Sequence analysis

Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence

Jacob A. Tennessen

Department of Zoology, 3029 Cordley Hall, Oregon State University, Corvallis, OR 97331, USA

Received on March 6, 2008; revised on April 21, 2008; accepted on April 22, 2008

Advance Access publication April 28, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Functional divergence among proteins is often assumed to be strongly influenced by natural selection, as inferred from the ratio of non-synonymous nucleotide divergence ($d_N$) to synonymous nucleotide divergence ($d_S$). That is, the more a mutation changes protein function, the more likely it is to be either selected against or selectively favored, and because the $d_N/d_S$ ratio is a measure of natural selection, this ratio can be used to predict the degree of functional divergence ($d_F$). However, these hypotheses have rarely been experimentally tested.

Results: I present a novel method to address this issue, and demonstrate that divergence in bacteria-killing activity among animal antimicrobial peptides is positively correlated with the log of the $d_N/d_S$ ratio. The primary cause of this pattern appears to be that positively selected substitutions change protein function more than neutral substitutions do. Thus, the $d_N/d_S$ ratio is an accurate estimator of adaptive functional divergence.

Contact: tennessj@science.oregonstate.edu

Supplementary information: Supplementary data, including GenBank Accession numbers, are available at Bioinformatics online.

1 INTRODUCTION

Although the neutral theory of molecular evolution (Kimura, 1983) has been the standard null hypothesis in evolutionary genetics, it is clear that natural selection does have an effect on the evolution of many genes (Hey, 1999). Much current research seeks to infer either positive selection or negative selection from sequence data (Nielsen, 2005). Often, such studies test whether rates of non-synonymous substitution per non-synonymous site ($d_N$) differ from rates of synonymous substitution per synonymous site ($d_S$) (Yang and Bielawski, 2000). Synonymous substitution is thought to be largely neutral, while non-synonymous substitution is influenced by selection. Thus, when the $d_N/d_S$ ratio is high between genes, functional divergence is assumed to be high, due to positive selection and/or relaxed selective constraint (e.g. Hughes and Friedman, 2004; Wilkinson et al., 2005). When the $d_N/d_S$ ratio is low, the functional properties of the gene products involved are thought to be conserved, because selective constraint is high and there is little or no positive selection (e.g. Sehgal and Lovette, 2003; Wilkinson et al., 2005).

However, the relationship between selection, the $d_N/d_S$ ratio, and functional divergence ($d_F$) has rarely been experimentally studied, and a tight correlation might not necessarily exist. First, selection does influence $d_S$ via effects on translational efficiency and mRNA folding, and this influence might not be truly negligible compared to the influence of selection on $d_N$ (Chamary and Hurst, 2005; Hey, 1999; Tennessen, 2005a). Second, not all functional changes affect fitness, so $d_F$ can theoretically be non-zero even under purifying selection and a low $d_N/d_S$ ratio (Dykhuizen and Hartl, 1980; Kimura, 1983). Third, a non-synonymous change that alters function could be followed by a compensatory substitution that restores the original function, resulting in a low net $d_F$ despite a high $d_N/d_S$ ratio (DePristo et al., 2005). Fourth, amino acid residues essential for intramolecular structure, such as cystines, might remain highly conserved and decrease the $d_N/d_S$ ratio, even while positive selection acts on surface residues influencing charge, hydrophobicity, shape or other mediators of intermolecular interaction. Finally, non-synonymous substitutions might change function in very specific ways that cannot be detected with general in vitro assays, resulting in no apparent functional change. Although the functional effects of amino acid substitutions have been modeled (DePristo et al., 2005; Rastogi et al., 2006), and the functional effects of particular positively selected substitutions have been evaluated for individual proteins (Antcheva et al., 2004; Frentiu et al., 2007; Trabesigner-Ruef et al., 1996; Zhang et al., 2004), data on functional effects have never been integrated with statistical evidence for selection over a set of proteins to determine significant trends.

