ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Gene 319 (2003) 127–135

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Detecting excess radical replacements in phylogenetic trees☆

Tal Pupko[a], Roded Sharan[b], Masami Hasegawa[c], Ron Shamir[d], Dan Graur[e,*]

[a] *Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel*
[b] *International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704, USA*
[c] *The Institute of Statistical Mathematics, 4-6-7 Minami Azabu, Minato, Tokyo 106-8569, Japan*
[d] *School of Computer Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel*
[e] *Department of Zoology, George S. Wise Faculty of Life Sciences, Ramat Aviv, Tel Aviv 69978, Israel*

Received 15 January 2003; received in revised form 19 June 2003; accepted 2 July 2003

Received by T. Gojobori

Abstract

There are a 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 excess of radical amino-acid replacements. Such excess is usually indicative of positive Darwinian selection, but may also be due to relaxed functional constraints or model misspecification. In our test, each amino-acid replacement is characterized in terms of a physicochemical distance, i.e., the degree of dissimilarity between the exchanged amino-acid residues. By using phylogenetic trees based on protein sequences, our test identifies statistically significant deviations of the mean physicochemical distance from the random expectation, either along a taxonomic lineage or across a subtree. The mean inferred distance is calculated as the average physicochemical 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 fast 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.

Keywords: Positive Darwinian selection; Maximum likelihood; Chemical distance; Molecular evolution; Dermaseptin-related peptides; Class I major-histocompatibility-complex glycoproteins; Carbonic anhydrase I

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 (for review, see Kimura, 1983). There are a few cases in which positive Darwinian selection was convincingly demonstrated at the molecular level (Stewart and Wilson, 1987; Seibert et al., 1995; Hughes and Yeager, 1998; Zhang et al., 1998; Yang, 2000; for reviews, see Yang and Bielawski, 2000; Fay and Wu, 2001). These cases are vital to our understanding of the link between sequence variability and adaptive evolution.

The most widely used method for detecting positive Darwinian selection is based on comparing synonymous and nonsynonymous substitution rates between nucleotide sequences (Nei and Gojobori, 1986; Hughes and Nei, 1988). Synonymous substitutions are assumed to be selectively neutral. If purifying selection operates, then the rate of synonymous 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., Lee et al., 1995;

Abbreviations: MHC, major-histocompatibility-complex; ASRV, among-site rate variation.

☆ Some results in this paper have appeared in the Proceedings of the First Workshop on Algorithms in Bioinformatics, 2001, *LNCS* 2149:142–155.

* Corresponding author. Tel.: +972-3-6408646; fax: +972-3-6409403.
E-mail addresses: talp@post.tau.ac.il (T. Pupko), roded@icsi.berkeley.edu (R. Sharan), hasegawa@ism.ac.jp (M. Hasegawa), rshamir@post.tau.ac.il (R. Shamir), graur@post.tau.ac.il (D. Graur).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00802-3

Zhang et al., 1997). One critical shortcoming of this method is that due to saturation, it may be impossible to estimate of the number of synonymous substitutions when the sequences under study are evolutionary distant. Estimation is often problematic even when closely related species are concerned. For example, saturation of substitutions at the third codon position is evident even when comparing cytochrome *b* sequences among species within the same mammalian order (Halanych and Robinson, 1999).

