

Assessing and Improving the Accuracy of Detecting Protein Adaptation with the TreeSAAP Analytical Software

David A. McClellan^{*‡} and David D. Ellison[†]

^{*}Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine 04575, USA

[†]Department of Biomedical Engineering

Johns Hopkins University, Baltimore, Maryland 21205, USA

[‡]Contact author: dmcclellan@bigelow.org

Abstract—The TreeSAAP software has been successfully used in a variety of protein studies for identifying and characterizing adaptation in terms of shifts in the physicochemical properties of amino acid replacements. It differentiates adaptive replacements from those that may have resulted from random mutation. The accuracy of TreeSAAP was tested using simulated protein-coding DNA data that was randomly generated using a bifurcating phylogeny to reflect a random pattern of mutation constrained only by the structure of the genetic code. A sampling of 1402 simulated amino acid replacements resulted in a default accuracy of 80.6%. More than 50% of the false-positive results were traced to just 11 of the possible single-step amino acid exchanges, each of which exhibited less than 50% accuracy. When these 11 exchanges are eliminated from the subsequent analysis, the accuracy of TreeSAAP is increased to nearly 90%. Further testing of this modified approach for adverse implications with empirical data is warranted.

Index Terms—Simulated protein-coding DNA sequences, TreeSAAP, analytical results, accuracy.

I. INTRODUCTION

Numerous methods for detecting natural selection acting on proteins have been proposed [22], [23], [24], [25]. The most common approach for detecting selection in protein-coding genes involves estimating rates of nonsynonymous (dN) and synonymous (dS) nucleotide substitution and calculating the rate ratio dN/dS . Neutral theory [9] predicts that these rates will be equal (i.e., $dN/dS = 1$) when unaffected by selection. Alternatively, $dN/dS > 1$ is indicative of positive selection, while $dN/dS < 1$ denotes purifying or negative selection. Although these conditions have resulted in accurate detection of different types of selection, it often fails to detect adaptation when it clearly has occurred [3], [19], and the accurate diagnosis of individual amino acid changes as adaptive or otherwise is mathematically impossible because multiple nonsynonymous substitutions are required to satisfy the condition $dN > dS$ in a statistically significant manner.

A new family of methods recently has been proposed that approaches the question of amino acid replacement diagnosis using changes in physicochemical amino acid properties [6], [21], [11]. The use of amino acid properties results in a high degree of resolution that is more sensitive than the dN/dS approach [12], [16], and allows for the detailed diagnosis of

individual amino acid changes in the context of the structure and function of proteins [1]. This alternative approach for detecting selection also exhibits great promise as a tool in the area of biomedical research [4], [14], [10].

TreeSAAP is a software program for detecting selection in protein-coding genes using physicochemical amino acid properties [20]. It has a demonstrated utility for characterizing the causes of selection in a wide variety of contexts. Briefly, it uses a multiple protein-coding DNA sequence alignment and a user-provided phylogenetic topology to estimate ancestral character-states. Substitution events are then inferred on each phylogenetic branch by directly comparing the hypothetical ancestral sequence at the basal end with the sequence at the terminal end. The ratio of inferred amino acid replacements to evolutionary pathways [11] is then calculated and used as a null hypothesis for the rate of random change for the sequence alignment. The overall pattern of inferred substitution is then compared to the distribution estimated using the null hypothesis to produce a series of z -scores to determine if certain magnitudes of change deviate from neutral expectations. When significantly more substitutions in higher magnitude categories are inferred than are expected by chance, that property is said to be adapting. A sliding window may be implemented to test the hypothesis that significant changes are localized in a particular part of the protein. Adaptive changes may then be mapped onto three-dimensional structures to qualitatively assess the degree to which adaptive changes are associated with functional domains and motifs [16], [14], [10].

Several studies have implemented both dN/dS and TreeSAAP approaches for studying positive selection. Examples include studies of cetacean cytochrome *b* [12], crustacean visual pigments [16], and eutherian mammal HADH2 [10]. In each case, more positively selected sites were identified with TreeSAAP than with dN/dS . Moreover, TreeSAAP results were largely consistent with the known biology of the proteins and served to enhance each study by providing information about the possible causes of selection. However, these results call into question the accuracy of TreeSAAP results; is it finding selection where there has not been any?

