A test for positive Darwinian selection

Tal Pupko [1] Roded Sharan [2], Ron Shamir [2] Dan Graur [3]

Keywords: protein evolution, positive selection, olfactory proteins, sequence analysis

1 Introduction.

The neutral theory of molecular evolution maintains that most sequence variation among genomes has no direct relevance to adaptation [1]. The rapidly-growing amount of genomic data makes it possible now to ask which proteins are undergoing adaptive evolution. This, in turn calls for better computational tools to detect positive selection. In this study, we present a new test for detecting positive Darwinian selection. The test identifies large deviation of the mean observed chemical distance between two proteins from the expected distance along a branch in a phylogenetic tree. The mean observed chemical distance is calculated as the average over all possible ancestral sequence reconstruction, weighted by their likelihoods. Our novel method of averaging over all possible ancestral amino-acid assignments eliminates possible bias that could have resulted if the calculation was based solely on the most likely ancestral sequence. We present an $O(n)$-time algorithm to perform this test for all branches of a phylogenetic tree with n leaves. We apply our test to study adaptive evolution in rat olfactory proteins.

2 A test for positive selection.

We use the term $d(a,b)$ for the Grantham's chemical distance [2] between amino-acid a and b. This chemical distance measures the differences between two amino-acids in terms of their volume, charge, and composition. We define the chemical distance between two sequences as the sum over all positions of the chemical distance between the pairs of amino-acids occupying the same position in the gapless alignment. The first step of our test is to align the sequences and construct the phylogenetic tree. We then compute the chemical distance along a branch in question, and test whether the mean observed chemical distance significantly deviate from its chance expectation. It is claimed that deviation of this random variable from it chance expectation is indicative of positive selection [3].

Since positions are assumed to be independent we restrict the subsequent description to a single site. Each branch (u,v) at the phylogenetic tree $\mathcal{T}$, partitions $\mathcal{T}$ into two subtrees. Let $L(u,v,a)$ denote the likelihood of the subtree which includes v, given that v is assigned amino-acid a. $L(u,v,a)$ is computed by the following recursive equation:

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

For a leaf v, we set $L(u,v,a)$=$1(a)$, i.e., 1 if amino-acid a is in v and 0 otherwise. $N(v)$ denote all the nodes connected to v, and $P_{ab}(t)$ is the replacement probability of amino-acid

[1] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan. E-mail: tal@ism.ac.jp

[2] Department of Computer Science, School of Mathenatical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: {roded,rshamir}@post.tau.ac.il

[3] Department of Zoology, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: graur@post.tau.ac.il

a by amino-acid b along a branch of length t. AA is the set of all 20 amino-acids. Let L denote the likelihood of the whole tree:

$$L = \sum_{a,b \in AA} f_{ab}(t(u,v)) \cdot L(u,v,b) \cdot L(v,u,a)$$

where (u,v) is any branch in $\mathcal{T}$, and $f_{ab}(t) = P_i \cdot P_{ij}(t)$ is the probability of observing i and j in the same position in two sequences of evolutionary distance t. P_i is the frequency of amino-acid i. The mean observed chemical distance for a given branch (x,y) in $\mathcal{T}$ is computed as follows:

$$E(d(x,y)) = \frac{1}{L} \sum_{a,b \in AA} \{d(a,b) \cdot f_{ab}(t(x,y)) \cdot L(x,y,b) \cdot L(y,x,a)\}$$

The variance $V(d(x,y))$ is computed similarly. Finally, the p value of the test for a branch (u,v) in $\mathcal{T}$ with $numberofsites > 30$ can be calculated using normal approximation for the distribution of $d(u,v)$. We note that if the test is done on all branches of the tree, correction for multiple test is needed. The algorithm for testing positive selection is summarized in Figure 1. For each branch $(u,v) \in \mathcal{T}$ the algorithm outputs the p-value of the test.

```
Traverse $\mathcal{T}$ bottom-up, computing for each branch $(u,v)$ and $a \in AA$ the value of $L(u,v,a)$, where $u$ is the parent of $v$.
Traverse $\mathcal{T}$ top-down, computing for each branch $(u,v)$ and $a \in AA$ the value of $L(v,u,a)$, where $u$ is the parent of $v$.
Forevery $(u,v) \in \mathcal{T}$ Calculate $E(d(u,v))$ and $V(d(u,v))$ and using the normal approximation output the $p$ value.
```

Figure 1: An algorithm for testing positive selection along a branch of a given tree

From the linearity of $E(d(u,v))$ and $V(d(u,v)$ and from the above recursion equation for $L(u,v,a)$ it is easy to show that:

Theorem 1 *The complexity of testing all the branches of a given phylogenetic tree with* n *leaves is* $O(n)$.

3 An empirical example.

Using this test we reevaluated positive selection in rat olfactory proteins [3]. We assumed the JTT stochastic model [4], and Maximum likelihood estimation of branch length. The result of our maximum-likelihood test roll-out positive selection in rat olfactory sequences, opposed to [3] that analyzed chemical differences between pairwise sequences.

References

[1] Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press. Cambridge.

[2] Grantham, R. 1974. Amino-acid differences formula to help explain protein evolution. *Science* 185:862–864.

[3] Hughes, A. S. 1999. *Adaptive evolution of genes and genomes.* Oxford University Press, New-York.

[4] Jones, W. R., Taylor, W. R., Thornton, J. M. 1974. The rapid enumeration of mutation data matrices from protein sequences. *Comput. Appl. Biosi.* 8:275–282.