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 (Stewart and Wilson, 1987; Zhang and Kumar, 1997). 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 amino-acid replacements (Hughes et al., 1990; Hughes, 1999). There are many measures in the literature aimed at quantifying the similarity or dissimilarity between two amino acids (e.g., Sneath, 1966; Grantham, 1974; Miyata et al., 1979; Wang et al., 1998). These so-called physicochemical distances are based on such properties of the amino acids, as hydrophobicity, polarity, charge, molecular volume, and chemical composition of the side chain. Amino acid replacements may be divided into conservative and radical replacements. A replacement of an amino acid by a similar one according to a certain similarity criterion is called conservative, whereas a replacement of an amino acid by a dissimilar one is called radical. In this method, radical and conservative replacements are counted separately for radical and conservative sites, respectively, 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, positive Darwinian selection is invoked. Using this method, positive selection was inferred for the antigen binding cleft of class I major-histocompatibility-complex (MHC) glycoproteins (Hughes et al., 1990) and rat olfactory proteins (Hughes and Hughes, 1993). This method for detecting positive selection has the advantage that distant protein sequences can be compared even when synonymous substitutions are saturated. Another desirable quality of this method is its flexibility with respect to the choice of the amino-acid characteristic used in the test. For example, if we suspect that replacements resulting in polarity changes might be advantageous, a test may be designed with radical replacements defined as those occurring between polar amino acids and nonpolar amino acids. If, on the other hand, changes in molecular volume are suspected to be under positive Darwinian selection, then the distance between two amino acids may be defined on the basis of the molecular volumes of the exchanged amino acids, and distances above or below a predetermined value will be considered radical or conservative, respectively. However, this method also has many shortcomings. First, no correction for multiple substitutions is applicable (Hughes, 1999). Second, the method treats replacements between different amino acids as equally probable. Third, the method ignores branch lengths, implicitly assuming independence of the replacement probabilities between amino acids and the evolutionary distance between the sequences under study. Finally, the phylogenetic signal is ignored, i.e., the test is applied to pairs of sequences rather than being used to test hypotheses across a phylogenetic tree.

Our test for detecting excess of radical replacements proposed in this study overcomes the shortcomings of the radical-conservative test. Our test incorporates a probabilistic framework for dealing with radical versus conservative replacements. It applies a novel method for averaging over ancestral sequence assignments weighted by their likelihood. This eliminates the bias that may result from assuming a specific ancestral sequence reconstruction. The rationale underlying our test is that the evolutionary acquisition of a new function requires a significant change in the biochemical properties of the amino-acid sequence (Hughes, 1999). To quantify this biochemical difference between two amino-acid sequences, we may use a physicochemical distance measure, e.g., Grantham's (1974) distance. Our test identifies large deviations of the mean physicochemical distance from the expected distance along a branch or across a subtree in a phylogenetic tree. If the inferred physicochemical 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 may be invoked.

In this paper we follow Hughes (1999), and work under the assumption that excess of radical replacements indicates positive Darwinian selection. However, there may be other explanations for such excess other than positive selection, such as model misspecifications and relaxed functional constraints (see Discussion).

Based on an assumed stochastic process, as well as on the tree topology and branch lengths, we calculate both the mean inferred physicochemical distance and its underlying distribution for the branch or subtree in question. The mean inferred physicochemical distance is calculated as the average physicochemical distance over all ancestral sequence reconstructions, weighted by their likelihood. The underlying distribution of this random variable is calculated using the JTT stochastic model (Jones et al., 1992), the tree topology and branch lengths, taking into account among-site rate variation. We provide a fast linear time algorithm to perform this test for all branches and subtrees of a phylogenetic tree.

For purposes of demonstration, we used our new method to reanalyze dermaseptin-related peptides from Hylidae (tree frogs). On the basis of excesses in radical amino-acid replacements, Duda et al. (2002) suggested that the propiece and mature parts of the sequences have been subjected to positive Darwinian selection throughout their evolutionary history. To validate our approach, we applied the new test to two control datasets: class I major-histocompatibility-complex (MHC) glycoproteins, and carbonic anhydrase I. These datasets were chosen since they were previously used as standard positive control (MHC) and negative control (carbonic anhydrase) for positive selection (Swanson et al., 2001).

