Detecting Signatures of Positive Selection against a Backdrop of Compensatory Processes

Peter B. Chi,[1,2] Westin M. Kosater,[2] and David A. Liberles*,[2]

[1]Department of Mathematics and Statistics, Villanova University, Villanova, PA

[2]Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA

***Corresponding author:** E-mail: daliberles@temple.edu.

Associate editor: Keith Crandall

Abstract

There are known limitations in methods of detecting positive selection. Common methods do not enable differentiation between positive selection and compensatory covariation, a major limitation. Further, the traditional method of calculating the ratio of nonsynonymous to synonymous substitutions (d*N*/d*S*) does not take into account the 3D structure of biomacromolecules nor differences between amino acids. It also does not account for saturation of synonymous mutations (d*S*) over long evolutionary time that renders codon-based methods ineffective for older divergences. This work aims to address these shortcomings for detecting positive selection through the development of a statistical model that examines clusters of substitutions in clusters of variable radii. Additionally, it uses a parametric bootstrapping approach to differentiate positive selection from compensatory processes. A previously reported case of positive selection in the leptin protein of primates was reexamined using this methodology.

***Key words:* positive directional selection, compensatory covariation, protein structure evolution.**

Introduction

Detecting lineage-specific changes in protein function, particularly those associated with positive directional selection, is an important goal in comparative genomics. The most common method for achieving this goal is through estimation of the ratio of nonsynonymous to synonymous nucleotide substitution rates in a lineage-specific manner (d*N*/d*S* or ω) (Yang 1998), referred to as the branches model. A limitation of this approach is that it averages across all sites in a protein-encoding gene, which often results in tests that are underpowered, particularly when selective pressures are localized to a few sites in the alignment. In other words, if a few sites are under heavy selective pressure resulting in high ω values, this may not be enough to outweigh low ω values from the remainder of the alignment, and thus selection would not be detected at all. A statistical method for detecting sitewise variation on a branch was introduced with the branch-site model (Yang and Nielsen 2002), in which a likelihood ratio test was devised to compare an alternative model where ω is allowed to vary at each amino acid site of the alignment, versus a null model in which it does not. However, although this model is an increase in biological realism, it still relies on the ubiquitous site-independence assumption that is also known to be biologically unrealistic.

Another class of approaches considers contiguous sets of genetic material. In primary sequence space, that is, DNA nucleotide sequences, one can examine windows of contiguous sites to detect regions of a protein that have undergone positive directional selection, whereas the remainder of the protein is under negative selection (Endo et al. 1996; Fares et al. 2002). However, this so-called "primary windowing" does not account for the tertiary structure of the resulting protein product that dictates which sites functionally interact with each other. Because proteins fold into 3D structures and selection acts upon functional sites in this context, tertiary windowing was introduced as a structure-aware alternative to finding functional regions under positive directional selection using the d*N*/d*S* statistic (Suzuki and Gojobori 1999; Berglund et al. 2005; Liang et al. 2006; Tusche et al. 2012). Additionally, because the d*N*/d*S* ratio is dependent upon proper calculation of the synonymous site substitution rate, it is subject to saturation and is only applicable over relatively short evolutionary distances (Anisimova and Liberles 2012). Structure-independent amino acid-based statistical methods have been developed to characterize selection, including rate shift methods (Gu and Vander Velden 2002; Penn et al. 2008) and mutation-selection models (Halpern and Bruno 1998; Spielman and Wilke 2016; Teufel et al. 2018). A protein structure-aware alternative that is applicable over much longer evolutionary periods and can be applied phylogenetically is desirable.

Early methods that rely on clustering statistics have been developed (Yu and Thorne 2006; Adams et al. 2017). Arnold and coworkers have established the general destabilizing effect of adaptive substitutions on protein structure, which can lead to compensation, observed as multiple substitutions

(Bloom and Arnold 2009). However, it is also known that mildly deleterious changes can fix as well, in particular in small effective population size organisms and that these changes will also result in compensation, an effect known by Pollock and Goldstein as the Stokes shift (Pollock et al. 2012; Goldstein and Pollock 2016). In large effective population size lineages, where the time to fixation is longer, compensatory changes can arise in the same genetic background as the original destabilizing change and these can fix together neutrally, an effect known as stochastic tunneling (Lynch 2010).

Beyond these effects, it is generally known that different parts of a protein undergo substitution at different rates, giving rise to a distribution of rates consistent with the gamma distribution when tested using standard model selection software (Abascal et al. 2005; Yang 2007; Grahnen et al. 2011; Echave et al. 2016). Biologically, one explanation of this observation is because position solvent accessible surface area (SASA) and the contact number of a position (i.e., the number of neighboring sites within a given distance of that position) are correlated with each other and are both known to be drivers of amino acid substitution rate. Specifically, the hydrophobic core of a protein evolves more slowly than the hydrophilic surface (Lesk and Chothia 1980; Chothia and Lesk 1982; Chi and Liberles 2016); thus, higher SASA and contact number would both be associated with faster substitution rates. A method that controls for this known biology to identify statistically unexpected patterns of amino acid substitution when positive selection is acting would be a valuable tool in the comparative genomic toolbox.

Here, we introduce such a method. The method relies upon the set of substitutions identified along the branch of a phylogenetic tree mapped onto a 3D protein structure, termed substitutional mapping (Bollback 2006; Monit and Goldstein 2018). As a starting point, we consider the method described by Yu and Thorne (2006) (henceforth referred to in this work as "YT06," after its authors' surnames and the year of publication) as it is a structure-aware method to detect clusters of amino acid substitution in proteins along a lineage. Specifically, it falls broadly into the category of tertiary windowing methods, by considering spheres around each amino acid residue in the context of the structure of the protein that it is in. Positive selection is detected if there is an increase in the number of substitutions that occur within each sphere than would be expected by chance. We introduce a few modifications to their method and provide a comparison of Type I Error rates and Power to detect spatial clustering naively, that is, without controlling for the effects mentioned above. Additionally, and more crucially, we devise a novel parametric bootstrap to explicitly account for the known biology that would give rise to clustering in the absence of positive selection. To accomplish this, we formulate a null model of substitutions that is consistent with varying rates due to SASA and contact distance and compensatory changes in the absence of positive selection and utilize this to construct a hypothesis test for positive selection against this backdrop. We demonstrate that our method, referred to in a likewise manner as "CKL20" throughout the remainder of this manuscript, is able to avoid incorrectly identifying this backdrop as positive selection while maintaining high power to detect positive selection under our simulation schemes.

