Pacific Symposium on Biocomputing 6:191-202 (2001)

ANALYZING SITE HETEROGENEITY DURING PROTEIN EVOLUTION

JEFFREY M. KOSHI
Biophysics Research Division, University of Michigan
Ann Arbor, MI 48109-1055 USA
Current address: Cereon Genomics, 45 Sidnet St., Cambridge MA 02139

RICHARD A. GOLDSTEIN
Dept. of Chemistry and Biophysics Research Division, University of Michigan
Ann Arbor, MI 48109-1055 USA

New computational models of the kinetics of natural site substitutions in proteins are described based on the underlying physical chemical properties of the amino acids. The corresponding reduction in the number of adjustable parameters allows us to analyze site-heterogeneity. Applying this evolutionary model to various data sets allows us to identify the important factors constraining molecular evolution, providing insight into the relationship between amino acid properties and protein structure.

Introduction

Despite the large and growing number of solved protein structures, we still do not understand the basic forces that determine a protein's three dimensional fold. The role of hydrophobicity has been emphasized by a number of researchers, but the extent of its effects and the importance of other factors such as side-chain volume and local structure propensity are still widely debated. A more detailed question is the effect of local environment on the importance of these factors. What amino acid characteristics are important in solvent-exposed locations as compared to solvent-buried positions, or for residues in alpha helices vs. coils?

Much has been learned through directed site-mutagenesis, where the effects of various substitutions on protein function and/or stability are examined. Creating, verifying, purifying, and characterizing these mutant proteins is a time-consuming process, however, limiting the number of substitutions that can be studied. Even worse, researchers are often interested in looking at key structural or active site residues that are unrepresentative of more general locations in the protein. With these problems, it is difficult to construct a data set of artificial substitutions large enough to analyze general tendencies.

Researchers have only recently been able to perform site-mutagenesis tests, but nature has been doing so for billions of years. In addition, all the experiments done by evolution were performed *in vivo*. The difficulty is in analyzing

nature's data base. Researchers have tried several methods to solve this problem based on the observation that evolution has held structure and function largely constant over geologic time scales and over widely varying sequences. It is likely that attributes preserved during evolution are the ones that are important in conserving structure and function. For instance, Scheraga, Nakai, Tomii, and their respective coworkers examined the correlations between the many amino acid properties, and investigated simple linear correlations of these properties with substitution rates.[1–3] We used a similar approach to analyze our previously-derived structure-dependent substitution matrices.[4,5]

Correlation analyses performed on substitution matrices have several limitations. One of these is the lack of a rigorous theoretical basis for the analysis. More fundamentally, the construction of these matrices generally assume that all locations in the protein are equivalent and that all prolines are equally-likely to mutate to alanine independent of position in the protein. In reality, the absolute and relative substitution rates will depend on many specific features of the given residue and location, including solvent exposure and secondary structure, tertiary contacts, and functional significance.[6–10] While there is a high degree of sequence plasticity, there are many locations under selective pressure to preserve specific physical-chemical properties or even amino acid identities. While models of evolution have been developed that include heterogeneity of substitution rates,[11] these models often tend to assume that the ratio between the various substitution rates at each location is relatively constant and that only the magnitude of the rates changes. In fact, a given amino acid change may represent a conservative substitution in some instances and a highly deleterious substitution in others.

One approach to deal with this distribution is to divide the protein into different site classes on the basis of local structure and surface accessibility, and calculate specific matrices for the various classes.[6,8,9,10] This ignores variations in the selective pressure at different locations that share local conditions so that different substitution rates due to functional considerations and packing constraints are averaged. An alternative approach is to consider the amino acid residues observed in each position in order to construct separate substitution models for each site.[12] The limited data available at each location makes it difficult to use these models to gain qualitative and quantitative understandings of the relationship between amino acid properties and protein structure and function. Recently an approach has been developed, which we call a Hidden States Model (HSM).[10,13–18] In this approach, each location in the protein is assumed to belong to one of a set of possible site classes, each corresponding to a separate substitution matrix. The identity of the site class describing any particular site is unknown (and thus "hidden"), and can only be deter-

mined probabilistically; each site class is assigned an *a priori* probability that any protein location would be in that class . We can use maximum-likelihood methods to optimize the entire set of substitution matrices and corresponding *a priori* probabilities. The problem with this approach is the explosion in the number of adjustable parameters that must be simultaneously determined.