Although TreeSAAP has demonstrated a great deal of

potential in several areas of molecular research, it has yet to be rigorously tested to determine how well it fulfills its stated purpose. The first step in this process is to determine a level of confidence in analytical results by assessing the degree to which it may produce false-positive results. We outline herein a strategy for evaluating and modifying the TreeSAAP analysis that will maximize resultant confidence levels.

II. MATERIALS AND METHODS

A. Data Simulation

We designed a new software program, SimSeq, for simulating realistic protein-coding DNA sequence alignments that have evolved under conditions of neutrality. First, it randomly generates a protein-coding DNA sequence of a given length that is consistent with a particular genetic code. It generates a user-defined number of amino acid changes using a variety of models and parameters. The user defines a value for κ (transition-to-transversion ratio) and ω (nonsynonymous-to-synonymous rate ratio) by which to model the probability of fixation within a user-defined population size. A user tree is used to “speciate” the sequences after a certain number of fixed mutations until all the tree branches are “filled” with substitutions. The resulting sequences are output in standard NEXUS format ready for subsequent analysis.

SimSeq was used to generate 16 aligned sequences 45,000 nucleotides in length using the universal genetic code and the following parameters: 3000 nonsynonymous substitutions; bifurcating user tree; $\kappa = 0.5$; $\omega = 1.0$; population of 10,000 individuals. These parameters were chosen to simulate neutral conditions and avoid detrimental effects of multiple hits. If TreeSAAP was completely accurate (i.e., it generates no false-positive results), it should not detect any positive selection in any of the physicochemical amino acid properties currently implemented in the software.

The randomness of the distribution of the simulated nonsynonymous substitutions was verified by testing the goodness of fit between a Poisson distribution and the distribution of the number of substitutions per 10-codon partition. This test begins with the calculation of an expected distribution for the number of times 0–6 substitutions are predicted to occur within a 10-codon partition across a sampling of the first 5000 codons of the simulated data. The critical value for the test statistic was estimated with a chi-square distribution.

B. Physicochemical Adaptation Analysis

TreeSAAP version 3.2 [20] was used to evaluate the sequences generated by SimSeq. All 31 default amino acid properties were used. These properties cover a broad spectrum of physical, chemical, energetic, and conformational characteristics of amino acids [17]. The bifurcating tree used to generate the sequences in SimSeq also was used in TreeSAAP. Dynamic estimation of κ , a fixed value of $\alpha = 0.0$, and general time-reversible model were used for ancestral sequence reconstruction. Three magnitude categories — conservative, moderate, and radical — were used to model the degree to which substitutions were destabilizing. Finally, a sliding-window of

11 codons was used to evaluate clustering among significant amino acid sites.

Further analysis was performed using the results of the TreeSAAP analysis. False-positive results were tallied by amino acid exchange and property to determine if certain properties generate more false results, or if certain exchanges are more likely to be falsely positive. The correlation between properties that generated similar false results was also evaluated.

C. Sliding-window and Magnitude Categories

The effect of alternative partitioning was evaluated by first simulating a smaller data set of 15,000 nucleotides and 1000 random nonsynonymous nucleotide substitutions, but otherwise the same as described above for the more general analysis.

A sliding-window analysis is often implemented in TreeSAAP to assess the degree to which significant results are locally clustered within the context of the codons within the window [16], [14], [10]. Often even single nonsynonymous nucleotide substitutions are found to be significant as the result of this type of analysis. Sliding-window (SW) sizes of 9, 11, 15, and 19 codons were used to evaluate the effect window size has on the generation of false-positive results.

The number of magnitude categories represents the number of times the distribution of single-step changes for a given amino acid property is subdivided for an analysis. The simulated data was analyzed using 3, 5, and 7 magnitude categories to evaluate the biases inherent to these alternative portioning strategies relative to the generation of false-positive results.

