A Chemical Distance Based Test for Positive Darwinian Selection

T. Pupko *, R. Sharan **, M. Hasegawa *, R. Shamir **, and D. Graur ***

Abstract. There are very few instances in which positive Darwinian selection has been convincingly demonstrated at the molecular level. In this study, we present a novel test for detecting positive selection at the amino-acid level. In this test, amino-acid replacements are characterized in terms of chemical distances, i.e., degrees of dissimilarity between the exchanged residues in a protein. The test identifies statistically significant deviations of the mean observed chemical distance from its expectation, either along a phylogenetic lineage or across a subtree. The mean observed distance is calculated as the average chemical distance over all possible ancestral sequence reconstructions, weighted by their likelihood. Our method substantially improves over previous approaches by taking into account the stochastic process, tree phylogeny, among site rate variation, and alternative ancestral reconstructions. We provide a linear time algorithm for applying this test to all branches and all subtrees of a given phylogenetic tree. We validate this approach by applying it to two well-studied datasets, the MHC class I glycoproteins serving as a positive control, and the house-keeping gene carbonic anhydrase I serving as a negative control.

1 Introduction

The neutral theory of molecular evolution maintains that the great majority of evolutionary changes at the molecular level are caused not by Darwinian selection acting on advantageous mutants, but by random fixation of selectively neutral or nearly neutral mutants [12]. There are very few cases in which positive Darwinian selection was convincingly demonstrated at the molecular level [10, 22, 34, 30, 23]. These cases are vital to understanding the link between sequence variability and adaptive evolution. Indeed, it has been estimated that positive selection has occurred in only 0.5% of all protein-coding genes [2].

The most widely used method for detecting positive Darwinian selection is based on comparing synonymous and nonsynonymous substitution rates between nucleotide sequences [17]. Synonymous substitutions are assumed to be selectively neutral. If only purifying selection operates, then the rate of synonymous

* The Institute of Statistical Mathematics 4-6-7 Minami Azabu, Minato ku, Tokyo, Japan. {tal,hasegawa}@ism.ac.jp.

** School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. {roded,rshamir}@post.tau.ac.il.

*** Department of Zoology, Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv, Israel. graur@post.tau.ac.il.

substitution should be higher than the rate of nonsynonymous substitution. In the few cases where the opposite pattern was observed, positive selection was invoked as the likely explanation (see, e.g., [33, 14]). One critical shortcoming of this method is that it requires estimating numbers of synonymous substitutions. Because of saturation, such estimation is virtually impossible when the sequences under study are evolutionarily distant. The estimation is problematic even if close species are concerned. For example, saturation of substitutions in the third position is evident even when comparing cytochrome *b* sequences among species within the same mammalian order [5].

Another method for detecting positive selection is searching for parallel and convergent replacements. It is postulated that such molecular changes in different parts of a phylogenetic tree can only be explained by the same selective pressure being exerted on different taxa that became exposed to the same conditions [23, 32]. This method is limited to the few cases in which the same type of positive Darwinian selection occurs in two or more unrelated lineages.

A third method of detecting positive selection is based on comparing conservative and radical nonsynonymous differences [9, 7]: Nonsynonymous sites are divided into conservative sites and radical sites based on physicochemical properties of the amino-acid side chain, such as volume, hydrophobicity, charge or polarity. Radical and conservative sites and radical and conservative replacements are separately counted, and the number of radical replacements per radical site is compared to the number of conservative replacements per conservative site. If the former ratio is significantly higher than the latter, then positive Darwinian selection is invoked. By using this method, positive selection was inferred for the antigen binding cleft of class I major-histocompatibility-complex (MHC) glycoproteins [9] and rat olfactory proteins [8]. This method for detecting positive selection has the advantage that distant protein sequences can be compared even when synonymous substitutions are saturated. Another virtue of this method is its flexibility with respect to the sequence characteristic tested. For example, if we suspect that polar replacements might be advantageous, a test can be applied with radical replacements defined as those occurring between amino-acids with polar and non-polar residues only. However, this method also has many shortcomings. First, no correction for multiple substitutions is applicable [7]. Second, each codon in a pair of aligned amino-acid is used twice: Once for estimating the number of radical and conservative sites, and once for estimating the number of radical and conservative replacements. Third, the method treats replacements between different amino-acids as equally probable. Fourth, the method ignores branch lengths, implicitly assuming independence of the replacement probabilities between the amino acids and the evolutionary distance between the sequences under study. Finally, the phylogenetic signal is ignored, i.e., the test is applied to pairwise sequence comparisons rather than testing hypotheses on a phylogenetic tree.

