MODELING EVOLUTION AT THE PROTEIN LEVEL USING AN ADJUSTABLE AMINO ACID FITNESS MODEL

MATTHEW W. DIMMIC*, DAVID P. MINDELL‡, AND RICHARD A. GOLDSTEIN*§
** Biophysics Research Division*
‡Department of Biology and Museum of Zoology
§Department of Chemistry
University of Michigan
Ann Arbor, MI 48109-1055

An adjustable fitness model for amino acid site substitutions is investigated. This model, a generalization of previously developed evolutionary models, has several distinguishing characteristics: it separately accounts for the processes of mutation and substitution, allows for heterogeneity among substitution rates and among evolutionary constraints, and does not make any prior assumptions about which sites or characteristics of proteins are important to molecular evolution. While the model has fewer adjustable parameters than the general reversible mtREV model, when optimized it outperforms mtREV in likelihood analysis on protein-coding mitochondrial genes. In addition, the optimized fitness parameters of the model show correspondence to some biophysical characteristics of amino acids.

1. Introduction

A great majority of phylogenetic analyses are performed using comparisons of DNA sequences between taxa. There are several methods of inferring a phylogenetic tree from available data. One is the maximum parsimony (MP) method[1]. This method has the advantage of being quick and exhaustive, but there are situations where it will repeatedly lead to the wrong tree[2]. Another method is maximum likelihood (ML) analysis[3,4]. This a more mathematically rigorous method which computes the probability of any set of sequences given a particular model and the tree. While ML analysis does not suffer from the same biases as MP, it is computationally more intensive and requires an explicit model for the process of molecular evolution. Several simple models have been developed[5], but the most general model in wide use is the general time-reversible (GTR) model. This model assigns a parameter to each of the possible nucleotide substitutions, leading to a model with 12 adjustable parameters.

Such models have had success in determining evolutionary relationships when the species being examined are fairly closely related. For distant relationships, however, these models can be inadequate. If the region is nonfunctional, then presumably a mutation is just as likely to be accepted into the population as the nucleotide which it replaced. In such cases, one would expect the rate of substitution to be close to the rate of mutation. But if the mutation rate is high then

the sequence will change relatively rapidly, and the probability of multiple mutations occurring at a site also increases. Eventually the sequence changes become "saturated", making it difficult to glean true evolutionary relationships.

Presumably one would then prefer to examine regions of the sequence which evolve slowly such as protein-coding regions, but the applicability of DNA models such as the GTR to these functional regions is questionable. While the mutation rate may still depend on such underlying biases as the transition-transversion ratio, the rate at which those mutations are accepted (the substitution rate) is affected by wholly unrelated evolutionary pressures on the gene product. For instance, a C→A mutation would have no effect in the third position of a CCC codon, but if it occurs in the first position then the amino acid product changes from proline (Pro) to threonine (Thr). Because proline has very different biophysical properties from threonine, this mutation may have some effect on the structure and/or function of the protein product, and therefore this mutation may be accepted at a rate different from that of a C→A mutation in the third position.

One solution to this bias has been to assume that different sites in the DNA are substituted at varying rates. Such rates can be specifically defined or pulled from a distribution, such as a gamma distribution[6]. Using variable rates, a C→A mutation in the first codon position might be accepted at a frequency much lower than one at the third codon position, reflecting the greater proportion of synonymous substitutions at the third position. Yet all third-position mutations are not created equal; in some cases a C→A is synonymous, in others it is not, and in some nonsynonymous cases it leads to the substitution of an amino acid with very different properties.

1.1 Amino-Acid Based Substitution Models

To avoid these problems some researchers have attempted to model the evolutionary process on the amino acid level. The earliest successful techniques were empirical methods developed by Eck and Dayhoff[7] in 1968; variations on these techniques have been used ever since[8]. Sequences with a high degree of identity are aligned in a parsimonious way (it is assumed that no double substitutions have occurred), and then the number and types of changes between the two sequences are tallied. After adjusting for the natural frequencies of the amino acids in the data set a substitution matrix is obtained, describing the probability for any amino acid to substitute for any other amino acid in a given amount of evolutionary time.

The fact that these empirical matrices (or updated versions of them) have been widely used for so long is a testament to their success in predicting evolutionary relationships, but they do make several simplifying assumptions. Because only closely-related sequences are included in the creation of these matrices, the information available in more distant relationships is not used. Also, these methods

make the assumption that all sites on the protein in all different proteins evolve at equal rates and with equal evolutionary pressures. Because different sites in the sequence are exposed to different environments and perform different functions in the folded protein, these assumptions are generally not valid.

