CORRELATING STRUCTURE-DEPENDENT MUTATION MATRICES WITH PHYSICAL-CHEMICAL PROPERTIES

J. M. KOSHI* and R. A. GOLDSTEIN*†
**Biophysics Research Division, †Department of Chemistry*
University of Michigan, Ann Arbor, MI 48109-1055, USA

We have investigated how structure-dependent mutation matrices derived in previous work correlate with various physical-chemical properties of the 20 naturally occurring amino acids. Among the properties we investigated were ΔG of transfer from water to octanol and cyclohexane, alpha helical and beta sheet propensity, size, and charge. We found that the ΔG of transfer to octanol had a high correlation with matrices for all categories of residues, especially the matrices for buried and exposed positions. This result suggests that octanol is a good model for understanding both the changes in stability resulting from substitutions of buried residues and changes in foldability resulting from varying exposed residues. We also found the correlations of the matrices with size and charge varied with the local environment, and that neither alpha helical nor beta sheet propensity had high correlations with most matrices. Thus, conservation of size and charge appear to be important in specific environments, and conservation of alpha helix and beta sheet propensity do not seem to be key factors.

1 Introduction

Current biochemical techniques have gathered a wide variety of information about the 20 amino acids commonly found in proteins. Parameters such as size and charge have been known for many years. More recently ΔG of transfer from water to solvents such as octanol and cyclohexane[1,2] and alpha helical and beta sheet propensity have been measured [3–5]. We still do not know, however, exactly how important each of these quantities are in determining the fitness of one amino acid over another in various local environments of protein structures, or more fundamentally, what attribute of the protein - foldability, stability, catalytic efficiency, etc. - is driving this local optimization [6,7]. Complicating this fact is that residues with few constraints may be evolving by random fixation of neutral or nearly neutral mutations [8,9].

Until recently, the most prevalent approach used to study the effects of specific mutations has been the creation of site mutations in various proteins, examining how alteration of one of the amino acids changes a specific property of the protein, such as stability or biochemical efficiency[11–17]. Such studies, however, are fraught with pitfalls. First, one must be careful to look only at mutations which do not significantly alter the native protein structure, preferably as verified by NMR or crystallography. It is also difficult to insure that

the property of interest is the only relevant one being altered, and that other factors, such as steric clashes, are not being introduced.

In order to reduce extraneous factors, several groups have looked at correlations between the change in ΔG of transfer from water to various solvents and mutations within a subset of amino acids - mainly the hydrophobic residues. The generality of these results is limited, however, as the number of mutations catalogued is only a small subset of all possible mutations. For instance, a study by Pielak *et al.* [10], analyzed data from 87 mutant proteins, but these mutants represented only 11 of the 380 possible amino acid substitutions. If one also wanted to differentiate between mutations in different secondary structures and surface accessibilities, the available data set becomes even more sparse. Thus, one must take care in extrapolating the conclusions found in studies like that of Pielak *et al.* beyond the specific questions they were addressing.

The limited number of site mutations that can be attempted and analyzed in the laboratory may also have significant affects on the results observed. Often, proteins can adapt to particular mutations, either through readjustments in the protein conformation, or through compensatory mutations [12,13,18,19]. With time, natural evolution may be able to opportunistically take advantage of compensable site mutations that the necessarily anecdotal approach of directed site mutagenesis might miss. In addition, directed site mutations are chosen based on some *a priori* biochemical intuition, and other possible interesting mutations may not be even attempted. Finally, it is not obvious how the changed behavior of the mutations will affect the proteins' performance in an *in vivo* context.

In contrast to the site mutation approach, we use the database provided by natural evolution. Such an approach must contend with all the random noise inherent in such a system, but it also has several advantages. First, we are not restricted to looking at only a single quantity like stability or function in our experiment: many parameters, like function, stability, or foldability, may constrain evolutionary changes, and our approach will encompass all of them. The database created through evolution is also vast, and we are certain that all our data is viable in real, *in vivo* systems.

Specifically, we incorporate evolutionary information with the use of optimal, structure-dependent mutations matrices derived in earlier work [20]. These matrices were created using phylogenetic trees by optimizing the probability that a given matrix fit the evolutionary transitions observed. And since the secondary structure of the proteins in the data set was known, we were able to create specific data sets for each local structure and derive mutation matrices optimized for each local environment [20]. In addition to enabling us to include

information found in local structure, use of these matrices allows us to look at all possible transitions. Thus, by using these structure-dependent mutation matrices, we hope to better understand the relationship between local structure and the constraints evolution places on mutations. To this end we have studied the applicability of various parameters, such as the ΔG of transfer from water to octanol and cyclohexane, size, charge, and secondary structure propensity, to model the tendencies seen in the mutation matrices for each local structure.

2 Theory

The goal of this study is to correlate changes in various parameters with our mutation matrices. To achieve this, it is necessary to define exactly how to compute the "distance" between the fitness of two amino acids, as measured by our mutation matrices, and how that value will correlate with changes in various physical-chemical parameters. In order to uncover these correlations, we have investigated two theoretical models for evolutionary transitions, one simple, and one more complex.

