PROTEINS: Structure, Function, and Genetics 32:289–295 (1998)

Models of Natural Mutations Including Site Heterogeneity

Jeffrey M. Koshi[1] **and Richard A. Goldstein**[1,2]*
[1]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **New computational models of natural site mutations are developed that account for the different selective pressures acting on different locations in the protein. The number of adjustable parameters is greatly reduced by basing the models on the underlying physical-chemical properties of the amino acids. This allows us to use our method on small data sets built of specific protein types. We demonstrate that with this approach we can represent the evolutionary patterns in HIV envelope proteins far better than with more traditional methods. Proteins 32:289–295, 1998.**

Key words: molecular evolution; protein evolution; mutation matrices; Metropolis kinetics; Boltzmann statistics

INTRODUCTION

There is increasing interest in characterizing protein evolution. Phylogetic analyses of specific proteins can tell us much about those proteins' structure and function. Ancestral protein sequences can be recreated using statistical methods, expressed in the laboratory, and characterized by standard biochemical techniques.[1–3] Relative mutation rates can provide insight into the relationship between protein properties and the characteristics of their constituent amino acids.[4] Maybe most importantly, the greater selective pressures acting on the expressed proteins at the amino acid level enable delineation of evolutionary relationships between organisms that may be too distant to be characterized using DNA sequences, allowing us to address basic questions of evolutionary biology.

The standard approach towards modeling natural site mutations in proteins is with a mutation matrix: a 20 × 20 array that represents the probability of any given amino acid changing to any other in a given length of evolutionary time. Most methods for deriving these matrices use the approach developed by Dayhoff, based on an analysis of corresponding amino acids in closely-related homologous proteins.[5] Variations of the original Dayhoff approach have been developed, including using blocks of aligned sequences and sequences aligned based on their three-dimensional structure.[6–12] Others have developed approaches that encode the tendency for important physical-chemical properties of the amino acids to be conserved during evolution.[13–18]

There are a number of limitations inherent in the use of mutation matrices. Their derivation involves the simultaneous determination of 380 adjustable parameters representing all possible amino acid mutations, and thus requires a large data base to set all of the variables without overfitting. Because of this, the mutation matrix approach is not well suited for specific proteins where only small sets of sequences are available. Even more importantly, the mutation-matrix approach assumes that mutations are a Markovian process, and that the probability of any given mutation (His → Gly, for example) is the same for all locations in the protein, whether at a solvent-exposed site in an alpha helix, in a buried turn, or even at a dimerization or catalytic site. As the relative fitness of a given pair of residues (His and Gly in this example) at these varying types of locations will be different, so will the mutation rates between these two types of residues. We partially addressed this issue in our earlier work by constructing mutation matrices specific for different secondary structure and surface accessibility classes.[19] Even this strategy is based on the assumption that all positions in a particular local environment, such as buried alpha helices, have the same fitness requirements and thus similar mutation rates. As a result, important deviations within local environments are averaged out. Also, secondary structure and surface accessibility are the easiest distinctions for us to observe; classes based on other characteristics might be more biologically relevant. Any attempt to treat the problem of site heterogeneity by simultaneously optimizing a set of mutation matrices where the

Grant sponsor: College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies; Grant sponsor: National Institutes of Health; Grant numbers: GM08270 and LM0577; Grant sponsor: NSF; Grant number: Equipment grant BIR9512955.
Jeffrey M. Koshi's present address is Theoretical Biology and Biophysics Division, Los Alamos National Laboratory, Los Alamos, NM 87545.
*Correspondence to: Richard A. Goldstein, Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055.
Received 15 September 1997; Accepted 6 April 1998

different types of locations were not identified a priori would result in an unacceptably large number of adjustable parameters.