New Approaches

We begin with a brief description of the aforementioned YT06 method, following the notation in the original manuscript: For a particular branch i of a phylogenetic tree, let N_i be the average number of the substitutions within each 10-Å sphere of every site of the protein. Then, via a permutation test where each of T iterations is random a shuffling of which sites are substituted, $N_{\bar{i}(S)}$ then represents the overall sample mean of all of the means from each iteration; that is,

$$N_{\bar{i}(S)} = \frac{\sum_{t=1}^{T} N_i^t}{T}. \qquad (1)$$

This is then used to standardize the original observed average count N_i from the data,

$$Z_i = \frac{N_i - N_{\bar{i}(S)}}{\sigma_{\tilde{i}(S)}}, \qquad (2)$$

where $\sigma_{\tilde{i}(S)}$ is the sample standard deviation of all N_i^t values. A nonparametric P value is obtained by calculating analogous Z values for each permuted N_i^t value and obtaining the proportion of these that are at least as extreme as the data Z_i value.

Permutation Test

Our new permutation test is structurally similar to the YT06 approach but contains three key differences: 1) Rather than a fixed radius of 10 Å, we allow for the radius to be specified by the user, with a default value of 7 Å based on our findings (see Results section). 2) In Yu and Thorne (2006), the statistic is based upon the count of substitutions in each sphere, whereas we consider the fraction of substitutions out of the total number of sites in each sphere. 3) Rather than considering spheres around every site of the protein, we only consider spheres around sites that were themselves substituted.

Parametric Bootstrap

Our greater contribution in this work is the development of our novel parametric bootstrap, which represents a greater shift in approach than the modifications to the YT06 permutation test described in the previous subsection. Here, our goal is to detect not just an increase in spatial clustering compared with that expected by pure chance, but rather, an increase beyond what would be expected due to the biological processes of SASA effects and compensatory processes. To accomplish this, our parametric bootstrap simulates a null model in which sites with higher SASA values have an increased rate of substitution, while simultaneously giving sites closer to existing substitutions an increased rate of substitution as well. A statistically significant P value would then only be observed if the data demonstrated an increase in clustering beyond these effects. Additionally, as opposed to testing for a mean shift, our test statistic here is based upon the 95th quantile. Further details are in the Materials and Methods section and the Discussion section.

Results

3D Protein Structure

Figure 1 shows a graphical representation of one of the protein structures under consideration, with a PDB ID of 2I0Q (Berman et al. 2000; Buczek and Horvath 2006) (see Materials and Methods section for further details). Under the naive null model, substitutions occur completely at random on the structure, shown in panel 1. Shown in the middle panel is the scenario for SASA + compensatory processes, which serves as our null hypothesis for our parametric bootstrap. The third scenario shows positive selection acting against a backdrop of SASA + compensatory processes. Our ultimate aim is to be able to detect this while simultaneously avoiding false positives due to clustering that is solely due to SASA effects and compensatory processes (middle panel of fig. 1).

Detection of Compensatory Clustering

For the detection of compensatory clustering, our proposed test is still a permutation test similar to that of the YT06 method (Yu and Thorne 2006), but with important modifications. To confirm that our proposed permutation test is a proper α-level test, we performed simulations under null scenarios to determine whether the Type I Error rates match with the α-level cutoff. Using the previously mentioned 2I0Q protein structure, we first simulated a naive null scenario in which there is the complete absence of spatial clustering; in other words, this null scenario does not account for varying rates due to SASA and compensatory changes, but rather simply permutes the substitutions completely at random across the protein structure. Simulations were performed with a branch length of ~0.05 (in other words, an expected number of substitutions of 0.05 per site). One aspect of the YT06 method as described in Yu and Thorne (2006) is that it uses a fixed radius of 10 Å, which the authors acknowledge to be somewhat arbitrarily chosen. We thus explore a variety of radius sizes in our test, at 6.5, 9, and 11.5 Å. Type I Error rates of CKL20 were all estimated to be in fact slightly below that of YT06 and closer to the nominal $\alpha = 0.05$ level, and there does not appear to be an association between Type I Error rate and radius size, as shown in figure 2. The 95% confidence intervals were calculated according to the length/coverage optimal (LCO) method, as it was demonstrated to have desirable statistical properties over the usual Wald method (Schilling and Doi 2014). This suggests that our permutation test is approximately a proper α-level test for any radius size, at least within the range of 6.5–11.5 Å that was investigated.

We also investigated the impact of radius choice on power performance in our new permutation test, shown in figure 3. In particular, we wanted to evaluate how our test would perform against that of the 10 Å fixed radius in YT06. For this test, we simulated data under the alternative hypothesis of clustering due to SASA effects and compensatory processes. First, using the 2I0Q structure previously mentioned, we notice maximum power attained at a radius of 7–7.5 Å and gains in power over the YT06 approach at all radii up to 14.5 Å. To see whether this trend would hold across different structures, we chose two proteins in different protein superfamilies (from 2I0Q and from each other), with PDB IDs of 1D4T (Poy et al. 1999) and 1AX8 (Zhang et al. 1997). Substitutions were simulated again at a branch length of ~0.05, for each structure, and results are shown in the middle and bottom panels of figure 3. Again, we note that in each case, maximal power is attained with our method around 7 Å, suggesting that the 10 Å of the YT06 method may not be the optimal radius size to use. Furthermore, power gains over the YT06 method are achieved in these two structures as well, as shown.