Genes encoding antimicrobial peptides (AMPs) are ideal for studying the relationship between nucleotide divergence and functional divergence (Tennessen, 2005b). These short, amphipathic, cationic molecules kill pathogens through a non-catalytic mechanism, either forming pores in the microbial membrane or carpeting the membrane until it is disrupted. Gene sequences of AMPs have been obtained from many taxa. Some AMPs clearly evolve adaptively, while in others evolution is selectively constrained (Tennessen, 2005b). Many studies have investigated the minimal concentrations necessary for different AMPs to kill various microbes, and results suggest remarkable functional diversity. Such extensive functional and genetic data, combined with substantial variation in selective pressures, provide an opportunity to examine the relationship between the $d_N/d_S$ ratio and functional change.

Here, I present a method to test whether the $d_N/d_S$ ratio is a predictor of functional divergence. If so, then phylogenetic
branches showing higher \( d_K/d_S \) ratios have been subject to more positive selection or less negative selection, and thus should show more divergence in function. I compare functional divergence with nucleotide divergence in animal AMP genes. By statistically incorporating functional data with sequence data, novel questions in molecular evolution can be addressed.

2 METHODS

I chose 71 animal AMPs from 26 families (Table 1; Supplementary Table 1; Supplementary Fig. 1). All of the AMPs within each family were identical in length and could be aligned without gaps. There is no evidence that all 71 AMPs are homologous, but there is clear homology within all families of AMPs. All comparisons are within families, not between, so homology among families is not relevant to this study. Comparisons included both orthologs and paralogs, which were treated equivalently because this analysis only depends on total genetic divergence and AMP function in vitro, not mode of evolutionary divergence or gene role in vivo.

I allowed AMPs to carry post-translational modifications, including amidation, cyclization and disulfide bonds, but only if all AMPs in a family carried the same modifications, or else if the modification were attached to an amino acid which differed within the family, such that it could be considered part of that amino acid difference. I did not include AMPs for which no nucleotide sequence was available. I only selected families for which antimicrobial activities of all AMPs against both the gram-negative bacterial species *Escherichia coli* and the gram-positive bacterial species *Staphylococcus aureus* had been reported in the same published study, apparently tested against the same strains under identical conditions. Strains and other test conditions were allowed to vary among the studies, though, including both radial diffusion assays and tests of bacterial growth in liquid media after incubation, as long as the study measured the lowest concentration necessary for antimicrobial activity, known as the minimal inhibitory concentration. Any family for which no activity could be demonstrated against either microbe by any AMP was excluded from analysis. I know of no other AMP families that fit my criteria.

I recorded the minimal inhibitory concentration of every AMP. When data were presented graphically, with concentration on the x-axis, I followed Lehrer et al. (1991) and estimated the minimal inhibitory concentration from the figure as the x-axis intercept. If an AMP had no demonstrated activity against a microbe at any concentration, I estimated minimal inhibitory concentration to be twice the highest concentration tested. I generated maximum likelihood phylogenies of AMP families using DNAML (Felsenstein, 1981). I used the Parsimony Ancestral States option in Mesquite (version 1.06; http://mesquiteproject.org) to estimate the ancestral minimal inhibitory concentrations at internal nodes in the phylogenies. In several cases, the predicted amino acid sequence at an internal node was identical to the sequence of an extant AMP (e.g. there were no non-synonymous substitutions along the branch leading to the AMP), and thus the minimal inhibitory concentration at the internal node could be accurately known. I selected 48 independent within-family comparisons between extant AMPs and/or internal nodes (Supplementary Fig. 1). No branch was included in more than one comparison. All comparisons included at least four nucleotide substitutions, because \( d_K/d_S \) cannot be estimated accurately if genetic divergence is small. Thus, comparisons were chosen to maximize the number of independent data points while combining adjacent branches into a single comparison when the number of substitutions was low.

For each comparison I measured the multiplicative difference in minimal inhibitory concentration, in micromolars, against both *E. coli* and *S.aureus* (Supplementary Table 1). I defined \( d_F \), or functional difference in antimicrobial activity, as the mean of the log multiplicative differences for these two bacterial species. For example, a 2-fold difference in minimal inhibitory concentration against both bacterial species would indicate \( d_F = 0.69 \). If the same AMPs were tested against multiple strains or under various conditions, I calculated the mean difference in the log minimal inhibitory concentration of the AMPs under all tests. The multiplicative difference in minimal inhibitory concentration was the ratio of the less effective AMP to the more effective AMP for each bacterial species separately. Thus, \( d_F \) is always positive and does not depend on which AMP is more effective or whether the same AMP is more effective against both microbes.

