Beyond Mutation Matrices: Physical-Chemistry Based Evolutionary Models

Jeffrey M. Koshi [1] **David P. Mindell** [2] **Richard A. Goldstein** [3]
jkoshi@umich.edu mindell@umich.edu richardg@umich.edu

[1] Biophysics Research Division, University of Michigan
Ann Arbor, MI, 48109-1055, USA

[2] Department of Biology and Museum of Zoology, University of Michigan
Ann Arbor, MI, 48109-1055, USA

[3] Department of Chemistry and Biophysics Research Division, University of Michigan
Ann Arbor, MI, 48109-1079, USA

Abstract

We describe a model for characterizing site mutations in evolving proteins. By representing the fitness of each of the amino acids as a function of the physical-chemical properties of that amino acid, and constructing mutation matrices based on Boltzmann statistics and Metropolis kinetics, we are able to greatly reduce the number of adjustable parameters. This allows us to include site heterogeneity in the model, as well as to optimize the model for specific protein types. We demonstrate the applicability of the model by investigating the phylogenetic relationship between various subtypes of HIV-1.

1 Introduction

A large number of methods have been developed for modeling the evolution of biological sequences. Phylogenetic reconstructions have usually been based on analyzing sequences at the DNA level for a number of important reasons. The small size of the alphabet of base-pairs resulting in only twelve possible transitions means that the statistics of base changes can be described with only a few parameters. Focusing on non-coding regions eliminates the need to complicate the model to account for selective pressure. Coding regions can be analyzed by gathering statistics about the substitution rate as a function of codon position, or of synonymous *vs.* non-synonymous substitutions. Simplicity also suggests that the models may be more universal, in that it is likely that the rate of different base changes will not be highly species-dependent.

In contrast, modeling evolution at the amino acid level contains numerous difficulties. There are a total of 380 possible transitions between amino acids, not counting insertions and deletions. Generally these are modeled with a mutation matrix, a 20 × 20 array that represents the probability of any amino acid mutating to any other in a given length of evolutionary time, with values fixed through an analysis of sets of homologous proteins. Most approaches for deriving mutation matrices rely on the approach originally derived by Dayhoff and Eck, based on the relative number of different amino acids aligned to each other in pairs of closely-related homologous proteins [2]. Others have used variations of the original Dayhoff approach, including the use of blocks of aligned sequences or the alignment of three-dimensional structures [1, 10, 12, 18, 19, 23, 25] In a contrasting approach, based on the fact that site-mutations tend to conserve important properties, a number of investigators have developed substitution matrices based on the physical-chemical differences between the amino acids [5, 6, 8, 21, 22, 24]. Such matrices have been useful at generating insight, but have not proven their use for quantitative applications [11].

As mentioned above, analysis of non-coding regions of DNA remove most of the need to consider selective pressure. In contrast, the situation for proteins is more complicated. The selective pressure at each location in the protein depends heavily on the characteristics of that location - whether that part of the amino acid chain is exposed to solvent or buried, is in one secondary-structure or another, is involved in important tertiary contacts, or has some functional significance such as being in a binding, dimerization, or catalytic site. The relative rates of amino acid substitutions will also depend on the nature of the protein, whether it is a membrane, globular, extra-cellular, intra-cellular, regulatory, or structural protein. These distinctions are generally ignored in the construction of amino acid substitution matrices. The probability of any amino acid mutating to any other is considered to be independent of the nature of the protein or the organism or the role of that particular amino acid. There has been some work developing specific mutation matrices as a function of local structure [14, 27, 28]. While these approaches address some issues of site heterogeneity, they still suffer from a number of inherent limitations. These approaches cannot deal with proteins of unknown structure, since the categorization of the sequences into different types of locations is based on prior knowledge. In addition, the division of the protein into different types of locations is based on criteria that may not be the most important, ignoring, for instance, functional constraints. Dividing the data base into different categories and finding appropriate mutation matrices for each category requires more data, exacerbating the limitations caused by considering only proteins of known structure. And the resulting model still assumes that all amino acids in a given stuctural type are under similar evolutionary pressures and will have similar rates of mutations.