Rather than develop a HSM for the substitution rates at various locations based on the identity of the amino acids and attempt to correlate the various substitution rates with changes in physical-chemical parameters, a number of investigators have constructed substitution rates as a direct function of the underlying properties of the amino acids. Two different approaches have been explored. One approach [11,18] is based on the fact that similar amino acids tend to replace each other more frequently than dissimilar amino acids.[19] Substitution rates can then be used to determine what defines "similar" and "dissimilar", that is, what properties nature considers sufficiently important to conserve. We have been pursuing a different approach based on concepts from structural biology, where we imagine that there is a propensity for different amino acids in different locations and that substitutions to amino acids with higher propensities would be favored.[15,16] In this approach, it is the relative propensities of the respective amino acids that matter rather than their similarities; evolution favors conservative substitutions because of a statistical propensity for changes between relatively high-propensity amino. Both of these methods greatly reduces the number of adjustable parameters so that multi site-class HSMs can be optimized for protein datasets of only modest size. In addition, the interpretation of the substitution rates are straight-forward in that the models are already based on the physical-chemical properties.

In earlier work, we showed that our models can better represent the evolutionary patterns of specific sets of proteins than traditional substitution matrices [16] and showed how these models could be used in phylogenetic analyses.[20] In this paper, we analyze what these substitution models can say about the nature of the selective pressure occurring at various locations in the protein. Optimizations were done over a general protein data set, and various subsets determined by secondary structure and surface accessibility. In agreement with earlier work, we found hydrophobicity to be an important factor in all local environments, especially in exposed positions. We also observed an interesting variations in the importance of hydrophobicity over different secondary structures, with exposed α-helices demonstrated the strongest dependence followed closely by exposed β-sheets. Most locations, especially turns and coils, preferred smaller residues in agreement with the idea that larger residues with more conformational flexibility are disfavored for entropic reasons.

Methods

We first review our model for site substitutions as described elsewhere.[16,20,17] We encompass the distribution of selective pressure at various locations in the protein by assuming that each location under consideration can be described by one of a number of possible site-classes $\mathcal{S}_k$. We do not know which location belongs to which particular site-class. Instead, we imagine that each location has an *a priori* probability $P(k)$ of belonging to site class $\mathcal{S}_k$. As all locations must belong to some site class, $\sum P(k) = 1$. The rates of substitution from amino acid $\mathcal{A}_i$ to amino acid $\mathcal{A}_j$ for locations in site class $\mathcal{S}_k$ are described by substitution matrix $M^k_{i,j}$. Each site class has its own distinct substitution matrix. The model consists of the set of substitution matrices and *a priori* probabilities for all of the various site classes.