Detection of Positive Selection

Here, we propose a new test to detect positive selection, and specifically to distinguish positive selection from SASA effects and compensatory processes known to occur in the absence of selection. Our null hypothesis, then, includes the presence of the aforementioned SASA effects and compensatory processes. As the YT06 method was not designed to take this into account, it will have an exorbitantly high Type I Error rate

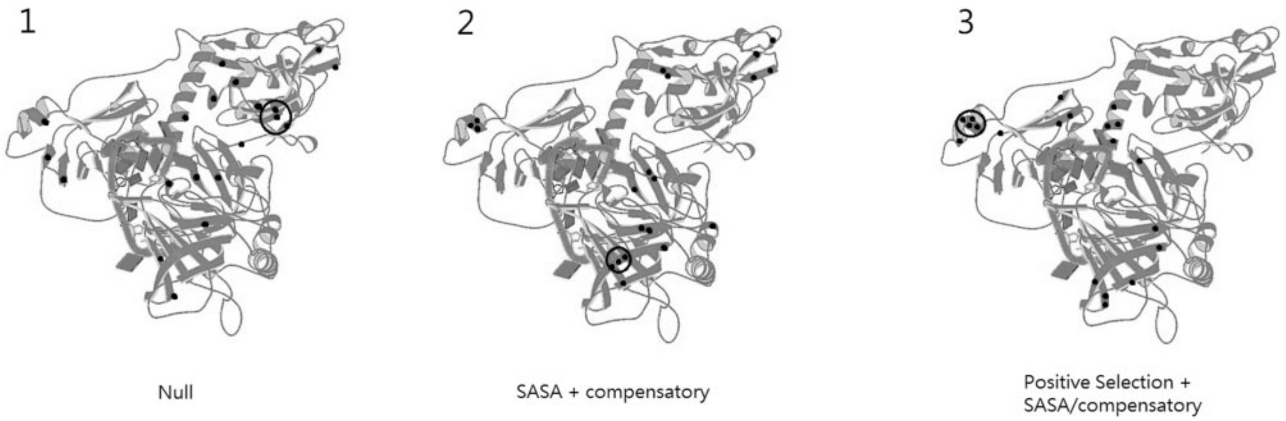

Fig. 1. Protein structure with PDB ID: 2I0Q. Substitutions are shown as black dots, and a potential sphere within which substitutions are to be counted is shown as a circle. Representations of three scenarios are shown in the following panels: 1) Null, in which substitutions occur completely at random with respect to structure. 2) SASA + compensatory, in which clustering may occur due to SASA effects and compensatory processes, as illustrated by an increase of substitutions on the surface of the protein. 3) Positive selection against a backdrop of SASA effects and compensatory processes, as illustrated by an even further increase in clustering beyond that shown in panel 2.

when clustering is occurring due to this reason. This is shown in comparison to the Type I Error rates of our proposed method, in figure 4. All estimated Type I Error rates from our method are fairly close to the α-level of 0.05. We note briefly though that unlike previously, there does appear to be a possible inverse relationship between Type I Error rate and radius size, with estimated Type I Error rates trending downwards as the radius size increases. However, the 95% confidence interval error bars are overlapping, thus suggesting that this trend may not be statistically significant.

Thus, given that our proposed test can avoid signatures of clustering that are due to forces aside from positive selection, our primary question of interest is whether it can then still adequately detect positive selection. Alternative hypothesis scenarios were simulated in which two, three, or four substitutions were deterministically chosen on the surface of the 2I0Q structure. These sites were selected iteratively, starting from the site in the structure with the highest SASA value, and then the three nearest sites to that original one, mimicking the evolution of a new binding site in proteins such as that proposed for leptin in mammalian species (Gaucher et al. 2003). The remainder of substituted sites were simulated due to SASA effects and compensatory processes. This was designed to mimic the overall suite of biological forces that may induce clustering.

Results are shown in figure 5, summarized in a receiver operating characteristic curve (ROC) manner, with Type I Error rates (i.e., 1-specificity) on the x-axis and power (i.e., sensitivity) on the y-axis. In the CKL20 method, we show power and Type I Error rates at a radius of 6.5 Å. When we compare this with the YT06 method, we observe that in the scenarios with three and four deterministically chosen sites, our method outperforms YT06, attaining an estimated power of 1. In the scenario with two deterministically chosen sites, although on a strict power scale it does not perform as well as YT06 (power of 0.297 vs. 0.599), we note that it is still further from the diagonal line than YT06 for two deterministic sites, indicating that it is overall a better discriminator of positive selection against the backdrop of SASA and compensatory processes. It is worth reiterating that this observation is occurring in spite of the fact that our proposed method is

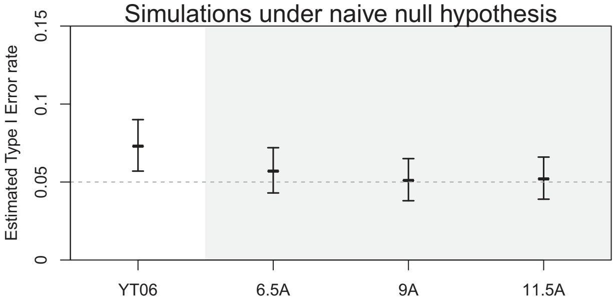

Fig. 2. Type I Error rates of permutation tests across three different radius sizes, for both versions of our test and the original method in Yu and Thorne (2006). The dotted horizontal line represents the α-level of 0.05. Error bars represent 95% confidence intervals. The grayed portion of the graph represents the CKL20 method, at three different radii. Simulations were performed with 1,000 iterations.

