A Likelihood Method for the Detection of Selection and Recombination Using Nucleotide Sequences

Nicholas C. Grassly and Edward C. Holmes
Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford

Different regions along nucleotide sequences are often subject to different evolutionary forces. Recombination will result in regions having different evolutionary histories, while selection can cause regions to evolve at different rates. This paper presents a statistical method based on likelihood for detecting such processes by identifying the regions which do not fit with a single phylogenetic topology and nucleotide substitution process along the entire sequence. Subsequent reanalysis of these anomalous regions may then be possible. The method is tested using simulations, and its application is demonstrated using the primate ß-globin pseudogene, the V3 region of the envelope gene of HIV-1, and argF sequences from Neisseria bacteria. Reanalysis of anomalous regions is shown to reveal possible immune selection in HIV-1 and recombination in Neisseria. A computer program which implements the method is available.

Introduction

The phylogenetic analysis of molecular sequence data using explicit evolutionary models and the maximum-likelihood approach was pioneered by Felsenstein (1981). This approach is now widely used and has seen many applications, from looking at viral divergence within a single patient (Holmes et al. 1992) to examining the relationships between mammalian orders (Novacek 1992). An explicit assumption of nearly all maximum-likelihood methods, and indeed of other methods of phylogenetic reconstruction from molecular sequence data, explicit or otherwise, is that the evolutionary process is independent and identical at each site. Clearly, in very many cases this assumption is invalid. Both intergenic (Huelsenbeck and Bull 1996) and intragenic (Dorit and Ayala 1995) variation in the reconstructed phylogeny have been demonstrated. Such "spatial" variation along sequences is hereafter referred to as spatial phylogenetic variation (SPV). SPV may reflect varying selective forces along the sequence (e.g., flavivirus envelope proteins: Gritsun, Holmes, and Gould 1995), some form of functional constraint (e.g., ADH evolution: Dorit and Ayala 1995), or recombination (e.g., HIV-1: Robertson et al. 1995; hepatitis B virus: Bollyky et al. 1996).

SPV not only violates an assumption of current phylogenetic algorithms but has been shown to cause such algorithms to select incorrect evolutionary topologies (Tateno, Takezaki, and Nei 1994; Gaut and Lewis 1995). Identification of SPV is thus not only important because of its biological relevance per se, but also is essential if an accurate phylogeny is to be constructed.

In cases where a priori ideas exist concerning the type of spatial variation, attempts can be made to address the particular phenomenon. Methods exist which try to diagnose recombination by looking at the compatibility of the "phylogenetic partitions" supported by the polymorphic sites along sequences (Stephens 1985; DuBose, Dykhuizen, and Ha tul 1988, Drouin and Dover 1990), by looking at changes in the most parsimonious topology along sequences (Hein 1993), or by using a maximum chi-square test (Maynard Smith 1992). Recently developed maximum-likelihood models of rate variation along sequences have been also described. Such models may allow autocorrelation of rates along sequences (Yang 1995; Felsenstein and Churchill 1996) or not (Yang 1993), and the discrete rate categories used may be chosen from a range of distributions (Yang 1995). However, no general methods exist which allow detection of specific regions showing "anomalous" evolutionary patterns, thus alerting the investigator to possible SPV.

In maximum-likelihood methods of phylogenetic reconstruction the parameters of a certain nucleotide substitution model (e.g., JC69 [Jukes and Cantor 1969], F84 [Felsenstein and Churchill 1996]) and the phylogenetic topology are estimated from the sequence data. Together these constitute the maximum-likelihood model, which, when applied to all sites in the sequences, can be termed "global." In this paper, a statistical method using site likelihoods is presented which identifies those regions of molecular sequences which do not appear to be evolving in the same way as the global maximum likelihood model. Reanalysis of such regions may then reveal a different evolutionary process operating, although short regions may not be very informative about the processes they are subject to. In some cases, reanalysis may reveal a mosaic of evolutionary rates along the sequences, permitting speculation as to the biological function of, and selective pressures on, the different regions. Alternatively, different topologies may be supported, which, if significant, may be the result of processes such as recombination or convergent evolution. If desired, superposition of these topologies produces a phylogenetic “network” which accurately reflects the global evolutionary process (Maynard Smith 1989).