The test for positive selection proposed in this study overcomes the shortcomings of the radical-conservative test. Our test incorporates a probabilistic framework for dealing with radical vs. conservative replacements. It applies a novel method for averaging over ancestral sequence assignments, weighted by their likelihood, thus eliminating bias which might result from assuming a specific ancestral sequence reconstruction. The rationale underlying our proposed test is that the evolutionary acquisition of a new function requires a significant change of the biochemical properties of the amino-acid sequence [7]. To quantify this biochemical difference between two amino-acid sequences, we define a chemical distance measure based on, e.g., Grantham's matrix [4]. Our test identifies large deviations of the mean observed chemical distance from the expected distance along a branch or across a subtree in a phylogenetic tree. If the observed chemical distance between two sequences significantly exceeds the chance expectation, then it is unlikely that this is the result of random genetic drift, and positive Darwinian selection should be invoked.

Based on the assumed stochastic process, the tree topology and its branch lengths, we calculate both the mean observed chemical distance and its underlying distribution for the branch or subtree in question. The mean observed chemical distance is calculated as the average chemical distance over all ancestral sequence reconstructions, weighted by their likelihood, thus, eliminating possible bias in a calculation based on a particular ancestral sequence reconstruction. The underlying distribution of this random variable is calculated using the JTT stochastic model [11], the tree topology and branch lengths, taking into account among site rate variation. We provide a linear time algorithm to perform this test for all branches and subtrees of a phylogenetic tree with n leaves.

In order to validate our approach, we applied it to two control datasets: Class I major-histocompatibility-complex (MHC) glycoproteins, and carbonic anhydrase I. These datasets were chosen since they were already used as standard positive control (MHC) and negative control (carbonic anhydrase) for positive selection [24]. For the MHC class I dataset, as reported in [9], we observe positive selection which favors charge replacements only when applying the test to the subsequences of the binding cleft ($P < 0.01$). In addition we observe positive selection which favors polarity replacements when using Grantham's polarity indices [4] ($P < 0.01$). When applying the test to the carbonic anhydrase dataset, no positive selection is observed.

The paper is organized as follows: Section 2 contains the notations and terminology used in the paper. Section 3 presents the new test for positive Darwinian selection. Section 4 describes the application of this test to the two control datasets. Finally, Section 5 contains a summary and a discussion of our approach.

2 Preliminaries

Let $\mathcal{A}$ be the set of 20 amino-acids. We assume that sequence evolution follows the JTT probabilistic reversible model [11]. For amino-acid sequences this model is described by a 20×20 matrix M, indicating the relative replacement rates of amino-acids, and a vector $(P_A, \ldots, P_Y)$ of amino-acid frequencies. For each branch of length t and amino-acids i and j, the $i \to j$ replacement probability, denoted by $P_{ij}(t)$, can be calculated from the eigenvalue decomposition of M [13].

(In practice, an approximation to $P_{ij}(t)$ is used to speedup the computation [19].) We denote by $f_{ij}(t) = P_i \cdot P_{ij}(t) = P_j \cdot P_{ji}(t)$ the probability of observing i and j in the same position in two aligned sequences of evolutionary distance t.

Let s be an amino-acid sequence. The amino-acid at position i in s is denoted by s_i. For two amino-acids $a, b \in \mathcal{A}$, we denote their *chemical distance* by $d(i, j)$. We assume we have a table of chemical distances between every pair of amino-acids. One such distance is Grantham's chemical distance [4]. (Other similar distance measures appear in [27, 16].) This chemical distance measures the difference between two amino-acids in terms of their volume, polarity and composition of the side chain. The choice of which distance measure to use, reflects the type of test we wish to perform. For example, Grantham's distance is appropriate when testing whether the replacements between the sequences under question are more radical with respect to a range of physicochemical properties (volume, charge and composition of the side chain). For testing whether polarity differences between sequences are higher than the random expectation, two distance measures are applicable: The first measure is based on dividing the set of amino-acids into 2 categories: Polar (C,D,E,H,K,N,Q,R,S,T,W,Y) and non-polar (the rest). The polarity distance between two amino-acids is then defined as 1 if one is polar and the other is not, and 0 otherwise [9]. The second polarity distance is defined as the absolute difference between the polarity indices of the two amino-acids, and yields real values [4]. For testing charge differences 3 categories of amino-acids are defined: Positive (H,K,R), negative (D,E) and neutral (all other). The charge distance between two amino-acids is defined as 1 if they belong to two different categories, and 0 if they belong to the same category [9].