Because of these issues, we have turned to simplified models of evolution. Rather than express mutation rates as a function of the identity of the amino acids, we express these rates as functions of the corresponding physical-chemical properties of these residues. The models represent the fitness of any particular type of amino acid by a simple functional form dependent on a set of physical-chemical properties such as hydrophobicity and size. A mutation matrix is then derived based on a Metropolis scheme, where upward changes in fitness are accepted at some maximum rate ν, and downward changes at ν times some exponentially decreasing function of the change in fitness. We can then use the estimation-maximization methods described in our previous work to calculate the likelihood that a given fitness function with its associated mutation matrix would produce the observed data, and find the optimal mutational model as a function of the physical-chemical properties.[19]

Because the mutation rates are a function of the parameters representing the fitness of the various amino acids, there are orders of magnitude fewer adjustable parameters than if we were to construct a traditional mutation matrix as a function of the amino acid identities. This allows us to extend our analysis to limited data sets, such as the evolutionary patterns of specific proteins. In addition, we can include site heterogeneity explicitly in the model. We consider that there are different types of locations, what we call site classes, each with its distinct fitness function and corresponding mutation matrix. While in principle these different site classes might correspond to locations with different secondary structures or functional significance, we do not need to define the nature of these site classes a priori. Nor do we need to assign different locations in the protein to specific site classes. Rather, there are adjustable probabilities that any location can be described by each site class. These probabilities are optimized simultaneously with the adjustment of the underlying fitness functions corresponding to the different site classes.

In order to demonstrate the validity of these models, it is necessary to show that the model optimized for one set of training proteins can describe the evolutionary process of a test set of proteins whose evolutionary history is completely disjoint. Fortunately, a convenient example exists in the proteins of HIV-1 and HIV-2. The simple models optimized over a HIV-1 envelope protein (*env*) data set can represent the evolutionary patterns of the HIV-1 *env* data better than a traditional mutation matrix optimized over the same data, even though the simple model involves many fewer adjustable parameters. More significantly, we find that a simple model optimized over the HIV-1 *env* data can more accurately describe the evolutionary events of the *env* proteins of HIV-2 than can a single standard mutation matrix optimized over a more general protein data set, or even a mutation matrix optimized over only the HIV-1 *env* proteins. In this case, the assumptions and approximations involved in the simple models, including the representation of the fitness as a simple function of a few physical-chemical parameters and the use of a Boltzmann and Metropolis formulations for developing relative mutation rates, are less drastic and more accurate than the Markovian assumptions implicit in the standard mutation matrix approach.

THEORY

Our model of site mutations consists of three different parts: the construction of simple functional forms that define how the fitness of the various amino acids depends upon that amino acid's physical-chemical properties, the calculation of the corresponding mutation matrices that encode how mutations would occur given these fitness values, and the optimization of the various parameters to fit the observed evolutionary data using estimation maximization. The presence of site-heterogeneity is incorporated directly into the theory through the use of multiple site classes. In contrast to standard mutation matrix approaches which ignore site heterogeneity, and our earlier work which divided locations in the protein into different sites on the basis of secondary structure and surface exposure,[19] we consider that each point in the protein is described by one of a number of different possible site classes, whose properties are not pre-defined but are rather determined during the optimization procedure. We do not assign individual locations to specific site classes, but instead consider that there is a probability that any site in the protein corresponds to any of the particular site classes. These probabilities are adjustable parameters of the model, determined during the optimization procedure. Different site classes correspond to different local evolutionary pressures, and thus to distinct mutation matrices. This approach would not be possible if the evolutionary data used in the optimization contained only pairs of homologous proteins; in this case, the various site classes would average. One of the strengths of our optimization procedure is that it uses arbitrarily large sets of homologous proteins with their associated phylogenetic tree. The optimization procedure can thus be informed by correlations in the multiple mutations that occur in the same locations, allowing us to unravel the various site classes.