Materials and Methods

Sequences

Biological Data

Sequences were extracted from the GenBank database and published alignments were used. Table 1 in-
icates the sequences analyzed. Because the primate \( \eta \)-globin gene is a pseudogene, it is reasonable to assume that it evolves in a neutral manner (Li, Gojobori, and Nei 1981; Goldman 1993). It has been used frequently in determining the phylogenetic relationships of the pri-
mates (e.g., Koop et al. 1986; Miyamoto, Slightom, and Goodman 1987; Holmes, Pesole, and Saccone 1989).

The question of SPV occurring within this gene has not been addressed by these analyses. The HIV-1 sequences are fascinating since they come from a single patient (Holmes et al. 1992) and present the patterns of diversity generated by the evolutionary pressures within this sin-
gle individual. The \( \text{argF} \) gene from \( \text{Neisseria} \) is a “housekeeping gene,” and is thought to be subject to little selective pressure (Houghton, O’Donovan, and Wild 1989; Zhou and Spratt 1992; Murata and Scach-
man 1996). However, SPV within this gene is important, since intragenic interspecies recombination has been demonstrated (Zhou and Spratt 1992). The \( \text{E. coli argF} \) gene was used to root trees for display purposes, but was not used in the likelihood analysis.

**Simulated Data**

To assess the accuracy of SPV detection, DNA se-
quences subject to SPV were obtained by simulating their evolution down phylogenies. Initially a single tree topology with branch lengths was generated under a neutral coalescent process (Hudson 1990) using the program Coal-Sim (Rambaut, personal communication). Sequences generated with specified base frequencies were evolved down this tree under the required substitu-
tion model with user-specified rates using the program SEQ EVOLVE (Rambaut et al. 1997). The substitution model used in these analyses was F84 (Felsenstein and Churchill 1996), which allows unequal base frequencies, and also a different rate for transitions and transversions determined by the transition : transversion \((t_v : t_v)\) ratio. (F84 is similar to HKY85 [Hasegawa, Kishino, and Yano 1985], but is mathematically more tractable).

To simulate SPV, partial sequences were evolved ei-
ther down different tree topologies or down the same to-
	opology but with different evolutionary rates. These partial
sequences could then be spliced together. Biological se-
quences subject to recombination have parts with different topologies, while those subject to varying selective forces have parts with different average divergence. Thus, the simulated data sets reflect these evolutionary processes.

In the case of recombination the alternative topol-
ologies used for the simulations were chosen in such a way as to reflect an ancient recombination event, or a recent recombination between either closely or distantly related ancestral taxa. This is illustrated in figure 1. Each 50-taxon tree had 50 sequences evolved down it at four different rates of evolution. Thus, for each of the three different types of recombination 10,000 sequences were generated. All recombinant sequences were 50 bp in length, while the rest of the sequence was 500 bp.

In the simulations involving different rates, four rates were used \((0.3, 0.4, 0.5, 0.6)\) together with four sizes of the faster evolving “anomalous” region. For each rate and size 50 sequences were generated. The length of the rest of the sequence was kept constant at 500 bp and evolved at a rate of 0.2. In total, 16 sets of 2,500 sequences were generated.

**Phylogenetic Reconstruction**

Maximum-likelihood phylogenies were recon-
structed using fastDNAml which uses the F84 sub-
titution model (Olsen et al. 1994). The \( t_v : t_v \) ratio was set to two.

