Using Physical-Chemistry-Based Substitution Models in Phylogenetic Analyses of HIV-1 Subtypes

Jeffrey M. Koshi,[1] *David P. Mindell,*† *and Richard A. Goldstein**‡

*Biophysics Research Division, †Department of Biology and Museum of Zoology, and ‡Department of Chemistry, University of Michigan

HIV-1 subtype phylogeny is investigated using a previously developed computational model of natural amino acid site substitutions. This model, based on Boltzmann statistics and Metropolis kinetics, involves an order of magnitude fewer adjustable parameters than traditional substitution matrices and deals more effectively with the issue of protein site heterogeneity. When optimized for sequences of HIV-1 envelope (*env*) proteins from a few specific subtypes, our model is more likely to describe the evolutionary record for other subtypes than are methods using a single substitution matrix, even a matrix optimized over the same data. Pairwise distances are calculated between various probabilistic ancestral subtype sequences, and a distance matrix approach is used to find the optimal phylogenetic tree. Our results indicate that the relationships between subtypes B, C, and D and those between subtypes A and H may be closer than previously thought.

Introduction

If our understanding of evolutionary processes were complete, we could model the change of discrete characters in phylogenetic analyses and readily obtain the most reasonable phylogeny. Our knowledge of evolution is limited, however, and models of evolutionary change are necessarily simplified. While many evolutionary studies have focused on modeling the evolution in DNA, for which the relative frequences of different base changes can be represented with simple models, there has been increasing interest in performing phylogenetic analyses on proteins. The hope is that purifying selective pressure acting at the protein level can reduce saturation effects, allowing the delineation of more distant evolutionary relationships.

Modeling evolution at the amino acid level has, however, proven problematic. The larger size of the alphabet of amino acids means the determination of a much larger set of relative substitution rates than is necessary for DNA sequences. More significantly, the assumption inherent in traditional approaches, including use of substitution matrices, is that the rate or probability for any particular pairwise substitution is considered to be the same for all sequence positions. This assumption ignores the fact that different parts of proteins evolve under different selective pressures. Maintaining functionality of the protein may require conservation of amino acid identity at key locations in catalytic or dimerization sites, while a more general residue characteristic, such as hydrophobicity or polarity, may be preserved at other locations. Site substitutions elsewhere in the protein may preserve charge or side chain size or flexibility. Variable sites may evolve by random fixation of neutral or nearly neutral mutations. A given substitution, for instance, from histidine to phenylalanine, would represent a conserved substitution in locations in which aromaticity was important, but it would represent a nonconservative substitution if charge or ligation was relevant. There have been numerous demonstrations of heterogeneity in absolute and relative rates of amino acid change depending on local structure and function within the protein, as well as between proteins (Kimura and Ohta 1973; Miyata, Miyazawa, and Yasunaga 1979; Wako and Blundell 1994*a*, 1994*b*; Koshi and Goldstein 1995; Thorne, Goldman, and Jones 1996). Any reasonable model of amino acid change must be able to encompass the heterogeneous nature of the selective pressure without increasing the number of adjustable parameters beyond that which can be determined by the available data. This is especially difficult since the nature of the variability may be complicated and difficult to identify a priori, especially for proteins of unknown structure or function.

While approaches such as protein profiles and hidden Markov models that include site heterogeneity have been developed for performing similarity searches and recognition of homologs (Taylor 1986; Gribskov, McLachlan, and Eisenberg 1987; Krogh et al. 1994; Tatusov, Altschul, and Koonin 1994; Yi and Lander 1994), these approached have generally not been extended to address more specific questions of phylogenetics. The construction of these models generally involves the determination of 20 parameters for each location in the protein, each representing the probability of one of the 20 amino acids occurring at that location. In contrast, extending this approach to the modeling of site substitutions would require determination of the value of 380 parameters at each location, representing the probabilities of all possible substitutions. This is beyond the limits of what can be done with currently available data. Heterogeneity of substitution rates has been included in some models of both amino acid and DNA evolution (Koshi and Goldstein 1995; Yang 1995; Van de Peer et al. 1996; Felsenstein and Churchill 1996; Thorne, Goldman, and Jones 1996). These models, however, have been based on more limited assumptions of homogeneity, assuming either that the relative rates are constant

[1] Present address: Theoretical Biology and Biophysics Division, Los Alamos National Laboratory, Los Alamos, New Mexico.

Key words: molecular evolution, HIV-1 subtype, Metropolis kinetics, Boltzmann statistics, HIV-1 evolution.

Address for correspondence and reprints: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055. E-mail: richardg@umich.edu.

Mol. Biol. Evol. 16(2):173–179. 1999

or that substitution rates are the same for all locations sharing the same local structure.

