

Detecting and Characterizing Adaptive Evolution in Chordate Proteins

David A. Liberles

in Comparative Genomics of Vertebrates: Concepts and Bioinformatic Tools,

published by INSERM

The genome of an organism has been shaped by many forces acting throughout its evolution. Gene duplication and recombination shape the gene and domain content of genomes. Within any gene family, gene pairs originating from a gene duplication event are termed paralogues, while those maintained through speciation and divergence events are termed orthologues. Both paralogues and orthologues are subject to the processes of mutation, drift, and selection.

Mutational events include indel events resulting in an insertion or deletion to a coding sequence, as well as point mutations altering the nucleotide at a specific position within a sequence. Interesting cases of selection have been detected involving indel events (Podlaha and Zhang, 2003), but this chapter will focus on the effects of simple point mutation.

Point mutation is a random process and mutations drift through a population dependent upon population size and structure, ultimately resulting in fixation throughout a population or loss from a population. In addition to this neutral process, selection occurs on nucleotide positions. Selection can occur at many levels, from genomic selection for GC content to mRNA selection for base pairing to translational selection for codon bias involving different tRNAs. This chapter will focus on events at the protein level, which occur on top of these other events.

Protein-level selective pressures that depart from neutrality can be either negative or positive. A negative selective pressure occurs when a residue (amino acid) is optimally suited for its functional role. In this case, mutations resulting in any other amino acid are eliminated from the population because such organisms are less fit. In practice, a continuum of selection between strongly negative (only 1 amino acid viable) to negative (for example only 5 hydrophilic amino acids like KREND are viable) to neutral (all 20 amino acids equally fit) exists.

A positive selective pressure occurs when an organism with a mutation has a higher fitness than those without and the mutation is swept to fixation through the population. The amino acid under positive selective pressure can represent an innovation that ultimately allows the organism to adapt to its environment in a new way. The substitution can also be a response to a slightly deleterious substitution elsewhere on the protein or on an interacting protein. For example, an optimized ionic contact between a Glu and a Lys can behave this way. If the Glu is mutated to Asp, the negative charge is retained, but is linked by a shorter chain to the backbone. A corresponding Lys to Arg mutation may restore the high affinity charge interaction. Both types of substitutions are under positive selective pressure and indistinguishable without closer examination.

Independent of protein function, one general observation that has been made is that single copy (those showing orthologous relationships to close outgroup species) and recently duplicated genes in a genome (those showing in-paralogous relationships to close outgroup species) do not behave the same way (see for example Wagner, 2000). Ohno (Ohno, 1970) famously proposed that gene duplication was a mechanism of generating evolutionary novelty. The mechanism behind this was that gene duplicates were freed from negative selective pressures, which enabled them to explore sequence space until one of the copies diverged sufficiently to no longer effectively carry out the original function. That copy was then free to continue to explore sequence space in the search for a new function, possibly involving positive selection and neofunctionalization. Lynch (Force et al., 1999) has characterized neofunctionalization, subfunctionalization, and pseudogenization as the alternative fates of duplicated genes.

With an interest in understanding neofunctionalization along individual

lineages, we now turn to the methodologies for detecting it. From the genetic code, some nucleotide positions redundantly encode the same amino acid. These positions, frequently third codon positions, are termed synonymous and their rate of substitution is termed the synonymous nucleotide substitution rate. Positions where substitution can alter the encoded amino acid are termed nonsynonymous positions and their rate of substitution is termed the nonsynonymous nucleotide substitution rate. The ratio of the rates of nonsynonymous to synonymous substitution (also known as K_a/K_s , dN/dS , or ω) is an indicator of selective pressures on a gene. When K_a is significantly larger than K_s , this is evidence for positive selection, when individuals with amino acid mutations were more fit than individuals without such mutations. In the context of a multiple sequence alignment and phylogenetic tree, K_a/K_s can be fit as a branch-specific parameter within the likelihood model (Yang, 1998). Alternatively, discrete ancestral sequences at internal branch points can be determined and K_a/K_s calculated pairwise between sequences along branches (Messier and Stewart, 1997).