2. Method

2.1. Definitions

We assume that sequence evolution follows the JTT probabilistic reversible model (Jones et al., 1992). We note, however, that other models may be used as appropriate. 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 $i \rightarrow j$ replacement probability, denoted by $P_{ij}(t)$, can be calculated from the eigenvalue decomposition of **M** (Kishino et al., 1990). 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 A be the set of 20 amino acids, and let s be an amino-acid sequence. The amino acid at position i in s is denoted by s_i. The physicochemical distance between two amino acids, $a,b \in A$, is denoted by $d(a,b)$. We assume that a table of distances (e.g., Sneath, 1966; Grantham, 1974; Miyata et al., 1979; Wang et al., 1998) is available for all possible pairs of amino acids. The choice of distance measure 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 three physicochemical properties: chemical composition, polarity, and molecular volume. For testing whether polarity differences between sequences are higher than the random expectation, at least two distance measures are applicable. One measure is based on dividing the set of amino acids into two categories: polar (C, D, E, H, K, N, Q, R, S, T, W, and Y), and nonpolar (the rest). The polarity distance between two amino acids is, then, defined as 1 if one amino acid is polar and the other is not, and 0 otherwise (Hughes et al., 1990). The second polarity distance is defined as the absolute difference between the polarity indices of the two amino acids, and yields real values (e.g., as in Grantham, 1974). For testing charge differences three categories of amino acids are defined: positive (H, K, and R), negative (D and 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 (Hughes et al., 1990).

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

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

Let 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 $(u,v) \in T$, we denote by $t(u,v)$ the length of the branch connecting nodes u and v.

2.2. A test for the detection of excess radical replacements

In this section we describe a new test for detecting excess of radical replacements, which in many cases is indicative of positive Darwinian selection. The input to the test is a set of aligned gapless sequences and a phylogenetic tree. We first present a version of the test for a pair of sequences. We then extend this method to test positive selection on specific branches of a phylogenetic tree. Finally, we generalize the test to subtrees (clades), and incorporate among-site rate variation.

2.3. Testing two sequences

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

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

Note that the expectation depends on t only, and not on the sequences themselves. Assuming that the distribution of the physicochemical distance in each position is identical, we obtain

$$E(D(s^1, s^2)) = \frac{1}{N}\sum_{i=1}^{N} E(d(s_i^1, s_i^2)) = E(d(s_1^1, s_1^2)) \tag{3}$$

The variance of the physicochemical distance at position i is:

$$V(d(s_i^1, s_i^2)) = E(d(s_i^1, s_i^2)^2) - E(d(s_i^1, s_i^2))^2 \tag{4}$$

$$V(d(s_i^1, s_i^2)) = \sum_{a,b \in A} d(a,b)^2 f_{ab}(t) - E(d(s_i^1, s_i^2))^2 \tag{5}$$

Assuming further that sequence positions are independent, we obtain:

$$V(D(s^1, s^2)) = \frac{V(d(s_1^1, s_1^2))}{N} \tag{6}$$

For practical values of N, $D(s^1,s^2)$ is approximately normally distributed with expectation $E(D(s^1,s^2))$ and variance

$V(D(s^1,s^2))$. This allows us to compute for each observed physicochemical distance d, the probability that it occurs by chance, i.e., its p value. If the observed physicochemical 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 a radical pattern of amino-acid replacements. Such a pattern can indicate positive or diversifying selection forces acting on these sequences.

2.4. Testing a specific branch

We shall now describe a pairwise test that is useful should one wish to test a statistical hypothesis on a specific branch of the phylogenetic tree. Suppose 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 by using, e.g., the maximum likelihood estimation of Pupko et al. (2000), and then check our hypothesis. Inferring ancestral sequences and then using these sequences as observations is a common practice, e.g., in Yang et al. (1995). However, treating inferred reconstructions as observations may lead to erroneous conclusions due to biases in the reconstruction. A more robust approach is to average over all possible reconstructions weighted by their likelihood.

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

Each branch $(u,v)\in T$ 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:

$$L(u,v,a) = \prod_{w\in(N(v)\setminus\{u\})} \left\{ \sum_{b\in A} P_{ab}(t(v,w)) \cdot L(v,w,b) \right\} \tag{7}$$

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 T is, thus:

$$P_T = \sum_{a,b\in A} f_{ab}(t(u,v)) \cdot L(u,v,b) \cdot L(v,u,a) \tag{8}$$

where (u,v) is any branch of T.

