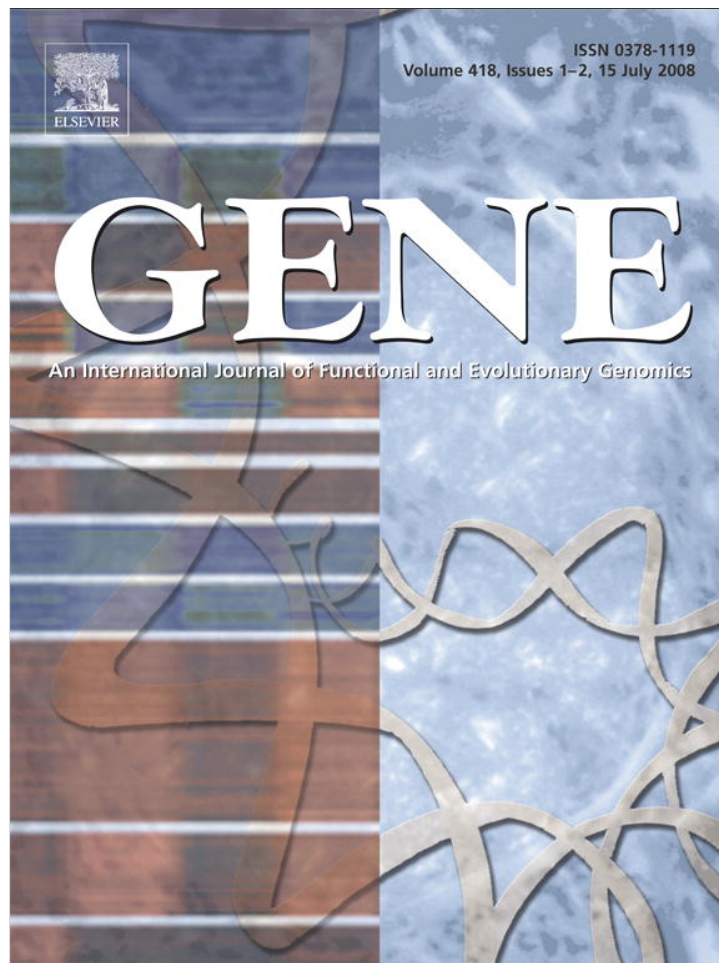


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Gene

journal homepage: [www.elsevier.com/locate/gene](http://www.elsevier.com/locate/gene)

## Characterizing positive and negative selection and their phylogenetic effects

Steven E. Massey<sup>a</sup>, Alexander Churbanov<sup>a,1</sup>, Shruti Rastogi<sup>a,b,2</sup>, David A. Liberles<sup>a,\*</sup>

<sup>a</sup> Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA

<sup>b</sup> Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway

### ARTICLE INFO

#### Article history:

Received 8 October 2007

Received in revised form 27 February 2008

Accepted 27 March 2008

Available online 8 April 2008

#### Keywords:

Systematics

Evolution

Phylogeny

Function

Structural constraint

Selection

Neutrality; Nonsynonymous substitution

Synonymous substitution

### ABSTRACT

Total evidence and the use of large datasets to overcome uncertainty are the state of the art in systematic analysis. This assumes that the only true phylogenetic signal is ancestry and that functional, structural, and other factors will not add an alternative signal. Using gene families, where individual codon positions were sorted into bins based upon average-pairwise dN/dS ratio, we show that standard, common phylogenetic methods that were designed for stochastic, neutral, site-independent processes, generate less robust phylogenetic signal for bins with strong negative or positive selection. This was true for phylogenetic reconstruction with parsimony, distance, and likelihood methods. Further, we present a case for the potential existence of systematic functional or structural signal that competes with ancestral signal. For the example of positive selection, we simulate the evolution of sequences through three dimensional lattice constructs with folding constraint and changing binding functionality and show that total evidence for these lattice genes presents trees with functional signal, but that the neutral synonymous sites in these genes show the true ancestral signal. In this case, sequence convergence is promoted by functional convergence.

© 2008 Elsevier B.V. All rights reserved.