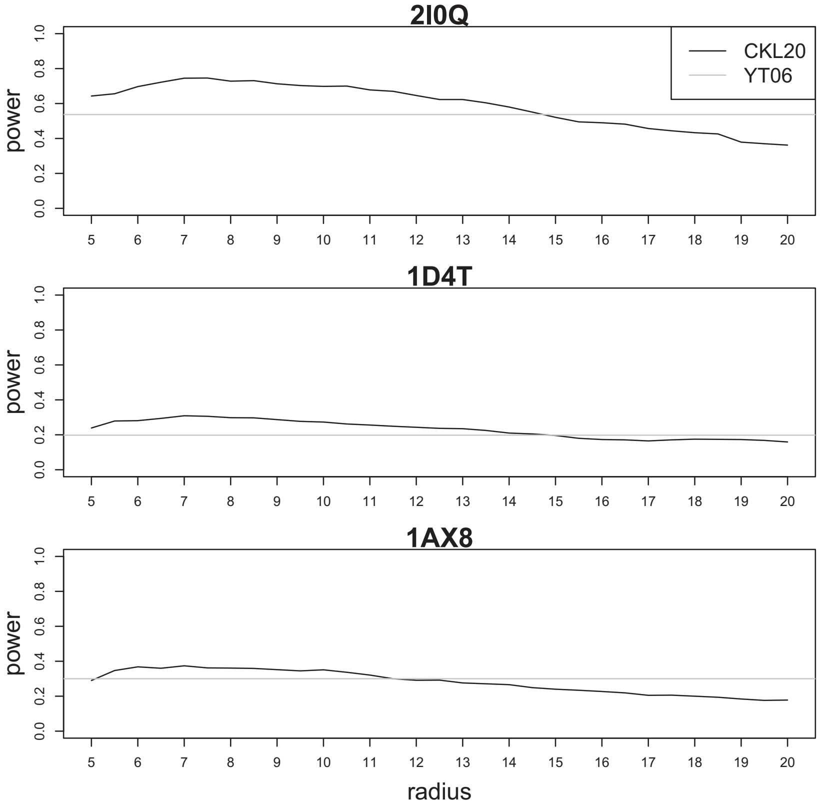

Fig. 3. Investigating power versus radius size to detect clustering due to SASA and compensatory processes, across three different structures. The gray horizontal line represents the power of the original method in Yu and Thorne (2006). Simulations were performed with 1,000 iterations.

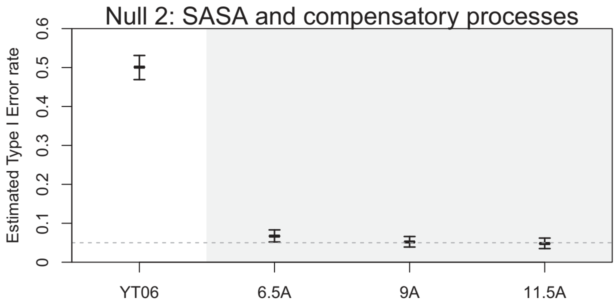

Fig. 4. Type I Error rates under the SASA + compensatory processes scenario are shown for YT06 and across three different radii for our method. The dotted horizontal line represents the α-level of 0.05. The grayed portion of the graph represents the CKL20 method, at three different radii. Error bars represent 95% confidence intervals.

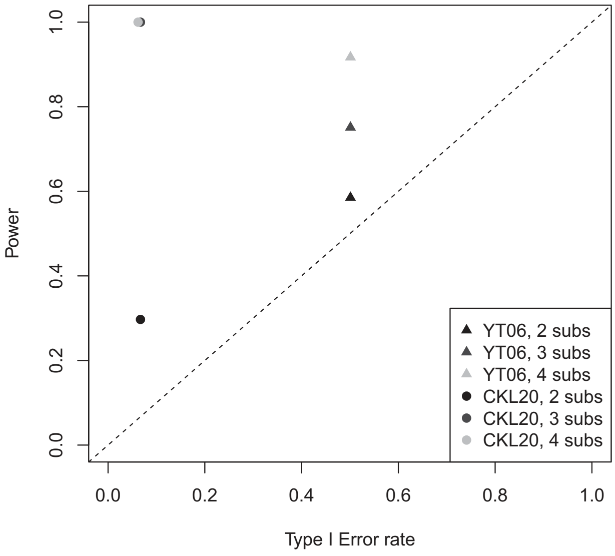

Fig. 5. An ROC-like graph illustrating the relationship between Type I Error rates and Power. The diagonal dotted line represents a test that has no discriminatory value.

fundamentally ignoring clustering due to SASA effects and compensatory processes and correctly treating these as noise, whereas the YT06 method is treating these as signal (as seen in its high Type I Error rate in fig. 4).

We note here that the choice of a radius of 6.5 Å is due to the fact that this radius provided the highest power in this scenario. Radius size should be considered as a tuning parameter, for which an optimal value must be decided upon. In practice, this should of course be done a priori. In order to provide some guidance as to how this should be done, we attempted to characterize optimal radius size with respect to power based on any signatures of the data (such as the distribution of pairwise distances between substituted sites in the null distribution of the parametric bootstrap) but were unable to obtain any conclusive information. For more detail, see supplementary figures S1 and S2, Supplementary Material online, in which we show that the distribution of pairwise distances appears to differ when the optimal radius is different, but not in a predictable manner. We also show results from power analyses across a range of radii for all three structures. We note also that in the 1D4T and 1AX8 structures, the power of our CKL20 method does in fact surpass that of YT06 at certain radii, as shown in supplementary figure S1, Supplementary Material online. Moreover, at all radii, the power of our CKL20 method is fairly close to that of YT06, while still avoiding false positives due to SASA effects and compensatory processes as designed. This suggests that in certain structures and perhaps at certain branch lengths, the CKL20 method may in fact be far superior to YT06 in the sense that it can both avoid false positives due to SASA effects and compensatory processes and also have greater ability to detect when positive selection is occurring within this backdrop.

Table 1. *P* Values from the YT06 Test, CKL20 Permutation Test, and CKL20 Parametric Bootstrap.