The expected outcome for both portioning strategies was an absence of significant results. The model used to simulate the evolution of the sequences randomly selects both the site and outcome of “mutation” under the constraint of the universal genetic code, so significant results do not represent selection, but the propensity a particular physicochemical property may have for generating false-positive results given the model and assumptions underlying TreeSAAP. The occurrence of false-positive results were tabulated under each sliding-window size and compared for each of the 31 properties currently included in TreeSAAP. Collective tabulations were also computed.

D. Calculation of Accuracy

Accuracy, a , was calculated with the following:

$$a = 1 - \frac{f}{s} \quad (1)$$

where f is the number of false-positive results and s is the total number of amino acid replacements being considered.

E. Correcting for Multiple Sampling

A Bonferroni statistical correction for multiple sampling is strongly suggested when employing a sliding-window analysis in TreeSAAP. However, whether or not such a correction will yield an appropriate level of confidence has not been determined. Accuracies were estimated using several p -value

thresholds to determine the correction most likely to yield the standard 95% confidence.

III. RESULTS

A. Overall Accuracy & Description of False-positive Results

A data set of 45,000 nucleotides (15,000 codons) was simulated as described. TreeSAAP recovered 2114, with 399 displaying false indications of destabilizing selection when $SW = 11$. This represents a initial accuracy of 81.1%. The discrepancy between the number of substitutions generated by SimSeq (3000) and those recovered by TreeSAAP (2114) is likely the result of multiple hits, which is dependant upon the stochastic model of SimSeq, and fallacies associated with ancestral character-state reconstruction, which is based on the general time-reversible likelihood model of molecular evolution.

Interestingly, all of the false-positive results generated by the general analysis of these simulated data would have been diagnosed as positive selection. No “negative selection” was ever detected regardless of the physicochemical property evaluated. The most negative z -score among a sampling of nearly 15,000 window analyses at 11 codons for each of the 31 properties was just -1.987 (*Bulkiness*), which would be significant at $p = 0.05$, but not after correcting for multiple sampling. In contrast, the most positive z -score for the same sampling was 8.508 (*Molecular volume*, *Molecular weight*, and *Partial specific volume*), which is highly significant. There was no significant correlation between the most positive z -score and the most negative z -score for each property (slope = -0.0097; $R^2 = 0.0206$), meaning that properties that perform poorly in the detection of positive selection may still accurately detect negative selection.

Most of the 31 properties were more accurate than the overall average, while a few faired deplorably (Table I). Eleven properties had accuracies less than the overall average, with *Bulkiness* scoring the worst. Of note are the three other properties describing the physical size of amino acid residues: *Molecular volume*, *Molecular weight*, and *Partial specific volume*, all of which scored among the least accurate properties.

Several properties appear correlated in terms of the occurrence of false-positives (Table II). *Molecular volume*, *Molecular weight*, and *Partial specific volume* seem to be correlated in their propensity to produce false-positive results. For example, *Molecular volume* produced 34 false-positives overall. Of these, 30 produced false-positives in *Partial specific volume* and 26 produced false-positives in *Molecular weight*. These three properties also are correlated with false-positives in *Helical contact area*, but to a lesser extent.

There was also a strong correlation between accuracy and the identity of the amino acids being exchanged (Table III). Collectively, the 11 exchanges with the worst accuracies (all less than 50% accuracy) accounted for 207 of the 399, or 51.9%, of all false-positive results. There is no correlation between the information in Tables I and III. Several sites were implicated by multiple physicochemical properties, but

TABLE I
ACCURACY OF TREE SAAP ANALYSIS AMONG RADICAL AMINO ACID REPLACEMENTS FOR THE DEFAULT PHYSICO-CHEMICAL PROPERTIES.