**Detection of Spatial Phylogenetic Variation**

To identify those regions evolving in an anomalous
way, a sliding window approach was taken (Tajima 1991; Dorit and Ayala 1995; Gritsun, Holmes, and Gould 1995). Using the maximum-likelihood phyloge-
y, the likelihoods for each site of the sequence data were calculated independently. Subsequently for win-
dows of varying sizes and positions along the sequence the following quantity could be calculated:

\[
Q = \frac{\sum_{i=s}^{i=n} \ln L_i}{s} \sum_{i=1}^{i=n} \ln L_i + \frac{\sum_{i=(sp+1)}^{i=n} \ln L_i}{n - s} \tag{1}
\]
The log likelihoods or "support" for each site ($\ln L_i$; Edwards 1992, pp. 31, 70-102) are summed over the region (which starts at $sp$) and this value is divided by the length of the region, $s$. This value is divided by a similar measure of the average log likelihood of the rest of the sequence (of length $n$) to give the final measure, $Q$. This gives a measure of the average likelihood of the window with respect to the rest of the sequence that is consistent over different window sizes.

The measure was calculated for window sizes from five nucleotides up to half the sequence length, and in all possible positions. Regions less than five bases long were deemed to be uninformative and subject to spurious patterns; hence, they were excluded from the analyses. The maximum values of $Q$ are associated with regions showing low likelihoods given the global maximum-likelihood model.

Although maximization of $Q$ may find a region of low average likelihood, the significance of this needs to be tested. In order to do this, a null distribution of this maximized function needs to be created. This is not possible analytically, owing to the complexity of the distribution; instead, a simulation approach was taken. Sequences were evolved down the maximum-likelihood phylogeny with the same parameters as the actual data, and a distribution of the maximized function for $N$ simulations for each region size was found. Such stochastic simulations represent a "Monte Carlo" approach which in the context of phylogenetics has recently been termed "parametric bootstrapping" (Huelsenbeck, Hillis, and Jones 1996). The value of $N$ in the simulations was set to 100, while in the analysis of biological data it was set to 1,000. It is widely accepted that $\geq 100$ null data points for Monte Carlo simulations give good results (Goldman 1993a).

Sampling theory suggests that the null distribution of maximized $Q$ for each region size should be normally distributed. This was tested using a Kolmogorov-Smirnov test and found to be the case. At the 5% significance level, 9% of the maximized $Q$ distributions for all region sizes for 50 aligned sequences 500 bases long were rejected as nonnormal, while for sequences 1,000 bases long, no distributions were nonnormal. These results are consistent with normality of the distribution given the expected type 1 (rejection of a true null hypothesis) error rate of the Kolmogorov-Smirnov test. The normality of the maximized $Q$ distribution enabled $Z$ values to be calculated for all values of $Q$ obtained from the actual data. The significance of anomalous regions (maximum $Q$) could then be tested at the desired level (e.g., 95% confidence limit; $\alpha = 0.05$). Because tests are carried out for each region size, from five nucleotides to half the length of the sequence, the problem of an inflated type 1 error rate at a given significance level occurs. In order to cope with this, the Bonferroni inequality may be used, whereby the level of $\alpha$ used for each test of significance is divided by the number of region sizes examined (Miller 1966). The $Z$ value corresponding to this adjusted $\alpha$ becomes asymptotic fairly quickly for increasing values of $n$ at $Z \approx 3$. Hence, in the simulations reported here, this $Z$ value was used to test for significance.

Owing to the autocorrelated nature of the $Z$ values (due to the use of overlapping windows), it is advisable to smooth their distribution (Lawrence, Hartl, and Ochman 1991). This was carried out using a 2D Gaussian filter (Bevington and Robinson 1992, pp. 235–238), over the $Q$ surface for different window sizes ($s$) and positions ($sp$). For each value of $Q$ in this surface, a smoothed value $Q'$ is calculated using the following equation:

$$Q' = (Q - 4c) + \sum_{i=1}^{4} cQ_i,$$

where $c$ is the smoothing constant and the four values of $Q_i$ represent the values of $Q$ at the four points adjacent to the value being smoothed. In all the results given in this paper, $c$ was set to the default value of 0.01.

