BURSTS OF NON-SYNONYMOUS SUBSTITUTIONS IN HIV-1 PHYLOGENETIC TREE REVEAL INSTANCES OF POSITIVE SELECTION AT CONSERVATIVE PROTEIN SITES

Bazykin G.A.*, Dushoff J.1, Levin S.1, Kondrashov A.2
1 Dept. of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08650, USA; 2 National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA
* Corresponding author: e-mail: gbazykin@princeton.edu

Key words: evolution, positive selection, HIV-1, bursts of substitutions

SUMMARY

Motivation: Darwinian positive selection can cause fixation of novel alleles in population. The conventional methods of identification of positive selection rely on excess of non-synonymous over synonymous substitutions. This approach is adequate for identification of selection affecting lineage for a substantial period of time. However, it is biologically realistic to expect that some selection events can be transient, affecting only one to a few successive substitution events at an otherwise conservative site.

Results: We analyzed pairs of successive substitutions within amino acid position in evolution of HIV-1 genes. Successive non-synonymous substitutions tend to be much more clumped than the random expectations and than synonymous substitutions. Strikingly, this effect is strongest in sites with low overall rate of non-synonymous substitutions. The observed bursts of amino acid-changing substitutions can not be explained by mutation biases, and are, therefore, due to positive selection.

INTRODUCTION

Positive selection which favors new alleles drives adaptive evolution and, thus, is of paramount importance (Fisher, 1930). However, positive selection always occurs on the background of pervasive negative selection which maintains status quo (Williams, 1966). Even at the simplest level of DNA and protein sequences, disentangling the two remains a major challenge.

We pursue a novel approach to detecting positive selection. Due to the nature of the genetic code, replacing amino acid X with amino acid Z often requires two or even three non-synonymous substitutions, since in only 75 out of 190 unordered amino acid pairs the members can be converted into each other by a single nucleotide substitution. Thus, when positive selection favors a particular amino acid replacement, in a large fraction of cases a clump of two or three successive nucleotide substitutions should occur (Gillespie, 1984), even if most of the time the site evolves slowly, due to negative selection.

Comparison of rat, mouse, and human orthologous proteins demonstrated that, at codon when rat and mouse differ by two nonsynonymous substitutions, both substitutions tend to occur, after rat-mouse divergence, in the same lineage, either rat or mouse (Bazykin et al., 2004). This result indicates that clumps of successive nonsynonymous substitutions may be common. However, since rat-and mouse are tightly related to each
other, codon at which they differ by two nucleotide substitutions are rare and must mostly come from the subset of rapidly-evolving codons.

Here we analyze the evolution of 4 genes of HIV-1, which collectively encode 10 proteins, using sets of many hundreds of genomes. These data make it possible to look for clumps of nonsynonymous substitutions both at rapidly evolving and at slowly evolving amino acid sites.

MATERIALS AND METHODS

Sequences and phylogenies. Alignments of nucleotide sequences of all full-length env, gag, pol, and nef protein coding regions from HIV-1 genomes of subtypes A-H were taken from the 2003 Los Alamos National Laboratory HIV-1 sequence database (Korber et al., 2000a). For each gene, a maximum parsimony tree was constructed by PAUP. Resulting trees were rooted using the consensus of consensus sequences for each subtype (Korber et al., 2000b). All the analysis was conducted using a set of Bioperl-based scripts (Stajich et al., 2002) which are available from the authors upon request.

Analysis of nucleotide substitutions. We used maximum parsimony to reconstruct the states of the codons at all internal nodes within each phylogeny. Then, for each codon, we inferred the edges of the tree at which each single-nucleotide substitution occurred, as follows. If a pair of successive nodes within a phylogeny differed at a nucleotide position, we assumed that exactly one nucleotide substitution occurred on the edge connecting these nodes. If the codons at successive nodes differed at more than one nucleotide site, the numbers of synonymous and non-synonymous substitutions were averaged over all possible orders of substitution events and rounded to the integer. For each codon, we estimated the number of non-synonymous substitutions and the number of synonymous substitutions on the whole tree.

For each codon, we treated synonymous and non-synonymous substitutions separately, so that synonymous substitutions were ignored when non-synonymous substitutions were considered and vice versa. If successive substitutions occurred along the path from the tree root to a leaf, for each substitution A, except the first one, the preceding substitution A’ at the same codon could be uniquely determined. We assume that two substitutions at a codon form a pair if there is a path from the root of the tree to at least one of the leaves, such that both the substitutions belong to this path, and there are no other substitutions on this path between them. In particular, two substitutions that occurred at the same codon on the same edge, revealed by the codons at two successive nodes differing from each other at two nucleotide sites, always constitute a pair. Substitutions A’ and A can occur either at the same site or at different sites of the codon.