Symbol	Description ^a	Calculated Accuracy
P_α	Alpha-helical tendency	90.1%
N_s	Average number of surrounding residues	93.9%
P_β	Beta-structure tendency	66.4%
B_1	Bulkiness	27.3%
B_r	Buriedness	85.7%
R_C	Chromatographic index	97.6%
P_c	Coil tendency	99.4%
c	Composition	71.3%
K^0	Compressibility	87.4%
pK'	Equilibrium constant (ionization COOH)	91.2%
C_a	Helical contact area	87.7%
h	Hydropathy	94.0%
pH_i	Isoelectric point	99.8%
E_l	Long-range non-bonded energy	62.7%
F	Mean r.m.s. fluctuational displacement	94.6%
M_v	Molecular volume	63.8%
M_w	Molecular weight	53.3%
H_{sc}	Normalized consensus hydrophobicity	94.6%
V^0	Partial specific volume	62.8%
P_r	Polar requirement	78.5%
p	Polarity	99.4%
α_c	Power to be at the C-terminal	77.0%
α_m	Power to be at the middle of the α -helix	51.6%
α_n	Power to be at the N-terminal	100.0%
μ	Refractive index	63.3%
E_{sm}	Short- & medium-range non-bonded energy	98.2%
R_o	Solvent accessible reduction ratio	93.0%
H_p	Surrounding hydrophobicity	87.2%
H_t	Thermodynamic transfer hydrophobicity	97.7%
E_t	Total non-bonded energy	97.1%
P_t	Turn tendency	100.0%

^aReferences for each property are included in Woolley et al. 2003.

this redundancy was taken into account when estimating the accuracies of individual amino acid exchanges.

Finally, there is a strong correlation between the generation of false-positive results and codon position. The base-exchange with the greatest false-positive occurrence was second-position transversions at 52.1%. The next greatest source of false-positive results was first-position transversions at 26.8%. Combined transitions, regardless of codon position, however, accounted for just 18.0% of all false-positives.

B. Sliding-window Evaluation Results

Sliding-window (SW) sizes of 9, 11, 15, and 19 codons were evaluated in terms of the incidence of false-positive results. Sliding-window sizes may be assessed in two ways: The number of false incidents (FI) over all properties, and number of falsely identified amino acid replacements (FR) regardless of property. The former is merely the sum of the falsely identified amino acid replacements for each property, while the latter takes into account the possibility of individual amino acid replacements affecting multiple properties. Only the latter may be used to calculate accuracy.

When $SW = 9$, $FI = 179$ and $FR = 126$, for an accuracy of 81.9% (out of 695 total reconstructed amino acid replacements). There is a decrease in accuracy (79.7%) when the SW is increased to 11 codons ($FI = 218$; $FR = 141$), but further increasing the window size more than corrects this loss in accuracy: For $SW = 15$, $FI = 176$ and $FR = 119$ for an accuracy

TABLE III
ACCURACY OF TREE SAAP ANALYSIS AMONG INDIVIDUAL SIMULATED
SINGLE-STEP AMINO ACID EXCHANGES.

Exchange	Overall	False-positives	Accuracy
G↔V	48	42	12.5%
E↔V	16	13	18.8%
G↔W	11	8	27.3%
D↔V	28	18	35.7%
I↔S	16	10	37.5%
I↔K	15	9	40.0%
G↔R	57	33	42.1%
S↔W	7	4	42.9%
L↔P	51	29	43.1%
K↔M	11	6	45.5%
A↔P	44	23	47.7%
I↔N	25	12	52.0%
C↔G	20	9	55.0%
C↔W	19	8	57.9%
C↔S	41	15	63.4%
C↔F	25	9	64.0%
H↔P	35	12	65.7%
I↔R	3	1	66.7%
L↔R	46	15	67.4%
A↔E	25	8	68.0%
P↔R	47	15	68.1%
P↔Q	19	6	68.4%
D↔H	23	7	69.6%
L↔S	14	4	71.4%
D↔Y	18	5	72.2%
H↔Y	24	6	75.0%
A↔D	20	5	75.0%
M↔R	12	3	75.0%
C↔Y	17	4	76.5%
C↔R	19	4	79.0%
E↔G	29	6	79.3%
L↔Q	27	5	81.5%
P↔T	44	8	81.8%
D↔G	16	2	87.5%
R↔W	16	2	87.5%
I↔L	45	5	88.9%
I↔T	27	3	88.9%
I↔V	27	3	88.9%
E↔K	21	2	90.5%
K↔T	33	3	90.9%
F↔Y	24	2	91.7%
R↔S	63	5	92.1%
A↔S	39	3	92.3%
F↔S	16	1	93.8%
A↔G	38	2	94.7%
H↔L	19	1	94.7%
S↔Y	22	1	95.5%
I↔M	34	1	97.1%
K↔N	41	1	97.6%
Total False-positives:		399	

Residues listed in the first column have been placed in alphabetical order.
All exchanges should be considered bi-directional.