As mentioned in the introduction, we represent the substitution rates encoded in $M^k_{i,j}$ as a function of the properties of amino acids $\mathcal{A}_i$ and $\mathcal{A}_j$, rather than their identities. We assume that the "fitness" $F_k(\mathcal{A}_i)$ of amino acid $\mathcal{A}_i$ for any location described by a particular site class $\mathcal{S}_k$ can be expressed as a simple linear or quadratic form of a set of physical-chemical parameters such as hydrophobicity, bulk, and local structure propensity. $F_{k,l}(\mathcal{A}_i)$, the contributions to the fitness function due to physical-chemical property l, are assumed to be either of the linear form $F_{k,l}(\mathcal{A}_i) = \alpha_{k,l}\, q_l(\mathcal{A}_i)$ or the quadratic form $F_{k,l}(\mathcal{A}_i) = \alpha_{k,l} \left(q_l(\mathcal{A}_i) - q^{\rm opt}_{k,l} \right)^2$, where $q_l(\mathcal{A}_i)$ represents the value of the physical-chemical parameter l for amino acid $\mathcal{A}_i$, and $\alpha_{k,l}$ and $q^{\rm opt}_{k,l}$ are parameters that depend upon the site class $\mathcal{S}_k$. The linear fitness function is appropriate when there is a general tendency for that physical-chemical factor to be either favored or disfavored at that site. The quadratic form would be appropriate when there is either an optimal parameter value (for positive $\alpha_{k,l}$) or most non-optimal value (for negative $\alpha_{k,l}$). The total fitness is the sum of the terms reflecting the various physical-chemical factors, as $F_k(\mathcal{A}_i) = \sum_l F_{k,l}(\mathcal{A}_i)$.

For the physical-chemical parameters $q_l(\mathcal{A}_i)$ in the above equations, we use the four orthogonal property indices developed by Scheraga and coworkers, correlated predominantly with alpha helical and turn propensity (α/turn), bulk-related factors (volume, molecular weight, etc.), beta sheet propensity, and hydrophobicity.[1] The α/turn index is negatively correlated with α-helical propensity and positively correlated with turn propensity - *i.e.* amino acids with high α-helical propensities tend towards negative values, and amino acids with high turn propensities tend towards positive values. The bulk-related and β-sheet propensity indices are positively correlated with their factors, so large residues such as Trp and high β-sheet propensity residues such as Val will have

large, positive values in their respective indices. The hydrophobicity index is negatively correlated with hydrophobicity, meaning *hydrophilic* residues have high positive values in this index.

We assume that the probability $P_k(\mathcal{A}_i)$ of any given amino acid $\mathcal{A}_i$ occurring at any location described by a site class k is given by a Boltzmann relation $P_k(\mathcal{A}_i) = e^{F_k(\mathcal{A}_i)} / \left(\sum_{i'} e^{F_k(\mathcal{A}_{i'})}\right)$where i' is an index over all amino acids. This expression can be considered a definition of the fitness $F_k(\mathcal{A}_i)$. We consider the substitution rate as equal to the product of a site-class dependent attempt rate ν_k and a relative probability of fixation in the population of the species. We consider that the relative probability of all favorable substitutions are constant while unfavorable substitutions to less-fit amino acids are accepted at an exponentially-decreasing function of the difference in fitness values. The value of M_{ij}^k corresponding to a substitution from amino acid $\mathcal{A}_i$ to $\mathcal{A}_j$ in a location described by site class $\mathcal{S}_k$ is then given by Metropolis kinetics:

$$M_{ij}^k = \begin{cases} \nu_k & | \quad F_k(\mathcal{A}_j) > F_k(\mathcal{A}_i) \\ \nu_k \, e^{(F_k(\mathcal{A}_j) - F_k(\mathcal{A}_i))} & | \quad F_k(\mathcal{A}_j) \leq F_k(\mathcal{A}_i) \end{cases} \tag{1}$$

The Metropolis scheme is the only kinetics scheme ensuring a Boltzmann distribution and detailed balance and where a favorable substitution is always accepted at the maximum rate.