One conceivable approach towards this situation is to imagine that there are a variety of site classes, depending upon unspecified and possibly currently unknowable considerations, each with a particular set of mutation rates defining a separate mutation matrix. The evolutionary patterns could then be modeled by the set of mutation matrices and the probabilities that any location in the protein would belong to one or the other of the site classes. When pairs of proteins were considered, the result would be identical to that of a single mutation matrix equal to a weighted sum of the various site-class specific mutation matrices. If, however, there were *sets* of homologous proteins, the correlations in the mutations between different pairs of the set could be used to both adjust the parameters in the model as well as to provide information about which locations belonged to which site class with what probability. There are, of course, practical problems involved with this approach. There would be an explosion in the number of adjustable parameters, presenting a difficult optimization problem and necessitating a large and extensive protein data base. It would certainly be difficult to imagine optimizing a model like this for specific types of proteins.

If, however, it were possible to greatly reduce the number of adjustable parameters, then an approach such as the one described above might be tractable. If the number of parameters could be sufficiently reduced, it might be possible to develop models for specific protein types, increasing our ability to recognize homologs and to perform phylogenetic analyses. There would be a sizable advantage to using protein rather than DNA sequences – as much as the evolution of DNA base-pairs is easy to understand because the selective pressure is weak, the same lack of selective pressure also results in a rapid saturation of the mutations, preventing the analysis of more distant evolutionary relationships. The enhanced selective pressure acting on proteins could conversely serve to make delineating these distant homologies possible.

How can we reduce the number of adjustable parameters to a reasonable level? To do this, we take advantage of two insights. The first is that the closer the functional form mirrors the nature of the data, the fewer adjustable parameters are necessary to achieve a required accuracy. The second insight goes back to the original construction of mutation matrices based on differences in physical-chemical properties. More recently, we investigated how these properties change during site-mutations [15, 16]. The properties of the amino acids are fixed and universal. If a mutation matrix could be constructed that is a function of these *properties* rather than of the amino acids themselves, then the large alphabet of amino acid types would not increase the number of adjustable parameters. The main challenge is to create such a matrix using insights from physical chemistry and evolutionary biology.

The second challenge is to create a way of optimizing the adjustable parameters for sets of homologous proteins. As mentioned earlier, in order to untangle the various differing site classes, we need to consider sets containing three or more homologous proteins, so that the presence of correlations in the mutations at each position can inform the optimization procedure. It is not appropriate to consider each pair of the set independently, as all of the sequences are coupled by their evolutionary heritage. Luckily, we already addressed this issue in earlier work, where we used estimation maximization to optimize mutation matrices based on larger sets of homologous proteins [14]. Our approach is to directly model the evolutionary process, considering the probability that the set of current sequences would result by summing over all possible evolutionary paths. The optimal mutation matrix is the matrix that maximizes this probability.

In this article, we describe site-class specific mutation matrices that represent the mutation rates with only three to five adjustable parameters [17]. This is done by considering that the fitness of any amino acid for any location in the protein is a simple functional form of the physical-chemical properties of that amino acid. We then use the Boltmann equation and Metropolis kinetics to convert these fitnesses into mutation matrices. As a result, we are able to include site-heterogeneity directly in the model, and still have a small enough set of adjustable parameters to optimize the model for a specific class of proteins, the envelope (*env*) proteins of HIV-1. We demonstrate that, in spite of having an order of magnitude fewer adjustable parameters than even a single mutation matrix, we are able to better represent the evolutionary data for this class of proteins [17]. We then demonstrate the applicability of this model by showing that our model better explain the mutations of the evolutionarily-distinct HIV-2 *env* proteins. Finally, we apply this model to explore the phylogenetic relationships between the subtypes of HIV-1.