The properties of proteins are not a function of the identities of the residues at each location but, rather, depend on the physical-chemical properties of these residues. The relative substitution rates can often be interpreted in terms of corresponding changes in these properties in a way that depends on the local context (Miyata, Miyazawa, and Yasunaga 1979; Koshi and Goldstein 1997). Motivated by this perspective, we recently developed a method for representing substitution matrices as a function of the physical-chemical properties of the amino acids (Koshi, Mindell, and Goldstein 1997; Koshi and Goldstein 1998). By using simple functions of small sets of such properties, the number of adjustable parameters to be determined is greatly reduced, allowing us to optimize substitution matrices for smaller data sets. In addition, we can explicitly include heterogeneity in substitution rates among sites so that different parts of the protein sequence are modeled as evolving under different constraints. The result is a simpler model, with greater sensitivity to protein structure and function, that can represent protein evolution with increased accuracy. While there have been other substitution models that use physical-chemical properties to assign distances between the amino acids (Goldman and Yang 1994; Schmidt 1995), these models have been constructed based on preconceived notions of what physical-chemical properties are important and have not addressed the issue that the similarities between amino acids will be context-dependent. In contrast, the model described below includes site heterogeneity in a natural way and allows the parameters in the model to be optimized by likelihood maximization based on data sets of homologous proteins.

As a demonstration of the use of our method in performing phylogenetic analyses, we use our model to reconstruct the phylogenetic relationships between the HIV-1 subtypes. Ten genetic subtypes of HIV-1 have been recognized based on monophyletic groups found in phylogenetic analyses and placed within a larger group, M (Louwagie et al. 1993; Korber et al. 1997). Subtypes A–E are the most common, with only a few known isolates belonging to the more recently recognized subtypes I and J. A small number of isolates comprise the distinctive outlier group O. Envelope protein sequences show about 70% similarity in comparisons between HIV-1 subtypes and about 80%–90% similarity in comparisons within subtypes. Resolution of phylogenetic relationships among HIV-1 subtypes is important in understanding the origin and spread of AIDS as well as in determining potential treatments, as subtypes may have phenotypic differences affecting transmission and susceptibility to alternative vaccines (Wolfs, Nara, and Goudsmit 1993; Birx et al. 1996). Previous analyses of HIV-1 subtype phylogeny have focused on DNA sequences using a variety of approaches (Louwagie et al. 1993; Myers 1994; Leitner et al. 1995, 1996: Delaporte et al. 1996; Van de Peer et al. 1996), and no consensus on topology has emerged. There has been a tendency to find close relationships between subtypes B and D and, to a lesser extent, between subtypes A, G, and H. In all cases, however, internodal distances between the subtypes appear similar, approximating a star phylogeny.

In this paper, we provide an analysis based on the sequences of the HIV-1 envelope (*env*) proteins. Our model is especially appropriate in this instance, as there seems to be a strong heterogeneity of substitution rates throughout the sequence (Starcich et al. 1986; Willey et al. 1986).

Materials and Methods

Our model of amino acid substitutions has two distinct parts: the inclusion of site heterogeneity by positing multiple types of sites, each changing according to a different substitution matrix, and the construction of simplified substitution matrices based on the underlying physical-chemical properties of the amino acids. It is the simplifications inherent in the construction of the substitution matrices that allow the site heterogeneity to be included without resulting in an unmanageable number of adjustable parameters. Our model has previously been described (Koshi, Mindell, and Goldstein 1997; Koshi and Goldstein 1998) and is summarized in the appendix. Briefly, we consider that any location in the protein can potentially belong to one of a number of ''site classes,'' possibly representing local structure or functional significance. Each site class is described by a different substitution matrix. No assumption is made about which locations correspond to which site classes; rather, each location in the protein has the same a priori probability of belonging to a given class. The propensity of any amino acid to be found at any site belonging to a particular site class is considered to be a simple adjustable function of the size and hydrophobicity of that amino acid, with two or four adjustable parameters. The substitution rate is an adjustable attempt rate times a relative fixation rate, which is equal to 1 if the resulting amino acid has a higher propensity for that site, or an exponentially decreasing rate for changes to amino acids with lower propensities. In this way, each substitution matrix can be defined by three to five adjustable parameters.