We define the *average chemical distance* between two sequences s^1 and s^2 of length N as the average of the chemical distances between pairs of amino-acids occupying the same position in a gapless alignment of s^1 and s^2:

$$D(s^1, s^2) = \frac{1}{N} \sum_{i=1}^{N} d(s_i^1, s_i^2)$$

Let $\mathcal{T}$ be an unrooted phylogenetic tree. For a node v, we denote by $N(v)$ the set of nodes adjacent to v. For an edge $(u, v) \in \mathcal{T}$ we denote by $t(u, v)$ the length of the branch connecting u and v.

3 A Test for Positive Darwinian Selection

In this section we describe a new test for detecting positive Darwinian selection. The input to the test is a set of gap-free aligned sequences and a phylogenetic tree for them. We first present a version of our test for a pair of known sequences. We then extend this method to test positive selection on specific branches of a phylogenetic tree under study. Finally we generalize the test to subtrees (clades) and incorporate among site rate variation.

3.1 Testing Two Known Sequences

Let s^1 and s^2 be two amino-acid sequences of length N and evolutionary distance t. The underlying distribution of $D(s^1, s^2)$ is inferred as follows. The expectation of the chemical distance at position i is:

$$E(d(s_i^1, s_i^2)) = \sum_{a,b \in \mathcal{A}} d(a,b) f_{ab}(t)$$

Assuming that the distribution of the chemical distance in each position is identical, we obtain

$$E(D(s^1, s^2)) = \frac{1}{N} \sum_{i=1}^{N} E(d(s_i^1, s_i^2)) = E(d(s_1^1, s_1^2))$$

The variance of the chemical distance at position i is:

$$V(d(s_i^1, s_i^2)) = E(d(s_i^1, s_i^2)^2) - E(d(s_i^1, s_i^2))^2 = \sum_{a,b \in \mathcal{A}} d(a,b)^2 f_{ab}(t) - E(d(s_i^1, s_i^2))^2$$

and assuming further that sequence positions are independent, we obtain

$$V(D(s^1, s^2)) = \frac{V(d(s_1^1, s_1^2))}{N}$$

For practical values of N, $D(s^1, s^2)$ is approximately normally distributed with expectation $E(D(s^1, s^2)$ and standard deviation $\sqrt{V(D(s^1, s^2))}$. This allows us to compute for each observed chemical distance d, the probability that it occurs by chance, i.e., its p-value. If the observed chemical distance is found above the 0.99 percentile of the normal distribution, we conclude that replacements in these two sequences significantly deviate from the expectation, and suggest positive selection to explain this phenomenon.

3.2 Testing a Tree Lineage

Here we first describe a general method to apply pairwise tests to a phylogenetic tree. Suppose that we wish to test a statistical hypothesis on a specific branch of the phylogenetic tree. Also suppose that we have a procedure to test our hypothesis on a pair of known sequences, like the procedure described above. In order to test our hypothesis on a specific branch, we could first infer the corresponding ancestral sequences (using, e.g., maximum likelihood estimation [20]) and then check our hypothesis. Inferring ancestral sequences and then using these sequences as observations was done in e.g., [31]. This approach, which treats estimated reconstructions as observations may lead to erroneous conclusions due to bias in the reconstruction. A more robust approach is to average over all possible reconstructions, weighted by their likelihood. By averaging over all possible ancestral assignments, we extend our test to hypothesis testing on a phylogenetic tree, without possible bias that results from reconstructing particular sequences at internal tree nodes.

We describe in the following how to apply our test to a specific branch connecting nodes x and y in a tree $\mathcal{T}$. Since we assume that different positions evolve independently we restrict the subsequent description to a single site.

Each branch $(u, v) \in \mathcal{T}$ partitions the tree into two subtrees. Let $L(u, v, a)$ denote the likelihood of the subtree which includes v, given that v is assigned the amino-acid a. $L(u, v, a)$ can be computed by the following recursion equation:

$$L(u, v, a) = \prod_{w \in (N(v) \setminus \{u\})} \{\sum_{b \in \mathcal{A}} P_{ab}(t(v, w)) \cdot L(v, w, b)\}$$

For a leaf v at the base of the recursion we have $L(u, v, a) = 1$, assuming amino-acid a in v, and $L(u, v, a) = 0$ otherwise.

The likelihood of $\mathcal{T}$ is thus:

$$P_{\mathcal{T}} = \sum_{a,b \in \mathcal{A}} f_{ab}(t(u, v)) \cdot L(u, v, b) \cdot L(v, u, a)$$

where (u, v) is any branch of $\mathcal{T}$.