We construct simple functional forms that define how $F_k(A_i)$, the fitness of amino acid A_i for any

location in site class k, depends upon the physical-chemical parameter of that amino acid. As the purpose of these models of evolution is to minimize the number of free parameters, the fitness functions examined in this paper are simple linear and quadratic forms. $F_{k,l}(A_i)$, the contributions to the fitness function for each physical-chemical property l, are of the form:

$$F_{k,l}(A_i) = \alpha_{k,l} q_l(A_i) \tag{1}$$

$$F_{k,l}(A_i) = \alpha_{k,l}(q_l(A_i) - q_{k,l}^{\text{opt}})^2, \tag{2}$$

where $q_l(A_i)$ represents the value of the physical-chemical parameter l for amino acid A_i, and $\alpha_{k,l}$ and $q_{k,l}^{\text{opt}}$ are parameters that depend upon the site class k. (Constants are omitted from the above expressions, as the mutation rates and relative populations are only functions of differences in fitness.) The linear fitness function models those situations where a physical-chemical parameter would be either favored or disfavored at a given location. The quadratic fitness function would be appropriate where there was a "best" parameter value, with fitness falling for both smaller and larger parameter values, or a "worst" value, with fitness increasing at both extremes. The total fitness for the amino acid in any site is assumed to be a simple sum of the terms reflecting the various physical-chemical factors

$$F_k(A_i) = \sum_l F_{k,l}(A_i). \tag{3}$$

Choosing what amino acid properties to consider in our model is not a trivial question, as hundreds of physical-chemical parameter scales for the 20 naturally occurring amino acids have been measured.[20–22] Many of these scales are highly correlated, however, making inclusion of all of them unnecessary. In particular, Scheraga and coworkers derived four orthogonal property indices that contained most of the variation observed over 180 different amino acid properties.[23] These factors correlated predominantly with alpha helical and turn propensity (α/turn), bulk-related factors (volume, molecular weight, etc.), beta sheet propensity, and hydrophobicity. Our study of evolutionary data was performed using these orthogonal parameters for the values of $q_l(A_i)$ in equations 1 and 2.

We assume that the probability $P_k(A_i)$ of any given amino acid A_i occurring at any location described by a site class k is given by a Boltzmann relation.

$$P_k(A_i) = \frac{e^{\beta F_k(A_i)}}{\sum_{i'} e^{\beta F_k(A_{i'})}} \tag{4}$$

where β is a free parameter, and i' is an index over all amino acids types. (As large fitness values are favorable, the sign of the exponential is opposite to the normal Boltzmann formula.) Boltzmann-like distributions of fitnesses have been observed when the fitness has an energetic interpretation.[24–30] Even in this case, β can not simply be identified with the reciprocal of the temperature, but expresses the distribution of energies among the various possible conformations.[24] An explanation recently presented for this phenomenon would apply to a wide range of fitness parameters, including parameters involving properties that were not strictly energetic in nature.[31] Alternatively, we can define the fitness of a particular amino acid in a given site class as the logarithm of the probability of that amino acid existing at that site, plus a normalization constant. In this case, equation 4 would be true by definition. As the fitness scale is arbitrary, we can set β in Equation 4 equal to one.

The evolutionary kinetics of our model are based on the Metropolis algorithm. We postulate that favorable or neutral mutations are accepted and fixed at some site-class dependent maximal rate ν_k, and unfavorable mutations are accepted at a rate of $\nu_k \times \exp(\Delta F_k)$, where ΔF_k is the difference in fitness values associated with the mutation for the given site class. We ignore differences in attempt rates due to differences in the number of nucleic-acid base-pairs necessary for changes at the amino-acid level—the strong correlation between number of required base changes and changes in physical chemical parameters makes the separation of these two effects problematic. The value of M_{ij}^k, the entry in the mutation matrix for site class k corresponding to a mutation from amino acid A_i to A_j, is then:

$$M_{ij}^k = \begin{cases} \nu_k & \mid \quad F_k(A_j) > F_k(A_i) \\ \nu_k e^{(F_k(A_j) - F_k(A_i))} & \mid \quad F_k(A_j) \le F_k(A_i). \end{cases} \tag{5}$$

While the use of Metropolis kinetics is an assumption, we note that the Metropolis scheme is the only kinetics scheme where all favorable mutations are accepted at the maximum rate, the Boltzmann distribution of fitnesses is maintained, and detailed balance is obeyed.