The complete substitution model is completely specified by the adjustable parameters defining the substitution matrix for each site class and the a priori probabilities. We used CLUSTAL W to align the sequences found within the various subtypes, and we used its implementation of the neighbor-joining method to construct phylogenetic trees for individual subtypes, with no assumed relationships between the subtypes. The model parameters were adjusted for the HIV-1 *env* proteins in subtypes A, C, and D by likelihood maximization. The midpoint of the longest branch was assumed to be the root. Only a central 780-residue region of the *env* protein was used. As shown in table 1, the evolutionary model based on simple kinetics including site heterogeneity was superior to either the Dayhoff PAM matrix or a single optimized substitution matrix, with the exception of the substantially different O subtype. After the model was optimized, the most likely distances between the roots of the various subtypes were calcu-

Table 1
Log-Likelihoods of the Data Representing the Logs of the Conditional Probabilities that the Current Sequences for Each of the Various Subtypes Would Result, Given a Particular Model for Evolution

Optimization Data Set	Number of Site Classes	HIV-1 Subtype								
		A	B	C	D	E	F	G	H	O
Traditional substitution matrices										
Dayhoff.....	1	−10,712	−15,035	−12,724	−7,634	−6,271	−4,944	−5,670	−1,595	−4,430
General	1	−10,848	−15,292	−12,894	−7.721	−6,347	−4,964	−5,732	−1,606	−4,464
HIV-1 *env*...	1	−10,378	−14,650	−12,452	−7,461	−6,160	−4,846	−5,574	−1,612	−4,370
Simple models										
HIV-1 *env*...	3	−10,966	−15,275	−12,841	−7,712	−6,389	−5,101	−5,772	−1,671	−4,776
	5	−10,519	−14,714	**−12,362**	**−7,422**	−6,236	−4,948	−5,622	−1,659	−4,677
	7	−10,420	**−14,504**	**−12,210**	**−7,360**	−6,163	−4,849	**−5,515**	−1,610	−4,558
	9	**−10,350**	**−14,420**	**−12,063**	**−7,302**	**−6,157**	**−4,803**	**−5,466**	**−1,568**	−4,534

NOTE.—Traditional substitution matrices assume that substitution rates are the same throughout the protein, that is, that there is a single ''site class.'' ''Dayhoff'' represents the matrix developed by Dayhoff, Schwartz, and Orcutt (1978). ''General'' and ''HIV-1 *env*'' represent substitution matrices optimized either over a general set of protein sequences or over the observed sequences of HIV-1 subtypes A, C, and D using methods described earlier (Koshi and Goldstein 1995). The ''Simple models'' represent the approach described in the text, where substitution rates are represented as a function of the physical-chemical parameters of the amino acids, and selective pressure heterogeneity is included by representing separate substitution rates for 3, 5, 7, or 9 different site classes. These models were also optimized over the A, C, and D subtypes. Bold numbers represent instances in which the simple models outperform all of the more standard models that ignore site heterogeneity.

lated based on the principle that the most likely evolutionary distance between the roots, given the observed sequences, is the distance that would maximize the probability that the observed sequences would result through the process of site substitutions. Optimal distances were calculated between all pairs of subtypes, and the best evolutionary tree was computed based on these distances in an exhaustive search, performed by inputing all possible trees into the Fitch routine from the PHYLIP package (Felsenstein 1993). The implementation of the more rigorous maximum-likelihood approach is currently in progress. In order to test the accuracy of our results, we performed a bootstrap test using 200 runs, each time selecting 100% of the data with resampling. The optimal evolutionary tree was then found for each of the 200 distance matrices generated by the bootstrap runs.

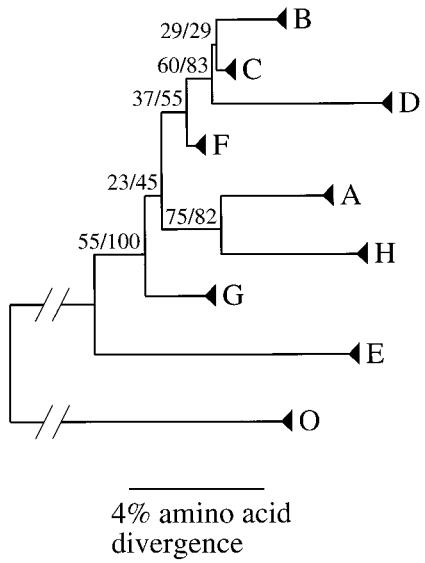

FIG. 1.—Phylogenetic hypothesis for HIV-1 subtypes based on *env* amino acids and a model accounting for evolutionary rate heterogeneity in different structural and functional regions of the protein (see text). Branches terminate at inferred ancestral sequences for each subtype. Each internal node is labeled with the percentage of 200 bootstrap runs that yielded a topology with that particular node being maintained when subtype O was specified as the outgroup (first) and when O was excluded (second).