### 1. Introduction

The recovery of species relationships from DNA sequence data is a common undertaking in modern systematic biology. A number of problems are known to complicate this undertaking, including the existence of gene duplicates and multiple speciation events over short evolutionary periods leading to lineage sorting. However, it is generally assumed that total evidence (including total molecular evidence) will recapitulate phylogenetic relationships in that a majority of the signal will automatically reflect ancestry. This has spurred phylogeny assessment using concatenated genome sequences to generate large datasets (Rokas et al., 2003; see also McInerney, 2006).

However, standard methods such as parsimony as well as models for distance and maximum likelihood are typically based upon assumptions including independent evolution at each site and the occurrence of stochastic (neutral) evolutionary processes. Further, it is assumed that

ancestry is the only true signal in gene families and that homoplasy is simply noise (De Pinna, 1991) rather than systematic in nature with the potential to dominate the ancestral signal. However, different selective pressures may have the ability to generate phylogenetic signal that is different from ancestry.

Positive diversifying selection is known to occur in a lineage-specific manner across gene families (Liberles et al., 2001; Roth and Liberles, 2006). Positive selection has two possible modes of action on phylogenies. One is an increase in rate causing long branch attraction (Philippe et al., 2000). The second is generating convergence or parallel evolution (homoplasy) through similar selective pressures. While it might be argued that there are multiple paths in response to any given selective pressure, recent work has suggested that such paths are more limited in number (Weinreich et al., 2006), making this hypothesis also plausible. Increasingly, it has been realized that the first mechanism is a problem, and masking of fast evolving sites has been suggested as a solution (Philippe et al., 2000; Townsend, 2007).

Negative selection also has two potential modes of action. The first mode is the inverse of positive selection, where a small number of changes provide little phylogenetic signal (Townsend, 2007). However, there is a second mode that is similar to the potential systematic problem with positive selection. Negative selection in protein coding genes is mediated by structural and functional constraint. Structural constraint violates site-independence assumptions, where deleterious changes can place selective pressures for compensatory changes (or reversion). When changes are observed, they are likely to be through a small number of

*Abbreviations:* dN, the non-synonymous substitution rate; dS, the synonymous substitution rate; GLP-1, Glucagon-Like Peptide-1; NCBI, National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>), houses a reference taxonomy; TAED, The Adaptive Evolution Database.

\* Corresponding author.

E-mail address: [liberles@uwyo.edu](mailto:liberles@uwyo.edu) (D.A. Liberles).

<sup>1</sup> Current address: Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA.

<sup>2</sup> Current address: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA.

compensatory paths. This small number of paths can lead to a systematic signal that is different from ancestry and would therefore also appear as homoplasy. In a simulation study using a heterotachy model that is site-independent but expected to behave similarly to these structural effects, systematic artifactual phylogenies were generated (Ruano-Rubio and Fares, 2007). Further, for both the positive and negative selection homoplasy mechanisms, it is expected that linkage effects will magnify the phylogeny estimation problem by compounding it with rate variation.

Given that such effects are possible, that phylogenies can show systematic signal for function (under positive and negative selection) and for ancestry (under neutrality), we sought to test what selective pressures showed the truest phylogenetic signals. Secondly, while the proposed systematic bias for negative selection has been shown in simulation to be possible (Ruano-Rubio and Fares, 2007), we sought to demonstrate that the proposed systematic bias for positive selection was also possible using a physical model where positive selection was driven through changes in binding interactions.

## 2. Methods

### 2.1. Evaluation of codon-specific selective pressures in a non-phylogenetic manner

Chordate gene families with multiple sequence alignments were taken from The Adaptive Evolution Database (TAED) (Roth et al., 2005). These multiple sequence alignments are available for download at <http://www.bioinfo.no/tools/TAED>. Previously, methods have been established for evaluating the ratio of non-synonymous to synonymous nucleotide substitution rates (dN/dS) at single sites (Suzuki and Gojobori, 1999). Unfortunately, such methods are phylogenetic in nature. Because the ultimate goal is to assess phylogeny to prevent any bias imposed by phylogenetic inference of such values, the options were the computationally-intensive iterative application of phylogeny-dependent methods with our analysis or the use of non-phylogenetic pairwise methodology. Using a classic counting method, the pairwise average of codon position-dependent dN/dS was calculated using the Nei and Gojobori (1986) method. This method will result in the over-counting of substitutions, but is not expected to bias the placement of codons into bins with a reliable ordering of selective pressures based upon dN/dS. With each codon in each gene assigned a dN/dS ratio based upon the pairwise average, the codons were assigned to a bin (0.0–0.2, 0.2–0.6, 0.6–0.9, 0.9–1.1, 1.1–2.0, >2.0). The codons in each bin were concatenated and multiple sequence alignments where each bin contained at least 10 codons were used to construct phylogenies. Bins with progressively higher dN/dS ratios were typically less populated.