* All single-step exchanges not listed here experienced 100% accuracy.

be assessed in more detail by a more comprehensive study. However, local sampling likely may generate extremely small expected frequencies for the most radical substitutions, such that even single mutations may produce statistically significant results. Often this may result in false-positive results.

Property correlations may compound the generation of false-positives for certain properties. When a property that has not been affected by positive selection exhibits some correlation with a property that has been strongly affected, false-positive results from the analysis of the unaffected property are much more likely even when correlations may be moderately poor.

Additionally, it is obvious from this study that partitioning strategy also has a significant effect on the generation of false-positive results. Both choice of sliding-window size and number of magnitude categories may affect the overall number of false-positives. Increasing the size of the sliding window increases local sample sizes, which makes a TreeSAAP analysis inherently more conservative, while maximizing authentic signal emerging from the genetic diversity among extant sequences. Decreasing the number of magnitude categories has the same effect. Implementing both of these strategies will likely reduce the overall effect of false-positive results, but different divergence levels likely will require different optimal combinations of window size and category number.

B. Improving Accuracy

There are two possible approaches for improving the accuracy of TreeSAAP results: (1) eliminate properties that produce a high level of false-positive results, or (2) flag specific amino acid exchanges as possibly false, while leaving them out of any subsequent analysis. We feel that the latter is preferable because it allows the user to make informed decisions about individual exchanges based on their proximity to active sites and functional motifs. Furthermore, many of the properties that performed poorly are tightly correlated, but not to the level that the majority of false-positive results would be eliminated. If the 11 exchanges that performed the worst are flagged as possibly adaptive and eliminated from subsequent analyses, results will likely improve in accuracy. With relation to the data simulated for this study, accuracy was improved from 80.3% to 89.8% by doing this. The frequency with which these 11 exchanges may occur in empirical data, however, remains to be determined as part of a future study.

Certain poorly performing physicochemical amino acid properties may also be substituted with properties that describe similar characteristics of amino acid residues that have greater accuracy. Even similar properties exhibit different accuracies. For example, four of the currently default properties in TreeSAAP may be used to describe amino acid residue size: *Bulkiness*, *Molecular volume*, *Molecular weight*, and *Partial specific volume*. Of these, accuracies vary from 27.3% (*Bulkiness*) to 63.8% (*Molecular volume*). Although none of these properties produce very high confidences in analytical results, unless a better descriptor for residue size can be found, *Molecular volume* may be tentatively used to assess selection on this genre of amino acid characteristics. As a result, the overall accuracy of an analysis may be improved. As mentioned above, however, different divergence levels and local amino acid residue compositions may have differential effects.

Finally, decreasing the p -value threshold may be the simplest method for increasing overall analytical accuracy. In most cases, a p -value threshold of 0.05 would yield 95% confidence, but TreeSAAP sliding-window analysis implements layer upon layer of multiple sampling and thus requires a statistical correction. In most circumstances, a Bonferroni correction will bring multiple-sampling analyses back in line

with statistical norms, but our results suggest that this may not be enough of a correction. The analysis of the smaller data set with three magnitude categories and a sliding-window of nine codons resulted in an initial accuracy of 81.9% when using a probability criterion of $p < 0.001$ as a Bonferroni correction. When this criterion is further decreased ($p < 0.0001$), accuracy is improved to 90.5%. Further adjustment of this criterion ($p < 0.00001$) improves accuracy to 95.1%. These results suggest that a p -value of 0.00001 is appropriate for 95% confidence. However, doing so may yield some false negative results. A simple solution to this problem may be to simply annotate all TreeSAAP results with the exact probability of positive selection for each site. Future versions of the TreeSAAP software will explore implementing this approach.

C. Confounding Variables

The general applicability of these results also remains to be seen. There are a number of variables that may influence the incidence of false-positive results. Most of these have to do with statistical significance, including divergence levels of the sequences in the alignment, the number of magnitude categories used to subdivide the data, and the size of the sliding window. Lower divergence levels would likely produce more false-positive results because this would correspondingly decrease the sampling of amino acid replacements. Increasing the number of magnitude categories may increase the number of false-positives because the sample size for each z -score test would proportionally decrease. Finally, smaller window sizes would likely produce more false positives because each inferred amino acid replacement would have a greater statistical influence.