Suppose that the data at the leaves of $\mathcal{T}$ is $\boldsymbol{w} = (w_1, \ldots, w_n)$. The *mean observed chemical distance* for a given branch $(x, y) \in \mathcal{T}$ can be calculated as follows:

$$\begin{aligned} D(x, y) &= \sum_{a,b \in \mathcal{A}} Pr(x = a, y = b | \boldsymbol{w}) \cdot d(a, b) \\ &= \frac{1}{P_{\mathcal{T}}} \sum_{a,b \in \mathcal{A}} \{d(a, b) \cdot f_{ab}(t(x, y)) \cdot L(x, y, b) \cdot L(y, x, a)\} \end{aligned}$$

It remains to compute the null distribution of this statistic. The expectation of $D(x, y)$ (with respect to all possible leaf-assignments) is as follows:

$$\begin{aligned} E(D(x, y)) &= \sum_{\boldsymbol{z} \in \mathcal{A}^n} Pr(\boldsymbol{z}) \sum_{a,b \in \mathcal{A}} Pr(x = a, y = b | \boldsymbol{z}) \cdot d(a, b) \\ &= \sum_{a,b \in \mathcal{A}} d(a, b) \sum_{\boldsymbol{z} \in \mathcal{A}^n} Pr(\boldsymbol{z}) \cdot Pr(x = a, y = b | \boldsymbol{z}) \\ &= \sum_{a,b \in \mathcal{A}} d(a, b) \cdot f_{ab}(t(x, y)) \end{aligned}$$

We conclude that $E(D(x, y))$ is the same as in the known-sequences case. For the variance of $D(x, y)$ we have no explicit formula. Instead, we evaluate $V(D(x, y))$ using parametric bootstrap [25]. Specifically, we draw at random many assignments of amino-acids to the leaves of $\mathcal{T}$ and compute $D(x, y)$ for each of them, thereby evaluating its variance. An assignment to the leaves of $\mathcal{T}$ is obtained as follows: We first root $\mathcal{T}$ at an arbitrary node r. We then draw at random an amino-acid for r according to the amino-acid frequencies. We next draw amino-acids for each child of r according to the appropriate replacement probabilities of our model, and continue in this manner till we reach the leaves.

Finally, since $D(x,y)$ is approximately normally distributed, we can compute a p-value for the test, which is simply $Pr(Z \geq \frac{D(x,y)-E(D(x,y))}{\sqrt{V(D(x,y))}})$ where $Z \sim Normal(0,1)$. Note, that if the test is applied to several (or all) branches of the tree, then the significance level of the test should be corrected in accordance with the number of tests performed, e.g., using Bonforroni's correction which multiplies the p-value by the number of branches tested.

The algorithm for testing the branches of a phylogenetic tree $\mathcal{T}$ is summarized in Figure 1. For each branch $(x,y) \in \mathcal{T}$ the algorithm outputs the p-value of the test for that branch. In the actual implementation we used $M = 100$.

PositiveSelectionTest($\mathcal{T}$):
- Root $\mathcal{T}$ at an arbitrary node r.
- Draw M assignments to the leaves of $\mathcal{T}$ using parametric bootstrap.
- Traverse $\mathcal{T}$ bottom-up, computing along the way for every $(u,v) \in \mathcal{T}, a \in \mathcal{A}$ the value of $L(u,v,a)$, where u is the parent of v.
- Traverse $\mathcal{T}$ top-down, computing along the way for every $(u,v) \in \mathcal{T}, a \in \mathcal{A}$ the value of $L(v,u,a)$, where u is the parent of v.
- **For** every $(x,y) \in \mathcal{T}$ **do**:
  - Calculate $D(x,y)$ and $E(D(x,y))$.
  - Evaluate $V(D(x,y))$.
  - **Output** the p-value for the branch (x,y).

Fig. 1. An algorithm for testing the branches of a phylogenetic tree $\mathcal{T}$.

Theorem 1. *For a given phylogenetic tree $\mathcal{T}$ with n leaves, the algorithm tests all branches of $\mathcal{T}$ in $O(n)$ time.*

Proof. Given $L(u,v,a)$ for every $(u,v) \in \mathcal{T}$ and every $a \in \mathcal{A}$, it is straightforward to compute $D(u,v)$ for all $(u,v) \in \mathcal{T}$ in linear time. The computation of $E(D(u,v))$ and $V(D(u,v))$ is clearly linear. The complexity follows.

3.3 Testing a Subtree

In this section we present an extension of our method to test subtrees of a given phylogenetic tree $\mathcal{T}$. This is motivated by the consideration that if a clade of contemporary sequences has undergone positive Darwinian selection, we cannot necessarily assume that this selection occurred solely along the branch leading to that clade. A reasonable scenario is that the selection was continuous and occurred along several or all branches of the subtree corresponding to this clade. In such a case, the test we have just described may not detect any significant positive selection along any specific branch. Hence, we are interested at testing for positive selection across subtrees as well.