I used the codeml package in PAML (version 3.1a; http://abacus.genie.ucl.ac.uk/software/paml.html) to calculate \( d_K \) and \( d_S \) along the branch(es) of each AMP comparison. When calculating \( d_K/d_S \), I added 0.01 to both the numerator and the denominator, in order to avoid a denominator of zero when there are no synonymous substitutions. I tested whether \( d_K/d_S \) values were significantly different than those expected under neutrality. For each AMP comparison, I ran both a model which estimates \( d_K/d_S \) and a model which constrains \( d_K/d_S \) to be one along those branch(es), and I used the log likelihoods to obtain the \( P \)-value for the test that \( d_K/d_S \) equals one.

I predicted that \( d_F \) is correlated with \( \log(d_K/d_S) \), because when \( d_K/d_S \) is high there is less selective constraint and/or more positive selection. Using simple linear regression, I tested if \( \log(d_K/d_S) \) along each phylogenetic branch predicts \( d_F \) along that branch. To investigate whether positive or negative selection was causing this correlation, I also identified AMP comparisons which PAML indicated as having significantly high or low \( d_K/d_S \) values and used a t-test to evaluate
if \( d_F \) differed between these AMP comparisons and the more neutrally evolving AMP comparisons.

3 RESULTS AND DISCUSSION

Functional divergence \((d_F)\) is positively correlated with log\((d_N/d_S)\) \((R^2 = 0.26, P < 0.001; \) Fig. 1\). In contrast, \(d_F\) is not correlated with non-synonymous divergence \((d_N)\) alone \((P > 0.1)\). Only when both \(d_N\) and \(d_S\), and therefore the evidence for selection, are included is there a relationship. In 24 of the 48 comparisons mean \(d_N/d_S\) > 1, and in the remaining 24 comparisons mean \(d_N/d_S < 1\) (Supplementary Table 1). The \(d_N/d_S\) ratio ranged from 0.004 to 22.5 (logarithmic mean = 1.16). Non-synonymous divergence, \(d_N\), ranged from 0.03 to 0.64 (mean = 0.19). Functional divergence, \(d_F\), ranged from 0 to 2.92 (mean = 0.85). Neither \(d_N/d_S\), \(d_S\), nor \(d_F\) differed significantly between orthologs and paralogs \((P > 0.1)\).

Among the 24 AMP comparisons in which \(d_N/d_S < 1\), in 5 the neutrality test \(P\)-value is <0.05, more than would be expected by chance alone \((P < 0.001)\). Among the 24 AMP comparisons in which \(d_N/d_S > 1\), in 5 the neutrality test \(P\)-value is <0.05, more than would be expected by chance alone \((P < 0.001)\). Therefore, both positive and negative selection are acting among these AMP families. This conservative test likely underestimates the number of families under positive selection. More powerful tests, which allow \(d_N/d_S\) to vary among codons and/or examine more AMPs within each family, have revealed evidence of positive selection in several of the families I used that did not show \(d_N/d_S\) to be significantly greater than one in any of my comparisons (Boniotto et al., 2003b; Lee et al., 2005; Patil et al., 2004; Semple et al., 2006; Tennessen and Blouin, 2007).

Natural selection is the only explanation for a significant regression equation, but the selection could be positive, negative or both. That is, it could be that positively selected substitutions are associated with larger functional changes than neutral substitutions are, so branches where more positive selection has acted have both a higher \(d_N/d_S\) ratio and greater functional divergence. Alternatively, negative selection might purify mutations that affect function substantially, so high \(d_N/d_S\) and high \(d_F\) only occur under relaxed selective constraint, for example, when a gene copy not needed for survival rapidly accumulates non-synonymous substitutions that degrade its function. In the five positively selected AMP comparisons shown by PAML to have \(d_N/d_S\) ratios significantly greater than one, mean \(d_F = 1.69\) is significantly higher than mean \(d_F\) in the 38 more neutrally evolving AMP comparisons in which \(d_N/d_S\) did not significantly differ from one \((mean = 0.79; \) \(P < 0.001\)). There is no difference in \(d_N\) between these two groups \((P > 0.1)\). In contrast, in the five negatively selected AMP comparisons shown by PAML to have \(d_N/d_S\) ratios significantly lower than one, mean \(d_F = 0.48\) is not significantly different than mean \(d_F\) in the 38 AMP comparisons in which \(d_N/d_S\) did not significantly differ from one \((P > 0.1)\). This result suggests that positive selection, not negative selection, is primarily responsible for the correlation between \(d_F\) and \(d_N/d_S\).