Additionally, there are other variables that may have an overall influence. For example, the tree structure may have an influence on the software's ability to reconstruct ancestral character states. The identity of the governing genetic code may also have an influence [13]. Finally, and not to be underestimated, is the influence of selection itself, which may naturally eliminate many of the exchanges that result in false-positive. Further study is required to determine the influence these and other potentially confounding variables may have on the accuracy of TreeSAAP results. It may also be possible to identify alternatives to properties that performed poorly in this study from among the over 500 properties in AAindex [7], [8]. The identification of a property or properties for describing amino acid size for replacing those currently included in TreeSAAP (i.e. *Bulkiness*, *Molecular volume*, *Molecular weight*, and *Bulkiness*) is especially important.

Several empirical and analytical variables may have a significant effect on TreeSAAP analytical accuracy. For example, inappropriate sequence sampling may result in substitutional saturation, as well as difficulties in alignment and phylogeny reconstruction, all of which may alter inferred patterns of amino acid replacements and change the ability of TreeSAAP to detect natural selection. If sampling is appropriate, inappropriate method and parameter choice for alignment and

phylogeny reconstruction may still produce adverse effects in TreeSAAP.

D. Other Observations

Upon preparation of the manuscript for this study, we found that three of the properties currently default in TreeSAAP, *Power to be at the C-terminal*, *Power to be at the middle of the α -helix*, and *Power to be at the N-terminal*, have been mislabeled in all studies but the study by Chou and Fasman that originally described them [2]. Subsequent studies [12], [18], [5] have inadvertently confused these properties, most likely due to the use of multiple descriptions for each by Chou & Fasman [2]. The values of the property that is now referred to as *Power to be at the C-terminal* correspond to the normalized values of what was originally referred to as *Frequency of residues in the N-terminal helix region* [2] or what has since become known as *Power to be at the N-terminal* [18]. The values of the property that is now referred to as *Power to be at the middle of the α -helix* correspond to the normalized values of what was originally referred to as *Frequency of residues in the C-terminal helix region* [2] or what has since become known as *Power to be at the C-terminal* [18]. Finally, the values of the property that is now referred to as *Power to be at the N-terminal* correspond to the values of what was originally referred to as *Average conformational parameter for the inner helix region* [2] or what has since become known as *Power to be at the middle of the α -helix* [18]. Corrections in the labels for these properties have been made in this paper, and will be made in subsequent versions of TreeSAAP.

V. CONCLUSIONS

The TreeSAAP software performs adequately (about 80% accuracy), but with just a few adjustments it can be made to perform significantly better (about 90% accuracy). However, confounding variables may cause this increase in accuracy to change in empirical data. Further study with both simulated and empirical data is required to determine the relative importance of these confounding variables to accurately identifying and characterizing selection. This study, however, suggests that just being cognizant of those amino acid replacements that result in the least accuracy, and using fewer magnitude categories, larger sliding-window sizes, and decreased p -values may allow users to more accurately interpret their analytical results.

ACKNOWLEDGMENT

Thanks to all the great undergraduate students at Brigham Young University who worked in the McClellan lab between 2001 and 2007. Although the work they did on the background for this project was pivotal to working out kinks in the TreeSAAP software, they were not directly involved in these results. The support received from the Department of Integrative Biology at Brigham Young University, and the BYU Office of Research and Creative Activities is also greatly appreciated. Finally, thanks to Bigelow Laboratory for Ocean Sciences for continued support for this and other research.