For a subtree $\mathcal{T}'$ of $\mathcal{T}$, we define the mean observed chemical distance $D(\mathcal{T}')$ as the average observed distance along its branches (i.e., the sum of the observed distance for each branch divided by the number of branches in $\mathcal{T}'$). Clearly, the expectation of $D(\mathcal{T}')$ is equal to the average expectation of the branches of $\mathcal{T}'$. The variance of $D(\mathcal{T}')$ can be evaluated using parametric bootstrap. We then use the normal approximation to compute a p-value for this test. We conclude:

Theorem 2. *For a given phylogenetic tree $\mathcal{T}$ with n leaves, the complexity of testing all its subtrees is $O(n)$.*

3.4 Introducing Among Site Rate Variation

The rate of evolution is not constant among amino-acid sites [28]. Consider two sequences of length N. Suppose that there are on average l replacements per site between these sequences. This means that we expect lN replacements altogether. How many replacements should we expect at each particular site? Naive models assume that the variation of mutation rate among sites is zero, i.e., that all sites have the same replacement probability. Models that take this Among Site Rate Variation (ASRV) into account assume that at the j-th position the average number of replacement is $lr[j]$, where each $r = r[j]$ is a rate parameter drawn from some probability distribution. Since the mean rate over all sites is l, the mean of r is equal to 1. Yang suggested the gamma distribution with parameters α and β as the distribution for r, and since the mean of the gamma distribution α/β, must be equal to 1, $\alpha = \beta$ [28], that is:

$$f(r;\alpha,\beta) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} e^{-\alpha r} r^{\alpha-1}$$

Maximum likelihood models incorporating ASRV are statistically superior to those assuming among site rate homogeneity [28]. They also help avoiding the severe underestimation of long branch lengths that can occur with the homogeneous models [15].

In this study we use the discrete gamma model with k categories whose means are $r_1, \ldots, r_k$ to approximate the continuous gamma distribution [29]. The categories are selected so that the probabilities of r falling into each category are equal. We thus assume that $Pr(r = r_i) = 1/k$.

The incorporation of the discrete gamma model in our test is straightforward. For each rate category i we calculate both the expected and observed chemical distance, given that the rate is r_i. This is equivalent to making the computation in the homogeneous case, where all branch lengths are multiplied by the factor r_i. The observed and expected chemical distance for each branch are then averaged over all rate categories.

4 Biological Results

In order to validate our approach, we applied it to two control datasets: Class I major-histocompatibility-complex (MHC) glycoproteins, and carbonic anhydrase I. We have chosen to analyze these datasets since they were already used as standard positive control (MHC) and negative control (carbonic anhydrase) for positive selection tests [24].

The datasets contain aligned sequences (all sequences are of the same length, and the best alignment is gapless). Phylogenetic trees were constructed using the MOLPHY software [1], with the neighbor-joining method [21] for MHC class I, and with the maximum likelihood method for carbonic anhydrase I. The reason for the use of two tree construction methods is that in the MHC case we are dealing with 42 sequences and, therefore, an exhaustive maximum likelihood approach is impractical. Branch lengths for each topology were estimated using the maximum likelihood method [3] with the JTT stochastic model [11], assuming that the rate is discrete gamma distributed among sites with 4 rate categories.

4.1 MHC Class I

The primary immunological function of MHC class I glycoproteins is to bind and "present" antigenic peptides on the surface of cells, for recognition by antigen-specific T cell receptors. MHC class I glycoproteins are expressed on the surface of most cells and are recognized by CD8-positive cytotoxic T cells, an essential step for initiating the elimination of virally infected cells by T-cell mediated lysis. These molecules are very polymorphic, and it was claimed that this polymorphism is the result of positive Darwinian selection that operates on the antigen-binding cleft [9]. Using pairwise comparisons of sequences, it was shown that the proportion of nonsynonymous differences in the antigen-binding cleft that cause charge changes was significantly higher than the proportion that conserve charge. This suggests that peptide binding is at the basis of the positive selection acting on these loci [7].

Following [9] we analyzed 42 human MHC class I sequences from three allelic groups: HLA-A, -B, and -C loci. Most of these sequences are not available in Genbank, and were copied from Parham et al. [18]. The length of each MHC class I sequence is 274 amino acids. The binding site is a subsequence of 29 residues [18]. The phylogenetic tree for MHC class I sequences is given in Figure 2. The α parameter found for this tree was 0.24. When our clade-based test was applied to the whole tree, no indication for positive selection was found. The respective z-scores are shown in Table 1.