Results and Discussion

Our optimal phylogenetic hypothesis for HIV-1 subtypes based on *env* amino acid sequences and our model incorporating evolutionary rate heterogeneity across functional site classes is shown in figure 1. The various subtypes are represented by triangles to indicate that tree branches connect subtype roots as represented by inferred ancestral sequences. Much of the uncertainty in the tree topology concerns the location of subtype O, due to its relative distinctiveness from the other subtype sequences. This is evident in the differences in bootstrap values given for nodes in which subtype O was designated as the outgroup (left number) and for those in which subtype O was excluded and no outgroup was specified (right number). The evolutionary distances between these roots and the most distant intrasubtype sequences are included in the tree in figure 2.

We found several differences between our tree and those proposed by others. Many phylogenetic hypotheses place subtype C farther from subtypes B and D than subtypes E or F (Louwagie et al. 1993; Leitner et al. 1995, 1996: Delaporte et al. 1996; Van de Peer et al. 1996). In contrast, our hypothesis places subtypes B, C, and D as a monophyletic group in 60% of the bootstrap runs in trees rooted with subtype O, and in 83% of the unrooted runs in which subtype O was excluded. The exact relationship of B, C, and D cannot be distinguished using our model, with ((B, C), D) and ((B, D), C) equally common in the bootstrap runs.

Our placement of subtype E is interesting not only because it is distant from B, C, and D, but also because

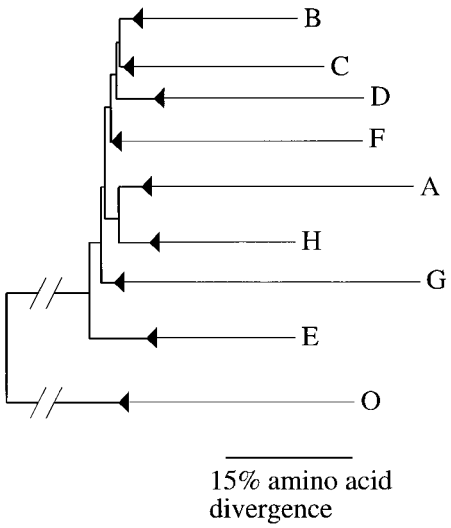

FIG. 2.—Optimal phylgeny for HIV-1 subtypes, as in figure 1, with branches (thin lines) indicating distances between inferred ancestral subtype sequences (dark triangles) and the most distant sequence for that same subtype.

it is distant from A. Subtype E appears to be a mosaic resulting from recombination between A and the original form of E, with the *env* gene being from the original E form (McCutchan et al. 1996). The results of our model suggest that this recombination involved relatively distant subtypes. We also found a sister relationship between subtypes A and H supported in 82% of the bootstrap analyses excluding subtype O.

In earlier work, we showed by comparison of log-likelihood scores that our simple models optimized over an HIV-1 *env* protein data set substantially outperformed more traditional substitution matrices on the highly dissimilar HIV-2 *env* protein data set, demonstrating the ability of the simple models to capture the evolutionary patterns found in different proteins evolving under similar selective pressure (Koshi and Goldstein 1998). Surprisingly, the simple models optimized on HIV *env* proteins from HIV-1 subtypes A, C, and D did not represent the evolutionary processes in HIV-1 subtype O as well as more traditional substitution matrices, as shown in table 1. It has been suggested that group O has only recently appeared in the human population (Gurtler 1996). Our results suggest that the evolutionary context may have an influence on the patterns of substitutions, especially for a protein evolving as rapidly as *env*.

Comparison with the known features of current subtype geographic distribution (Burke and McCutchan 1997) indicate a general lack of correlation between geographic proximity and phylogenetic relatedness. For example, subtypes B, C, and D form a monophyletic group in figure 1 but are found in disparate locales worldwide. This lack of correlation suggests a history of HIV1 subtype dispersal via human travel and migration rather than dispersal through geographically contiguous populations. Further phylogenetic analyses using additional proteins and additional isolates will help in assessing the reliability of our *env*-based tree.

The approach described here, using simple models to delineate amino acid substitution patterns by incorporating information about their physical-chemical properties, is a general approach that can be used in a wide variety of phylogenetic analyses, including maximum-likelihood formalisms. The utility of this approach stems from the benefits of incorporating a more specific model of evolutionary change in the phylogenetic analyses based on observed differences among HIV-1 *env* sequences, with fewer unsupported assumptions of rate homogeneity across amino acid sequence regions and functional classes. Our analyses modeling evolutionary change among amino acids support close phylogenetic relationships among HIV-1 subtypes B, C, and D and between subtypes A and H, whereas previous subtype analyses based on nucleic acids and more assumptions of rate homogeneity have been less resolved.

Acknowledgments

We would like to thank Darren Taverna for his assistance in deriving the models of amino acid substitution used in this work. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, NIH Grants GM08270 and LM0577, and NSF equipment grant BIR9512955.

