

USING EVOLUTIONARY INFORMATION TO STUDY PROTEIN STRUCTURE

RICHARD A. GOLDSTEIN

Siena Biotech, via Fiorentina 1, 53100 Siena, Italy

E-Mail: rgoldstein@sienabiotech.it

Received: 26th July 2002 / Published: 15th May 2003

ABSTRACT

The genomic data available to computational biologists represents the product of the complex processes of evolution. In particular, the forces of mutation, duplication, and selection have acted to sculpt modern protein sequence and structure in the context of changing functional requirements. Just as crystallographers are able to determine protein structures through an analysis of X-ray diffraction patterns, we wish to read the evolutionary history of proteins in order to understand their structures, functions, and interactions. To this end, we have been developing models of natural site substitutions that are informed by the protein structure and function and the resulting variations in selective pressures, even when the structure and function of the protein are unknown. By phrasing the substitution process in terms of the underlying properties of the constituent amino acids we can build models that are both much more accurate and more interpretable. The model is applied to a large set of globular proteins as well as a set of G-protein coupled receptors, identifying general structural and functional features of these biomolecules.

INTRODUCTION

The various genome projects have produced a plethora of gene sequences encoding proteins for which we have little information. While there are extensive experimental efforts to characterize the structure, function, and other characteristics of these proteins, there still remains a substantial backlog. In addition, many proteins of major interest are resistant to many of these experimental techniques. This has helped to spur the development of techniques to predict the

characteristics of these proteins based only on sequence information. Often we have multiple sequences of related proteins from different organisms.

It has been long recognized that these multiple sequences provide us with a valuable opportunity, that a set of related sequences convey more information than just a single example. The challenge has been to extract meaningful information from these multiple sets.

Given an alignment, it is possible to identify the conserved residues, to characterize the amount of sequence variation at each location using such concepts as “sequence entropy”, and to look for correlated changes between different locations in the protein. These approaches generally treat the observed protein sequences as a random sampling from the space of all possible sequences. This, of course, is false. One obvious problem is the uneven distribution of proteins among different organisms, depending upon the relative importance of the species to individual scientific investigators. More insidiously, homologous proteins are related by a phylogenetic structure that can induce confounding tendencies in the data. For example, Figure 1 shows a phylogenetic pattern where two substitutions occurred in different branches of the tree. In this simple example, there is a complete correlation between the third and seventh positions, even though this does not represent the effect of compensatory substitutions. One approach to handling these complications is to model the evolutionary process explicitly. This is the approach that we take here.

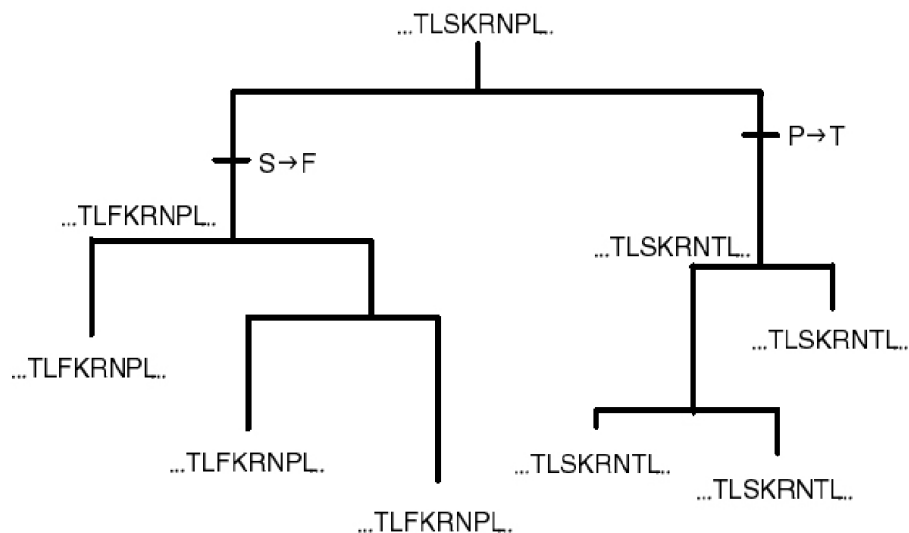


Figure 1. Example of an evolutionary trajectory producing an artificial correlation between sequence locations.