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

$$D(x,y) = \sum_{a,b\in A} P_r(x=a, y=b \mid \bar{w}) \cdot d(a,b) \tag{9}$$

$$D(x,y) = \frac{1}{P_T} \sum_{a,b\in A} \{ d(a,b) \cdot f_{ab}(t(x,y)) \cdot L(x,y,b) \cdot L(y,x,a) \} \tag{10}$$

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

$$E(D(x,y)) = \sum_{\vec{z}\in A^n} P_r(\vec{z}) \sum_{a,b\in A} P_r(x=a, y=b \mid \vec{z}) \cdot d(a,b) \tag{11}$$

$$E(D(x,y)) = \sum_{a,b\in A} d(a,b) \sum_{\vec{z}\in A^n} P_r(\vec{z}) \cdot P_r(x=a, y=b \mid \vec{z}) \tag{12}$$

$$E(D(x,y)) = \sum_{a,b\in A} d(a,b) f_{ab}(t(x,y)) \tag{13}$$

For the variance of $D(x,y)$ we have no explicit formula. Instead, we evaluate $V(D(x,y))$ using parametric bootstrap (Swofford et al., 1995). Specifically, we draw at random many assignments of amino acids to the leaves of T and compute $D(x,y)$ for each of them, thereby evaluating its variance. An assignment to the leaves of T is obtained as follows: We first root 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 descendent of r according to the corresponding branch length and the appropriate replacement probability, 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 $P_r\ (Z \geq [D(x,y) - E(D(x,y))]/\sqrt{V(d(x,y))})$ where $Z \sim \text{Normal}(0,1)$. 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., by using Bonferroni's correction, which in our case will mean that the significance level should be divided by the number of tested branches. The application of the test to all branches of a given phylogenetic tree takes linear time in total.

2.5. Testing a subtree

In this section we present an extension of our method to test subtrees of a given phylogenetic tree T. This is motivated by the consideration that if a clade of contemporary sequences has undergone a change in selection pressures, we should not necessarily assume that this selection occurred solely along the branch leading to that clade. Rather, it is possible that the selection continuous and occurred along several or all branches of the subtree corresponding to the clade under study. In such a case, the test we have just described may not detect any significant change in selection intensity along any specific branch. Hence, we are interested in testing for changes in selection intensities across subtrees as well.

For a subtree T' of T, we define the mean inferred physicochemical distance $D(T')$ as the average inferred distance along its branches (i.e., the sum of the inferred distance for each branch divided by the number of branches in T'). Clearly, the expectation of $D(T')$ equals to the average expectation of the branches of T'. The variance of $D(T')$ can be evaluated using parametric bootstrap. We, then, use the normal approximation to compute a p value for this test.

2.6. Incorporating among-site rate variation in the test

The rate of evolution varies among amino-acid sites. 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 among-site rate variation (ASRV) into account assume that at the jth position the average number of replacement is lr_j, where each $r=r_j$ is a rate parameter drawn from some probability distribution. Maximum likelihood models incorporating ASRV were found to be an important factor in the fitting of models to sequence data (Yang, 1996). They also help us avoiding the severe underestimation of long-branch lengths that can occur in the homogeneous models (Lie and Goldman, 1998).

Yang (1993) suggested that site rates are independently and identically distributed according to a gamma distribution with parameters $\alpha=\beta$ (Yang, 1993). 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 (Yang, 1994). The categories were selected so that the probabilities of r falling into each category are equal, i.e., $\Pr(r=r_i)=1/k$.

The incorporation of the discrete gamma model in our test is straightforward. For each rate category i we calculate both the expected and inferred physicochemical 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 a factor of r_i. The inferred and expected physicochemical distances for each branch are then averaged over all rate categories.