	YT06	CKL20 Perm Test	CKL20 Par Boot
Hominoid	0.740	0.309	0.631
Macaca	0.262	0.077	0.552

Application to Empirical Data

The protein structure with PDB ID 1AX8 (investigated in fig. 3) is the human leptin protein, which has been well studied and is known to be linked to obesity (Caro et al. 1996; Mantzoros 1999). Particularly, evidence of positive adaptive selection has been found in leptin on the branch of the evolutionary history leading to apes (termed hominoids) as well as the lineage leading to rhesus macaque (Benner et al. 2002; Siltberg and Liberles 2002; Gaucher et al. 2003). Three different lines of evidence were used in these studies. One study examined increases in the clade-specific alpha value of the gamma distribution, whereas the other two studies relied upon d*N*/d*S* with tertiary windowing or with structural partitioning. We compare those results with analysis based upon YT06 and CKL20. Here, we use the aligned set of sequences used by Gaucher et al. (2003) and shown in figure 3 of their manuscript, which, in addition to hominoids includes the rhesus monkey (*Macaca mulatta*), cat, dog, sheep, and several other mammals. We reanalyze the homonoid branch, and also the branch leading to the rhesus macaque. Ancestral sequences were reconstructed using the aaml program within the PAML suite (Yang 2007), and then sites in which substitutions were inferred to have occurred were analyzed using the YT06 method, the CKL20 permutation test (radius of 7.0 Å), and the CKL20 parametric bootstrap (radius of 6.5 Å). These radii were selected from the a priori optimization for power as shown previously. The *P* values from these three tests for the two branches of interest are shown in table 1.

As seen in table 1, neither branch showed significant support for nonrandom clustering, although the permutation test of CKL20 for the Macaca lineage presented a *P* value of 0.077. From this result, it is likely that many of the changes on both branches were not driven by positive selection and that any positive selection that did occur either was reflected by

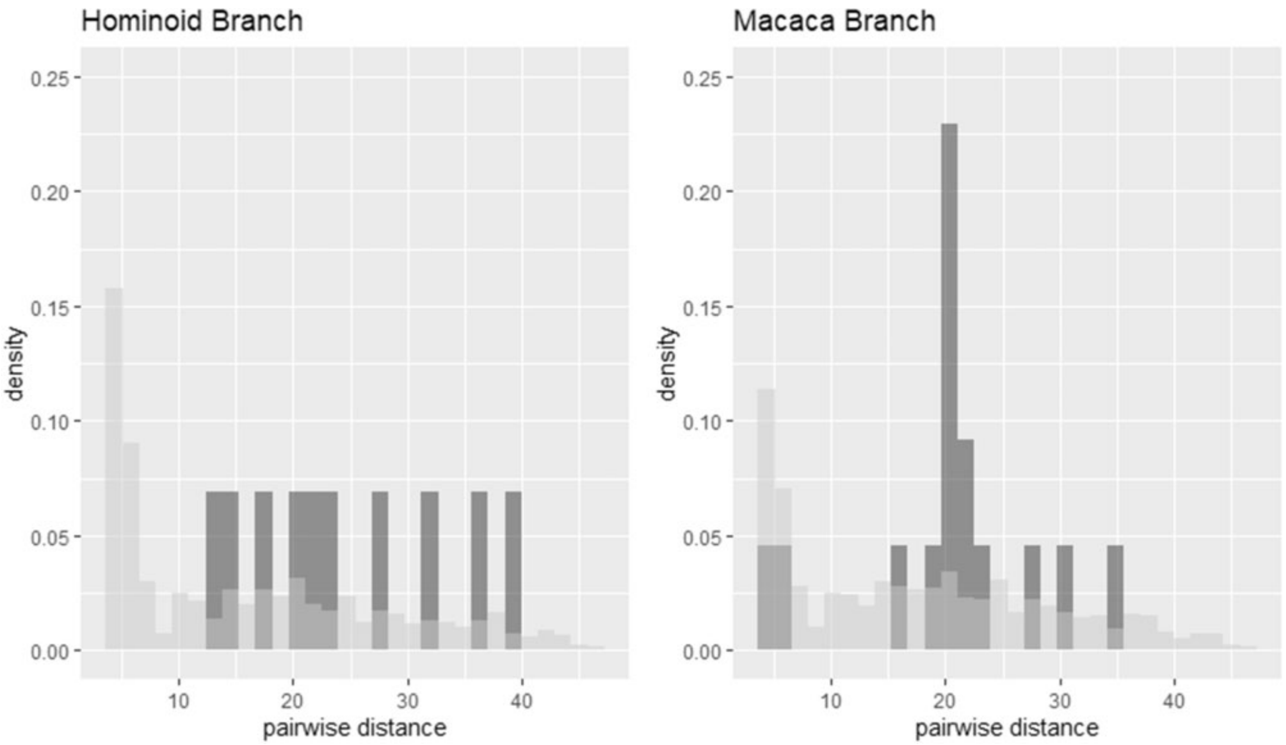

FIG. 6. Distributions of pairwise distances in the data, and that of simulated data with a strong signal for positive selection. The branch data from the alignment are shown in dark gray, whereas the simulated distribution is shown in lighter gray.

changes that either were spatially separated over a larger binding interface or were small in number (1 or 2 substitutions).

To investigate whether a statistically significant signal is to have been expected by the CKL20 parametric bootstrap if multiple changes driven by positive selection were tightly clustered, we examine the distribution of pairwise distances observed in our data and compare that with the distribution of pairwise distances formed by a strong signal for positive selection. Specifically, we simulate three sites in proximity that are all substituted, followed by the remainder of substitutions being simulated by probabilities according to their SASA and compensatory processes (similar to what was performed for fig. 5). In this case, the three deterministically chosen sites are positions 117, 118, and 119, based on the finding by Gaucher et al. (2003) that site 118 may be a strong target for functional change. The distributions are shown in figure 6, in which we observe that the amino acid substitutions observed on both focal branches were much more distant from each other than in the simulated case of strong selection at contiguous residues. It should further be noted that the lack of a significant signal in either the YT06 or the CKL20 permutation test may be driven by the large geometric space of the binding interface suggested by Gaucher et al. (2003).