The standard method to model the site substitutions that occur during evolution is through a “substitution matrix”, a 20 x 20 matrix representing the probability that one amino acid would be replaced by another in a given length of evolutionary time. Standard approaches generally

use a single substitution matrix for all locations in the protein, implicitly assuming that all locations in the protein can be represented by the same model, that is, are under similar selective pressure. This is, of course, unrealistic. It has been shown that substitution rates vary with surface accessibility, secondary structure, and functional significance. One method to approach this problem is to subdivide the various locations in proteins according to their local structure, constructing and using structure-dependent substitution matrices (1, 2, 3, 4). This approach still assumes that all locations with the same local structure are under similar selective pressure, ignoring differences based on the inevitable “coarse graining” of the structural classifications as well as selective pressure due to function.

Recently we developed an approach, which we call a Hidden States Model (HSM), for dealing with these differences in selective pressure (5, 6, 7, 8). In this model, each location is assumed to belong to one of a set of possible “site classes”, each corresponding to a single substitution matrix. The various substitution matrices are unknown, as is the site class to which each location belongs. Instead, each location in the protein has the same set of a priori probabilities for belonging to each site class. The a priori probabilities as well as the set of substitution matrices are determined based on a set of related proteins through a maximum likelihood formulation. The result of this procedure is a set of site classes with corresponding substitution matrices, as well as the ability to calculate the a posteriori probability that any given location is a member of any particular site class. This then provides us with information regarding 1) which locations are under related selective pressure, 2) what is the nature of this selective pressure, and 3) when is the selective pressure different for different subsets of proteins.

The central challenge in this approach is the total number of parameters that must be adjusted in the optimization process. We deal with this situation by representing the entire substitution matrix with a biologically-inspired reduced set of parameters. In general, we consider the local propensity, or fitness, of each amino acid for any location described by a given site class, and then represent the probability of substitution of one amino acid for another in terms of the differences in these local fitnesses. The functional forms of this representation can be quite general, with additional parameters that can be optimized based on the observed data.

In this paper, we first investigate the nature of the substitution process at different types of locations in a set of globular proteins. We then demonstrate the application of these models for understanding the selective pressure acting on one particular set of proteins, G-protein coupled receptors (GPCRs).

METHODS

We first recap our model for site substitutions, as described elsewhere (5, 7, 8). We first consider that there are a number of different site classes, which characterize locations in the protein under similar selective pressure. As described above, the model does not assign locations to site classes; instead we define an unknown prior probability $P(k)$, that any given location belongs to site class k . As all locations must belong to a site class, $\sum_k P(k) = 1$.

We need to reduce the number of adjustable parameters that characterize each particular substitution matrix. In order to do this, we first consider that there is a relative fitness $F_k(A_i)$ of amino acid A_i for any location described by a particular site class k , related to the logarithm of the probability of finding such an amino acid at this location described by this site class. The instantaneous rate of substitution Q_{ij}^k from amino acid A_i to A_j at site class k is then reflected by the relative changes in fitness. In this paper, we use a few different models. In our analysis of a general set of globular proteins, we use so-called Metropolis kinetics, where advantageous substitutions ($\Delta F \equiv F_k(A_j) - F_k(A_i) \geq 0$) are accepted at a maximum site-class dependent rate v_k , while disadvantageous substitutions ($\Delta F < 0$) are accepted with a probability that decreases exponentially with the resulting change in fitness.

$$Q_{ij}^k = \begin{cases} v_k e^{\Delta F} & \left| \begin{array}{l} \Delta F < 0 \\ \Delta F \geq 0 \end{array} \right. \\ v_k & \end{cases} \quad (1)$$

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

In addition, as we were most interested in modeling the general nature of the selective pressure at different locations, we further parameterized the fitness of each amino acid at a given site class as a function of the physical-chemical properties of the amino acids:

$$F_k(A_i) = \sum_l \alpha_k^l (q_l(A_i) - \phi_k^l)^2 \quad (2)$$

where $q_l(A_i)$ represents the value of physical-chemical property l of amino acid A_i , and α_k^l and ϕ_k^l represent site-class specific adjustable parameters. In this study, we used the four

orthogonal property indices developed by Scheraga and coworkers (9). The first property is positively correlated with turn propensity and negatively correlated with α -helix propensity; the second is positively correlated with size and bulk, the third is positively correlated with β -sheet propensity, and the last is negatively correlated with hydrophobicity, meaning hydrophilic residues have high positive values in this index.