3. Results

3.1. Illustration of the method on dermaseptin-related peptides

We illustrate our method by applying it to a dataset of amphibian antimicrobial peptides called dermaseptin-related peptides (DRPs), in which a significant excess of radical over conservative charge changes was found in the propiece domain by Duda et al. (2002). The two sequences of the propiece domain shown below are DRP PD-3-6 and the inferred ancestor of DRP PD-3-6 and DRP AA-2-5:

DRP PD-3-6 = “EAEKREEENEEKQEDDDESEKKR”,
ANCESTRAL NODE = “EEEKREEENEEKQEDDDQSEKKR”.

We first compute the genetic distance between the two sequences. There are two amino-acid replacements (E → A at position 2 and Q → E at position 18). The aligned sequence length is 23 amino acids. Hence, the genetic distance is 2/23 = 0.087. The genetic distance underestimates the real distance, since it does not correct for multiple amino-acid replacements. Indeed, the maximum likelihood distance we found was 0.089. We now compute the observed physicochemical distance (Eq. (1)), where the criterion is charge changes. Since glutamic acid (E) is acidic, i.e., charged, whereas alanine (A) and glutamine (Q) are not charged, the sum of the physicochemical distances is 2. Dividing by the number of positions we obtain an observed physicochemical distance of 0.087.

Calculating the expectation based on Eq. (3) yields the value of 0.011. The variance computation based on Eq. (6) yields 0.00048. Thus, we obtain a z value of 3.44. It, thus, seems that the fraction of radical changes out of the total number of changes is more than three standard deviations from the random expectation. Clearly, this is a statistically significant result supporting the conclusions of Duda et al. (2002).

3.2. Test cases

We next applied our test to two datasets, the MHC class I glycoproteins serving as a positive control, and the housekeeping gene carbonic anhydrase I serving as a negative control. Phylogenetic trees were constructed using the MOLPHY software (Adachi and Hasegawa, 1996), with the neighbor-joining method (Saitou and Nei, 1987) for MHC class I, and with the maximum likelihood method for carbonic anhydrase I. The reason for the use of two tree-reconstruction 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 (Felsenstein, 1981) with the JTT stochastic model (Jones et al., 1992) under the assumption that the rate is discrete-gamma distributed among sites with four rate categories.

3.3. Class I MHC proteins

The primary immunological function of class I MHC 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 all nucleated cells and are recognized by CD8-positive cytotoxic T cells, thus initiating an essential phase in 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 (Hughes et al., 1990). Using pairwise comparisons of sequences, it was shown that the proportion of amino-acid replacements in the antigen-binding cleft that cause charge changes was significantly higher than the proportion that do not affect electrical charge. This finding indicates that peptide binding may be the target of positive selection (Hughes, 1999).

Following Hughes et al. (1990), we analyzed 42 human MHC class I sequences from three allelic groups: HLA-A, -B, and -C loci. These sequences are the same as in Hughes et al. (1990), except for the omission of two sequences: HLA-B7.2, which is identical to HLA-B7.1, and HLA-CX52, which can only be aligned with gaps. The length of each MHC class I sequence is 274 amino acids. The binding cleft, a subregion