Discussion

Starting with the set of amino acid substitution that have been mapped to a lineage of a phylogenetic tree using existing methods, we present an approach aimed at differentiating positive directional selection from neutral evolution and from compensatory processes. Our overall aim in this work is to present a novel parametric bootstrap methodology to detect positive selection in the presence of SASA considerations and compensatory processes. As both processes can result in an increase in clustering of substitutions in the 3D space of a protein, methods that do not account for the SASA and compensatory processes will result in highly inflated Type I Error rates, as demonstrated here with YT06. Conversely, our proposed method avoids signatures of clustering due to SASA and compensatory processes by design; the question, then, is whether it would still have enough power to pick up any signal of positive selection. Our results show that it does.

However, some questions remain. Specifically, although we have determined that a radius of 10 Å may not always be optimal with regard to maximizing power, we have not determined any specific patterns that could be used to estimate what it should be a priori for any given protein structure. Further work could attempt to characterize this, perhaps by obtaining a larger sample of protein structures across an array of protein families, and investigating power trends by radius size on each one.

One of the major modifications to the YT06 method that we propose is to construct our statistic based upon the 95th quantile of the proportions of substituted sites within each sphere, as opposed to a *z*-score statistic as proposed in Yu and Thorne (2006). Specifically, we do this in our parametric bootstrap test to distinguish between positive directional selection versus SASA effects and compensatory processes. Our biological rationale for doing so is because although data patterns according to the alternative hypothesis should indeed result in a mean shift, this is an indirect consequence of the mechanism that actually occurs. In other words, suppose that selection acts upon some neighboring sites to produce a functional change. Then, spheres that involve these selected sites should produce an inflation, specifically in the upper tail of the overall distribution since this signal will be most pronounced in a few spheres that would be expected to have a high proportion of substituted sites in them. In this way, we believe that we are capturing the signal directly, as opposed to a side-effect of the actual signal.

Conversely, our permutation test for clustering (e.g., shown in figs. 2 and 3) does still rely on a mean shift rather than the 95th quantile. In this scenario, our biological rationale for using the mean instead of the 95th quantile is that the data-generating process is not quite as precisely targeted as it is when selection is acting to produce a functional change as described above. Thus, a mean shift will indeed be a more direct consequence of the data patterns expected under the alternative hypothesis. For more details, refer to the Materials and Methods section and Supplementary Material online.

Positive selection acts on multiple residues arrayed at a binding interface (Ames et al. 2016) or in an enzyme active site (Lai et al. 2012). The distance dependence of those multiple residues under selection for function in contrast with the background distribution of distances under selection for compensatory processes will dictate the power and performance of our proposed parametric bootstrap test. This was noted in the analysis of cluster distributions on primate lineages of leptin. In the example studied here, our method did not detect significant evidence for clustering in leptin. On both lineages that were studied, signatures of positive selection did not correlate with structural clustering according to the methods developed here. Ultimately, power will relate to the number of changes driven by selection, their structural relationship to each other, and the total background changes on the same branch. Further studies are suggested to evaluate the geometric relationship of multiple residues selected for the same adaptive feature.

Across the three folds studied in this work, the distribution of residue distances (see supplementary fig. S2, Supplementary Material online) varied with the optimal radius in our permutation test for clustering, but not in a predictable way. Further exploration of this across different levels of hierarchy in CATH (Dawson et al. 2017) remains an avenue for further investigation. Currently, our three structures of 2I0Q, 1D4T, and 1AX8 were chosen simply to represent one from each of the main superfamilies of "Mainly Alpha," "Mainly Beta," and "Alpha Beta." One aspect of protein stability that was not accounted for in this study was the presence of negative design in protein structures and selective pressures for folding and binding specificity (Noivirt-Brik et al. 2009; Liberles et al. 2011). Along these lines, the null distribution for our parametric bootstrap functions as a very simple force field, one that is much more computationally tractable than previous approaches (Grahnen et al. 2011) but that captures the most important factors in sequence evolution (Chi et al. 2018). This method fundamentally captures the epistatic process in generating a null distribution for positive directional selection, something that is important in differentiating compensatory processes from selection that occurs at a higher level of biological organization. Epistasis is understood to give rise to conditional selection at a local level that broadly leaves function unchanged (Eguchi et al. 2019). Capturing underlying nondirectional and nonselective processes that can masquerade as positive directional selection is ultimately necessary for accurately identifying it.

Materials and Methods

Ascertainment of Structural Information

Protein structural information, including 3D Euclidean coordinate values for each atom in the protein, was obtained from the Protein Data Bank at rcsb.org (Berman et al. 2000), for the structures with PDB IDs 2I0Q, 1D4T, and 1AX8 (Zhang et al. 1997; Poy et al. 1999; Buczek and Horvath 2006). Proteins were chosen from different superfamilies to represent a range of different features that may be present in the structure.

Within each structure, each amino acid's putative location was represented by the coordinates of its respective central carbon atom for our analyses. The null distribution generated by our parametric bootstrap (see below) relies on knowledge of SASA; thus, solvent accessibility of each amino acid in the structure was calculated with the DSSP program, via the online interface at mrs.cmbi.umcn.nl (Kabsch and Sander 1983; Touw et al. 2015). In order to obtain the relative solvent accessibility for each amino acid, the maximum possible solvent accessibility of each amino acid type was assigned according to Tien et al. (2013). Each amino acid's solvent accessibility was thus divided by its maximum possible solvent accessibility to obtain the final SASA values used for analyses.

Permutation Test

Implementation of the YT06 method was written in R according to the description in Yu and Thorne (2006), for inference on one branch of a tree, as described above in the New Approaches section. Our novel permutation test, described textually there as well, proceeds according to the pseudocode outlined in Algorithm 1.