With $P_k(A_i)$ given by equation 4, the set of mutation rates M_{ij}^k given by equation 5, a set of values for $P(k)$ representing the probability of any given location actually belonging to site class k, and a phylogenetic tree representing the evolutionary relationship among a set of aligned homologous proteins, we can calculate the probability of the observed current sequences resulting through evolution using methods described in an earlier paper.[19] Consider a corresponding location n in an set of four aligned homologous proteins, related phylogenetically as shown in

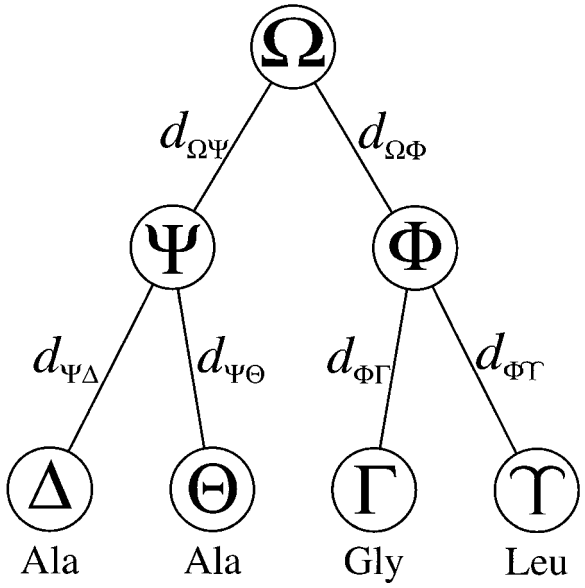

Fig. 1. Example evolutionary relationship between four current sequences, represented as nodes Δ, Θ, Γ, and Υ, and the sequences of their ancestral proteins, represented by nodes Ω, Ψ, and Φ. $d_{\Omega\Psi}$ represents the evolutionary distance between nodes Ω and Ψ.

Figure 1. The current residues in the four homologs at this location are represented by $\{A_n\}'$, in this case consisting of two alanines, one glycine, and one leucine, at nodes Δ, Θ, Γ, and Υ, respectively. (The prime indicates that the set of amino acids only represents the amino acids at the root of the tree.) We do not know the identity of amino acids A_Ω, A_Ψ, and A_Φ at this location in the ancestral proteins Ω, Ψ, and Φ. We thus have to consider all possibilities for these nodes explicitly. The conditional probability of residues $\{A_n\}'$ resulting through evolution if location n corresponded to site class k with corresponding mutation matrix $M^k_{i,j}$ can be expressed as

$$P(\{A_n\}'|M^k_{i,j}) = \sum_{A_\Omega, A_\Psi, A_\Phi} P(A_\Omega) \times M^k_{A_\Omega,A_\Psi}(d_{\Omega\Psi}) M^k_{A_\Psi,\mathrm{Ala}}(d_{\Psi\Delta}) M^k_{A_\Psi,\mathrm{Ala}}(d_{\Psi\Theta}) \times M^k_{A_\Omega,A_\Phi}(d_{\Omega\Phi}) M^k_{A_\Phi,\mathrm{Gly}}(d_{\Phi\Gamma}) M^k_{A_\Phi,\mathrm{Leu}}(d_{\Phi\Upsilon}) \quad (6)$$

where $M^k_{A_\Omega,A_\Psi}(d_{\Omega\Psi})$ is the probability that at such a location amino acid A_Ω would mutate to A_Ψ in evolutionary time $d_{\Omega\Psi}$ between nodes Ω and Ψ, computed by taking mutation matrix $M^k_{i,j}$ to the appropriate power.

As we do not know which site class k location n belongs to, and thus which is the appropriate mutation matrix $M^k_{i,j}$ to use, we must calculate $P(\{A_n\}'|M^k_{i,j})$ for each of the specific site classes, multiply by the probability $P(k)$ that the location can be described by site class k, and sum over all possible classes.

$$P(\{A_n\}') = \sum_k P(\{A_n\}'|M^k_{i,j})P(k) \quad (7)$$

Summing the logarithm of this probability over all locations provides us with a measure of the log likelihood for the entire database of sets of homologous proteins to have arisen given the model.