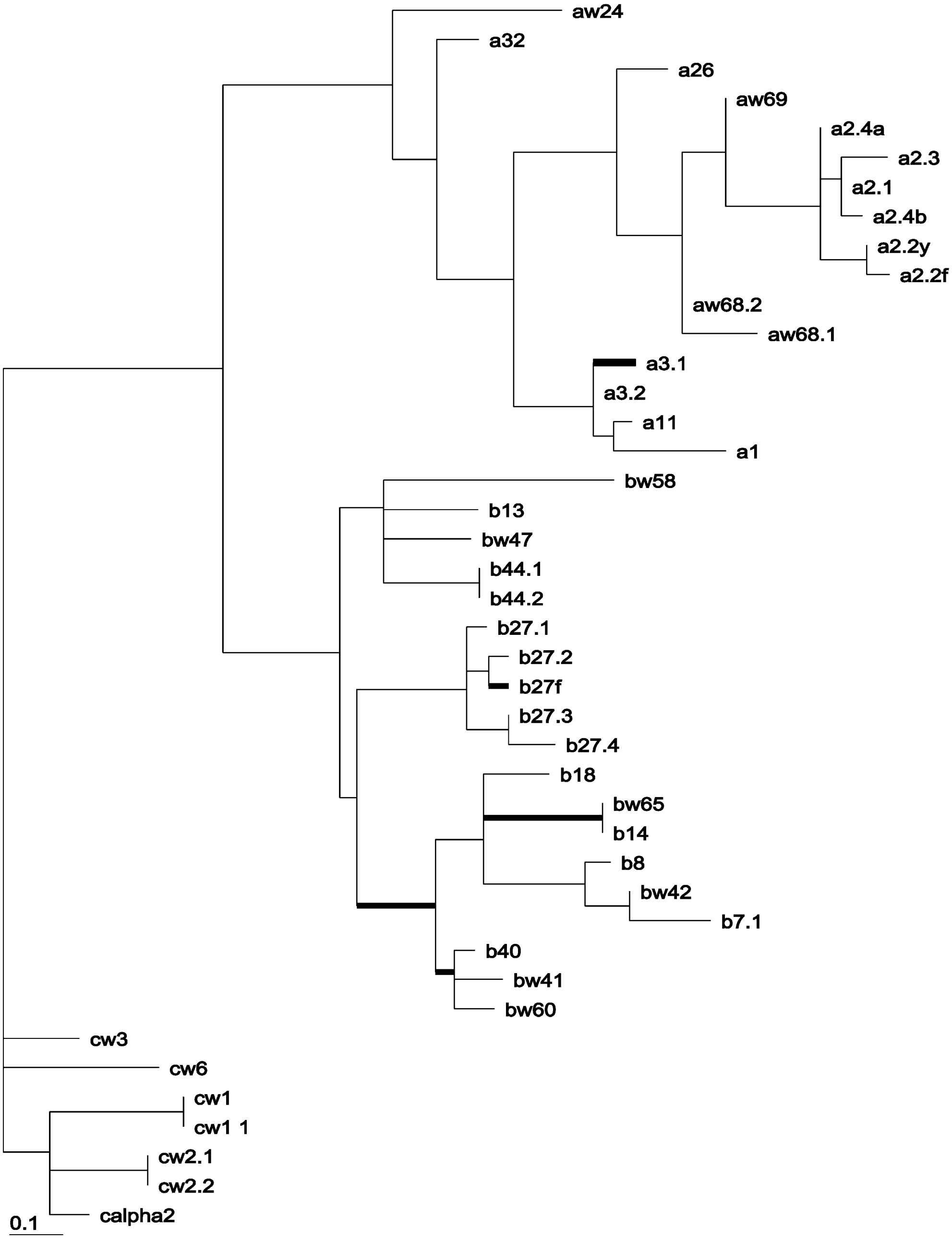

Fig. 1. A phylogenetic tree for MHC class I sequences. Species labels are as in Hughes et al. (1990). Tree topology was estimated using whole sequences. Branch lengths were estimated for the cleft subsequences only. Each branch was subjected to the positive selection test on the cleft subsequences. Branches in boldface indicate positive selection at $p < 0.01$.

of the antigen recognition site, consists of 29 residues (Parham et al., 1988). The phylogenetic tree for MHC class I sequences is given in Fig. 1. The maximum likelihood estimate of the α parameter of the gamma distribution found for this tree was 0.24. When our test was applied to the whole tree, no indication for positive selection was found. The relevant 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 than that for the rest of the sequence, the branch lengths that were computed by using whole sequences are most certainly underestimated. This underestimation can result in falsely inferring positive selection where there is none. To overcome this problem, branch lengths were reestimated on the basis of the binding site part of the sequence only. Statistically significant excesses of polar and charge replacements were also found with these new estimates ($p < 0.01$). The corresponding z scores are shown in Table 1. We note that using the 0/1 polarity measure as in Hughes et al. (1990), we found no evidence for positive selection. On the other hand, when we used Grantham's (1974) polarity indices, significant deviations from the random expectations were observed (Table 1). We conclude that there is a significant excess in all types of replacements affecting polarity rather than an excess in replacements affecting charge only as has been reported by Hughes et al. (1990).