For the analysis of the G-Protein Coupled Receptors, we used a more general function of the form:

$$Q_{ij}^k = v_k e^{-\lambda_k (\Delta F)^2} \frac{\beta_k e^{\Delta F / 2}}{\beta_k e^{|\Delta F / 2|} + e^{-|\Delta F / 2|}} \quad (3)$$

where v_k again characterizes the overall substitution rate for site class k , and λ_k and β_k are parameters of the function. Note that this model is equivalent to the Metropolis scheme under the conditions $\lambda_k = 0$ and $\beta_k \gg 1$. In contrast to the case for the general set of globular proteins, we left the values of $F_k(A_j)$ as independently adjustable parameters.

To determine the substitution matrix M , representing the possible substitutions from amino acid A_i to A_j for any particular amount of evolutionary time t , the Q matrix is exponentiated:

$$M_{(k)}(t) = e^{tQ_{(k)}} \quad (4)$$

The model involves a large number of adjustable parameters. We will notate the parameters for site class k , including the prior probability $P(k)$, as $\{\theta\}_k$. For the study of the large set of globular proteins, this includes $P(k)$, v_k , and α^l_k and ϕ^l_k for the four different physical-chemistry parameters (the values of $q_l(A_i)$ for the twenty amino acids are measured, not adjustable, parameters). For the GPCR study, these parameters include $P(k)$, v_k , λ_k and β_k , and the twenty fitness parameters of the amino acids $F_k(A_i)$ (as the fitness values are relative, one of these parameters is set to zero). We will notate the entire set of all parameters, including the parameters for all of the site classes, as Θ .

These parameters are adjusted in order to maximize the log likelihood, that is, the log of the conditional probability that the observed data would result if the model were correct. At each location l , we first calculate the probability $P(D_l | \theta_k, T)$ of the observed amino acids at that

location, D_l , resulting from the evolutionary dynamics if the location was assigned to site class k with model parameters θ_k . Given the evolutionary tree topology and branch lengths T . Since each location can be represented by any of the site classes and each site class has distinct parameters θ_k we have to sum over all possible site classes to calculate the total likelihood for that location, L_l :

$$L_l = \sum_k P(D_l | \theta_k, T) P(k) \quad (5)$$

The log-likelihood for the entire set of proteins is calculated as the sum of the log of this likelihood at each location in the alignment.

While we do not know to which site class a location belongs *a priori*, following optimization of the model we can calculate *a posteriori* probabilities. The conditional probability that a location l belongs to site class k is given by:

$$P(k | D_l) = \frac{P(D_l | \theta_k) P(k)}{\sum_k P(D_l | \theta_k) P(k)} \quad (6)$$

DATASETS

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 (10), all with 6 to 11 homologs of 30% or greater sequence identity listed in the HSSP database (11). The average number of homologs for each protein was 10.5. A multiple alignment and phylogenetic tree was created for each set using the program ClustalV (12). The sequence, structure, and surface accessibilities were found by use of the DSSP program on the corresponding PDB files (13, 14). Residues were considered exposed if greater than 18% of their surface area was exposed to solvent.

Models with two site classes were optimized where $F_k(A_i)$ was a function of all four of Scheraga's orthogonal indices. Separate analyses were performed for buried and exposed residues. 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.

For the GPCR project, we selected a group of 185 amine-binding proteins, obtaining the multiple alignment from GPCRdb (15). We used PHYLIP (16), which uses a parsimony approach to calculate the best tree from a given set of data. Resulting trees were optimized for

their branch lengths using PAML (17). A model consisting of 5 site classes was optimized. We then calculated the posterior probabilities of the site classes for each location. In order to interpret the selective pressure described by each site class, we calculated the correlation coefficient between the fitness values and physicochemical properties of amino acids. These properties were derived from the AAindex database (18), which contains 434 different amino acid indices. We avoided indices related to spectroscopic methods and selected 145 physicochemical indices (see supplement for AAindex database codes of the used indices). Solvent accessibility calculations for rhodopsin were done using the publicly available software GETAREA 1.1 (19).

RESULTS

General properties for globular proteins

Representations for the selective constraints on exposed locations is shown in Figure 2. The optimization resulted in two distinct site classes, one site class representing the majority of sites (represented by the relative size of the pie charts for the two site classes), with a faster rate of variation (larger v_k), with the fitness of the amino acids primarily determined, unsurprisingly, by the hydrophilicity. In addition, there was a preference for small residues as well as a slight preference for residues with high turn propensity. The less common site class, conversely, had a slower rate of variation (smaller v_k), and had a strong preference for hydrophobic residues.