The model is defined by the various parameters in the fitness functions, the maximum mutation rate ν_k, and the various site-class probabilities $P(k)$. Bayes theorem can be used to demonstrate that, in the absence of other information, the most likely value of the various parameters in a model given a set of data are the parameters that maximize the likelihood that the data would be produced by that model. By adjusting these parameters to maximize the total probability, we can find the optimal sets of these parameters. This was performed using a sequential quadratic programming algorithm[32] 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 $Q = \log[P(\mathrm{Model})] - \log[P(\mathrm{Random})]$, where $\log[P(\mathrm{Model})]$ is the log of the probability that the given model would produce the data, and $\log[P(\mathrm{Random})]$ is a constant representing the probability that the data would result from purely neutral drift where all mutations were equally likely.

RESULTS AND DISCUSSION

We demonstrate our model with the application of simple evolutionary models to data sets consisting of *env*, *rev*, and *tat* sequences from HIV-1 and *env* sequences from HIV-2. The proteins varied from approximately 150 residues to over 1000, with the number of examples ranging from 20 to 100. Best results were obtained with fitness functions dependent on the hydrophobicity and bulk-related indices. The fewer number of parameters per site class allowed us to look at models with up to 11 site classes (4 linear, 7 quadratic fitness functions) with 57 variables. These models were optimized over several HIV-1 data sets, as well as over a general protein data set. The Q values for the various models are presented in Table 1.

A single traditional mutation matrix optimized over the 57 HIV-1 *env* proteins was, unsurprisingly, better able to model the HIV-1 *env* proteins than a similar mutation matrix optimized over a larger and more comprehensive data set. Interestingly, simple models with more than 7 different site classes optimized over the HIV-1 *env* proteins were better able to model the evolutionary process than a traditional matrix optimized over the same data set, in spite of the approximate factor of 10 fewer adjustable parameters. In this case, the inclusion of site heterogeneity

TABLE I. *Q* Values for Mutation Matrices and Simple Models, Optimized Using Estimation Maximization*

Optimization data set	Number of site classes	Number of adjustable parameters	HIV-1 ENV	HIV-1 REV	HIV-1 TAT	HIV-2 ENV
Traditional mutation matrices						
Gen	1	380	1665	281	201	2179
HIV-1 ENV	1	380	2249	311	222	2578
Dayhoff	1	380	1384	252	174	1858
Simple models						
HIV-1 ENV	2	9	*1764*	120	162	*2248*
HIV-1 ENV	3	15	*2096*	184	*218*	**2713**
HIV-1 ENV	5	25	*2192*	250	**290**	**3026**
HIV-1 ENV	7	35	**2294**	249	**303**	**3164**
HIV-1 ENV	9	47	**2350**	276	**323**	**3276**
HIV-1 ENV	11	57	**2475**	267	**334**	**3382**
Gen	2	9	642	126	104	796
Gen	3	15	152	40	−8	47
Gen	5	25	1342	189	*220*	1727
Gen	7	35	1243	154	165	1835

*The matrices and models were optimized either for a comprehensive general protein data set ("Gen") or for a data set consisting of the *env* proteins of HIV-1. Numbers in italics represent those models achieving better *Q* scores than the optimized Gen mutation matrix, and bold face numbers represent those models with *Q* scores superior to those from the mutation matrix optimized over the HIV-1 *env* data set. Results using the Dayhoff PAM 20 matrix ("Dayhoff") are shown for comparison.[33] Note that Q scores depend on the number of homologs in each data set, and thus numbers across columns are not comparable.

in the model more than made up for the limitations of the simplified assumptions. The use of simple models optimized over general data sets did not outperform single mutation matrices, indicating that the ability to focus on specific proteins is also significant.

Also presented in Table 1 are the results of our simple models optimized for HIV-1 *env* on the HIV-1 *tat* and *rev* proteins. Models optimized over HIV-1 *env* were better able to describe the HIV-1 *tat* data than any traditional mutation matrix. This was not the case for the HIV-1 *rev* protein, where even 11 site models optimized over HIV-1 *env* were not able to outperform the mutation matrices. These differing results for HIV-1 *rev* and *tat* might reflect structural or functional differences between these proteins—that as far as amino acid fitness is concerned, there are more similarities between *env* and *tat* than between these proteins and *rev*.