My results indicate that the \(d_N/d_S\) ratio can be used to help predict functional divergence. They support the hypothesis that the \(d_N/d_S\) ratio is an accurate predictor of natural selection, and the hypothesis that the probability of natural selection acting on a mutation is correlated with the degree to which the mutation affects protein function. These results might not be surprising, but they nevertheless represent an important step towards an experimentally validated comprehensive theory of molecular evolution. Even if natural selection acts on synonymous sites in antimicrobial peptide genes (Tennen, 2005a), any effects of selection on \(d_S\) do not completely disrupt the signature of selection in the \(d_N/d_S\) ratio. In addition, even if some function-altering substitutions are neutral, natural selection is more likely to act on mutations with greater functional impact. Furthermore, although compensatory substitutions might make \(d_F\) lower than the additive functional change conveyed by all non-synonymous substitutions (DePristo et al., 2005), net \(d_F\) is still greater under positive selection than under neutrality.

Finally, although most of these AMPs have probably evolved to target pathogens other than the human microbes \(E.coli\) and \(S.aureus\), and thus many of the functional changes induced by adaptive non-synonymous substitutions would be more accurately measured with assays using different microbes, clearly \(E.coli\) and \(S.aureus\) are acceptable proxies for measuring \(d_F\). Indeed, these results indicate that AMPs which have adaptively diverged due to coevolution with one type of pathogen are more likely to differ in their ability to kill other, novel pathogens that an organism might encounter. Overall, the use of the \(d_N/d_S\) ratio is justified for inferring the action of natural selection on genes, and thus the degree of functional divergence.

Remarkably, the correlation is observed despite multiple sources of noise in the data. The peptides in this study were tested under different conditions, antimicrobial activity was measured in slightly different ways, antimicrobial activity was sometimes roughly estimated from graphical displays,
minimal inhibitory concentration had to be estimated at internal nodes, and the formula for the calculation of \( d_{S} \) was somewhat arbitrarily chosen. The relationship between \( d_{S} \) and \( d_{N}/d_{S} \) is strong enough to be detectable in spite of all of this variation. As research in molecular biology proceeds, it is likely that cleaner, more extensive datasets will soon appear. Ideally, future applications of this method will provide additional conditions for all measurements of function, or its analogs such as in vitro activity or protein structure. This method should also be applied to more datasets in which \( d_{N}/d_{S} \) is consistently less than one, to test whether variation in negative selection alone can cause a correlation between \( d_{N}/d_{S} \) and functional divergence. This method might have practical uses as well, for proteins such as antimicrobial peptides where the identification of diverse functional variants could lead to therapeutic applications. For example, if multiple homologous gene sequences are obtained from one or more species, a frequent practice (e.g. Boniotto et al., 2003a,b; Lai et al., 2002; Patrzykat et al., 2003), the two sharing the highest pairwise \( d_{N}/d_{S} \) ratio are likely to code for the most functionally divergent peptides, and these could by synthesized and studied.

In conclusion, functional divergence among antimicrobial peptides is positively influenced by \( d_{N}/d_{S} \) and therefore by natural selection. Evidence suggests that this is at least partly due to increased positive selection with higher \( d_{N}/d_{S} \), not merely relaxed functional constraint. The method presented here, or variations of it, can and should be applied to many different types of genes to thoroughly investigate the relationship between natural selection and the evolution of protein function.

ACKNOWLEDGEMENTS

The author would like to thank his advisor, M.S. Blouin; and C.D. Criscione, H. Araki, K. Georgitis and A. Hughes for conceptual and statistical help.

Funding: National Science Foundation Graduate Research Fellowship.

Conflict of Interest: none declared.

REFERENCES