Finally, we tested specific branches in the tree to find those branches that contribute the most to the excess of charge replacements. Branches with p values smaller than 0.01 appear in boldface in Fig. 1. 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, all the high scoring branches are located in the subtrees corresponding to the A and B alleles, matching the findings of Hughes et al. (1990), who reported positive selection for these alleles only.

Table 1
z scores for each of the tests performed on the MHC class I dataset

Dataset	Grantham distance	Charge	Polarity according to Grantham	Polarity according to Hughes et al.
Whole	− 1.30	0.01	− 1.25	1.10
Cleft	**9.38**	**9.32**	**13.23**	**5.79**
Cleft and cleft-based branch lengths	1.08	**3.14**	**2.78**	0.01

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 ($p < 0.01$) appear in boldface.

3.4. Carbonic anhydrase I

This dataset comprises of six sequences of the carbonic anhydrase I house-keeping gene, for which there is no evidence of positive selection (Swanson et al., 2001). The carbonic anhydrase I sequences are the same as in Swanson et al. (2001), except that amino-acid sequences were used instead of the nucleotide sequences. Sequence accession numbers are as follows: 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 amino-acid sequences, no evidence for positive selection was found. This was true, irrespective of the distance measures we used: Grantham ($z = 0.01$), polarity ($z = -0.49$), polarity indices ($z = -1.04$), and charge ($z = -1.73$).

4. Discussion

Natural selection may favor amino acid replacements that change certain properties of amino acids (Hughes, 1999). 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 the identification of large deviations of the mean inferred physicochemical distance between two proteins from the expected distance.

Two variants of the test were presented: The first is a statistical test of a single branch in a phylogenetic tree. The second looks for selection in a clade, e.g., a subtree or the entire tree. If selection is suspected to have operated in a certain lineage, say due to a specific adaptation to a certain environment, then the branch-specific test should be used. If, on the other hand, selection is suspected to be continuous, as for instance, in the case of the allele-diversity-promoting selection in MHC, then the clade-based test should be used.

We validated our method on two datasets: carbonic anhydrase I sequences served as a negative control, and the cleft of MHC class I proteins 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 can be used to detect positive Darwinian selection only in those cases where the selection gives rise to excess of radical replacements. Also, the method relies heavily on an assumed stochastic model of evolution. In many cases, one must carefully construct the model such that the effects of factors other than selection on the neutral expected degree of physicochemical dissimilarity are minimized. In particular, one must pay attention to possible extreme biases in amino-acid composition and codon usage that have been previously shown to affect the ratio of radical to conservative amino-acid replacement (Dagan et al., 2002). If such extreme biases are not evident in the data, our method is expected to perform well. In addition, it is important to estimate branch lengths under realistic models, taking into account among-site rate variation. Finally, 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 Thorne et al., 1996). 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 (Hughes et al., 1990; Swanson et al., 2001) 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 (Huelsenbeck et al., 2000). Simulation studies are most probably needed to measure the sensitivity of our test to different assumptions regarding the stochastic process and the phylogenetic tree, as well as the robustness and power of our test. These factors will be better understood when more real datasets are available for analysis.

Acknowledgements

We thank Hirohisa Kishino and Yoav Benjamini for their suggestions regarding the statistical analysis. Tal Pupko was supported by a postdoctoral fellowship from the Japan Society for the Promotion of Science. Roded Sharan was supported by a Fulbright grant. Ron Shamir was supported in part by the Israel Science Foundation (grant no. 565/99). Dan Graur wishes to thank the “Klum mit Gurnisht” Section of the Israel Science Foundation for their consistent support in the last 17 years.