APPENDIX

Inclusion of Site Heterogeneity

One of the aspects of amino acid substitutions that must be included in any reasonable model is the presence of variable relative and absolute rates in different structural and functional regions of the protein. In order to encompass this site heterogeneity, we consider that any location in the protein can belong to one of a number of ''site classes'' $\mathcal{S}_k$, each represented by a distinct substitution matrix $M^k_{i,j}$ describing the rate of replacement of amino acid $\mathcal{A}_i$ by amino acid $\mathcal{A}_j$ in a given length of evolutionary time. The specifications of the various types of site classes, as well as the assignment of various locations in the proteins to the different site classes, are not performed in advance. Rather, every location in the protein has the same a priori probability $P(k)$ for being a member of site class $\mathcal{S}_k$. As all locations must be in some site class, $\Sigma_k P(k) = 1$. The substitution matrices $M^k_{i,j}$ and probabilities $P(k)$ for all of the different site classes represent the parameters optimized based on analysis of the data set of homologous sequences.

Construction of Physical-Chemistry-Based Substitution Matrices

With 20 different amino acids, the complete specification of a substitution matrix would require the determination of 380 adjustable parameters, representing all possible substitutions. Setting the values for individual substitution matrices for each of a set of site classes would result in an unacceptably large number of adjustable parameters. For this reason, we developed sim-

pler substitution matrices that represented the probability of substitutions in terms of the underlying physical-chemical properties of the amino acids, rather than in terms of their identities (Koshi, Mindell, and Goldstein 1997; Koshi and Goldstein 1998). This was done using a model of propensities based on a Boltzmann distribution and substitution rates based on Metropolis kinetics, as described below.

Each amino acid $\mathcal{A}_i$ is considered to have a particular propensity value $F_k(\mathcal{A}_i)$ denoting its tendency to occur in any particular location described by site class $\mathcal{S}_k$. We would expect that amino acids with strong propensities for particular sites would be common, while amino acids with weak propensities for these sites would be rarer, although not completely absent. We formally define the relative propensities of the different amino acids in a location belonging to a given site class as equal to the logs of the relative probabilities that these amino acids would be found in that particular location. Inverting this definition of propensity results in an expression for $P_k(\mathcal{A}_i)$, the probability that amino acid $\mathcal{A}_i$ would be found in that location, as an exponential function of the propensity, similar in form to the Boltzmann distribution

$$P_k(\mathcal{A}_i) = \frac{e^{F_k(\mathcal{A}_i)}}{\sum_{i'} e^{F_k(\mathcal{A}_{i'})}} \tag{1}$$

The propensity of any amino acid in any location is a function of the physical-chemical properties of that amino acid, such as its hydrophobicity and size. In particular, we assume that the propensity can be expressed either as a linear function of these properties, so that

$$F_k(\mathcal{A}_i) = \alpha_k^H H_i + \alpha_k^B B_i, \tag{2}$$

or, with a quadratic dependence on these parameters

$$F_k(\mathcal{A}_i) = \alpha_k^H (H_i - H_0^k)^2 + \alpha_k^B (B_i - B_0^k)^2, \tag{3}$$

where H_i and B_i represent the hydrophobicity and bulk of amino acid $\mathcal{A}_i$ as characterized by the indices developed by Scheraga and co-workers (Kidera et al. 1985), and α_k^H, α_k^B, B_0^k, and H_0^k represent adjustable parameters. By using fixed physical-chemical characteristics for the amino acids, the propensities of all of the amino acids for each particular site class can be modeled by determining only two (α_k^H and α_k^B) or four (α_k^H, α_k^B, H_0^k, and B_0^k) parameters. Note that the propensity values are specific to each particular site class such that the same amino acid with given values for hydrophobicity and size would have different propensities for locations characterized by different site classes. Once the parameters defining $F_k(\mathcal{A}_i)$ are fixed, we can calculate the propensity for any particular amino acid to be in a location belonging to a particular site class.

The relative substitution rates are modeled with Metropolis kinetics (Metropolis et al. 1953), for which favorable substitutions to a more fit amino acid proceed with site-class-specific maximum substitution rate ν_k, and unfavorable substitutions occur with a rate of ν_k times an exponentially decreasing function of the resulting propensity change.

$$M_{i,j}^k = \begin{cases} \nu_k & \mid F_k(\mathcal{A}_j) > F_k(\mathcal{A}_i) \\ \nu_k e^{(F_k(\mathcal{A}_j) - F_k(\mathcal{A}_i))} & \mid F_k(\mathcal{A}_j) \le F_k(\mathcal{A}_i), \end{cases} \tag{4}$$

which represents the rate of substitutions as the product of an attempt rate ν_k, representing the probability that a random mutation with no effect on survivability would be fixed in the population, times an acceptance rate that decreases as the mutation becomes more unfavorable. Metropolis kinetics are the only possible means of maintaining a Boltzmann distribution of propensities while obeying detailed balance whereby all favorable mutations are accepted at the maximum rate. While other arrangements are possible, where favorable mutations have probabilities that depend on the resulting difference in propensities, the latter approximation involves the assumption that substitutions are dominated by neutral drift.