Recently efforts have been made to address these shortcomings. For example, the gamma distribution (either continuous or discrete) has been used to account for the different observed rates of evolution at different locations[9]. But in reality variability is not restricted to rate alone, as the various pressures under which each site evolves may lead to completely different substitution probabilities. For example, conservation of an amino acid's size may be extremely important in the interior of a globular protein, yet on the exterior the size may be relatively unconstrained. In the absence of any other pressures, one would expect the rate of Gly→Arg substitution (small-to-large) to be higher on the exterior of the protein while a Gly→Ala substitution (small-to-small) might occur with similar rates independent of location. Yet Gly is also known to be critical in the formation of certain secondary structure such as a β-turn. If a site were in such a position, then both Gly→Arg and Gly→Ala would be unfavorable. Models which deal with heterogeneity of absolute rates alone cannot take such pressures into account.

Another method of dealing with site heterogeneity has been to assign different models based upon the protein's secondary structure[10,11]. For instance, one substitution matrix might be used on external beta sheets, while another might be used for interior alpha helices. While this technique has promise (especially if the protein structure is known), it assumes that local structures which are important to crystallography are also important to evolution, which may not necessarily be the case.

In this paper we extend and explore a model for amino acid substitutions in protein sequences. The model is designed to incorporate the following properties:

- An explicit separation of the amino acid mutation and substitution processes.
- Reversibility.
- Heterogeneous evolutionary pressures among sites.
- Amenability to likelihood analysis.
- As few assumptions about the underlying evolutionary pressures as possible while still remaining computationally tractable.

The model is primarily based upon the models explored by Koshi *et al.*[12], with changes made to generalize the model by remove any assumptions about those biophysical parameters of the amino acids important to evolution.

2. Model

2.1 The Amino Acid Mutation/Substitution Process

Assume for a moment that all sites in an amino acid sequence are subject to the same evolutionary pressures and constraints. Since amino acids have different biophysical characteristics and some will be more advantageous than others, each amino acid A_n can be considered to have a relative fitness $F(A_n)$. This represents the functional fitness of the protein with residue A_n at a particular site in the sequence relative to the functional fitness of a protein which has any other amino acid at that position. For example, if charge is important to a protein's function, then a charged amino acid such as Glu (E) would be more fit than a hydrophobic amino acids such as Ile (I), and therefore be represented with a higher relative fitness $F(E) > F(I)$.

It is assumed that, through the course of evolution, any mutations to an amino acid with a high relative fitness will have a greater probability of acceptance than a mutation to an amino acid with a lower relative fitness. This model assumes that the probability Q_{ij} of substituting amino acid i with amino acid j is a function of these relative fitnesses in a Metropolis-like scheme:

$$Q_{ij} = \begin{cases} \nu & \text{if } \Delta F_{ji} > 0 \\ \nu e^{\Delta F_{ji}} & \text{if } \Delta F_{ji} < 0 \end{cases} \tag{1}$$

where ν is the average rate at which mutations occur and

$$\Delta F_{ji} = F(A_j) - F(A_i) \tag{2}$$

Using this scheme, mutations occur with a certain rate; if the mutation is favorable then it is always accepted, while if the mutation is unfavorable it is tolerated with a decaying exponential probability. While this acceptance scheme is admittedly *ad hoc*, it has proven to be reasonable in previous evolutionary models[12], and can be improved upon in future studies. The collection of Q_{ij}'s is the substitution rate matrix **Q**, where the diagonal elements are equal to the sum of the off-diagonal elements of the row (so each row sums to zero). To determine the substitution matrix **M** for any particular amount of evolutionary time t, the substitution rate matrix is exponentiated:

(3)

$$\mathbf{M}(t) = e^{t\mathbf{Q}}$$

The values for the relative fitnesses are not assumed, but are set as adjustable parameters in the likelihood maximization scheme, described below. The fitnesses are also used to calculate the prior probabilities for each of the amino acids:

$$P(A_n) = \frac{e^{F(A_n)}}{\sum_i e^{F(A_i)}} \tag{4}$$

This is analogous to a Boltzmann distribution, and preserves reversibility in the model. Note that the assumption of reversibility is not a requirement of the model, but is helpful for purposes of comparison.