A program which carries out the above analysis, called Partial Likelihoods Assessed Through Optimisation (PLATO), was written in C. The speed of this program depends mainly on the length of the sequences analyzed, rather than on the number of taxa. For the HIV-1 dataset of 78 sequences of 234 bp, it took 14.9 s on a 170 MHz hypersparc processor, while for 50 sequences 1,000 bases long it took 627.9 s. PLATO and

\[ \text{FIG. 1.—Trees used to generate sequences subject to recombination. Tree } A \text{ is the standard tree generated using a neutral coalescent model. Tree } B \text{ is that used for regions subject to an ancient recombination event; two deep-rooted lineages (ancestors) are swapped with one another at a single point in time. Trees } C \text{ and } D \text{ were used for more recent recombination events (recent lineages are swapped): } C \text{ for recombination between closely related ancestors, and } D \text{ for recombination between more distantly related ancestors. The numbers on the tips of the trees are representatives of the taxa numbers. The triangles on the time arrows represent the times of the "recombination" events.} \]
the programs SEQEVOLVE and SPOT are available on the World Wide Web at http://evolve.zoo.ox.ac.uk/. Alternatively, they can be obtained by anonymous ftp from evolve.zoo.ox.ac.uk/packages/. They have been compiled for both the Macintosh and Power Macintosh, and the source files are available for compiling on UNIX systems.

Results
Recombination Simulations

Figure 2 shows the mean accuracy with which ancient recombination and recent recombination between distantly related taxa can be detected at different rates of evolution, and the mean Z value associated with this detection. In addition, the percentage of recombination events not detected is shown. Recent recombination between closely related taxa was not detected at any of the rates used. The accuracy figures represent the average distance of the boundaries of the detected region from the actual known boundaries of the simulated recombinant region in base pairs. Thus, a small figure means detection was accurate, zero being perfectly accurate. For evolutionary rates of more than 0.5 (substitutions per unit time) the two types of recombinant region were generally detected with accuracies of about four bases either side. As the evolutionary rate increased (up to 1.5), the detection became more accurate. Improved accuracy was matched by a larger Z value. For all rates used, recent recombination between divergent taxa was detected with greater accuracy than ancient recombination.

Heterogeneous Rate Simulations

Figure 3A shows the mean accuracy of, and figure 3B the mean Z value associated with, detection of different-sized regions evolving at different rates with respect to the rest of the DNA sequence. Figure 3C shows the percentage of faster-evolving regions not detected for each scenario. Regions evolving at lower rates than the rest of the sequence do not tend to have a lower likelihood, even when rates of evolution are high (since transition probabilities for a base staying the same never become particularly low, even for long branches). Thus, such regions (e.g., invariant regions) are not used in the simulations. Other methods should be used to detect invariant sites (e.g., Hasegawa, Kishino, and Yano 1985; Goldman 1993b). When the anomalous region is evolving at a faster rate, increasing the ratio of this rate to the rest of the sequence improves the accuracy of detection and the chance of it being detected (fig. 3A and C). Such improvement also follows increased region size. Thus increased variation in the rates of evolution along actual DNA sequences means smaller anomalous regions can be detected, and the existence of larger anomalous regions means even small differences in rate can be detected. The relationship between the Z values and rate/size is shown in figure 3B to be the inverse of that for the accuracy values (average distance of the boundaries of the detected region from the actual region). Thus a higher Z value is associated with more accurate detection of anomalous regions.

Hominoid \( \psi \)-Globin

Analysis of bases 1 to 1000 of the \( \psi \)-globin pseudogene revealed no SPV. This is both expected and reassuring, since pseudogenes are usually assumed to evolve in a neutral, spatially homogeneous manner (Li, Gojobori, and Nei 1981; Hasegawa, Kishino, and Yano 1989).
HIV-1 V3 Region