The ability to represent the evolutionary patterns of HIV-1 *env* may represent identification of the salient patterns, or over-fitting and memorization of the evolutionary history of this particular data set. These possibilities can be distinguished by testing the resulting optimized model on a second data set with a distinct evolutionary history. The *env* proteins of HIV-2 are under similar evolutionary pressures as the *env* proteins of HIV-1, yet the post-divergence evolutionary trajectory of these proteins is completely different from HIV-2, with these two sets averaging only 35–45% sequence identity. The mutation matrix optimized over the HIV-1 *env* proteins was better able to model the evolutionary data of the HIV-2 *env* proteins than were the more generic matrices. This effect was even more dramatic when the simple models optimized over the HIV-1 *env* proteins were applied to the *env* proteins of HIV-2, where the reduced number of parameters lowered the danger of over-fitting of the data. Even the simple 3-site model outperformed the HIV-1 *env* mutation matrix. This indicates the importance of including site-heterogeneity in a way impossible with a single mutation matrix.

The performance of the model optimized and tested over the HIV-1 *env* data set should keep improving as the number of site classes and parameters are increased. Conversely, the performance of the HIV-1 *env* optimized model as tested over the disjoint HIV-2 *env* data-set should only improve until the number of adjustable parameters are large enough for the algorithm to start to memorize patterns specific to the HIV-1 *env* proteins. As shown in Table 1, this point has not been reached even with an 11 site model. This suggests that even better results might be obtained with more complicated models that are, unfortunately, currently beyond our computational resources.

CONCLUSIONS

Site heterogeneity has been included in the construction of protein-specific profiles and Hidden Markov models that encode the relative probability

of any amino acid occurring at a given location in a particular family of proteins.[34–38] These profile methods have shown promise in the detection of proteins that are related either by evolution or structure. Because the relative rate of change from one amino acid to another is not included in these approaches, these profiles are unsuited to answering more specific questions about evolutionary relationships, such as phylogenetic or ancestral reconstruction. A profile for a protein of length N would require, in the most general sense, the adjustment of $20 \times N$ parameters (neglecting insertions and deletions.) In contrast, a location-specific model for mutation matrices would require the determination of $380 \times N$ parameters, representing the probability of all possible changes in all different positions. This much larger number of adjustable parameters necessitates the types of simplifications described in this paper.

In using a protein-specific data set of HIV proteins, we have demonstrated that our simple model of evolution can be successfully applied to small data sets. This was shown by the results of models optimized for the HIV-1 *env* protein on the HIV-2 *env* protein. All models more complex than 3 site classes showed a higher probability of producing the data than any mutation matrix, even one optimized for the HIV-1 *env* protein. These results show that the use of a single mutation matrix is overlooking a very important factor: site-heterogeneity. Our models, which can encompass site-heterogeneity, prove more likely to reproduce the observed data. These models of evolution also have an advantage in that the number of adjustable parameters can be tuned to best balance the needs of specificity and generalization.

ACKNOWLEDGMENTS

We would like to thank Darin Taverna for his work in deriving the models of amino acid substitution used in this work, and Kurt Hillig for computational assistance.