2.1 Calculating the Likelihood (without site heterogeneity)

At each site *s* in the amino acid sequence, the likelihood L_s can be represented as the probability of the data given the model's parameters θ and the evolutionary tree topology and branch lengths *T*:

$$L_s = P_s(Data \mid \theta, T) \tag{5}$$

The calculation of $P_s(Data \mid \theta, T)$ follows that of most likelihood schemes[4], and is not shown here. The likelihood Ω for the entire sequence is the product of these likelihoods, equivalent to the sum of their logs. Therefore

$$\Omega = \sum_s \mathrm{Log}(L_s) = \sum_s \mathrm{Log}\left[\prod_s P_s(Data \mid \theta, T)\right] \tag{6}$$

2.2 Incorporating site heterogeneity

Now we will relax the assumption that each site is under the same evolutionary constraints. To account for site heterogeneity, it is assumed that there is a set Θ of *k* site classes, each with its own set of fitness parameters $F_k(A_n)$ and its own mutation attempt rate ν_k. If we knew all the biophysical characters of amino acids which were important to evolution in a protein, then we might be able to assign each site to a site class. For instance, one site class might represent all sites where charge is important for the protein's function; in this site class, Glu would have a higher fitness than Ala. Another site class might represent all sites where small bulk is preferred; in this site class Ala would have a higher fitness than Glu.

Previously, variations of this model have been used to explore pre-assigned site classes[10]. For this analysis we are interested in a general model, where each site's site class is unknown[12]. Instead, each site has a certain probability of being represented by each site class. This probability is the likelihood function calculated using the parameters for that site class. Each site class, in turn, has a prior probability of representing any site $P(\Theta_k)$. The likelihood at each site is just the sum over all the probabilities of site classes at that site:

$$L_s = \sum_k P_s(\, Data \,|\, \Theta_k, T\,)\, P(\,\Theta_k\,) \qquad (7)$$

with the likelihood for the tree calculated as the product of the likelihood at each site.

2.3 Optimizing the parameters

By using a maximum likelihood formalism, we can optimize the model by adjusting the parameters to optimize the likelihood function. In this model there are 21 adjustable parameters per site class: the 20-member vector of fitnesses $\mathbf{F}_k$ (with one held constant because the fitnesses are relative), the mutation attempt rate ν_k, and the prior for that site class $P(\Theta_k)$. Since the priors must add to 1, in any optimization one of the site classes' priors will be dependent on the others.

$$\Theta_k = \{\quad \mathbf{F}_k, \upsilon_k, P(\,\Theta_k\,)\quad \} \qquad (8)$$

By adjusting the parameters to increase the likelihood, a maximum likelihood estimate for the parameters can be obtained.

3. Methods

In order to test the utility of our amino acid fitness models, we compare our results with the mtREV model of Adachi and Hasegawa[13]. In the mtREV model, each of the substitution probabilities are represented by an adjustable parameter, leading to a total of 189 adjustable parameters (the model is reversible). It can be seen as the most general single-site class reversible model, representing a good basis for comparison. The mtREV model was optimized by Adachi and Hasegawa over most of the mitochondrial protein-coding sequences from a tree of 20 vertebrate species, and for reasons of comparison its parameters were held constant for this study.

We optimized our models using a training set of mitochondrial protein-coding genes. As mentioned above, each of our models has 20 adjustable parameters per site class, plus *k*-1 additional adjustable parameters representing all but one of the site class priors $P(\Theta_k)$. The training tree and sequences used are those given in Mindell *et al.*[14], representing 16 taxa over several families of bird, reptile, amphibian, fish, and mammal. The sequences represent the protein-coding portion of the mitochondrial genome, with ND4, ND5, and ND6 not used. Because the optimization process assumes a certain tree topology we chose to use a relatively well-determined tree, and so the following modifications were made to the Mindell *et al.* dataset: the falcon and suboscine songbird data were removed, as this monophyletic relationship is currently controversial. The turtle and platypus data also were not used, and two fish species were added to the tree as an outgroup.

Optimal branch lengths of the tree were obtained from PAML[15] by applying the mtREV model to a subset of the proteins on the full training tree. Because current phylogenetic packages cannot optimize branch lengths for our multiple-site-class models, the fitness models used the same branch lengths as those obtained for mtREV. Optimization of the parameters of the fitness models was performed using an EM algorithm[19] combined with the downhill simplex method[16] of Nelder and Mead. The simplex was started at random vertices, with restarts after convergence.