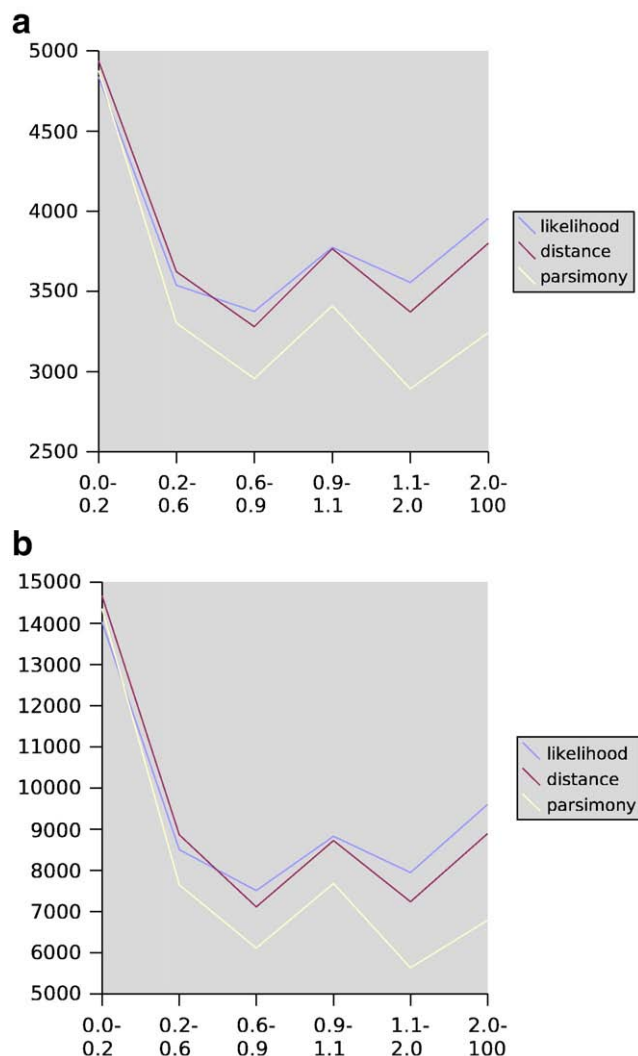
The program used for the pairwise average dN/dS ratios by codon position over a multiple sequence alignment is available for download from <http://www.wyomingbioinformatics.org/~achurban/>.

### 2.2. Phylogeny estimation

Phylogenies were constructed using programs in the Phylip package (Felsenstein, 1989). The programs protml, protpars, and protdist were used, with neighbor joining trees constructed in the distance method. From each gene independently, trees were constructed for each bin using all three methods. Reconstruction was performed at the protein level to avoid the confounding effects of saturation at the DNA level, a potential problem with the Chordate dataset.

### 2.3. Evaluating phylogenetic signal fidelity using gene tree/species tree reconciliation

Given gene bin trees and the reference species relationships from NCBI (Wheeler et al., 2007), the inferred number of gene duplication and loss events using soft parsimony as well as the most parsimo-



**Fig. 1.** While gene duplication and loss occur at appreciable rates, these processes are not expected to add systematic signal and a phylogeny reconstruction with fewer inferred duplication and loss events is expected to be a better estimate of evolutionary history than one with more duplication and loss events. In this figure, multiple sequence alignments from gene families were subdivided into sets of codons evolving under different selective pressures and gene trees were independently calculated for all gene families for all bins. The total number of inferred duplication and loss events for each bin as a sum across gene families was calculated and is presented comparatively across bins. a. The total number of inferred duplication events from Softparsmap (Berglund-Sonnhammer et al., 2006) for all dN/dS bins across gene families using various phylogeny reconstruction methods is shown. b. The total number of inferred gene loss events from Softparsmap (Berglund-Sonnhammer et al., 2006) for all dN/dS bins across gene families using various phylogeny reconstruction methods is shown. The general trend observed is that bins showing dN/dS ratios around neutrality (0.6–2.0) have the truest ancestral signal. The dN/dS ratios reported reflect an average of pairwise values from a multiple sequence alignment.

nious root from this characterization were determined using Softparsmap (Berglund-Sonnhammer et al., 2006). This data was plotted to generate Fig. 1.

