H, Tame JRH, Lu G, Zhang R, Gu X: **The crystal structure of a high oxygen affinity species of haemoglobin (bar-headed goose haemoglobin in the oxy form).** *J Mol Biol* 1996, **255**:484-493.
- Nei M, Gu X, 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.
- Crandall KA, Kelsey CR, Imamichi H, Lane HC, 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.
- Wigger M: *Receptor-Assisted Combinatorial Synthesis. RACS - A New Approach for Combinatorial Chemistry*. Zurich: PhD Thesis 12929, Swiss Federal Institute of Technology, 1998.
- 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.
- Duret L, Mouchiroud D, Guoy M: **HOVERGEN: a database of homologous vertebrate genes.** *Nucleic Acids Res* 1994, **22**:2360-2365.
- Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28**:263-266.
- Page RDM, Holmes EC: *Molecular Evolution, A Phylogenetic Approach*. Oxford: Blackwell Sciences, 1998.
- Peltier MR, Raley LC, Liberles DA, Benner SA, Hansen PJ: **Evolutionary history of the uterine serpins.** *J Exp Zool* 2000, **288**:165-174.
- Endo T, Ikeo K, Gojobori T: **Large-scale search for genes on which positive selection may operate.** *Mol Biol Evol* 1996, **13**:685-690.
- The Adaptive Evolution Database (TAED) 2.1** [<http://www.sbc.su.se/~liberles/TAED.html>]
- EraGen Biosciences** [<http://www.eragen.com>]
- Bode W, Renatus M: **Tissue-type plasminogen activator: variants and crystal/solution structures demarcate structural determinants of function.** *Curr Opin Struct Biol* 1997, **7**:865-872.
- Nakashima KI, Nobuhisa I, Deshimaru M, Nakai M, Ogawa T, Shimohigashi Y, Fukumaki Y, Hattori M, Sakaki Y, Hattori S, 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.
- Hughes AL, Nei M: **Patterns of nucleotide substitution at Major Histocompatibility Complex class-I loci reveals overdominant selection.** *Nature* 1988, **335**:167-170.
- Lee SJ, McPherron AC: **Myostatin and the control of skeletal muscle mass.** *Curr Opin Genet Dev* 1999, **9**:604-607.
- Benner SA, Trabesinger N, Schreiber D: **Post-genomic science: Converting primary structure into physiological function.** *Adv Enzyme Regul* 1998, **38**:155-180.
- Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst Zool* 1971, **20**:406-416.