As the physical-chemical parameters $\{q_l(\mathcal{A}_i)\}$ for all of the amino acids are fixed, the model is completely defined by the *a priori* probabilities $\{P(k)\}$, the fitness parameters $\{\alpha_{k,l}\}$ and $\{q_{k,l}^{\text{opt}}\}$, and the maximum substitution rates $\{\nu_k\}$. The various substitution matrices are calculated using equation 1, and the entire set of parameters optimized as described before. [6,16] The log-likelihood of the data given the model is calculated by considering each location n in a set of aligned sequences separately and calculating $P(\{\mathcal{A}_n\}|M_{i,j}^k)$, the probability of observing present-day amino acids $\{\mathcal{A}_n\}$ at that location in the various protein sequences, given that this location belongs to site class $\mathcal{S}_k$. As we do not know the identity of the site class specific for each location, we can calculate $P(\{\mathcal{A}_n\})$, the overall probability of the observed amino acids being observed at site n, by multiplying the conditional probabilities $P(\{\mathcal{A}_n\}|M_{i,j}^k)$ by the *a priori* probability $P(k)$ and summing over all possible classes, as $P(\{\mathcal{A}_n\}) = \sum_k P(\{\mathcal{A}_n\}|M_{i,j}^k)\, P(k)$. Summing the logarithm of this probability over all locations provides us with a measure of the log likelihood for the database of observed sequences given the model. The parameters of the model were then optimized for the dataset using a sequential quadratic programming algorithm [21] 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(Model)] is the log of the probability that the given model would produce the data, and log[P(Random)] is a constant representing the probability that the data would result from purely neutral drift with no selective pressure.

Results

One use of our simple models is to determine what amino acid indices contribute the most to the fitness functions. For this purpose, a general protein data set was constructed by selecting 42 proteins of length greater than 80 residues from the list constructed by Hobohm and Sander,[22] all with 6 to 11 homologs of 30% or greater sequence identity listed in the HSSP database.[23] The average number of homologs for each protein was 10.5. A multiple alignment and unrooted phylogenetic tree was created for each set using the program ClustalV.[24] The sequence, structure, and surface accessibilities were found by use of the DSSP program on the corresponding PDB files.[25,26] π-helices were included with α-helices, 3_{10}-helices and bends were included with turns, while β-bridges were included with coils. Residues were considered exposed if greater than 18% of their surface area was exposed to solvent.

Models with two site classes were optimized where F_k was a function of all four of Scheraga's orthogonal indices. These models used quadratic fitness functions for each index in each site class, separate ν_k values for each site class, and two $P(k)$ values, one for each site class. As the two $P(k)$ values must sum to one, there were a total of nineteen adjustable parameters. This process was carried out for the total ensemble of data points, as well as independently for subsets of the data based on secondary structure and solvent accessibility. These models, although they have 20 times fewer parameters than our substitution matrices, seemed to encompass most of the details captured by our matrices, achieving from 51 to 74% of the Q value of the more complete substitution matrix optimized over the same data set. In each case, we calculated how much each physical chemical parameter contributed to the variance of the fitness values of the different amino acids for each of the site classes. The data set was broken into fifths, and parameters optimized separately for each subset; the values, reported in Table I, represent the mean as found from these 5 trials.

For exposed residues, hydrophobicity seemed to be the most important constraint. The most populated site class for exposed residues had a large fraction of the total variance dependent on preserving hydrophilicity, with the bulk-related index a distant second. As we have suggested earlier, the reason for the importance of conserving hydrophilicity in exposed residue positions is likely the reverse hydrophobic effect, that is, the tendency of the protein to

Table 1: Importance of different parameters for various site classes. Percentage of variances from fitness functions depending on the amino acid properties listed, for the data sets on the left. This was done for each data set with a 2 site class model, the site class and percentage occupancy denoted in the 2nd and 3rd columns respectively. ν_k values represent the maximal acceptance rate of mutations for that site class and data set, in arbitrary units. Bold faced numbers are those variances that contributed over 25% of the total. A plus (+) indicates a positive correlation (i.e. either a quadratic function with a positive maxima or a negative minima) with that index in 3 of 5 trials, two symbols (++) indicates a positive correlation in 4 of 5 trails, and a (+++) indicates a positive correlation in all 5 trials. Similarly, a minus (−) implies a negative correlation (i.e a quadratic function with a negative maxima or positive minima) in 3 or 5 trials, (−−) in 4 of 5, and (− − −) in all 5.