The NCBI reference species tree used for this analysis can be downloaded from <ftp://ftp.ncbi.nih.gov/pub/taxonomy/>.

### 2.4. Simulating the evolution of proteins under negative and positive selection

For this study, we have constructed DNA genes encoding a protein folding into a  $4 \times 4 \times 4$  cubic lattice, where each fold is a self-avoiding walk inside a fixed cubic lattice. The proteins used were 64 codons

long with a predefined active site for the binding of a ligand of 7 amino acids. Ligand binding sites were adjacent to the active site in a fixed orientation. An A ligand initially bound stably at the beginning of the simulation, while a B ligand did not. The folding energy of the molecule was calculated using the formula,

$$E = \sum_{i \neq j} g(A_i, A_j) U_{ij}$$

where  $\gamma(A_i, A_j)$  is the contact potential between residue type  $A_i$  at position  $i$  and residue type  $A_j$  at position  $j$ , and  $U_{ij}$  is equal to one if residues  $i$  and  $j$  are not adjacent in sequence but are on adjacent lattice sites, and zero otherwise. The value of  $\gamma(A_i, A_j)$  was obtained from the symmetric interaction matrix given by Miyazawa and Jernigan (1985). Binding energies were also calculated in a similar way, as previously described (Williams et al., 2001; Rastogi and Liberles, 2005; Rastogi et al., 2006). The molecules with folding energy  $< -kT$  and binding energy  $< -0.25kT$  were considered stably folded and functional.

The lattices were evolved under constant populations of size 1000 and a mutation rate of  $10^{-4}$  per site per generation. Under the branches with positive selection, individuals binding the new ligands had a selective advantage of 5% for population in the next generation. In every generation the individuals were picked randomly provided they folded stably and functioned. Simulations were replicated 10 times with different seed lattice and ligand sequences.

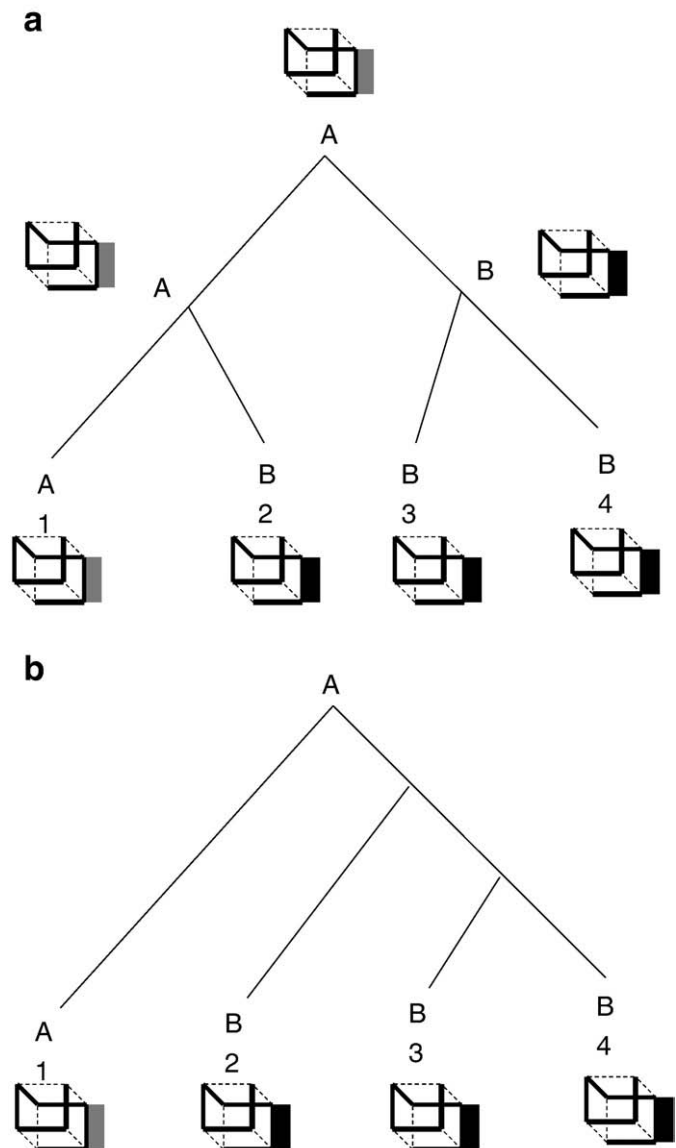
### 2.5. Testing the tree topology of simulated proteins