References

Adachi, J., Hasegawa, M., 1996. MOLPHY: Programs for Molecular Phylogenetics Based on Maximum Likelihood, Version 2.3. Institute of Statistical Mathematics, Tokyo, Japan.

Dagan, T., Talmor, Y., Graur, D., 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol. Biol. Evol. 19, 1022–1025.

Duda, T., Vanhoye, D., Nicolas, P., 2002. Roles of diversifying selection and coordinated evolution in the evolution of amphibian antimicrobial peptides. Mol. Biol. Evol. 19, 858–864.

Fay, J., Wu, C.-I., 2001. The neutral theory in the genomic era. Curr. Opin. Genet. Dev. 11, 642–646.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Halanych, K., Robinson, T., 1999. Multiple substitutions affect the phylogenetic utility of cytochrome *b* and 12S rDNA data: examining a rapid radiation in Leporid (Lagomorpha) evolution. J. Mol. Evol. 48, 369–379.

Huelsenbeck, J., Rannala, B., Masly, J., 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288, 2349–2350.

Hughes, A., 1999. Adaptive Evolution of Genes and Genomes. Oxford Univ. Press, New York.

Hughes, A., Hughes, M., 1993. Adaptive evolution in the rat olfactory receptor gene family. J. Mol. Evol. 36, 249–254.

Hughes, A., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170.

Hughes, A., Yeager, M., 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. 32, 415–435.

Hughes, A., Ota, T., Nei, M., 1990. 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.

Jones, D., Taylor, W., Thornton, J., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge.

Kishino, H., Miyata, T., Hasegawa, M., 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31, 151–160.

Lee, Y., Ota, T., Vacquier, V., 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. 12, 231–238.

Lie, P., Goldman, N., 1998. Models of molecular evolution and phylogeny. Genome Res. 8, 1233–1244.

Miyata, T., Miyazawa, S., Yasunaga, T., 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12, 219–236.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Parham, P., Lomen, C., Lawlor, D., Ways, J., Holmes, N., Coppin, H., Salter, R., Won, A., Ennis, P., 1988. Nature of polymorphism in HLA-A, -B, and -C molecules. Proc. Natl. Acad. Sci. U. S. A. 85, 4005–4009.

Pupko, T., Pe’er, L., Shamir, R., Graur, D., 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17, 890–896.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Seibert, S., Howell, C., Hughes, M., Hughes, A., 1995. Natural selection on the *gag*, *pol*, and *env* genes of human inummodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 12, 803–813.

Sneath, P., 1966. Relations between chemical structure and biological activity in peptides. J. Theor. Biol. 12, 157–195.

Stewart, C., Wilson, A., 1987. Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. Cold Spring Harbor Symp. Quant. Biol. 52, 891–899.

Swanson, W., Yang, Z., Wollner, M., Aquadro, C., 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. U. S. A. 98, 2509–2514.

Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1995. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics, 2nd ed. Sinauer, Sunderland, MA, pp. 407–514.

Thorne, J., Goldman, N., Jones, D., 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. 13, 666–673.

Wang, H., Dopazo, J., Carazo, J., 1998. Self-organizing tree growing network for classifying amino acids. Bioinformatics 14, 376–377.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11, 367–372.

Yang, Z., 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51, 423–432.

Yang, Z., Bielawski, J., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15, 496–503.

Yang, Z., Kumar, S., Nei, M., 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141, 1641–1650.

Zhang, J., Kumar, S., 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol. Biol. Evol. 14, 527–536.

Zhang, J., Kumar, S., Nei, M., 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol. Biol. Evol. 14, 1335–1338.

Zhang, J., Rosenberg, H., Nei, M., 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. U. S. A. 95, 3708–3713.