Analysis of these sequences from a single patient (p82; Holmes et al. 1992) led to the identification of an anomaly in the region which corresponds to the amino acids RAFYTTGE (sequence corresponding to the consensus for subtype B) with a Z value of 4.94 (P << 0.01). This region lies between amino acids 322 and 329 inclusively of the envelope gene (following Ratner et al. 1985). This region is part of the principal antibody-mediated neutralizing domain of the HIV-1 env gene (expressed on the virion surface) (Javaherian et al. 1989), and so is expected to be subject to the selective pressures of immune surveillance within the human host. Such selection will cause a more rapid rate of evolution within this region and hence a low (anomalous) likelihood. However, more importantly, this region corresponds almost exactly to the cytotoxic T-lymphocyte (CTL) epitope for the human leukocyte antigen (HLA) type B-27 mediated immune response (Jardetzky et al. 1991). This epitope lies between amino acids 321 and 329 following Ratner et al. (1985). The HLA type of patient p82 has been reported (Zhang 1993), and one of the two B-locus alleles present in this patient is B-27. Thus, identification of the anomalous region together with knowledge of the patient’s HLA type strongly suggests that there is an important role of CTL-mediated killing of HIV-1 particles, via recognition of the HLA-B27 epitope, in this patient.

Neisseria argF Gene

Analysis of the argF gene from Neisseria isolates led to the simultaneous detection of several anomalous regions. With Z values of 3.41, 3.87, and 7.30, three anomalous regions were detected. Following the numbering system of Zhou and Spratt (1992), these three regions correspond to nucleotides 302-314, 355-359, and 466-479, respectively. These regions fall within, and indeed nearly define, the recombinant region previously detected in N. meningitidis using the maximized chi-square test of Maynard Smith (Maynard Smith 1992; Zhou and Spratt 1992). Just these three regions are detected as being anomalous since they are the polymorphic regions that fall within the recombinant section (see fig. 4A). Clearly, regions lacking significant polymorphism cannot be informative, or indeed anomalous (as previously mentioned, invariant sites never have a particularly low likelihood). Further subdivision of the recombinant region, i.e., the hypothesis of multiple recombination events within this region, was not significant when assessed using the maximized chi-square test. Figure 4B illustrates the conflicting trees obtained when using the entire sequence length, the recombinant
region, and the sequences with the recombinant region removed. When the phylogeny is constructed from just the recombinant region, the *N. meningitidis* isolates show clear grouping with *N. cinerea* as opposed to *N. gonorrhoeae* in the other phylogenies. This indicates that the recombinant region in the two *N. meningitidis* strains may derive from a commensal isolate closely related to *N. cinerea* (Zhou and Spratt 1992). The maximum-likelihood phylogenies for the non-recombinant region and the entire sequence length are congruent with the accepted phylogeny derived from DNA-DNA hybridization studies (Guibourdenche, Popoff, and Riou 1986), and group *meningitidis* with *gonorrhoeae* (cf. incongruence found by Zhou and Spratt (1992) using the entire sequence length and distance methods of phylogenetic reconstruction).

At a lower significance level with Z value 3.01 a further anomalous region was detected, corresponding
to nucleotides 728–818. This overlaps a region (803–833) detected as divergent by the maximized chi-square test on *N. meningitidis* and *N. gonorrhoeae* by Zhou and Spratt (1992). They presupposed that such divergence is a result of recombination, but failed to find homology of the region with any of the 10 other *Neisseria* species examined. Inspection of the aligned sequences reveals that in fact the entire region from 728 to 818 is highly polymorphic across all species examined, not just within the *N. meningitidis* lineage. In addition, the reconstructed phylogeny for this region (results not shown) is the same as the species phylogeny (fig. 4Biii). Thus, such polymorphism may simply be a result of possible near neutrality (due to possible lack of constraint on helices 7 and 8 and beta sheet 8; Murata and Scachman 1996; Nguyen et al. 1996).

**Discussion**

The simulation of evolutionary processes such as recombination and variable rates of evolution along a sequence provides important information about the efficiency and accuracy of detection of these processes by the method presented here. It reveals that recombinant regions are identified fairly accurately (±~4 bases either side) when the recombination event is either an ancient one, or a recent one between distantly related taxa (fig. 2). However, recent recombination events between closely related taxa were not detected. This lack of detection reflects the difficulty in determining the exact patterns of divergence of recent lineages in a phylogeny when the length of sequence data available is small. It is possible that as the size of the recombinant region increases such recombination events will become detectable.