Sequences obtained from the simulation were phylogenetically reconstructed using the same programs in the Phylip package (Felsenstein, 1989) as above. Because the bins of low non-zero dN/dS were not populated in this simulation due to the small number of sequences and the drive for positive selection along two independent branches, trees were also obtained for synonymous substitutions as a control.

## 3. Results and discussion

In this paper, we sought to test the assumption of total evidence that the majority of signal from a multiple sequence alignment will generate the true ancestry. With the hypothesis that both negative selection and positive selection can generate a systematic signal that is non-ancestral in nature, we sought to characterize the phylogenetic signal generated by negatively selected, neutrally evolving, and positively selected sites. We subsequently directly test the total evidence assumption with simulated data involving convergent positive selection on different lineages.

The underlying ancestral signal in a gene tree is driven by the species process, as reflected in the species tree. One problem with the accepted species tree in evaluating gene trees is that processes like gene duplication will generate a true signal that is not compatible with the species tree (see for example, Scannell et al., 2007). However, while chordate gene families at close distance, will have been subjected to lineage-specific duplication and a discrete number of whole genome duplication events, the majority of signal in such gene families is from speciation events (Hughes and Liberles, in press). Further, the vast weight of whole genome duplication in chordates was very ancient and per duplication retention rates for smaller scale duplication events are expected to be much smaller (Maere et al., 2005). Therefore, while the true number of duplication events is significantly above zero, it is expected that errors in topology will increase the measured numbers of duplication and loss events above the true number. Therefore, minimizing the count of duplication and loss events in the context of gene tree/species tree reconciliation is expected to yield a measure of gene tree accuracy. This is used as a metric to evaluate the robustness of gene tree



**Fig. 2.** A. Sequences folded in a lattice model which were selected to maintain the fold and to bind to a peptide were simulated along this tree. Two independent branches had positive selective pressure to switch the peptide bound from A (grey) to B (black). B. Phylogenetic reconstruction of the tree using distance, parsimony, and likelihood methods all generated a tree where sequences clustered by function rather than ancestry. A tree based solely upon synonymous substitutions recovered the ancestral tree, shown in A.

signal that is independent of the sequence evolution data from which the gene trees were generated.

Multiple sequence alignments from TAED Chordate gene families (Roth et al., 2005) were subjected to codon-specific average-pairwise (non-phylogenetic to avoid bias) calculation of the ratio of non-synonymous to synonymous nucleotide substitution rate ratios (dN/dS; a measure of selection). Codons from each multiple sequence alignment were concatenated and used for phylogeny estimation. Using the counts of duplication and loss from reconciliation for trees reconstructed using likelihood, distance, and parsimony at the protein level for each bin for each multiple sequence alignment, the signal with the fewest duplication and fewest loss events (inferred to be the truest phylogenetic signal) was obtained for the most neutrally evolving sites (Fig. 1). This is perhaps not surprising given that the phylogenetic models applied were designed for stochastic neutral-type changes. These are the most common phylogenetic models.

Those bins under stronger negative and positive selection showed less robust ancestral signal under the minimized duplication and loss event metric. The loss of ancestral signal was stronger for negative than positive selection and the mode of loss (small numbers of substitutions vs. homoplasy) could not be readily established from the dataset. Both mechanisms (Ruano-Rubio and Fares, 2007; Townsend, 2007) have been established in the literature for negative selection. For positive selection, only the rate increase mechanism has been well established (Philippe et al., 2000; Townsend, 2007).

Next, we sought to establish that homoplasy driven by positive selection could establish this trend. While it is well known that homoplasy can confound phylogenies, it has not been established that this can be systematically caused by positive selection. This depends upon the available mutational paths to generate a new function from a given starting point and therefore on the ability to generate simulated sequences that are neither biased in showing homoplasy nor random in their mapping between function and sequence. To achieve that, protein-encoding DNA sequences with a defined fold in a three dimensional lattice model and a defined binding function that could be changed were simulated based upon established methodology (Williams et al., 2001; Rastogi and Liberles, 2005; Rastogi et al., 2006). Here lattices that initially bound to a peptide ligand designated A (Fig. 2) were subjected on independent lineages to positive selective pressure to alternatively bind to a ligand designated B in a population genetic model.