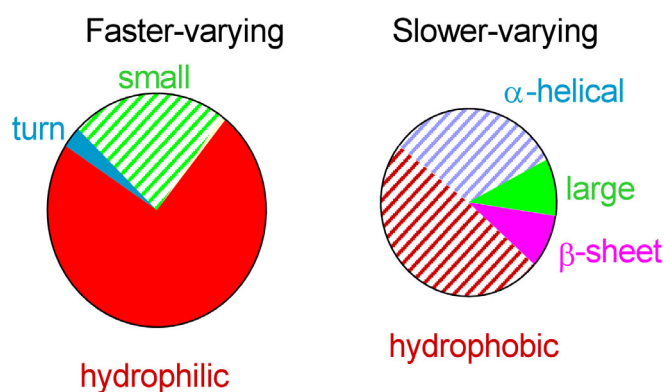


Figure 2. Pie charts representing the various contributions to the selective pressures acting on surface locations belonging to the two site classes. The relative sizes of the charts represents the percentage of the surface locations assigned to these classes. The color scheme represents the various Scheraga factors, including hydrophilicity (red), α -helix or turn propensity (blue), bulk (green), and β -sheet propensity (magenta). Solid colors represent a positive correlation with the Scheraga factor, while a striped pattern represents a negative correlation.

Helical propensity was also important, with smaller preferences for bulkier residues and residues with a larger sheet propensity.

The situation for buried residues is portrayed in Figure 3. The faster-varying locations actually occupied a minority of the buried locations, and predictably preferred hydrophobic residues, although this preference was less strong than the tendency for faster-varying exposed locations for hydrophilic residues. Equally strong was a tendency for residues with propensity for β sheets, as well as a moderate preference for residues with α -helical preferences, as well as large residues. The larger group of locations were in a slower-varying site class with a strong tendency towards small residues, and smaller preferences for hydrophilicity, turn propensity, and β -sheet propensity.

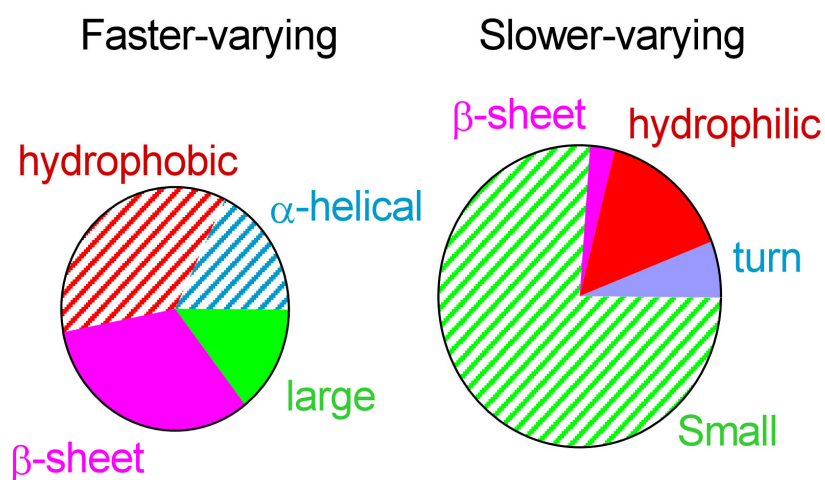


Figure 3. Pie charts representing the various contributions to the selective pressures acting on buried locations belonging to the two site classes. The relative sizes of the charts represents the percentage of the surface locations assigned to these classes. The color scheme is as for Figure 2.

G-Protein Coupled Receptors

The various parameters for the five site-class model obtained for the amine-binding GPCRs are listed in Table 1.

By mapping the locations in the amine-binding proteins to the known structure of Bovine bacteriorhodopsin, we can identify which locations are assigned to different site classes. Locations which are likely to reside in the membrane are largely assigned to site classes 1, 2, and 3, while loop locations are almost entirely assigned to classes 4 and 5. In addition, transmembrane locations in site classes 3 and 4 are generally in areas exposed to the membrane, while locations in site classes 1 and 2 generally face into the interior of the protein.