2 Theory

2.1 The model

As mentioned in the introduction, we consider that the protein consists of different types of locations under different forms of evolutionary pressure, categories that we call "site classes". Each site class $\mathcal{S}_k$ is characterized by an individual mutation matrix $M^k_{i,j}$ describing the probability that an amino acid $\mathcal{A}_i$ would mutate to amino acid $\mathcal{A}_j$ in a given period of evolutionary time. The probability that any location in the protein would belong to site class $\mathcal{S}_k$ is equal to $P(k)$. As all locations belong to *some* site class,

$$\sum_k P(k) = 1 \tag{1}$$

We make no attempt to define the site classes *a priori*, letting them be defined by the optimization procedure. In this way we do not need to know or postulate anything about the structure or function of the proteins.

We consider that the physical-chemical properties of each amino acid $\mathcal{A}_i$ can be represented by a set of factors $\{\theta^\gamma_i\}$. These factors could presumably be anything, from size to charge to hydrophobicity to α-helical propensity. We further postulate that $F_k(\mathcal{A}_i)$, the fitness of any amino acid $\mathcal{A}_i$ in site class $\mathcal{S}_k$, can be represented as a simple site-class dependent function of these factors. We assume for simplicity that these factors enter into the fitness in an additive way, so that we can write

$$F_k(\mathcal{A}_i) = \sum_\gamma F_{k,\gamma}(\theta^\gamma_i) \tag{2}$$

where $F_{k,\gamma}(\theta^\gamma_i)$ represents the contribution to the fitness due to physical-chemical factor θ^γ_i describing amino acid $\mathcal{A}_i$. Again, for simplicity we assume that the specific functional forms for $F_{k,\gamma}(\theta^\gamma_i)$ are either linear in the physical-chemical factor

$$F_{k,\gamma}(\theta^\gamma_i) = \alpha_{k,\gamma} \;\; \theta^\gamma_i \tag{3}$$

or quadratic

$$F_{k,\gamma}(\theta_i^\gamma) = \alpha_{k,\gamma} \left(\theta_i^\gamma - \theta_{k,\gamma}^{\text{opt}}\right)^2 \tag{4}$$

where $\alpha_{k,\gamma}$ and $\theta_{k,\gamma}^{\text{opt}}$ are parameters that depend upon the site class $\mathcal{S}_k$. As can be seen, the fitness of *all* of the amino acids for sites in this particular site class can be defined once the values of $\alpha_{k,\gamma}$ and $\theta_{k,\gamma}^{\text{opt}}$ are specified.

It is not a trivial problem to decide what physical-chemical factors to include in our model. As many of the measurable amino acid properties are highly correlated, we take advantage of the work of Kidera *et al.*, who derived four orthogonal indices that encompassed most of the variation over a set of 180 different amino acid properties [13]. These factors correlated predominantly with α-helical propensity, bulk, β-sheet propensity, and hydrophobicity. We can simplify the model further by considering our earlier work, that showed that hydrophobicity and size tended to be more conserved than secondary-structure propensity [16]. We then consider that $F_k(\mathcal{A}_i)$ only depends upon the hydrophobicity and the size, as characterized by the appropriate Kidera factors, θ_i^H and θ_i^B, respectively. As a result, the fitness can be defined by only two or four adjustable parameters, depending upon whether a linear fitness function (equation 3) or quadratic fitness function (equation 4) is used.

We assume that the probability that any amino acid is found in a location characterized by a given site class is a Boltzmann-function of the fitness

$$P^k(\mathcal{A}_i) = \frac{e^{F_k(\mathcal{A}_i)}}{\sum_{i'} e^{F_k(\mathcal{A}_{i'})}} \tag{5}$$

(As large fitness values are favorable, the sign of the exponential is opposite to the normal Boltzmann formula.) Alternatively, the Boltzmann function can be inverted and the fitness function defined in terms of the probability for a given amino acid to be found in that location.