Simulating SASA Effects and Compensatory Processes

Our permutation test is intended for detecting clustering that may occur due to rate variation that may arise due to SASA, and compensatory processes dictating that sites near substituted sites have an increased probability of themselves substituting. To simulate this, we first obtain SASA information for the structure as described above. Then, the discrete gamma model as originally described in Yang and Nielsen (2002) and frequently used to model rate variation was used to obtain putative varying rates of each site of the structure.

Specifically, we used the gamma distribution with $\alpha = 0.80$ and $\beta = 1.0$, discretized into four categories. Accordingly, the 0.125, 0.375, 0.625, and 0.875 quantiles from this distribution were obtained, and then scaled so that their true mean is equal to 1. Each site was then mapped to one of these gamma quantile values based on its SASA value, dependent upon which quartile the SASA value was in.

```
Algorithm 1 CKL20 Permutation Test
  Input:
    Cdata: Central carbon atoms from PDB
    subs: Site positions of substitutions
    radius: Sphere radius size in Å
    reps: Number of iterations
  Output:
    Permutation test P value
  Perform:
  num.neighbors <- number of other residues
    within radius for each residue
  data.fractions <- fraction of residues
    within radius of each substitution
    that are also substituted
  data.mean <- mean(data.fractions)
  for i in 1: reps do
    permuted.data <- Cdata with
      substituted site positions shuffled at
      random
    Calculate perm.fractions in the
      same manner as data.fractions,
      but on permuted.data
    perm.mean[i] <- mean(perm.fractions)
end for
return: sum(perm.mean $\geq$ data.mean)/reps
```

In other words, the sites with the smallest 25% of all SASA values were given the smallest gamma quantile value; the sites with the next highest 25% of all SASA values were given the second gamma quantile value, and so on. These quantile values were then scaled to sum to 1 across all sites, so that they could be used directly as probabilities of substitution. These probabilities were then used to make a single draw from a multinomial distribution with *n* equaling the number of sites, and probability vector equal to the scaled quantile values. This draw represents the first substitution in the protein.

Next, compensatory processes were simulated by considering the distance from existing substitutions. The smallest distance to any existing substitution was obtained for every site on the protein. The goal here is to mimic biophysical interactions between residue sites that might cause sites near other substituted sites to have an increased probability of themselves substituting. To mimic this, we calculate the square of the reciprocal of each distance, and then these values were scaled to add to 1. The scaled quantile values and scaled were then added together, with the product subtracted, mimicking the probability of the union of two independent events. These probabilities were then again used to make another single draw from a multinomial distribution, to obtain the next substitution. After each substitution, the minimum distance to any existing substitution was recalculated, and substitutions proceeded until the desired branch length was obtained. For further details, see pseudocode in Algorithm 2 in the following subsection and also the actual code supplied in the Supplementary Material online.

```
Algorithm 2 CKL20 Parametric Bootstrap
  Input:
    Cdata: Central carbon atoms from PDB
    subs: Site positions of substitutions
    radius: Sphere radius size in Å
    B: Number of bootstrap iterations
  Output:
    Parametric bootstrap P value
  Perform:
  n.subs <- length(subs)
  gam[1:4] <- $\Gamma^{-1}_{\alpha=0.8,\beta=1}(0.125, 0.375, 0.625, 0.875)$
  scaled.gam <- gam/sum(gam)
  sasa.prob <- scaled.gam values mapped to each site
    based on its SASA value quartile
  for i in 1: B do
      subs[1] <- rmultinom(1, sasa.prob)
      for j in 2: n.subs do
        min.dist <- vector of distances from each site to
            nearest substituted site
        dist.prob <- min.dist/sum(min.dist)
        new.prob <- sasa.prob + dist.prob
            - sasa.prob*dist.prob
        subs[j] <- rmultinom(1, new.prob)
    end for
    boot.fractions <- fraction of residues within radius of
      each substitution that are also substituted
    boot.95th[i] <- quantile(boot.fractions , 0.95)
  end for
  data.fractions <- fraction of residues within radius of
    each substitution that are also substituted
  data.95th <- quantile(data.fractions , 0.95)
  return: sum(boot.95th $\geq$ data.95th)/B
```

Parametric Bootstrap

The core of our novel parametric bootstrap is its null distribution generated by the same simulation described above and outlined in the pseudocode shown in Algorithm 2. To generate the null distribution for the test for positive directional selection that avoids inflated Type I Errors caused by the signals of SASA and compensatory processes, this simulation is performed repeatedly ($B = 1{,}000$ in our trials). Then, our test statistic is the 0.95 quantile of the proportion of substituted sites within each sphere.Rather than using the most simple order statistic estimator of the 0.95 quantile (e.g., with 20 data values, the 19th order statistic would be the estimate of the 0.95 quantile), we use a bias-reduced quantile estimator described as *Definition 7* in Hyndman and Fan (1996) and designated as type = 7 in the quantile function in R. This is in fact the default setting in the quantile function and has advantages over other quantile estimators; its definition can be found in the quantile function documentation.

In this manner, we generate the sampling distribution under the null hypothesis for the 0.95 quantile of the proportion of substituted sites in each sphere for a given structure. The *P* value is then the proportion of this null distribution that is at least as extreme as the 0.95 quantile from the data distribution.

Software

An R package called evolclustR is currently under preparation for submission to the Comprehensive R Archive Network, and all code used to run simulations in this manuscript are available at github.com/peterbchi/evolclustR. In parallel, python code for the same tasks is also under development and is available at github.com/wes-kosater/Py-evolclustR.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We acknowledge Laura Landweber, Tristan Stark, Andrew Ritchie, and Amanda Wilson for helpful discussions. This work was partially supported by the National Science Foundation (Grant No. 1817413). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104–2105.

Adams J, Mansfield MJ, Richard DJ, Doxey AC. 2017. Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function. *Bioinformatics* 33:1338–1345.