Table 1. Overall substitution rates and properties of preferred amino acids for the five site-class model optimized on the set of amine-binding GPCRs

site class (<i>k</i>)	Rate	Preferences
1	0.01	alpha-helical propensity
2	0.12	Hydrophobic
3	0.41	Hydrophobic, membrane
4	0.97	Flexible, buried
5	2.64	Polarizable

DISCUSSION

Both the buried and exposed locations can be divided into two different site classes, a faster-varying set of locations and a slower-varying set. It is not surprising that the faster varying locations on the exterior prefer hydrophilic residues, while the faster-varying locations on the interior prefer hydrophobic. It is surprising, however, that for the slower-varying sites in both contexts these preferences are reversed. It is likely that the faster-varying sites are under less purifying selective pressure than the sites that vary more slowly. While most locations in the inside would be under some selective pressure to remain hydrophobic, the other, more specialized forms of selective pressure acting on some locations may favor conservation of such things as particular hydrogen bond or ionic bond formations. In these locations, this specialized selective pressure may well favor hydrophilic residues in a way that “trumps” the more general forms of selective pressure felt by more average locations in the protein. These locations would have slower substitution rates as well as more complex forms of the selective pressure. Similarly, these locations may be involved in specific packing or aromatic stacking interactions, so the preference for larger residues might be explainable. Conversely, the locations on the protein exterior that change slowly might be under more specific forms of selective pressure that prefer hydrophobic residues. In both instances, the needs of stabilizing a specific conformation may result in a tendency for specific locations to have selective pressure opposite in form to that of other, seemingly similar locations. One interesting point to note is that, for buried locations, the majority of sites are slower varying. Another observation is that the dominant hydrophobic selective pressure is on faster-varying external locations to remain hydrophilic. This provides further evidence for the reverse hydrophobic effect, that is, the need to avoid stabilizing alternative conformations where these particular locations are buried (20).

The analysis of the GPCRs demonstrate that we can obtain specific structural information from sets of aligned sequences, even identifying trans-membrane residues facing into the protein

interior or out into the lipid membrane. All this information is gathered as a result of the optimization procedure, with no *a priori* knowledge about structure or function. As such, it is a powerful way of generating important information about the new proteins whose sequences are becoming available.

ACKNOWLEDGMENTS

This contribution represents the work of a number of different investigators in my lab, including Jeffrey Koshi, Darin Taverna, Matthew Dimmic, and Orkun Soyer, as well as a collaborator, Richard Neubig. Financial support was provided by NIH Grants GM08270 and LM0577, NSF equipment grant BIR9512955, and a grant from the University of Michigan Program in Bioinformatics.

REFERENCES

- [1] Wako, H. & Blundell, T. (1994). 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.
- [2] Wako, H. & Blundell, T. (1994). 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.
- [3] Koshi, J. M., & Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein Engineering* **8**:641-645.
- [4] Goldman, N., Thorne, J. L., Jones, D. T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Evol.* **263**:196-208.
- [5] Koshi, J. M. & Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins* **32**:289-295.
- [6] Koshi, J. M. & Goldstein, R. A. (2001). Analyzing site heterogeneity during protein evolution. Pacific Symposium on Biocomputing. **6**:191-202.
- [7] Dimmic, M. W., & Goldstein, R. A. (2000). Modeling evolution at the amino acid level using a general fitness model, in Pacific Symposium on Biocomputing 2000, (Altman, Dunker, Hunger, Lauderdale, and Klein, eds), World Scientific, Singapore, pps. 18-29.
- [8] Soyer, O., Dimmic, M. W., Neubig, R. R., Goldstein, R. A., Modeling protein evolution with applications to the understanding of G-Protein Coupled Receptors. *Proteins*, submitted.

- [9] Kidera, A., Konishi, Y., Oka, M., Ooi, T., Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Prot. Chem.* **4**:23-55.
- [10] Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**:522-524.
- [11] Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**:56-68.
- [12] Higgins, D. G., Bleasby, A. J., Fuchs, R. (1992). Clustal V: Improved software for multiple sequence alignment. *CABIOS* **8**:189-191.
- [13] Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopoly.* **22**:2577-2637.
- [14] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. (1997). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**: 535-542.
- [15] Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F., Vriend, G. (1998). GPCRDB: an information system for G protein coupled receptors. *Nucl. Acid Res.* **26**:275-279.
- [16] Felsenstein, J. (1993). PHYLIP - Phylogeny Inference Package. *Cladistics* **5**:164-66.
- [17] Yang, Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J. Molec. Evol.* **39**:306-314.
- [18] Kawashima, S., Ogata, H., Kanehisa, M. (1999). AAindex: Amino acid index database. *Nucleic Acid Res.* **27**:368-369.
- [19] Fraczekiewicz, R. & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comp. Chem.* **19**:319-333.
- [20] Koshi, J. M. & Goldstein, R. A. (1997). Mutation matrices: correlations and implications. *Proteins* **27**:336-344.