It is reasonable to expect that mutations would obey detailed balance, that is, $P^k(\mathcal{A}_i)M_{i,j}^k = P^k(\mathcal{A}_j)M_{j,i}^k$. If we in addition assume that all favorable mutations are accepted at a constant rate ν_k, it can be shown that site-mutations must obey Metropolis kinetics [20], where unfavorable mutations are accepted at an exponentially-decreasing function of the change in fitness

$$M_{ij}^k = \begin{cases} \nu_k & | \quad F_k(\mathcal{A}_j) > F_k(\mathcal{A}_i) \\ \nu_k \, e^{(F_k(\mathcal{A}_j) - F_k(\mathcal{A}_i))} & | \quad F_k(\mathcal{A}_j) \leq F_k(\mathcal{A}_i) \end{cases} \tag{6}$$

Once the fitness functions are set by $\alpha_{k,\gamma}$ and $\theta_{k,\gamma}^{\text{opt}}$, the mutation matrix only requires fixing one more parameter, the maximum fixation rate ν_k. The evolutionary model is then completely defined by the set of $\alpha_{k,\gamma}$ and $\theta_{k,\gamma}^{\text{opt}}$ values, by the values of ν_k, and the values of $P(k)$ provided that they satisfy equation 1. For instance, the evolutionary model used for the phylogenetic work described below has a total of nine site classes; four with linear dependences on hydrophobicity and bulk, and five with quadratic dependences on these factors. The resulting model has a total of 45 adjustable parameters, compared with 380 for a single traditional mutation matrix that neglects site-heterogeneity.

2.2 Optimization and testing

We use our previously-developed estimation-maximization method to set the various parameters in the model [14, 17]. We construct a phylogenetic tree for each set of homologous proteins using the program ClustalW [26]. We then analyze each location in the set of homologs separately.

Consider a typical protein location l, with four homologous proteins, as shown in Figure 1, where the current sequences at that location are represented by $\{\mathcal{A}_l\}'$, in this case consisting of two alanines, one glycine, and one leucine, at nodes D, E, F, and G, respectively. (The prime indicates that the set of amino acids only represents the amino acids at the root of the tree.) If we knew the identity of the residues at the other nodes in the tree, we could easily calculate the probability of that particular set of mutations necessary for the current-day residues to exist at the leaves of the tree. As we do not know the identities of these other residues, we have to sum over all of the possibilities at each position. If there were only a single mutation matrix we could write the likelihood of $\{\mathcal{A}_l\}'$ given mutation matrix $M_{i,j}$ as

$$\begin{aligned} P(\{\mathcal{A}_l\}'|M_{i,j}) = & \sum_{\mathcal{A}_A,\mathcal{A}_B,\mathcal{A}_C} P(\mathcal{A}_A) \\ & \times M_{\mathcal{A}_A,\mathcal{A}_B}(d_{AB})\ M_{\mathcal{A}_B,\mathrm{Ala}}(d_{BD})\ M_{\mathcal{A}_B,\mathrm{Ala}}(d_{BE})\ M_{\mathcal{A}_A,\mathcal{A}_C}(d_{AC})\ M_{\mathcal{A}_C,\mathrm{Gly}}(d_{CF})\ M_{\mathcal{A}_C,\mathrm{Leu}}(d_{CG}) \end{aligned} \tag{7}$$

where $M_{\mathcal{A}_A,\mathcal{A}_B}(d_{AB})$ is the probability that amino acid $\mathcal{A}_A$ would mutate to $\mathcal{A}_B$ in evolutionary time d_{AB}, where d_{AB} is the time between nodes A and B, computed by taking mutation matrix $M_{i,j}$ to the appropriate power.

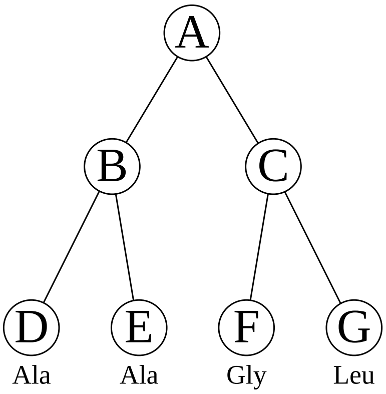

Figure 1: Example evolutionary relationship between four current sequences, represented as nodes D, E, F, and G, and their root sequences, represented by nodes A, B, and C.