	site			Percentage of variance accounted by			
data set	class	% occ	ν_k	α-helix,turn	bulk-related	β-sheet	hydrophobicity
exposed	1	65	2.09	3 (+++)	23 (− − −)	1 (−−)	**74** (+++)
	2	35	0.94	**32** (−)	10 (+)	9 (++)	**49** (− − −)
buried	1	56	0.85	6 (+++)	**76** (−−)	3 (+++)	15 (+++)
	2	44	1.61	18 (− − −)	14 (+++)	**32** (+++)	**35** (− − −)
exposed	1	59	2.99	0.3 (− − −)	6 (− − −)	2 (−−)	**92** (+++)
α-helix	2	41	1.05	**47** (− − −)	4(−)	9 (+++)	**40** (+)
exposed	1	53	2.04	1 (−)	17 (− − −)	1 (−−)	**82** (+++)
β-sheet	2	47	1.09	15 (−−)	**26** (−)	22 (+++)	**36** (−)
exposed	1	65	2.52	5 (+++)	**53** (− − −)	6 (− − −)	**36** (+++)
turn	2	35	0.81	22 (+)	**29** (−−)	3 (+++)	**46** (−)
exposed	1	62	2.20	9 (+++)	**36** (− − −)	1 (− − −)	**54** (+++)
coil	2	38	0.62	**45** (++)	24 (− − −)	1 (+)	**29** (−−)
buried	1	44	2.18	22 (− − −)	0.5 (−)	**50** (+++)	**27** (− − −)
α-helix	2	56	0.85	0.2 (− − −)	**79** (−−)	0 (++)	20 (+++)
buried	1	56	1.02	7 (−)	**70** (− − −)	8 (+++)	15 (−)
β-sheet	2	44	1.94	9 (−−)	**36** (+)	12 (+++)	**43** (− − −)
buried	1	60	1.04	5 (+)	**63** (−−)	9 (+)	23 (−)
turn	2	40	0.76	**33** (+)	19 (−−)	**28** (−)	20 (−)
buried	1	58	1.05	9 (−)	**52** (−−)	11 (++)	**28** (−)
coil	2	42	0.76	**49** (+)	13 (−)	4 (−)	**34** (−−)

avoid stabilizing alternative conformations in which these residues are buried.[5] Although hydrophobicity was also an important factor for the second site class, the preference was for *hydrophobic* residues. In addition, this hydrophobicity factor was significantly modified by an almost equally strong preference for residues with high α-helical propensity (the α/turn index is negatively correlated with α-helical propensity).

The importance of hydrophobicity differed for different exposed secondary structures. In the site class where hydrophilic residues were most strongly preferred, exposed α-helices showed the greatest dependence on hydrophobicity, with the importance for exposed β-sheets somewhat lower, and even lower importance for exposed coils and turns. This might reflect the relative importance of hydrophobicity in maintaining these various exposed secondary structures. This conclusion agrees with our previous conclusions and those of other groups that secondary structure propensity appears to be of little importance compared to patterns of hydrophobicity in maintaining structures, especially in α-helices. [5,27,28] In all cases except exposed α-helices, one exposed site class showed a preference for hydrophobic residues. For exposed α-helices,

the second site with a weaker dependence on hydrophobicity showed a strong tendency for large α-helix propensity (corresponding to a negative value of the α/turn index). In general, exposed locations showed at most moderate concern for β-sheet propensity compared to these other factors, even in β-sheets. Coils, and to a lesser extent turns, showed a propensity for amino acids with a strong turn propensity, and a corresponding tendency to avoid residues with α-helical propensity, indicative of the need for the protein to interrupt regular structures.

The situation for buried residues was somewhat different, with hydrophobicity competing with bulk and β-sheet propensity for importance. This demonstrates the importance of packing interactions in the protein interior, as well as the greater tendency for β-sheets to be buried, where their structure cannot be constrained by patterns of hydrophobicity. Surprisingly, buried α-helices exhibited a strong *positive* dependence on β-sheet propensity. This preference of buried α-helices for residues with high β-sheet propensity would mean these two structures would exert similar selective pressures on their constituent amino acids. This fact could help explain why secondary structure prediction algorithms have such a difficult time differentiating between buried α-helices and β-sheets.