With Metropolis kinetics, the substitution matrix is determined by the propensity function with the inclusion of only one more adjustable parameter, ν_k, in addition to the two or four parameters that characterize the propensity function represented by equations (2) and (3). The model is then completely specified by the parameters representing the propensity functions for each site class, the maximum substitution rate for each site class, and the set of values of $P(k)$ subject to the constraint on their sum. As an example, a model with nine different site classes was used in the phylogenetic analysis of HIV-1 *env* proteins described below: four site classes with propensities that depend linearly on hydrophobicity and bulk, and five with a quadratic dependence on these properties. Since the linear site classes have substitution matrices that are each specified by three adjustable parameters, the substitution matrices for the quadratic site classes are specified by five adjustable parameters, and there are eight a priori probabilities (one fewer than the number of site classes), there are a total of 45 adjustable parameters, approximately an order of magnitude fewer parameters than the 380 needed for a single standard substitution matrix neglecting site heterogeneity. With fewer adjustable parameters, we can optimize these matrices for specific protein types.

Optimization of the Model

These adjustable parameters are set using likelihood maximization, as described in earlier work (Koshi and Goldstein 1995) using a quadratic programming algorithm (Gill et al. 1986). For a set of homologous proteins, we first construct a sequence alignment and evolutionary tree using CLUSTAL W (Thompson, Higgins, and Gibson 1994). We then calculate the probability for each possible set of substitutions resulting in the currently observed sequences. Consider a single location l in the multiple alignment. The amino acids in the homologous proteins are represented by $\{\mathcal{A}\}_l'$, where the prime sign indicates knowledge of only the currently existing sequences. We do not know the identity of the amino acid at the internal nodes of the phylogenetic tree, so we must sum explicitly over all possibilities at these positions in order to calculate $P(\{\mathcal{A}\}_l' \mid M_{i,j}^k)$, the condi-

tional probability that these amino acids would result given that the location can be characterized as belonging to site class $\mathcal{S}_k$ with substitution matrix $M^k_{i,j}$.

We do not know a priori what site class this location belongs to. We must therefore consider all possibilities explicitly. The total probability of residues $\{\mathcal{A}\}'_l$ resulting from the model with the set of substitution matrices $\{M^k_{i,j}\}$ is obtained by calculating the probability that these residues would result if the site belonged to site class $\mathcal{S}_k$ times the probability that site $\mathcal{S}_k$ is the appropriate classification, equal to $P(\{\mathcal{A}\}'_l \,|\, M^k_{i,j})P(k)$, and summing over all possible site classes to yield

$$P(\{\mathcal{A}\}'_l \,|\, \{M^k_{i,j}\}) = \sum_k P(\{\mathcal{A}\}'_l \,|\, M^k_{i,j})P(k). \qquad (5)$$

The optimal model, then, is the model with the set of parameters that maximizes this probability multiplied over all locations in all sets of the homologous proteins.

As mentioned above, there is no attempt to characterize the locations in the protein as belonging to one site class or another. The properties of these various sites emerge during the optimization procedure. It is possible, however, to calculate a posteriori the probability $P(k \,|\, \{\mathcal{A}\}'_l)$ that a given location is a member of site class $\mathcal{S}_k$ based on the amino acids found at that location, $\{\mathcal{A}\}'_l$,

$$P(k \,|\, \{\mathcal{A}\}'_l) = \frac{P(\{\mathcal{A}\}'_l \,|\, M^k_{i,j})P(k)}{\sum_{k'} P(\{\mathcal{A}\}'_l \,|\, M^{k'}_{i,j})P(k')}. \qquad (6)$$

One of the advantages of our method is that the small number of adjustable parameters relative to substitution matrices allows us to focus on the patterns specific for specific types of proteins under their unique selective pressure. In earlier work, we optimized the model for the test set of HIV-1 *env* proteins and tested these models on the evolutionary relationships of the patterns of the HIV-2 *env* proteins (Koshi, Mindell, and Goldstein 1997; Koshi and Goldstein 1998). Our results showed that the inclusion of site heterogeneity increased the log-likelihood of the test data significantly relative to the standard Dayhoff substitution matrix or a single traditional substitution matrix optimized for the test set.