Ultimately, averaging K_a/K_s over a whole gene is likely to be a conservative measure for detecting positive selective pressures. In a protein undergoing neofunctionalization, sites responsible for proper folding will likely be under negative selective pressure, while those, for example, in a binding pocket evolving to bind a new ligand will be under positive selective pressure. One method to partition the sequence is known as primary sequence windowing, where K_a/K_s is calculated over a series of contiguous residues in primary sequence (Endo et al., 1996; Fares et al., 2002). A second method partitions the sites into those that are absolutely invariable through evolutionary time and those that are potentially variable. Then, K_a/K_s is calculated for those residues that are potentially variable (Siltberg and

Liberles, 2002). Ultimately, it is the interaction of structure and function that is undergoing selection. A third method, called tertiary sequence windowing, calculates Ka/Ks in spheres of a given radius around each amino acid position (Berglund and Liberles, in preparation).

Another approach for detecting positive selective pressures is to look for a shift in the substitution model. A specific likelihood ratio test is designed to increase the number of parameters and ask for a correspondingly significant improvement in the likelihood score. The most common parameter to look for is α , the shape parameter of the Γ distribution of amino acid rates across sites (Gaucher et al., 2002). The optimum substitution matrix describing a site can also be tested (Soyer et al., 2003).

These methods can be applied to detect neofunctionalization in large scale in chordate gene families. HOVERGEN (Duret et al., 1994) and the Master Catalog (Benner et al., 2000) are two collections of gene families that can be used in such an analysis. Our first approach to detecting neofunctionalization used the Master Catalog as a starting point (Liberles et al., 2001). More recently, we have begun using our own gene families.

A simple procedure to build appropriate gene families is to start with an all against all BLAST search to eliminate proteins not closely related to each other. The cutoff one uses can be defined depending upon if the family definition should be based upon species taxonomy or sequence distance. If one is interested in sequence distance as a criterion, one can then filter the BLAST hits with a defined PAM distance and pairwise length threshold. Different algorithms are available for single and complete linkage clustering to generate families. Once a set of families has been identified, protein-based multiple sequence alignments and phylogenetic trees can be

calculated. The exact procedure depends upon the scientific questions to be addressed with the gene families.

In calculating Ka/Ks over all 5305 chordate gene families in the Master Catalog, we identified 643 branches from 280 gene families representing 63 branches of the NCBI taxonomy (Benson et al., 2004) after a gene tree to species tree mapping (Liberles et al., 2001). These were collected in a database called The Adaptive Evolution Database or TAED. As more chordate genomes are sequenced, the gene families will fill out with shorter branch lengths between genes and more accurate information. Many of the positively selected genes in the initial TAED calculation were immune system genes and genes involved in reproduction, two classic examples of the evolutionary arms race. Three mammalian genes from the original TAED calculation that we have examined in more detail are leptin, plasminogen activator, and myostatin.

Leptin has been identified as the obesity gene in mice. Leptin deficient mice are obese and treatment with leptin cures obesity. However, mice did not seem to be a good model organism for the behavior of leptin in humans (Benner et al., 1998). The leptin tree in TAED offered some indication of primate-specific evolution. This was also seen in a shift in α between primates and other Eutherian mammals in the leptin gene family (Siltberg and Liberles, 2002). Further, the extracellular domain of the leptin receptor showed similar patterns of positive selection to leptin (Benner et al., 1998). Evidence for transposon-mediated novel expression of leptin in placenta and another insertion event in its receptor during primate evolution has also emerged (Bi et al., 1997; Kapitonov and Jurka, 1999). Using structure, a novel binding site in leptin has been proposed (Gaucher et al., 2003). Further, the binding interface of leptin with its receptor as modeled

computationally (Hiroike et al., 2000) has undergone radical substitution. All of this points to some neofunctionalization of leptin in primates and a partially different functional role from what it has in mice.

Plasminogen activator has undergone multiple rounds of gene duplication between *Desmodus rotundus* (common vampire bat) and its closest nonsanguivorous relative, *Carollia perspicillata* (California short tailed fruit bat) (Kemi, Savolainen, and Liberles, unpublished observations). In TAED, positive selection is apparent from the cow outgroup to the gene duplicates, as well as along the branches separating the duplicates from each other (Liberles et al., 2001). The vampire bat paralogues, which are expressed in saliva, enable vampire bats to prevent blood clotting when lapping upon blood from a bite. Ongoing sequencing and analysis aims to understand the divergent functional roles of the different duplicates.