For our model with a variety of site classes, each characterized by mutation matrix $M^k_{i,j}$, equation 7 is summed over all of the different site classes

$$P(\{\mathcal{A}_l\}') = \sum_k P(\{\mathcal{A}_l\}'|M^k_{i,j})\ P(k) \tag{8}$$

$P(\mathcal{A}_A)$ in equation 7, now given by the site-class specific $P^k(\mathcal{A}_A)$, is computed using equation 5. The total probability that all of the sets of homologous proteins would arise from our evolutionary model is computed by taking the product of $P(\{\mathcal{A}_l\}')$ over all locations l in all of the sets of homologous proteins. The optimal model, then, is the model that maximizes this total probability. Optimization of the adjustable parameters defining the model was performed using a sequential quadratic programming algorithm [7] from the NAG software package (Numerical Algorithms Group Ltd, Oxford, UK). The ability of a given model to represent the data is presented as a Q value, defined by $Q = \log[P(\text{Model})] - \log[P(\text{Random})]$, where $\log[P(\text{Model})]$ is the log of the probability that the given model would produce the data, and $\log[P(\text{Random})]$ is the probability that the data would result from purely neutral drift where all mutations were equally likely.

In all such models, especially those adopted to limited data sets, it is important to understand if we are learning mutational patterns characteristic of a broader set of proteins or simply memorizing the particular proteins used in the optimization. It is then important to construct an equivalent set of proteins of the same class, but with a completely disjoint evolutionary, to verify the model. The probability for this test set can be computed using equation 8, and the Q values for different models compared.

2.3 Phylogenetic reconstruction

The sets of mutation matrices can find applications in all of the typical applications for mutation matrices, including homolog recognition, ancestral reconstruction, and the construction of phylogenetic trees. We concentrate on the latter application.

Although our model could be easily adapted to maximum-likelihood methods, we initially implemented a distance-matrix method. $P(\mathcal{A}_{l,x}, \mathcal{A}_{l,y}|d_{x,y})$, the probability that residues $\mathcal{A}_{l,x}$ and $\mathcal{A}_{l,y}$ would occur at location l in two proteins, x and y, if the two sequences were separated by a distance $d_{x,y}$ is

$$P(\mathcal{A}_{l,x}, \mathcal{A}_{l,y}|d_{x,y}) = \sum_k P^k(\mathcal{A}_{l,x})\ \ M_{\mathcal{A}_{l,x},\mathcal{A}_{l,y}}(d_{x,y})\ \ P(k) \tag{9}$$

The total probability for the two sequences to be separated by distance $d_{x,y}$ is obtained by taking the product of $P(\mathcal{A}_{l,x}, \mathcal{A}_{l,y}|d_{x,y})$ over all locations in the aligned set of proteins. The most likely distance can then easily be obtained by finding the value of $d_{x,y}$ that optimizes this probability.

In the application described below, we want to find the distance between subclades comprised of multiple members. Rather than compute the most likely distance between all members of the various subclades, we instead find the optimal distance between the roots of these subclades. The extension of equation 9 is straight-forward, assuming the phylogenetic relationship between the members of the subclade are known, and a root node identified. First, $P(\{\mathcal{A}\}'_{l,x}|k, \mathcal{A}_r)$ is computed, representing the probability that current-day amino acids $\{\mathcal{A}\}'_{l,x}$ would be found at position l of the leaves of the tree for subtype x given that the root sequence to the subtype contained $\mathcal{A}_r$ at that location *and* the location could be described by site class $\mathcal{S}_k$. As an example, for the set of sequences represented in Figure 1,

$$\begin{aligned} P(\{\mathcal{A}\}'_{l,x}|k, \mathcal{A}_r) = & \sum_{\mathcal{A}_B,\mathcal{A}_C} M^k_{\mathcal{A}_r,\mathcal{A}_B}(d_{AB})\ M^k_{\mathcal{A}_B,\mathrm{Ala}}(d_{BD}) \\ & \times M^k_{\mathcal{A}_B,\mathrm{Ala}}(d_{BE})\ M^k_{\mathcal{A}_r,\mathcal{A}_C}(d_{AC})\ M^k_{\mathcal{A}_C,\mathrm{Gly}}(d_{CF})\ M^k_{\mathcal{A}_C,\mathrm{Leu}}(d_{CG}) \end{aligned} \tag{10}$$