4. Results & Discussion

A summary of the results is shown in **Table 1**. All the multi-site-class fitness models performed better than mtREV on the training set. Whether or not mtREV can statistically be rejected in favor of the fitness models using simple likelihood ratio tests is less clear, since the mtREV and fitness models are not nested. But consider that the largest (5 site class) fitness model tested has 86 fewer adjustable parameters than the mtREV model, yet the fitness model was able to exceed the mtREV by nearly 900 log-likelihoods.

Table 1: Calculated –log-likelihood values (-Ω) for each model on each dataset. Values which are not in the same column cannot be compared.

Model	*training sequences*	*test sequence* **(ND6)**	*test sequence* **(ND6+new branch lengths)**
mtREV	38234	3662.7	3730.6
fitnesses, 1 site class	41941	3767.5	3843.7
fitnesses, 3 site classes	37939	3673.5	3687.4
fitnesses, 5 site classes	37342	3620.1	3637.7

Table 2: Optimized parameter values for the five site class model. The $F(A_n)$ values are normalized to Ala at 0.00, and the $P(\Theta_k)$ probabilities sum to 1. One mutation rate ν_5 was held constant at 0, yielding a site class which best represents conserved sites.

	Site Class				
Parameter	**1**	**2**	**3**	**4**	**5**
$P(\Theta_k)$	0.28	0.22	0.21	0.20	0.09
ν_k(x10^3)	3.90	1.27	0.42	0.057	0
$F(A_n)$:					
Ala	+.000	+0.00	+0.00	+0.00	+0.00
Arg	-5.67	-4.02	-8.77	+3.44	+4.30
Asn	-2.24	-1.67	-7.66	-2.67	+5.69
Asp	-4.34	-2.46	-12.05	+3.34	+3.15
Cys	-3.77	-5.28	+0.29	-3.43	-1.14
Gln	-3.34	-2.65	-21.13	+3.54	+2.40
Glu	-4.74	-2.48	-2.34	+3.55	-5.02
Gly	-3.56	-1.07	-1.70	+0.86	+6.72
His	-3.18	-2.68	-0.44	+3.64	-6.50
Ile	+1.39	-4.33	+2.02	-0.77	+5.20
Leu	+1.50	-3.30	+3.14	+4.13	-1.35
Lys	-3.11	-2.25	-5.37	+3.28	-1.13
Met	+1.03	-3.30	+2.33	-6.97	-7.13
Phe	-0.70	-3.40	+3.14	-1.73	-9.49
Pro	-1.84	-1.34	-6.28	+4.46	-1.29
Ser	-0.17	-0.36	-0.24	+2.60	+5.19
Thr	+0.51	-0.36	+0.86	+2.06	+5.44
Trp	-4.10	-4.78	+0.46	+3.95	+2.10
Tyr	-2.68	-3.19	+2.71	+1.98	+1.42

The success of the fitness models on the training set is encouraging, but not conclusive considering that they were optimized over that dataset, while mtREV was optimized over a different set of taxa. To examine the robustness of the models it is useful to compare them on a separate test set. Because mtREV was designed for mitochondrial proteins, another mitochondrial sequence was used which presumably has similar evolutionary pressures as the training set. Neither the mtREV model nor our fitness models used the ND6 mitochondrial gene during optimization, and so the ND6 sequences from the training taxa were used for testing. Because ND6 is on the heavy strand of the mitochondrial genome, its amino acid frequencies differ from those of the other genes, and so the equilibrium frequencies of the amino acids in the mtREV model were adjusted to the frequencies found on this gene. In this case only the 5-site-class model outperformed the more general mtREV on the test sequence, although if the mtREV frequencies are left unadjusted then the 3-site-class model also outperforms mtREV (data not shown).