Myostatin is a negative regulator of skeletal muscle development that has been implicated in a double muscling phenotype in cattle and mice (Lee and McPherron, 1999). Myostatin in TAED showed positive selective pressure separating cow and sheep (Liberles et al., 2001). Interestingly, it also showed positive selective pressure following gene duplication in teleost fish. In the artiodactyl tree, myostatin was subsequently sequenced from five additional ruminant artiodactyls and positive selection pinpointed to three separate branches between 10 and 24 million years ago (Tellgren et al., 2004). These branches may overlap with a period in evolution when ruminants first populated Africa and increased in body mass.

As gene and genome sequencing in various chordate species increases, comparative genomics offers the power to detect positive selection and neofunctionalization, ultimately providing hints into molecular mechanisms for lineage-specific adaptation. Simultaneously, the methodology to detect important

changes is progressing. This combination allows a framework for a better understanding of the evolution of chordate genomes.

Supplementary Information

Some interesting websites for using various methods and databases online:

-Ka/Ks, gene families, other functionality in the future: <http://www.bioinfo.no/>

-current home of TAED: <http://www.sbc.su.se/~liberles/TAED2002/> (to be moved to the above address in 2004)

-HOVERGEN: <http://pbil.univ-lyon1.fr/databases/hovergen.html>

-NCBI: <http://www.ncbi.nlm.nih.gov/>

-a server for likelihood ratio tests:

<http://www.daimi.au.dk/~compbio/rateshift/protein.html>

References

Benner SA, Trabesinger N, and Scheiber D. 1998. Post-genomic science: Converting primary sequence into physiological function. *Adv. Enzyme Regul.* 38:155-190.

Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, and Knecht L. 2000. Reconstructed evolutionary biology involving rectified databases- An evolutionarily grounded approach to functional genomics. *Res. Microbiol.* 151:97-106.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. 2004. GenBank: Update. *Nucl. Acids Res.* 32:D23-D26.

Bi S, Garilova O, Gong DW, Mason MM, and Reitman M. 1997. Identification of a placental enhancer for the human leptin gene. *J. Biol. Chem.* 272:30583-30588.

Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: A database of homologous vertebrate genes. *Nucl. Acids Res.* 22:2360-2365.

Endo T, Ikeo K, and Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13:685-690.

Fares MA, Elena SF, Ortiz J, Moya A, and Barrio E. 2002. A sliding window-based method to detect selective constraints in protein coding genes and its application to RNA viruses. *J. Mol. Evol.* 55:509-521.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531-1545.

Gaucher EA, Gu X, Miyamoto MM, and Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* 27:315-321.

Gaucher EA, Miyamoto MM, and Benner SA. 2003. Evolutionary, structural, and biochemical evidence for a new interaction site of the leptin obesity protein. *Genetics* 163:1549-1553.

Hiroike T, Higo J, Jingami H, and Toh, H. 2000. Homology modeling of human leptin/leptin receptor complex. *Biochem. Biophys. Res. Comm.* 275:154-158.

Kapitonov VV and Jurka J. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *J. Mol. Evol.* 48:248-251.

Lee SJ and McPherron AC. 1999. Myostatin and the control of skeletal muscle mass. *Curr. Opin. Genet. Dev.* 9:604-607.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, and Benner SA. 2001. The Adaptive Evolution Database (TAED). *Genome Biol.* 2(8):research0028.1-research0028.6.

Messier W and Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151-154.

Ohno S. 1970. *Evolution by gene duplication*. New York: Springer-Verlag.

Podlaha O and Zhang J. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc. Natl. Acad. Sci., USA* 100:12241-12246.

Siltberg J and Liberles, DA. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. *J. Evol. Biol.* 15:588-594.

Soyer OS, Dimmic MW, Neubig RR, and Goldstein RA. 2003. Dimerization in aminergic G-protein-coupled receptors: Application of a hidden-site class model of evolution. *Biochem.* 42:14522-14531.

Tellgren Å, Berglund AC, Savolainen P, Janis CM, and Liberles DA. 2004. Myostatin rapid sequence evolution in ruminants predates domestication. submitted.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci., USA* 97:6579-6584.

Yang ZH. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568-573.