Use of the Model for HIV-1 Subtype Phylogenetic Analysis

We were interested in constructing a phylogenetic tree outlining the relationship between previously constructed rooted phylogenetic trees for the various subtaxa. The most likely distances between the roots of the various subtypes were calculated by maximizing the probability that the observed sequences would result through the process of site substitutions. This was done by first computing $P_{j,x}(\{\mathcal{A}\}'_{j,x} \,|\, k,\ \mathcal{A}_r)$, the conditional probability that observed amino acids $\{\mathcal{A}\}'_{j,x}$ would be found at position j in subtype isolate x given that the ancestral sequence for the subtype contained $\mathcal{A}_r$ at that location and the location could be described by site class $\mathcal{S}_k$. $P_{j,x}(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y},\ k)$, the conditional probability that observed sequences of subtype x at location j would be given by $\{\mathcal{A}\}'_{j,x}$ and observed sequences of subtype y at the same location would be given by $\{\mathcal{A}\}'_{j,y}$ if the distance between the two ancestral sequences were $d_{x,y}$, and, again, if the site belonged to site class $\mathcal{S}_k$, is calculated by considering the probability of any pair of residues occurring in the two ancestral sequences times the probability that the observed sequences in the two subtypes would consequently result, expressed as

$$\begin{aligned} P(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y},\ k) &= \sum_{r,r'} P_{j,x}(\{\mathcal{A}\}'_{j,x} \,|\, k,\ \mathcal{A}_r)P_{j,y}(\{\mathcal{A}\}'_{j,y} \,|\, k,\ \mathcal{A}'_r) \\ &\quad \times P_k(\mathcal{A}_r)M^k_{r,r'}(d_{x,y}), \end{aligned} \qquad (7)$$

where $P_k(\mathcal{A}_r)$ is given by equation (1) and $M^k_{r,r'}(d_{x,y})$ is equal to the probability that amino acid $\mathcal{A}_r$ would mutate to amino acid $\mathcal{A}_{r'}$ in evolutionary distance $d_{x,y}$ in a location described by site class $\mathcal{S}_k$. As emphasized above, we do not identify a priori which locations in the protein correspond to the different site classes. For this reason, $P(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y})$, the conditional probability that the observed sequences would result in these two subtypes in any site class given distance $d_{x,y}$, is obtained through a weighted sum over all possible site classes

$$P(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y}) = \sum_k P(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y},k)P(k). \qquad (8)$$

Note that this equation is quite different from what would be obtained if the substitution matrices themselves were averaged over site classes at the beginning of the calculation.

The conditional probability of observed sequences resulting for subtypes x and y given distance $d_{x,y}$ between ancestral nodes is estimated as the product of $P(\{\mathcal{A}\}'_{j,x},\ \{\mathcal{A}\}'_{j,y} \,|\, d_{x,y})$ over all locations j. This probability was maximized in order to find the most likely value of $d_{x,y}$. Optimal distances were calculated between all pairs of subtypes, and the best evolutionary tree was computed based on these distances in an exhaustive search, performed by inputing all possible trees into the Fitch routine from the PHYLIP package (Felsenstein 1993).

LITERATURE CITED

BIRX, D. L., T. VANCOTT, N. MICHAEL, J. MCNEIL, N. STAMATOS, B. GILLIAM, R. DAVIS, J. CARR, and F. E. MCCUTCHAN. 1996. Summary of track A: basic science. AIDS **10**:S85–S106.

BURKE, D. S., and F. E. MCCUTCHAN. 1997. Global distribution of human immunodeficiency virus − 1 clades. Pp. 119–126 *in* V. T. DEVITA JR., S. HELLMAN, and S. A. ROSENBERG, eds. AIDS: etiology, diagnosis, treatment, and prevention. Lippincott-Raven, Philadelphia.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. P. 345 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

DELAPORTE, E., W. JANSSENS, M. PEETERS et al. (17 co-authors). 1996. Epidemiological and molecular characteristics of HIV infection in Gabon, 1986–1994. AIDS **10**:903–910.

FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FELSENSTEIN, J., and G. A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

GILL, P. E., S. J. HAMMARLING, W. MURRAY, M. A. SAUNDERS, and M. H. WRIGHT. 1986. User's guide for MPSOL. Version 4.0. Department of Operations Research, Stanford University Systems Optimization Laboratory technical report 86-2.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GRIBSKOV, M., A. D. MCLACHLAN, and D. EISENBERG. 1987. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA **84**:4355–4358.

GURTLER, L. G., L. ZEKENG, J. M. TSAGUE, A. VAN BRUNN, Z. E. AFANE, J. EBERLE, and L. KAPTUE. 1996. HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV. Arch. Virol. **11**:195–202.