We estimate the distance l between A’ and A as the sum of the lengths of edges between them, assuming that substitutions always occur at the middle of an edge. If A’ and A occurred within the same edge, l = 0 for them. For each codon, we calculated the distances within the pairs of successive synonymous and of non-synonymous substitutions, and compared them with the distances obtained in simulated evolution for the same total number of substitutions on the phylogeny.

Simulations of sequence evolution. To find the expected numbers of pairs of successive substitutions at a codon and the distance between substitutions in such pairs, we simulated random substitution events in each codon on the phylogeny. Simulations were then done for each amino acid position in each gene separately. First, we counted the total number of substitutions of particular type (synonymous or non-synonymous) at the given codon on the phylogenetic tree. Next, in each of the 10,000 simulation trials, we distributed the same number of events randomly over the edges of the actual phylogenetic tree, with the probabilities weighted by the lengths of the corresponding branches. We then counted the numbers of pairs and distances within pairs of successive substitutions in simulation.
RESULTS

We built the phylogenetic trees of 343, 218, 193, and 674 full-length sequences of env, gag, pol, and nef genes from HIV-1 genomes. Let us consider how non-synonymous substitutions are distributed relatively to each other. Among the 7424 observed pairs of successive non-synonymous substitutions, 3022 (40.7 %) were reversals, 836 (11.3 %) were non-reversing substitutions at the same site, and 2108 (28.4 %) substitutions occurred in different nucleotides of the same codon.

Non-synonymous substitutions display reduced evolutionary distances between the members of a pair, relative to random expectation and to the distances between successive synonymous substitutions. In contrast, the average distance between successive synonymous substitutions was close to that predicted in the simulation (Fig. 1). Clumping of successive non-synonymous substitutions is the strongest at codons with the smallest total number of non-synonymous substitutions, i.e., the ones that are highly conserved (Fig. 1).

![Figure 1](image.png)

**Figure 1.** Mean distance between successive non-reversing substitutions, plotted against the total number of non-synonymous (a) and synonymous (b) substitutions in each site of the env HIV-1 gene. Solid lines indicate mean distances between successive substitutions in each sliding 30 site window. Dashed lines indicate mean distances in simulations.

DISCUSSION

Our results indicate that while the distribution of synonymous substitutions over the phylogenetic tree is consistent with the molecular clock model, the non-synonymous substitutions occur in bursts, forming pairs of rapid successive substitutions. This clumping of non-synonymous substitutions is not an artifact of phylogenetic reconstruction, of mutations or sequencing errors spanning multiple adjacent nucleotides in some of the sequences, or of our approach to inference of positions of individual substitutions (analysis not shown).

Therefore, the non-synonymous substitutions actually tend to occur in bursts. Although different codons have intrinsic differences in the rate of non-synonymous evolution due to the structure of the genetic code and peculiarities of the substitution matrix, these differences can not explain the observed clumping (data not shown). Apparently the only explanation of the clumping of amino acid-changing nucleotide substitutions that we are left with is episodes of positive selection alternating with periods of negative selection acting on individual codons.

Bursts of substitutions can be caused by a single environmental change that creates a new fitness landscape for the given locus (Gillespie, 1984). Even at a single codon, up to three successive substitutions can be facilitated by selection after a single change in fitness due to the structure of the genetic code. Indeed, after the change of fitness landscape, the new preferred amino acid can be reachable by no less than two or even
three nucleotide substitutions, and each of these substitutions can be facilitated by selection. The observed clumping of successive substitutions in a single nucleotide can be explained by assuming that the intermediate codon has lower fitness than the final variant. We observe strongest clumping of non-synonymous substitutions in the most conservative codons, with the lowest total number of substitutions. Apparently, positive selection plays a major (and previously unobserved) role in the evolution of important and generally conservative amino acids which cannot be replaced by random drift. New substitutions that do occur in this site can also be expected to be functionally important, and get fixed rapidly under positive selection. Our results indicate a new set of sites in HIV-1 proteins, most of which are conservative, that evolve under positive selection.

ACKNOWLEDGEMENTS

GB gratefully acknowledges fellowships from the Pew Charitable Trusts award 2000-002558 and the Burroughs Wellcome Fund award 1001782, both to Princeton University. We thank Mikhail Gelfand, Sergey Kryazhimskiy and Joshua Plotkin for valuable discussions.

REFERENCES

Korber B. et al. eds. (2000a) HIV Immunology and Sequence Databases, Los Alamos National Laboratory, Los Alamos, NM. Available at http://lanl.hiv.org