$P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y}, k)$, the probability that current sequences of subtype x at location l would be given by $\{\mathcal{A}\}'_{l,x}$ and current sequences of subtype y at the same location would be given by $\{\mathcal{A}\}'_{l,y}$ given a distance $d_{x,y}$ between the two root sequences if this location belongs to site class $\mathcal{S}_k$, is calculated by summing over all possible amino acids that could be in the two root sequences, yielding

$$\begin{aligned} & P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y}, k) = \\ & \sum_{\mathcal{A}_r,\mathcal{A}'_r} P(\{\mathcal{A}\}'_{l,x}|k, \mathcal{A}_r)\ \ P(\{\mathcal{A}\}'_{l,y}|k, \mathcal{A}'_r)\ \ P^k(\mathcal{A}_r)\ \ M^k_{r,r'}(d_{x,y}) \end{aligned} \tag{11}$$

Again, we can do a weighted sum over all possible site classes in order to get the total probability irrespective of site class, $P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y})$.

$$P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y}) = \sum_k P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y}, k)\ P(k) \tag{12}$$

We can then take the product of $P(\{\mathcal{A}\}'_{l,x}, \{\mathcal{A}\}'_{l,y}|d_{x,y})$ over all locations in the pair of proteins, and find the distance $d_{x,y}$ that maximizes this product, in order to obtain the most likely evolutionary distance. Distances obtained between all of the subtypes can then be used to generate a phylogenetic tree, using for instance the Fitch routine from the Phylip package [4].

3 Results

3.1 Simple Models for Specific Data Sets

With the reduced number of parameters, we can create specific mutation models for specific sets of proteins. To demonstrate this, we constructed data sets of envelope (*env*) proteins from HIV-1 and

Type of	Optimization	Test data set	
Model	Data set	HIV-1 *env*	HIV-2 *env*
Dayhoff mutation matrix [3]		1384	1858
mutation matrix	General	1665	2179
mutation matrix	HIV-1 *env*	2249	2578
2 site classes	HIV-1 *env*	1764	2248
3 site classes	HIV-1 *env*	2096	**2713**
5 site classes	HIV-1 *env*	2192	**3026**
7 site classes	HIV-1 *env*	**2294**	**3164**
9 site classes	HIV-1 *env*	**2350**	**3276**
11 site classes	HIV-1 *env*	**2475**	**3382**

Table 1: Q values for mutation matrices and simple models, calculated over a data set consisting of *env* proteins from HIV-1 or HIV-2. The various models were either optimized over a general protein data set ("Gen") or a data set consisting only of *env* proteins of HIV-1. Higher numbers correspond to a higher likelihood that the model would generate the current sequences. Bold faced numbers indicate those models with Q scores superior to any of the mutation matrices, including those optimized over the HIV-1 *env* data set. As the Q score is dependent on the number of homologs in the set as well as the length of the proteins, numbers in different columns cannot be compared. Results using the Dayhoff PAM 20 matrix are shown for comparison [3]

HIV-2. We constructed a number of different models with up to 11 site classes (with four linear and seven quadratic fitness functions), all depending upon hydrophobicity and bulk. Even the most complicated model had only 57 adjustable parameters. The model was then optimized for the HIV-1 *env* proteins. The resulting model was then compared with other models in representing the data from the HIV-2 *env* protein evolution. The results, represented as Q values, are presented in Table 1 [17].

As expected, the mutation matrices optimized over the HIV-1 *env* proteins were able to out-perform the more general mutation matrices, over both the HIV-1 and the HIV-2 *env* test sets. As shown, even with many fewer parameters, the simple model with seven or more site classes had higher Q values over the HIV-1 test set than any of the mutation matrices, even those optimized for this test set. And the reduced number of parameters reduced memorization and enhanced generalization, so that even the three site class model was able to out-perform all of the matrices over the HIV-2 *env* data base. This indicates that the rather simplistic assumptions made in the model – the assumption that the fitness is a simple function of hydrophobicity and size, and that the mutation rates follow simple Metropolis kinetics – is better able to model the evolution than the more complicated mutation matrices which consider amino acids individually yet ignore site heterogeneity.