KIDERA, A., Y. KONISHI, M. OKA, T. OOI, and H. A. SCHERAGA. 1985. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem. **4**:23–55.

KIMURA, M., and T. OHTA. 1973. Mutation and evolution at the molecular level. Genet. Suppl. **73**:19–35.

KORBER, B., B. FOLEY, T. LEITNER, F. MCCUTCHAN, B. HAHN, J. W. MELLORS, G. MYERS, and C. KUIKEN. 1997. Human retroviruses and Aids 1997: a compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Lab, Los Alamos, N.M.

KOSHI, J. M., and R. A. GOLDSTEIN. 1995. Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. Protein Eng. **8**:641–645.

———. 1997. Mutation matrices and physical-chemical properties: correlations and implications. Proteins **27**:336–344.

———. 1998. Mathematical models of natural amino acid site mutations. Proteins **32**:289–295.

KOSHI, J. M., D. P. MINDELL, and R. A. GOLDSTEIN. 1997. Beyond mutation matrices: physical-chemistry based evolutionary models. Pp. 80–89 *in* S. MIYANO and T. TAKAGI, eds. Genome informatics 1997. Universal Academy Press, Tokyo.

KROGH, A., M. BROWN, I. S. MIAN, K. SJÖLANDER, and D. HAUSSLER. 1994. Hidden Markov models in computational biology. J. Mol. Biol. **235**:1501–1531.

LEITNER, T., D. ESCANILLA, S. MARQUINA, J. WAHLBERG, C. BROSTROM, H. B. HANSSON, M. UHLEN, and J. ALBERT. 1995. Biological and molecular characterization of subtypes D, G, and A/D recombinant HIV-1 transmission in Sweden. Virology **209**:136–146.

LEITNER, T., G. KOROVINA, S. MARQUINA, T. SMOLSKAYA, and J. ALBERT. 1996. Molecular epidemiology and MT-2 cell tropism of Russian HIV type 1 variants. AIDS Res. Hum. Retroviruses **12**:1595–1603.

LOUWAGIE, J., F. E. MCCUTCHAN, M. PEETERS et al. (11 coauthors). 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS **7**:769–780.

MCCUTCHAN, F. E., M. O. SALMINEN, J. K. CARR, and D. S. BURKE. 1996. HIV-1 genetic diversity. AIDS **10**:S13–S20.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER. 1953. Equation of state calculations for fast computing machines. J. Chem. Phys. **21**: 1087.

MIYATA, T., S. MIYAZAWA, and T. YASUNAGA. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12**:219–236.

MYERS, G. 1994. Tenth anniversary perspectives on Aids–HIV: between past and future. AIDS Res. Hum. Retroviruses **10**: 1317–1324.

SCHMIDT, W. 1995. Phylogeny reconstruction for protein sequences based on amino acid properties. J. Mol. Evol. **41**: 522–530.

STARCICH, B. R., B. H. HAHN, G. M. SHAW, P. D. MCNEELY, S. MODROW, H. WOLF, E. S. PARKS, W. P. PARKS, S. F. JOSEPHS, and R. C. GALLO. 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell **45**:637–648.

TATUSOV, R. L., S. F. ALTSCHUL, and E. V. KOONIN. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA **91**:12091–12095.

TAYLOR, W. 1986. The classification of amino acid conservation. J. Theor. Biol. **119**:205–218.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

THORNE, J. L., N. GOLDMAN, and D. T. JONES. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

VAN DE PEER, Y., W. JANSSENS, L. HEYNDRICKX, K. FRANSEN, G. VAN DER GROEN, and R. DE WACHTER. 1996. Phylogenetic analysis of the env gene of HIV-1 isolates taking into account individual nucleotide substitution rates. AIDS **10**: 1485–1494.

WAKO, H., and T. BLUNDELL. 1994*a*. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J. Mol. Biol. **238**:682–692.

———. 1994*b*. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J. Mol. Biol. **238**:693–708.

WILLEY, R. L., R. A. RUTLEDGE, S. DIAS, T. FOLKS, T. THEODORE, C. E. BUCKLER, and M. A. MARTIN. 1986. Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome retrovirus. Proc. Natl. Acad. Sci. USA **83**:5038–5042.

WOLFS, T. F. W., P. L. NARA, and J. GOUDSMIT. 1993. Genotypic and phenotypic variation of HIV-1: impact on AIDS pathogenesis and vaccination. Chem. Immunol. **56**:1–33.

YANG, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics **139**:993–1005.

YI, T. M., and E. S. LANDER. 1994. Recognition of related proteins by iterative template refinement. Protein Sci. **3**: 1315–1328.

TONY DEAN, reviewing editor

Accepted October 9, 1998