REFERENCES

1. Pauling, L., Zuckerkandl, E. Chemical paleogenetics: Molecular "restoration studies" of extinct forms of life. Acta Chem. Scand. 17:S9–S16, 1963.
2. Jermann, T.M., Optiz, J.G., Stackhouse, J., Benner, S.A. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature (London) 374:57–59, 1995.
3. Koshi, J.M., Goldstein, R.A. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. 42:413–420, 1996.
4. Koshi, J.M., Goldstein, R.A. Mutation matrices and physical-chemical properties: Correlations and implications. Proteins 27:336–344, 1997.
5. Dayhoff, M.O., Eck, R.V. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Vol. 3. Dayhoff, M.O., Eck, R.V. (eds.). Silver Spring, MD: National Biomedical Research Foundation, 1968:33–41.
6. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219:555–565, 1991.
7. Henikoff, S., Henikoff, J.G. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89:10915–10919, 1992.
8. Jones, D.T., Taylor, W.R., Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282, 1992.
9. Luthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
10. McLachlan, A.D. Tests for comparing related amino-acid sequences. J. Mol. Biol. 61:409–424, 1971.
11. Overington, J., Donnelly, D., Johnson, M.S. Šali, A., Blundell, T.L. Environment-specific amino-acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1:216–226, 1992.
12. Risler, J.L., Delorme, M.O., Delacroix, H., Henaut, A. Amino acid substitutions in structurally related proteins. J. Mol. Biol. 204:1019–1029, 1988.
13. Feng, D.F., Johnson, M.S., Doolittle, R.F. Aligning amino-acid sequences: A comparison of commonly used methods. J. Mol. Evol. 21:112–125, 1985.
14. Fitch, W.M. An improved method of testing for evolutionary homology. J. Mol. Biol. 16:9–16, 1966.
15. Grantham, R. Amino acid difference formula to help explain protein evolution. Science 185:862–864, 1974.
16. Miyata, T., Miyazawa, S., Yasunaga, T. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12:219–236, 1979.
17. Miyazawa, S., Jernigan, R.L. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng. 6:267–278, 1993.
18. Rao, J.K.M. New scoring matrix for amino acid residue exchange based on residue characteristic physical parameters. Int. J. Pep. Prot. Res. 29:276–281, 1987.
19. Koshi, J.M., Goldstein, R.A. Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. Protein Eng. 8:641–645, 1995.
20. Nakai, K., Kidera, A., Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. 2:93–100, 1988.
21. Tomii, K., Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9:27–36, 1996.
22. Ladunga, I., Smith, R.F. Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. Protein Eng. 10:187–196, 1997.
23. Kidera, A., Konishi, Y., Oka, M., Ooi, T., Scheraga, H.A. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Protein Chem. 4:23–55, 1985.
24. Pohl, F.M. Empirical protein energy maps. Nat. New Biol. 234:227–279, 1971.
25. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.
26. Finkelstein, A.V., Ptitsyn, O.B., Kozitsyn, S.A. Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structures with experiment. Biopolymers 16:497–524, 1977.
27. Serrano, L., Sancho, J., Hirshberg, M., Fersht, A.R. Alpha-helix stability in proteins. J. Mol. Biol. 227:544–559, 1992.
28. MacArthur, M.W., Thornton, J.M. Influence of proline residues on protein conformation. J. Mol. Biol. 218:397–412, 1991.
29. Bryant, S.H., Lawrence, C.E. The frequence of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 9:108–119, 1991.
30. Rashin, A.A., Iofin, M., Honig, B. Internal cavities and buried waters in globular proteins. Biochem. 25:3619–3625, 1986.
31. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Y. Boltzmann-like statistics of protein architectures. Subcell Biochem. 24:1–26, 1995.

32. Gill, P.E., Hammarling, S.J., Murray, W., Saunders, M.A., Wright, M.H. User's guide for MPSOL (version 4.0). Stanford, CA: Department of Operations Research, Stanford University Report SOL 86-2, 1986.
33. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Vol. 5, suppl. 3. Dayhoff, M.O. (ed.). Washington, D.C.: National Biomedical Research Foundation, 1978:345.
34. Taylor, W. The classification of amino acid conservation. J. Theor. Bio. 119:205–218, 1986.
35. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.
36. Krogh, A., Brown, M., Mian, I.S., Sjölander, K., Haussler, D. Hidden markov models in computational biology. J. Mol. Biol. 235:1501–1531, 1994.
37. Tatusov, R.L., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. U.S.A. 91:12091–12095, 1994.
38. Yi, T.M., Lander, E.S. Recognition of related proteins by iterative template refinement. Protein Sci. 3:1315–1328, 1994.