The last column of **Table 1** shows the results when the branch lengths are re-optimized in PAML, this time using the mtREV model but a larger dataset of

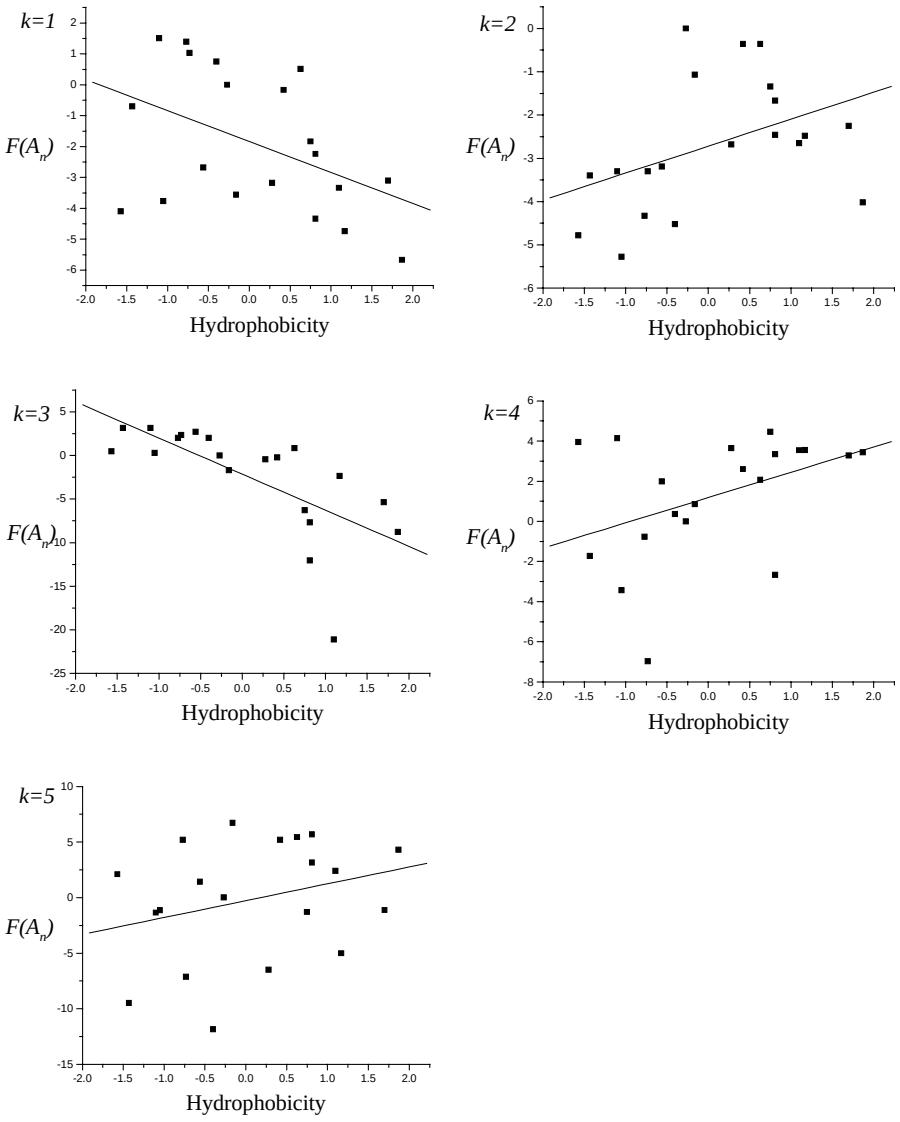

Figure 1: Optimized parameter values for each site class *k* in the 5-site class model plotted against biophysical indices of Kidera *et al*[17].

mitochondrial genes, including ND6. Using these branch lengths, each of the multiple-site-class models outperforms mtREV. Oddly, the results on the mtREV become significantly worse with these branch lengths, despite the inclusion of the test sequence in the branch length optimization. More definitive tests may be performed once it becomes possible to optimize tree branch lengths using the fitness models.

While the fitness-based models primarily concentrate on using our knowledge of amino acids to predict phylogenies, the physical basis for these models may allow us to make general statements about properties of amino acids as well. **Table 2** shows the optimized values obtained for the model's parameters on the training set. To examine whether the site classes in the 5-site-class model had any correlation with known biophysical properties, we plotted them against two sets of indices[17] for amino acid characteristics: bulk and hydrophobicity. While most correlations with bulk were not strong, several of the site classes showed significant correlation with hydrophobicity (see **Figure 1**). Site class #3 showed substantial negative correlation with amino acid hydrophobicity ($R=-0.68$, $P<0.001$), and site class #4 showed some positive correlation ($R=0.42$, $P=0.065$). By analyzing the posterior probabilities that any particular sequence site is represented by a particular site class, one could presumably draw inferences regarding the influence of protein structure and function on the protein's evolution.