The accuracy of detection of recombination improves as rates of evolution along the tree are increased. This is because there is more change, and hence more information, along branches. It is somewhat surprising that this is the case for ancient recombination, where evolution subsequent to the recombination event may be expected to obscure that event. However, such "saturation" may not be important in the case reported here because of the long branch separating the recombinant lineages (see fig. 1B). It would be interesting to look at ancient recombination events for trees not produced by the constant-population-size, neutral coalescent process used here. Because such trees may have many deep divergences (for instance, bushlike trees occur when sampling from an exponentially growing population; Slatkin and Hudson 1991), ancient recombination events could occur between lineages more closely related than in figure 1B, and may illustrate the problem of saturation as defined here.

For the heterogeneous rate simulations, as the ratio of the rate of evolution of the anomalous region to that of the rest of the sequence is increased, the accuracy of detection is increased, and smaller regions can be detected. This is expected, given the continual decrease in the likelihood of the anomalous region (under the global maximum-likelihood model) as its divergence increases.

When SPV is detected, improved accuracy is matched by an increased Z value. This means that the Z value associated with detection of an anomalous region gives a good estimate of how accurate the detection of this region is likely to be. Thus, for example, in the case of detection of recombination in *Neisseria*, the Z values (see fig. 4A) should give a good idea as to how close the identified regions are to the actual anomalous regions.

In the case of *Neisseria* it may be asked why only three components of the recombinant region are seen to be anomalous, and why not the whole region? The reason for this is that the method presented here looks for regions of maximum Z value. Type 1 statistical errors resulting from this process are avoided because it is the maximized $Q$ that is stored for each region size in the null simulations, and because over the different region sizes the Bonferroni corrected significance level is used (Miller 1966). In *Neisseria*, although the entire recombinant region may have a significantly larger Q value than expected for a model without SPV, it is the region(s) of maximum anomaly (Z value) that is/are detected. If regions are chosen which are significantly, but not maximally, anomalous, then large regions may be selected simply because they contain a small but anomalous region. Thus, in the method presented here, even if a region is significantly unlikely for a null, non-SPV model, it is not said to be anomalous if it overlaps a region with a higher Z value. Rather, that region is said to be anomalous. (The three regions in the *Neisseria* sequences detected as being anomalous are revealed in sequence alignments as being the most polymorphic parts of the recombinant region, and hence the most informative. It is for this reason that they reveal the recombination event and can be said to be anomalous.)

Once SPV has been discovered, it is clearly desirable to reanalyze anomalous regions in order to determine the underlying evolutionary cause. This is particularly important when biological knowledge concerning the nucleotide sequence is sparse. For sizeable regions, as in the case of recombination in *Neisseria*, maximum-likelihood phylogenetic reconstruction of the region is possible. This may result in a different topology, suggesting recombination or convergent evolution (as in figs. 1 and 4B). Alternatively, different maximum-likelihood substitution parameters may be estimated from the region (e.g., different overall rate of substitution, or different $t$: $v$ ratio). To test exactly which parameter(s) is/are responsible for the anomaly requires an assessment of the confidence limits about these parameters. Methods to estimate confidence sets about the maximum-likelihood topology have been developed (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989; Deby and Abele 1995), and estimation of confidence limits about substitution parameters and branch lengths is fairly simplistic. However, such approaches are rarely taken. It would be interesting to identify the important parametric changes, since these changes may give detailed information about the evolutionary process. For instance, a higher $t$: $v$ ratio may reflect a negative, conservative, selective pressure on the region (and vice ver-
difficult in some cases. The exact limit of an informative
the information it contains decreases, making reanalysis
difficult in some cases. The exact limit of an informative
the information it contains decreases, making reanalysis
difficult in some cases. The exact limit of an informative
the information it contains decreases, making reanalysis


PAUL M. SHARP, reviewing editor

Accepted November 19, 1996