In order to more completely unravel the site heterogeneity for hydrophobicity and the bulk-related indices, we further investigated models with 11 site-classes, each with a different fixed optimal parameter. Each site-class was quadratic in the same single parameter, with q_k^{opt} values spaced at regular intervals and fixed α values. ν was considered a global parameter representing the maximum substitution rate for all site classes. Thus, the only variables in this model were the single average ν value and the site class probabilities $P(k)$. Figure 1 shows these results for different surface accessibility and secondary structure classes.

A small fraction of exposed locations, 17%, preferred hydrophobic residues. (The reduced fraction compared with the population of exposed locations in site class 2 from Table 1 is likely due to the incorporation of locations with strong α-helical propensity into this latter site class.) In contrast, buried locations had a much less stringent requirement, as shown in Figure 1, where buried residues have a significant population preferring slightly hydrophilic residues, indicative of the greater complexity of selective constraints in the protein interior. This heterogeneity of locations can explain why we found hydrophobicity more conserved in exposed locations than in buried locations in earlier work.[5]

There was a general tendency against bulky residues in all secondary structures and surface accessibilities. One possible explanation for this result is

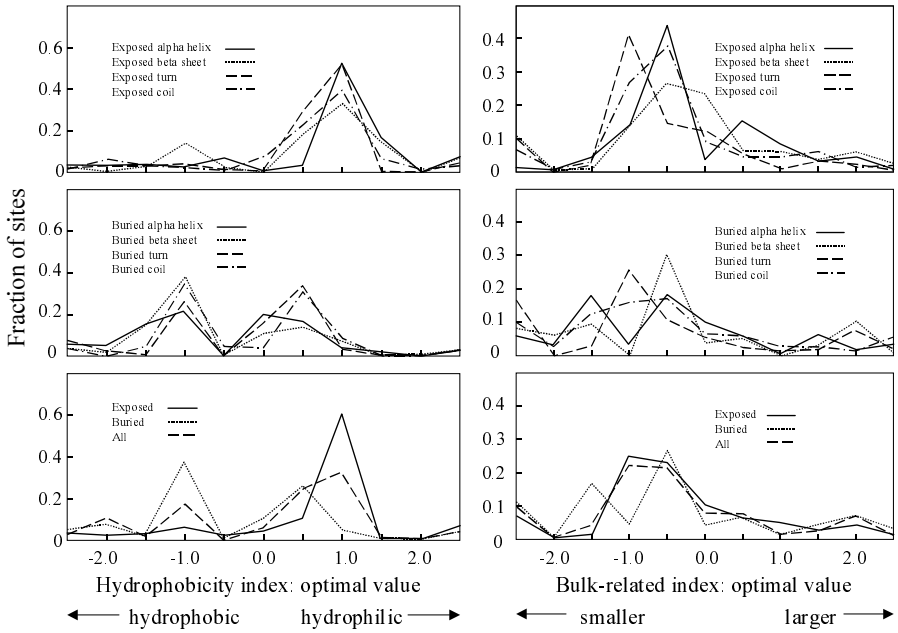

Figure 1: The fraction of locations in the protein classified as a member of the particular site class with a given value of optimal hydrophobicity or bulk index $q_k^{\rm opt}$, for various subsets of the general protein data base.

that during folding large side chains experience a steep loss of conformational entropy. Large side chains with more conformational entropy would have a greater tendency to destabilize the folded state. There is greater variation in the optimal size for buried residues as would be expected given the heterogeneity of packing interactions. Interestingly, both buried α-helices and β-sheets show a bimodal distribution within the negative range of the bulk-related index (preferring smaller residues). This suggests these buried secondary structures have at least two distinct environments preferring small residues - one with a preference for very small residues, and another preferring only moderately small residues.