3.2 Phylogenetic reconstruction

In order to demonstrate the use of our simple model, we applied it to the phylogenetic analysis of the various subtypes of the HIV-1 virus. Over 10 subtypes have been found, with the majority classified as part of the major group M, the remainder in the rarer outlier group O. M-group subtypes A-E are the most common. Envelope protein sequences from different subtypes are approximately 30% different, while intrasubtype variation is on the order of 10-20%. Variation in the sequences of some of the other proteins is more limited. Understanding the phylogeny of these subtypes is important in understanding how AIDS spreads, especially as treatment options may depend upon phenotypic factors that differ between subtypes. A variety of phylogenies have been proposed, but no tree structure has proven definitive.

One major advantage of our approach is our ability to customize evolutionary models for particular classes of proteins. Optimizing the model, however, requires a phylogenetic tree. We took advantage of the fact that the relationship between sequences within each subtype is more clearly defined, and can be reconstructed with reasonable accuracy by standard methods. Our model could then be optimized to describe the evolutionary patterns observed *within* each of the subtypes, and used to explore the relationship *between* subtypes. We aligned the sequences and generated a phylogenetic tree for each of the subtypes using the program ClustalW [26]. The midpoint of the longest node was taken as the root of that subtype. A nine-site model was then optimized for the sequences in subtypes A, C, and D, by maximizing the total probability that these sequences would result given our model. For verification, we tested how this model would match the evolutionary patterns of subtypes B, E, F, G, H, and O. With the exception of subtype O, in all cases our model provided higher probabilities than any of the more traditional mutation matrices derived from a general data set. The fact that our model was noticeably poorer for subtype O was in itself interesting, as it is believed that O may have started to infect humans more recently than other subtypes [9]. If it is true that one of the pressures driving variation is the need to evade the immune system, it may be that the evolutionary process for these proteins might be somewhat different depending upon how long they have infected a particular host.

The distances between the roots of the various subtypes were calculated as described above, and then input into the Fitch program of the Phylip package [4]. The resulting optimal phylogenetic tree is shown in Figure 2, with the subtypes represented by triangles to reflect the fact that these are the roots of the subtypes, which subsequently diverge into the currently known members.

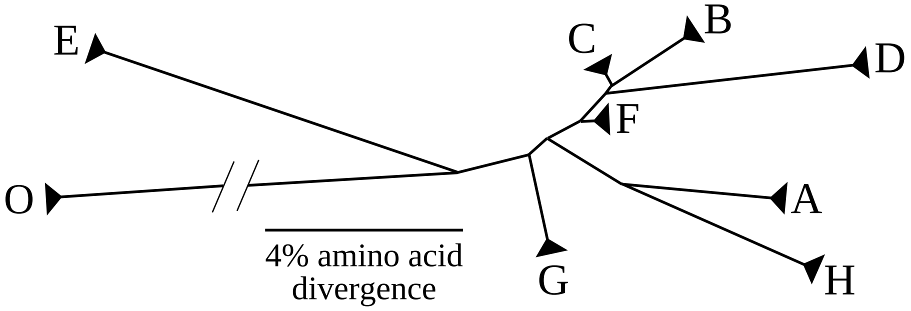

Figure 2: Unrooted tree describing the phylogenetic relationship of the HIV-1 subtypes as computed using our site-dependent model.

A number of interesting results are seen in our phylogenetic reconstruction. While previous trees tended to group subtypes B and D together, the close relationship of subtype C to this group is surprising. Similarly, the placement of E further from the other subtypes is in contrast to what has been proposed based on other models. These conclusions are preliminary, and can be explored further with maximum-likelihood models and analyses of other HIV-1 protein sequences.

4 Conclusion

In this paper, we describe a method for modeling protein evolution that explicitly takes into account the physical-chemical properties of the constituent amino acids. By doing this, we can greatly reduce the number of adjustable parameters, allowing us to develop models for specific proteins and include the effects of site-heterogeneity. The result is a model that better represents the evolutionary patterns, even with many fewer parameters. These matrices can have a wide range of possible uses. In this paper, we develop one application – the reconstruction of the evolutionary relationship between the HIV-1 subtypes. Our results suggest interesting differences from other phylogenetic trees that have been proposed previously.