As observed in the trees recapitulated from the amino acid sequences (this result was repeated and was observed in approximately 50% of simulations under the specified conditions using distance, likelihood, and parsimony (5/10 simulations)), the sequences with similar evolved binding functions group together. Therefore, independent bursts of positive selection generated similar molecular solutions and this functional signal systematically swamped the ancestral signal in phylogenetic analysis. While it may be argued that lattice models are smaller than most domains and this restricts the number of possible solutions to a given functional problem relative to real proteins, similar restrictions have been observed in the natural evolution of real proteins (Weinreich et al., 2006). Similar anecdotal observations of phylogenetic placement potentially driven by function have also been reported for GLP-1 (Skovgaard et al., 2006) and opsin proteins (Taylor et al., 2005). Lastly, it should be noted that the observed evolutionarily accessible solutions in the lattice are not the only theoretically possible solutions to the selection criteria.

From the gene trees showing functional signal in Fig. 2, the small number of sequences (4) precluded the use of the protein-level site-stripping method. However, construction of the tree at the DNA level using synonymous substitutions recapitulated the true ancestral signal.

So, given the observations of this study, how should one proceed with phylogenetic analysis? A flowchart summarizing our recommendations is presented in Fig. 3. For positive selection, filtering of the fastest sites has already been proposed (Philippe et al., 2000). It may be that positive selection does not always correlate with the globally fastest evolving sites and more sophisticated approaches (e.g. Galtier, 2001; Rodrigue et al., 2006) may be needed to detect these. At the negative selection level, the obvious solution is to also filter the slowest evolving sites (Townsend, 2007). Again, if the problem is driven by structural constraint, models that are robust to protein structure (Parisi and Echave, 2001; Kleinman et al., 2006) or at least that relax assumptions of site-independence (Stern and Pupko, 2006) would be desirable. One limitation of such models is that the relationship between thermodynamics and fitness is controversial (Taverna and Goldstein, 2002; DePristo et al., 2005; Sasidharan and Chothia, 2007), but needs to be built into the phylogenetic model. The observations here should be of concern to the systematics community, but improved model development is currently an exciting field and promises to continue to be one into the future.

#### 4. Conclusions

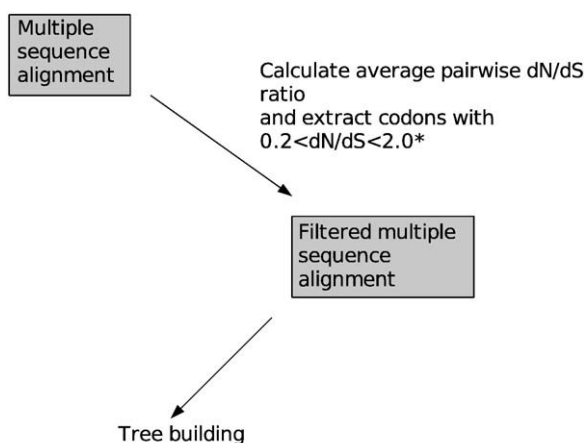
1. Strong negative and positive selection can lead to gene tree phylogenies that infer an increased number of duplication and loss events compared to those generated from more neutral characters.
2. Using a lattice model construct to simulate sequence evolution, we have shown that parallel convergent evolution of binding functionality can result in gene trees that show clustering based upon binding function rather than ancestry. We further show that the neutral synonymous site signal in these simulated genes retains the signal from ancestry. The sets of conditions under which this functional clustering will occur will depend upon factors like effective population size governing the ratio of selected versus neutral changes as well as factors from the protein fold like the ratio of the size of the protein to the size of the functional patch and the fraction of shell residues needed for proper folding and orientation of the binding/functional region.

#### Acknowledgments

Christian Roth assisted with the generation of early versions of simulated data working closely with Shruti Rastogi. Funding for this project was provided by an INBRE (NIH) grant to University of Wyoming and by FUGE, the Norwegian Functional Genomics Research Platform.