Discussion

Our study demonstrates that the hydrophobicity index was the dominant factor for exposed positions while the bulk-related index was dominant for buried

residues. This observation held over all secondary structures. The importance of hydrophobicity in exposed positions has not been as emphasized as has its role in the hydrophobic protein core. This may be an indication that the reverse hydrophobic effect has more of a key role than previously thought. Another interesting observation was the differing importance in the hydrophobicity index over exposed secondary structures. Exposed α-helices showed the strongest contribution of hydrophobicity, followed by exposed β-sheets and finally by turns and coils. This difference may reflect the importance of hydrophobicity in maintaining exposed secondary structures. Almost all types of locations showed heterogeneity in hydrophobicity, with some exposed locations preferring hydrophobic residues, especially in β-sheets, and with a large fraction of internal locations preferring hydrophilic residues.

Exposed α-helices showed a preference for residues with high α-helical propensity, while buried α-helices had a preference for residues with high β-sheet propensity, suggesting the importance of considering the location of the protein surface when predicting local structure. Our results also showed both buried *and* exposed locations tend to disfavor large, bulky residues, consistent with the idea that bulky residues are disfavored due to their large loss in side-chain entropy upon folding.

Acknowledgments

We would like to thank Darin Taverna for his work deriving the substitution models used in this work, and Kurt Hillig for computational assistance. Financial support was provided by NIH Grants GM08270 and LM05770 and NSF equipment grant BIR9512955.

References

1. 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.*, 4:23–55, 1985.
2. K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engin.*, 2:93–100, 1988.
3. K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engin.*, 9:27–36, 1996.
4. 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, Singapore, 1995.

5. J. M. Koshi and R. A. Goldstein. Mutation matrices and physical-chemical properties: Correlations and implications. *Proteins*, 27:336–344, 1997.
6. J. M. Koshi and R. A. Goldstein. Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. *Protein Engin.*, 8:641–645, 1995.
7. M. Kimura and T. Ohta. Mutation and evolution at the molecular level. *Genet. Suppl.*, 73:19–35, 1973.
8. 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.
9. 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.
10. J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, 13:666–673, 1996.
11. Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.*, 15:1600-1611, 1998.
12. A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15:910-917, 1998.
13. Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.
14. J. Felsenstein and G. A. Churchill. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.
15. J. M. Koshi, D. P. Mindell, and R. A. Goldstein. Beyond mutation matrices: Physical-chemistry based evolutionary models. In S. Miyano and T. Takagi, editors, *Genome Informatics 1997*, pages 80–89. Universal Academy Press, Tokyo, 1997.
16. J. M. Koshi and R. A. Goldstein. Models of natural mutations including site heterogeneity. *Proteins*, 32:289–295, 1998.
17. M. W. Dimmic, D. P. Mindell, and R. A. Goldstein. Modeling evolution at the protein level using an adjustable amino acid fitness model. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. Klein, editors, *Pacific Symposium on Biocomputing 2000*, pages 18-29. World Scientific, Singapore, 1999.

18. Z. Yang. Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. Klein, editors, *Pacific Symposium on Biocomputing 2000*, pages 81-92. World Scientific, Singapore, 1999.
19. E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving genes and proteins*, pages 97-116. Academic Press, New York, 1965.
20. J. M. Koshi, D. P. Mindell, and R. A. Goldstein. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. *Mol. Biol. Evol.*, 16:173–179, 1999.
21. 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.
22. U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–524, 1994.
23. C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.
24. D. G. Higgins, A. J. Bleasby, and R. Fuchs. Clustal V: improved software for multiple sequence alignment. *CABIOS*, 8:189–191, 1992.
25. W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopoly.*, 22:2577–2637, 1983.
26. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.
27. M. W. West and M. H. Hecht. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.*, 4:2032–2039, 1995.
28. H. Xiong, B. L. Buckwalter, H. Shieh, and M. H. Hecht. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Nat. Acad. Sci. USA*, 92:6349–6353, 1995.