When we applied our test to the binding site only, positive selection was found with very high confidence ($P < 0.001$). The respective z-scores are shown in Table 1. However, it might be argued, that when only the binding site part of the sequence is analyzed, the branch lengths estimated for the whole sequences are irrelevant. Since it is known that the rate of evolution in the binding site is faster relative to the rest of the sequence, the branch lengths estimated from the whole sequences are underestimated. This underestimation can result in a false

Dataset/Distance	Grantham	Charge	Polarity	
			Grantham	Hughes et al. [9]
Whole	-1.30	0.01	-1.25	1.10
Cleft	**9.38**	**9.32**	**13.23**	**5.79**
Cleft & cleft-based lengths	1.08	**3.14**	**2.78**	0.01

Table 1. A list of z-scores for each of the tests performed on the MHC class I dataset. The first row contains scores with respect to whole sequences. The second row contains results with respect to the binding cleft subsequences, with branch lengths as for the whole sequences. The third row contains results with respect to the binding cleft subsequences, with branch lengths reestimated on this part of the sequence only. Significant z-scores (p-value< 0.01) appear in bold-face.

positive conclusion of positive selection, since we expect in this case an excess of radical replacements. To overcome this problem, branch lengths were reestimated on the binding site part of the sequence only. Significant excess of polar and charge replacements were found also with these new estimates ($P < 0.01$). The corresponding z-scores are shown in Table 1. We note, that using the 0-1 polarity distance of [9], we found no evidence for positive selection. On the other hand, when we used Grantham's polarity indices [4], significant deviations from the random expectations were observed (see Table 1). The latter distance measure is clearly more accurate since it is not restricted to 0-1 values. We conclude that there is a significant excess in both charge and polar replacements, and not only in charge replacements, as reported in [9].

Finally, we tested specific branches in the tree to find those branches which contribute the most to the excess of charge replacements. Branches whose corresponding p-value was found to be smaller than 0.01 appear in bold-face in Figure 2. We note, that since we have no prior knowledge of which branches are expected to show excess of charge replacements, these p-values should be scaled according to the number of branches tested. Nevertheless, these high scoring branches lie all in the subtrees corresponding to the A and B alleles, matching the findings of Hughes et al. who report positive selection for these alleles only [9].

4.2 Carbonic anhydrase I

This dataset comprises of 6 sequences of the carbonic anhydrase I house-keeping gene, for which there is no evidence of positive selection [24]. The carbonic anhydrase I sequences were the same as in [24], except that amino-acid sequences were used instead of nucleotide sequences. Sequence accession numbers are: JN0835 (*Pan troglodytes*), JN0836 (*Gorilla gorilla*), P00915 (*Homo sapiens*), P35217 (*Macaca nemestrina*), P48282 (*Ovis aries*) and P13634 (*Mus musculus*). The maximum likelihood estimate of the α parameter for this dataset was 0.52.

When analyzing carbonic anhydrase I sequences, no evidence for positive selection was found. This was true, irrespective of the distance measures we used: Grantham (z-score$= 0.01$), Grantham's polarity (z-score$=-1.04$), Hughes et al. polarity (z-score$= -0.49$), and charge (z-score$= -1.73$).

5 Discussion

Natural selection may act to favor amino acid replacements that change certain properties of amino acids [7]. Here we propose a method to test for such selection. Our method takes into account the stochastic model of amino-acid replacements, among site rate variation and the phylogenetic relationship among the sequences under study. The method is based on identifying large deviations of the mean observed chemical distance between two proteins from the expected distance. Our test can be applied to a specific branch of a phylogenetic tree, to a clade in the tree or, alternatively, over all branches of the phylogenetic tree. The calculation of the mean observed chemical distance is based on a novel procedure for averaging the chemical distance over all possible ancestral sequence reconstructions weighted by their likelihood. This results in an unbiased estimate of the chemical distance along a branch of a phylogenetic tree. The underlying distribution of this random variable is calculated using the JTT model, taking into account among site rate variation. We give a linear time algorithm to perform this test for all branches and subtrees of a given phylogenetic tree.

Two variants of the test are presented: The first is a statistical test of a single branch in a phylogenetic tree. Positive selection along a tree lineage can be the result of a specific adaptation of one taxon to some special environment. In this case, the branch in question is known a priori, and the branch-specific test should be used. Alternatively, if the selection constraints are continuous, as for example, the selection that promotes diversity among alleles of the MHC class I, the test should be applied to all the sequences under the assumed selection pressure - a clade-based test.