Ames RM, Talavera D, Williams SG, Robertson DL, Lovell SC. 2016. Binding interface change and cryptic variation in the evolution of protein–protein interactions. *BMC Evol Biol.* 16(1):40.

Anisimova M, Liberles DA. 2012. Detecting and understanding natural selection. In: Cannarozzi GM, Schneider A, editors. Codon evolution. Chapter 6. Oxford: Oxford University Press. pp. 73–97.

Benner SA, Caraco MD, Thomson JM, Gaucher EA. 2002. Planetary biology: paleontological, geological, and molecular histories of life. *Science* 296(5569):864–868.

Berglund A, Wallner B, Elofsson A, Liberles DA. 2005. Tertiary windowing to detect positive diversifying selection. *J Mol Evol.* 60(4):499–504.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res.* 28(1):235–242.

Bloom JD, Arnold FH. 2009. In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci U S A.* 106(Suppl 1):9995–10000.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinf.* 7(1):88.

Buczek P, Horvath MP. 2006. Structural reorganization and the cooperative binding of single-stranded telomere DNA in *Sterkiella nova*. *J Biol Chem.* 281(52):40124–40134.

Caro JF, Sinha MK, Kolaczynski JW, Zhang PL, Considine RV. 1996. Leptin: the tale of an obesity gene. *Perspect Diabetes* 45(11):1455–1462.

Chi PB, Kim D, Lai JK, Bykova N, Weber CC, Kubelka J, Liberles DA. 2018. A new parameter-rich structure-aware mechanistic model for amino acid substitution during evolution. *Proteins* 86(2):218–228.

Chi PB, Liberles DA. 2016. Selection on protein structure, interaction, and sequence. *Protein Sci.* 25(7):1168–1178.

Chothia C, Lesk AM. 1982. Evolution of proteins formed by β-sheets: I. Plastocyanin and azurin. *J Mol Biol.* 160(2):309–323.

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. 2017. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* 45(D1):D289–D295.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet.* 17(2):109–121.

Eguchi Y, Bilolikar G, Geiler-Samerotte K. 2019. Why and how to study genetic changes with context-dependent effects. *Curr Opin Genet Dev.* 58–59:95–102.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol.* 13(5):685–690.

Fares MA, Elena SF, Ortiz J, Moya A, 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(5):509–521.

Gaucher EA, Miyamoto MM, Benner SA. 2003. Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein. *Genetics* 163(4):1549–1553.

Goldstein RA, Pollock DD. 2016. The tangled bank of amino acids. *Protein Sci.* 25(7):1354–1362.

Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. 2011. Biophysical and structural considerations for protein sequence evolution. *BMC Evol Biol.* 11(1):361.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence in a protein family. *Bioinformatics* 18(3):500–501.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol.* 15(7):910–917.

Hyndman RJ, Fan Y. 1996. Sample quantiles in statistical packages. *Am Stat.* 50(4):361–365.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.

Lai J, Jin J, Kubelka J, Liberles DA. 2012. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. *J Mol Biol.* 422(3):442–459.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol.* 136(3):225–230.

Liang H, Zhou W, Landweber LF. 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Res.* 34(Web Server):W382–W384.

Liberles DA, Tisdell MD, Grahnen JA. 2011. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. *Proc Biol Sci.* 278(1714):1930–1935.

Lynch M. 2010. Scaling expectations for the time to establishment of complex adaptations. *Proc Natl Acad Sci U S A.* 107(38):16577–16582.

Mantzoros CS. 1999. The role of leptin in human obesity and disease: a review of current evidence. *Ann Intern Med.* 130(8):671–680.

Monit C, Goldstein RA. 2018. SubRecon: ancestral reconstruction of amino acid substitutions along a branch in a phylogeny. *Bioinformatics* 34(13):2297–2299.

Noivirt-Brik O, Horovitz A, Unger R. 2009. Trade-off between positive and negative design of protein stability: from lattice models to real proteins. *PLoS Comput Biol.* 5(12):e1000592.

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol.* 4(11):e1000214.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci U S A.* 109(21):E1352–E1359.

Poy F, Yaffe MB, Sayos J, Saxena K, Morra M, Sumegi J, Cantley LC, Terhorst C, Eck MJ. 1999. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol Cell* 4(4):555–561.

Schilling MF, Doi JA. 2014. A coverage probability approach to finding an optimal binomial confidence procedure. *Am Stat.* 68(3):133–145.

Siltberg J, Liberles DA. 2002. A simple covarion-based approach to analyze nucleotide substitution rate. *J Evol Biol.* 15(4):588–594.

Spielman SJ, Wilke CO. 2016. Extensively parameterized mutation-selection models reliably capture site-specific selective constraint. *Mol Biol Evol*. 33(11):2990–3002.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol*. 16(10):1315–1328.

Teufel AI, Ritchie AM, Wilke CO, Liberles DA. 2018. Using the mutation-selection framework to characterize selection on protein sequences. *Genes* 9(8):409.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* 8(11):e80635.

Touw WG, Coos B, Jon B, te Beek TAH, Krieger E, Joosten RP, Gert V. 2015. A series of PDB related databases for everyday needs. *Nucleic Acids Res*. 43(D1):D364–D368.

Tusche C, Steinbrück L, McHardy A. 2012. Detecting patches of protein sites of influenza A viruses under positive selection. *Mol Biol Evol*. 29(8):2063–2071.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15(5):568–573.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19(6):908–917.

Yu J, Thorne JL. 2006. Testing for spacial clustering of amino acid replacements within protein tertiary structure. *J Mol Evol*. 62(6):682–692.

Zhang F, Basinski MB, Beals JM, Briggs SL, Churgay LM, Clawson DK, DiMarchi RD, Furman TC, Hale JE, Hsiung HM, et al. 1997. Crystal structure of the obese protein leptin-E100. *Nature* 387(6629):206–209.