REFERENCES

- [1] S. Chamala, W.A. Beckstead, M.J. Rowe, and D.A. McClellan. Evolutionary selective pressure on three mitochondrial SNPs is consistent with their influence on metabolic efficiency in Pima indians. *International Journal of Bioinformatics Research and Applications*, 3(4):504–522, 2007.
- [2] P.Y. Chou and G.D. Fasman. Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2):211–222, 1974.
- [3] K.A. Crandall, C.R. Kelsey, H. Imamichi, H.C. Lane, and N.P. Salzman. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *MBE*, 16(3):372–382, 1999.
- [4] M.T.W. Ebbert, W.A. Beckstead, T.D. O’Connor, M.J. Clement, and D.A. McClellan. Pharmacogenomics: Analysing SNPs in the CYP2D6 gene using amino acid properties. *International Journal of Bioinformatics Research and Applications*, 3(4):471–479, 2007.
- [5] M.M. Gromiha and P.K. Ponnuswamy. Relationship between amino acid properties and protein compressibility. *JTB*, 165:87–100, 1993.
- [6] A.L. Hughes, T. Ota, and M. Nei. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class i major-histocompatibility-complex molecules. *MBE*, 7(6):515–524, 1990.
- [7] S. Kawashima and M. Kanehisa. AAindex: Amino acid index database. *NAR*, 28(1):374, 2000.
- [8] S. Kawashima, P. Pokarowski, M. Pokarowski, A. Kolinski, T. Katayama, and M. Kanehisa. AAindex: Amino acid index database, progress report 2008. *NAR*, 36(Database issue):D202–D205, 2008.
- [9] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK, 1983.
- [10] A.T. Marques, A. Antunes, P.A. Fernandes, and M.J. Ramos. Comparative evolutionary genomics of the HADH2 gene encoding $\alpha\beta$ -binding alcohol dehydrogenase/17 β -hydroxysteroid dehydrogenase type 10 (ABAD/HSD10). *BMC Genomics*, 7:202, 2006.
- [11] D.A. McClellan and K.G. McCracken. Estimating the influence of selection on the variable amino acid sites of the cytochrome b protein functional domains. *MBE*, 18(6):917–925, 2001.
- [12] D.A. McClellan, E.J. Palfreyman, M.J. Smith, J.L. Moss, R.G. Christensen, and J.K. Sailsbery. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *MBE*, 22(3):437–455, 2005.
- [13] D.A. McClellan, D.G. Whiting, R.G. Christensen, and J.K. Sailsbery. Genetic codes as evolutionary filters: Subtle differences in the structure of genetic codes result in significant differences in patterns of nucleotide substitution. *JTB*, 226:393–400, 2004.
- [14] D.S. Osorio, A. Antunes, and M.J. Ramos. Structural and functional implications of positive selection at the primate angiogenin gene. *BMC Evolutionary Biology*, 7:167, 2007.
- [15] M. Perez-Losada, R.P. Viscidi, J.C. Demma, J. Zenilman, and K.A. Crandall. Population genetics of *Neisseria gonorrhoeae* in a high-prevalence community using a hypervariable outer membrane porB and 13 slowly evolving housekeeping genes. *MBE*, 22(9):1887–1902, 2005.
- [16] M.L. Porter, T.W. Cronin, D.A. McClellan, and K.A. Crandall. Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins. *MBE*, 24(1):253–268, 2007.
- [17] M. Prabhakaran. The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochemistry Journal*, 269:691–696, 1990.
- [18] M. Prabhakaran and P.K. Ponnuswamy. The spatial distribution of physical, chemical, energetic and conformational properties of amino acid residues in globular proteins. *JTB*, 80:485–504, 1979.
- [19] P.M. Sharp. In search of molecular darwinism. *Nature*, 385:111–112, 1997.
- [20] S. Woolley, J. Johnson, M.J. Smith, K.A. Crandall, and D.A. McClellan. TreeSAAP: Selection on amino acid properties using phylogenetic trees. *Bioinfo*, 19(5):671–672, 2003.
- [21] X. Xia and W.-H. Li. What amino acid properties affect protein evolution? *JME*, 47:557–564, 1998.
- [22] Z. Yang and J.P. Bielawski. Statistical methods for detecting molecular adaptation. *TREE*, 15(12):496–503, 2000.
- [23] Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *MBE*, 19(6):908–917, 2002.
- [24] Z. Yang, R. Nielsen, N. Goldman, and A.-M.K. Pedersen. Codon substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, 2000.
- [25] Z. Yang and W.J. Swanson. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *MBE*, 19(1):49–57, 2002.