We validated our method on two datasets: Carbonic anhydrase I sequences served as a negative control, and the cleft of MHC class I sequences as a positive control. MHC class I sequences were previously shown to be under positive selection pressure, acting to favor amino-acid replacements that are radical with respect to charge.

There are, however, some limitations to our method. The method relies heavily on an assumed stochastic model of evolution. If this model underestimates branch lengths, one might get false positive results. It is for this reason that it is important to estimate branch lengths under realistic models, taking into account among site rate variation. Furthermore, if the test is applied to specific parts of the protein, such as an alpha helix, a replacement matrix that is specific for this part might be preferable over the more general JTT model used in this study (see [26]). One might claim that if excess of, say, polar replacements is found, it should not be interpreted as indicative of positive selection, but rather, as an indication that a more sequence-specific amino-acid replacement model is required. In MHC class I glycoproteins, however, other lines of evidence [9, 24] suggest positive Darwinian selection.

In the future, we plan to make the test more robust by accommodating uncertainties in branch lengths and topology. This can be achieved by Markov-Chain Monte-Carlo methods [6]. The sensitivity of our test to different assumptions regarding the stochastic process and the phylogenetic tree will be better understood when more datasets are analyzed.

Acknowledgements

We thank Hirohisa Kishino and Yoav Benjamini for their suggestions regarding the statistical analysis. The first author was supported by a JSPS fellowship. The second author was supported by an Eshkol fellowship from the Ministry of Science, Israel. The fourth author was supported in part by the Israel Science Foundation (grant number 565/99). This study was also supported by the Magnet Da'at Consortium of the Israel Ministry of Industry and Trade and a grant from Tel Aviv Univeristy (689/96).

References

1. J. Adachi and M. Hasegawa. Molphy: programs for molecular phylogenetics based on maximum likelihood, version 2.3. Technical report, Institute of Statistical Mathematics, Tokyo, Japan, 1996.
2. T. Endo, K. Ikeo, and T. Gojobori. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.*, 13:685–690, 1996.
3. J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.
4. R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.
5. K.M. Halanych and T.J. Robinson. Multiple substitutions affect the phylogenetic utility of cytochrome b and 12 rDNA data: Examining a rapid radiation in Leporid (Lagomorpha) evolution. *J. Mol. Evol.*, 48(3):369–379, 1999.
6. J.P. Huelsenbeck, B. Rannala, and J.P. Masly. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288:2349–2350, 2000.
7. A.L. Hughes. *Adaptive Evolution of Genes and Genomes.* Oxford University Press, New-York, 1999.
8. A.L. Hughes and M.K. Hughes. Adaptive evolution in the rat olfactory receptor gene family. *J. Mol. Evol.*, 36:249–254, 1993.
9. 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. *Mol. Biol. Evol*, 7:515–524, 1990.
10. A.L. Hughes and M. Yeager. Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.*, 32:415–435, 1998.
11. D.T. Jones, W.R. Taylor, and J.M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8:275–282, 1992.
12. M. Kimura. *The neutral theory of molecular evolution.* Cambridge university press, Cambridge, 1983.
13. H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.*, pages 151–160, 1990.

14. Y.H. Lee, T. Ota, and V.D. Vacquier. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.*, 12:231–238, 1995.
15. P. Lio and N. Goldman. Models of molecular evolution and phylogeny. *Genome research*, 8:1233–1244, 1998.
16. T. Miyata, S. Miyazawa, and T. Yashunaga. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, 12:219–236, 1979.
17. M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3:418–426, 1986.
18. P. Parham, C.E. Lomen, D.A. Lawlor, J.P. Ways, N. Holmes, H.L. Coppin, R.D. Salter, A.M. Won, and P.D. Ennis. Nature of polymorphism in HLA-A, -B, and -C molecules. *Proc. Natl. Acad. Sci. USA*, 85:4005–4009, 1998.
19. T. Pupko and I. Pe'er. Maximum likelihood reconstruction of ancestral amino-acid sequences. In S. Miyano, R. Shamir, and T. Takagi, editors, *Currents in Compuational Molecular Biology*, pages 184–185. Universal Academy Press, 2000.
20. T. Pupko, I. Pe'er, R. Shamir, and D. Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, 17(6):890–896, 2000.
21. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*, 4:406–425, 1987.
22. S.A. Seibert, C.Y. Howell, M.K. Hughes, and A.L. Hughes. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.*, 12:803–813, 1995.
23. C.B. Stewart and A.C. Wilson. Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symp. Quant. Biol.*, 52:891–899, 1987.
24. W.J. Swanson, Z. Yang, M.F. Wolfner, and C.F. Aquadro. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. USA*, 98:2509–2514, 2001.
25. D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B.K. Mable, editors, *Molecular systematics, 2nd Ed.*, pages 407–514. Sinauer Associates, Sunderland, MA, 1995.
26. J.L. Thorne, N. Goldman, and D.T. Jones. Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, 13:666–673, 1996.
27. H.C. Wang, J. Dopazo, and J.M. Carazo. Self-organizing tree growing network for classifying amino acids. *Bioinformatics*, 14:376–377, 1998.
28. Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.
29. Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–314, 1994.
30. Z. Yang. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.*, 51:423–432, 2000.
31. Z. Yang, S. Kumar, and M. Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650, 1995.
32. J. Zhang and S. Kumar. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.*, 14:527–536, 1997.
33. J. Zhang, S. Kumar, and M. Nei. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol. Biol. Evol.*, 14:1335–1338, 1997.
34. J. Zhang, H.F. Rosenberg, and M. Nei. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA*, 95:3708–3713, 1998.