Acknowledgments

We would like to thank Darin Taverna for his work in deriving the 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, and NIH Grants GM08270 and LM0577, and NSF equipment grant BIR9512955.

References

[1] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219:555–565, 1991.

[2] M. O. Dayhoff and R. V. Eck. A model of evolutionary change in proteins. In M. O. Dayhoff and R. V. Eck, editors, *Atlas of Protein Sequence and Structure*, volume 3, pages 33–41. National Biomedical Research Foundation, Silver Spring, Maryland, 1968.

[3] M. O. Dayhoff, R. M Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, suppl. 3, page 345. National Biomedical Research Foundation, Washington, D.C., 1978.

[4] J. Felsenstein. Phylip (phylogeny inference package) version 3.5c. Distributed by the author, 1993. Department of Genetics, University of Washington, Seattle.

[5] D. F. Feng, M. S. Johnson, and R. F. Doolittle. Aligning amino-acid sequences: A comparison of commonly used methods. *J. Mol. Evol.*, 21:112–125, 1985.

[6] W. M. Fitch. An improved method of testing for evolutionary homology. *J. Mol. Biol.*, 16:9–16, 1966.

[7] P. E. Gill, S. J. Hammarling, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for MPSOL (version 4.0). *Department of Operations Research, Stanford University*, Report SOL 86-2, 1986.

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

[9] L. G. Gurtler, L. Zekeng, J. M. Tsague, A. van Brunn, Z. E. Afane, J. Eberle, and L. Kaptue. HIV-1 subtype O: Epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV. *Archives of Virology*, 11:195–202, 1996.

[10] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci., U.S.A.*, 89:10915–10919, 1992.

[11] M. S. Johnson and J. P. Overington. A structural basis for sequence comparisons. *J. Mol. Biol.*, 233:716–738, 1993.

[12] D. T. Jones, W. R. Taylor, and J. M Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.

[13] A. Kidera, Y. Konishi, M. Oka, T. OOi, and H. A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.*, 4:23–55, 1985.

[14] J. M. Koshi and R. A. Goldstein. Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. *Protein Engineering*, 8:641–645, 1995.

[15] J. M. Koshi and R. A. Goldstein. Correlating mutation matrices with thermodynamic and physical-chemical properties. In L. Hunter and T. Klein, editors, *Pacific Symposium on Biocomputing '96*, pages 488–499. World Scientific, 1995.

[16] J. M. Koshi and R. A. Goldstein. Mutation matrices and physical-chemical properties: Correlations and implications. *Proteins*, 27:336–344, 1997.

[17] J. M. Koshi and R. A. Goldstein. Mathematical models of natural amino acid site mutations. *J. Mol. Biol.*, submitted, 1997.

[18] R. Luthy, A. D. McLachlan, and D. Eisenberg. Secondary structure-based profiles: Use of structure conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, 10:229–239, 1991.

[19] A. D. McLachlan. Tests for comparing related amino-acid sequences. *J. Mol. Biol.*, 61:409–424, 1971.

[20] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations for fast computing machines. *J. Chem. Phys.*, 21:1087, 1953.

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

[22] S. Miyazawa and R. L. Jernigan. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Engineering*, 6:267–278, 1993.

[23] J. Overington, D. Donnelly, M. S. Johnson, Andrej Šali, and T. L. Blundell. Environment-specific amino-acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci.*, 1:216–226, 1992.

[24] J. K. M. Rao. New scoring matrix for amino acid residue exchange based on residue characteristic physical parameters. *International Journal of Peptide and Protein Research*, 29:276–281, 1987.

[25] J. L. Risler, M. O. Delorme, H. Delacroix, and A. Henaut. Amino acid substitutions in structurally related proteins. *J. Mol. Biol.*, 204:1019–1029, 1988.

[26] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[27] H. Wako and T. Blundell. 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.

[28] H. Wako and T. Blundell. 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, 1994.