#### References

- Berglund-Sonnhammer, A.C., Steffansson, P., Betts, M.J., Liberles, D.A., 2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol.* 63, 240–250.
- De Pinna, M.C.C., 1991. Concepts and test of homology in the cladistic paradigm. *Cladistics* 7, 367–394.
- DePristo, M.A., Weinreich, D.M., Hartl, D.L., 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* 6, 678–687.
- Felsenstein, J., 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164–166.
- Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* 18, 866–873.
- Hughes, T., Liberles, D.A., in press. A characterisation of the effects of speciation and whole genome duplication on the distribution of gene family sizes and its application to detecting large scale duplication. *J. Mol. Evol.*
- Kleinman, C.L., Rodrigue, N., Bonnard, C., Philippe, H., Lartillot, N., 2006. A maximum likelihood framework for protein design. *BMC Bioinformatics* 7, 326.
- Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., Benner, S.A., 2001. The Adaptive Evolution Database (TAED). *Genome Biology* 2(8) research0028.1-0028.6.



\*An alternative to this site stripping procedure for negatively selected residues is to use models that relax the site-independence assumption or that use protein tertiary structure.

**Fig. 3.** A flow diagram representing an alternative direction forward based upon the results of this study is shown. An alternative to site stripping is the ultimate development of better, more realistic evolutionary models that are motivated by protein biochemistry.

- Maere, S., et al., 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci., USA* 102, 5454–5459.
- McInerney, J.O., 2006. On the desirability of models for inferring genome phylogenies. *Trends Microbiol.* 14, 1–2.
- Miyazawa, S., Jernigan, R.L., 1985. Estimation of effective inter-residue contact energies from protein crystal structures – quasi-chemical approximation. *Macromol.* 18, 534–552.
- Nei, M., Gojobori, T., 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.
- Parisi, G., Echave, J., 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.* 18, 750–756.
- Philippe, H., et al., 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond. B* 267, 1213–1221.
- Rastogi, S., Liberles, D.A., 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* 5, 28.
- Rastogi, S., Reuter, N., Liberles, D.A., 2006. Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys. Chem.* 124, 134–144.
- Rodrigue, N., Philippe, H., Lartillot, N., 2006. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol. Biol. Evol.* 23, 1762–1775.
- Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804.
- Roth, C., Liberles, D.A., 2006. A systematic search for positive selection in higher plants (embryophytes). *BMC Plant Biol.* 6, 12.
- Roth, C., Betts, M.J., Steffansson, P., Sælensminde, G., Liberles, D.A., 2005. The Adaptive Evolution Database (TAED): a phylogeny-based tool for comparative genomics. *Nucleic Acids Res.* 33, D495–D497.
- Ruano-Rubio, V., Fares, M.A., 2007. Artfactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. *Syst. Biol.* 56, 68–82.
- Sasidharan, R., Chothia, C., 2007. The selection of acceptable protein mutations. *Proc. Natl. Acad. Sci., USA* 104, 10080–10085.
- Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M., Wolfe, K.H., 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci., USA* 104, 8397–8402.
- Skovgaard, M., Kodra, J.T., Gram, D.X., Knudsen, S.M., Madsen, D., Liberles, D.A., 2006. Using evolutionary information and ancestral sequences to understand the sequence-function relationship in GLP-1 agonists. *J. Mol. Biol.* 363, 977–988.
- Stern, A., Pupko, T., 2006. An evolutionary space-time model with varying among-site dependencies. *Mol. Biol. Evol.* 23, 392–400.
- Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328.
- Taverna, D.M., Goldstein, R.A., 2002. Why are proteins marginally stable? *Proteins* 46, 105–109.
- Taylor, S.D., de la Cruz, K.D., Porter, M.L., Whiting, M.F., 2005. Characterization of the long-wavelength opsin from Mecoptera and Siphonaptera: does a flea see? *Mol. Biol. Evol.* 22, 1165–1174.
- Townsend, J.P., 2007. Profiling phylogenetic informativeness. *Syst. Biol.* 56, 222–231.
- Weinreich, D.M., Delaney, D.F., DePristo, M.A., Hartl, D.L., 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312, 111–114.
- Wheeler, D.L., et al., 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35, D5–D12.
- Williams, P.D., Pollock, D.D., Golstein, R.A., 2001. Evolution of functionality in lattice proteins. *J. Mol. Graph. Model.* 19, 150–156.