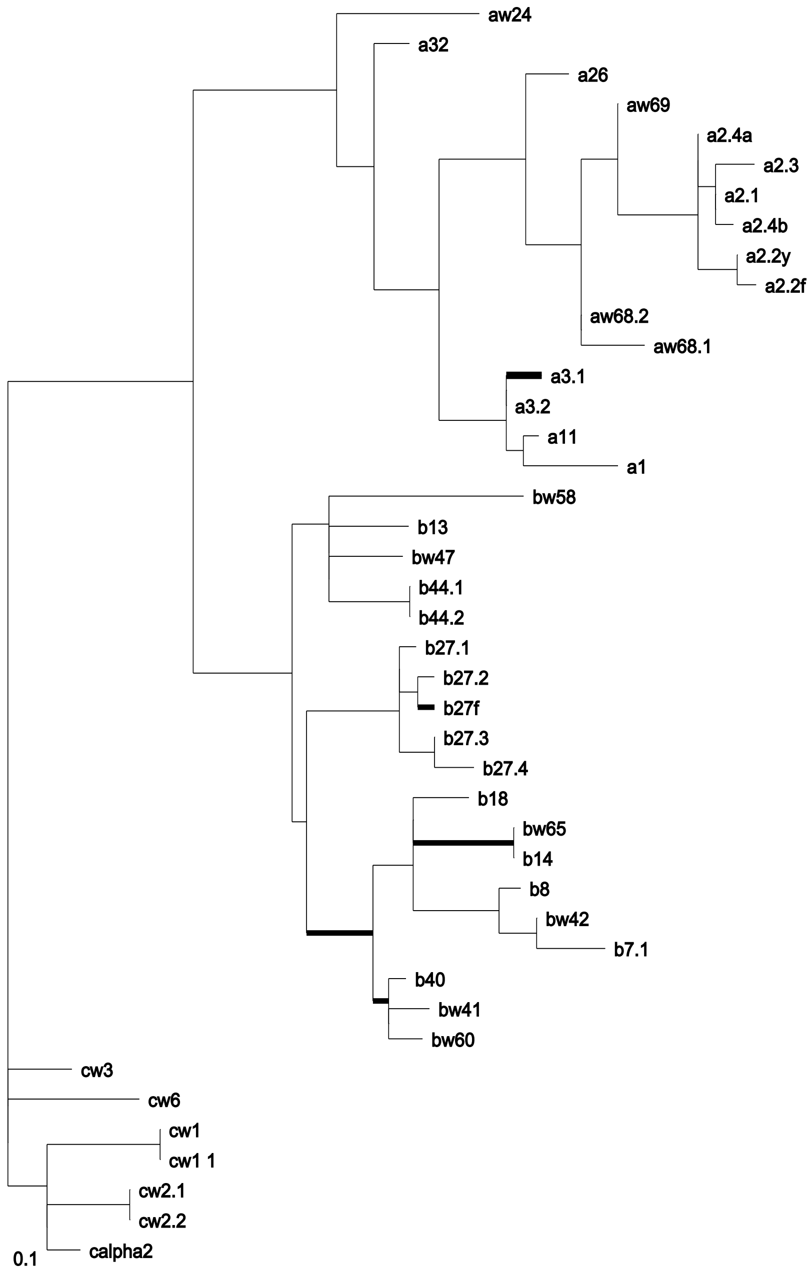

Fig. 2. A phylogenetic tree for MHC class I sequences. Species labels are as in [9]. The tree topology was estimated by using whole sequences. Branch lengths were estimated for the cleft subsequences only. Each branch was subjected to the positive selection test on the cleft subsequences. Branches in bold-face indicate p-value< 0.01.