The lack of correlation with biophysical properties among other site classes does not necessarily imply that the parameters have no physical meaning, but it does imply that setting the parameters as simple functions of a few biophysical characteristics[18] may not adequately capture the selective pressures at work on the protein. Another possibility is that any important biophysical characteristics are "mixed" into the various site classes during the optimization scheme, and so it may be useful to test schemes which attempt to keep these properties as separate as possible.

5. Conclusions and future work

Because coding regions of DNA are under very specific evolutionary constraints, it seems natural to instead model evolution in these regions using amino acid sequences. Here we have presented a model for amino acid substitution which encapsulates several desirable evolutionary traits: explicit modeling of the mutation process, a substitution process based upon the evolutionary constraints of the amino acids, and heterogeneity among sites in the protein. It appears that the incorporation of these characteristics allows the amino acid fitness models to perform substantially better than the widely-used mtREV model, even though the number of adjustable parameters in the new model is much lower. Presumably the trend in increasing

likelihood would continue with the addition of more site classes; it is not until a 10-site-class model is used that the number of adjustable parameters exceeds that of the mtREV model. In addition, the resulting parameters show some correlation with biophysical characteristics of amino acids, indicating they may be useful in determining the constraints on the evolution of a particular data set.

While these results are promising, more comparisons need to be performed using other data sets, and it may be helpful to analyze these optimized models using statistical tests such as Monte Carlo simulation. The idea that each amino acid has a certain fitness for each site leads naturally to more biologically realistic substitution models than the Metropolis scheme used here, and the authors are currently implementing one such model, as well as faster methods of optimizing these models. Once the robustness and versatility of these simple fitness-based models is determined, they show promise for general applicability in phylogenetic analyses of protein-coding sequences.

Acknowledgments

Thanks to Michael Newton for helpful discussions on optimization algorithms. Financial support was provided by the Horace H. Rackham School of Graduate Studies, NIH Grants GM08297 and LM0577, NSF Grant 9726427, and NSF equipment grant BIR9512955.

References

1. W. M. Fitch, "On the problem of discovering the most parsimonious tree" *Am. Nat.* **111**, 223 (1977)
2. J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading" *Syst. Zool*. **27**, 401 (1978)
3. L.L. Cavalli-Sforza and A.W.F. Edwards, "Phylogenetic analysis: models and estimation procedures" *Am. J. Hum. Genet.* **19**, 233 (1967)
4. J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach" *J. Mol. Evol.* **17**, 368 (1981)
5. A. Zharkikh, "Estimation of evolutionary distances between nucleotide sequences" *J. Mol. Evol.* **39**, 315 (1994)
6. Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods" *J. Mol. Evol.* **39**, 306 (1994)
7. R. V. Eck and M. O. Dayhoff in *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation (1966)
8. D.T. Jones, W.R. Taylor, and J.M. Thornton, "The rapid generation of mutation data matrices from protein sequences" *Comp. Appl. Biosci.* **8**, 275 (1992)

9. Z. Yang, N. Nielsen, and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution" *Molecular Biology and Evolution* **15**, 1600 (1998)
10. J.M. Koshi and R.A. Goldstein, "Context-dependent optimal substitution matrices" *Protein Engineering* **8**, 641 (1995)
11. N. Goldman, J. L. Thorne, and D. T. Jones, "Assessing the impact of secondary structure and solvent accessibility on protein evolution" *Genetics* **149**, 445 (1998)
12. J. M. Koshi, D. P. Mindell, and R. A. Goldstein, "Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes" *Molecular Biology and Evolution* **16**(2), 173 (1999)
13. J. Adachi and M. Hasegawa, "Model of amino acid substitution in proteins encoded by mitochondrial DNA" *J. Mol. Evol.* **42**, 459 (1996)
14. D.P. Mindell, M.D. Sorenson, D.E. Dimcheff, M. Hasegawa, J.C. Ast, and T. Yuri, "Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes" *Syst. Biol.* **48**(1), 138 (1999)
15. Z. Yang, "PAML: A program package for phylogenetic analysis by maximum likelihood" *CABIOS* **13**, 555 (1997)
16. W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, in *Numerical Recipes in C: The Art of Scientific Computing Second Edition*, (Cambridge University Press, Cambridge UK, 1997)
17. 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. Prot. Chem.* **4**(1), 23 (1985)
18. J. M. Koshi and R. A. Goldstein, "Mathematical models of natural amino acid site mutations" *Proteins* **32**, 289 (1998)
19. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society B*, **39**, 1-38.