In the simple case, we need to make several major assumptions. The first is that the mutational matrix distance between two amino acids is only correlated with the parameter of interest, or alternatively that the effects of the other parameters average out to zero. The second is the use of Metropolis algorithm to describe how evolutionary transitions occur. The third, and largest assumption is that the fitness function, $f(a_i)$, where a_i corresponds to one of the 20 amino acids, is a simple linear function of the parameter ($f(a_i) = mq_i + p$, where q_i is the parameter value associated with amino acid a_i), and is the same for all possible residue positions where we find that mutation. We recognize the simplicity of this model, but using our structure-dependent mutation matrices, this may be a plausible model, as all residues of a particular type have similar behavior in each local environment . We can show that the parameter difference should be correlated with $\log(M_{a_1 a_2} M_{a_2 a_1})$, where $M_{a_1 a_2}$ corresponds to the mutation matrix entry for a transition from amino acid a_1, with parameter value q_1 to amino acid a_2, with parameter value q_2. Assuming $f(a_1) > f(a_2)$, the Metropolis algorithm states:

$$P(a_1 \rightarrow a_2) = \kappa P(a_1) \exp\left(-m|\Delta q|/d\right) \tag{1}$$

where $P(a_1 \rightarrow a_2)$ is the probability of a transition from an amino acid a_1 with parameter value q_1 to an amino acid a_2 with parameter value q_2, κ is the intrinsic rate of mutations, $P(a_1)$ is the probability of finding amino acid a_1, $|\Delta q|$ is the absolute value of the difference in parameter value between q_1

and q_2, and d is some "mutational temperature". $P(a_2 \rightarrow a_1)$ is simply equal to $\kappa P(a_2)$, as this mutation results in an increase in fitness, and, given the chance, will always occur. Thus,

$$P(a_1 \rightarrow a_2)P(a_2 \rightarrow a_1) = \kappa^2 P(a_1)P(a_2) \exp\left(-m|\Delta q|/d\right) \tag{2}$$

$$|\Delta q| = \frac{-d}{m} \log\left(\frac{P(a_1 \rightarrow a_2)P(a_2 \rightarrow a_1)}{\kappa^2 P(a_1)P(a_2)}\right) \tag{3}$$

But since by definition:

$$M_{a_1 a_2} \equiv \frac{P(a_1 \rightarrow a_2)}{P(a_1)} \tag{4}$$

This implies:

$$|\Delta q| = \frac{-d}{m}\left(\log(M_{a_1 a_2} M_{a_2 a_1}) - 2\log\kappa\right) \tag{5}$$

The model just described made the assumption that all residue positions under consideration had the same linear fitness function. While this may be a reasonable first approximation, it is probably a bad model for parameters like size, where mutations to overly large or small residues will both be destabilizing. The optimal value of the parameter, q_{opt}, will vary with different residue positions, and we must explicitly consider a distribution of q_{opt} values.

To this end, we chose a Gaussian dependence of q for the fitness function, and for the distribution of q_{opt} values. With these choices, we can show that the parameter difference, $|\Delta q|$, should correlate with the quantity $\log(M_{a_1 a_2} M_{a_2 a_1}/[P(a_1)P(a_2)]^K)$, where K ranges from 1 to 2, as follows:

For an individual position within the data set we wish to know the probability that we will find a transition from amino acid a_1 to a_2, with a parameter change of q_1 to q_2. We model this probability, $P(a_1 \rightarrow a_2)$, as proportional to both the intrinsic mutation rate κ, and to the probabilities of finding amino acids a_1 and a_2: $P(a_1 \rightarrow a_2) = \kappa P(a_1)P(a_2)$. Now, averaging over all residue positions, and assuming the intrinsic mutation rate is constant over all sites, we get:

$$\overline{P(a_1 \rightarrow a_2)} = \kappa \overline{P(a_1)P(a_2)} \tag{6}$$

Using the definition of $M_{a_1 a_2}$ in Eqn. 4,

$$M_{a_1 a_2} = \kappa \frac{\int P(q_1|q_{opt})P(q_2|q_{opt})P(q_{opt})d_{q_{opt}}}{\int P(q_1|q_{opt})P(q_{opt})d_{q_{opt}}} \tag{7}$$

where we are explicitly modeling $P(a_1)$ and $P(a_2)$ as functions of q_{opt} by writing them as conditional probabilities: $P(q_1|q_{opt})$ is the probability of finding an amino acid with parameter value q_1, given that the optimum point of the distribution is q_{opt}. Note that q_{opt} is being allowed to vary over residue positions (thus the integral over q_{opt}). Considering the numerator of this fraction:

$$\int P(q_1|q_{opt})P(q_2|q_{opt})P(q_{opt})d_{q_{opt}}$$

$$= \int d_{q_{opt}} \exp\left(\frac{-(q_1-q_{opt})^2}{2\Gamma^2} - \frac{-(q_2-q_{opt})^2}{2\Gamma^2} - \frac{q_{opt}^2}{2\gamma^2}\right) \quad (8)$$

$$= C\exp\left(\frac{-\Gamma^2(q_1^2+q_2^2)-\gamma^2(q_1-q_2)^2}{2\Gamma^2(\Gamma^2+2\gamma^2)}\right) \quad (9)$$

where we have explicitly written $P(q_1|q_{opt})$, $P(q_2|q_{opt})$, and $P(q_{opt})$ as Gaussian probability distributions, and C is simply a constant in the equation, with a value of: $C = \Gamma(\gamma\sqrt{\pi})/(\Gamma^2+2\gamma^2)^{1/2}$ The denominator of the fraction from equation 7 is:

$$\overline{P(a_1)} = C'\exp\left(\frac{-q_1^2}{2(\Gamma^2+\gamma^2)}\right) \quad (10)$$

where C' is equal to : $C' = (\Gamma\gamma\sqrt{(2\pi)})/(\Gamma^2+\gamma^2)^{1/2}$ The product of the two transition probabilities is then:

$$M_{a_1a_2}M_{a_2a_1} = \kappa^2\left(\overline{P(a_1)P(a_2)}\right)^2 / \overline{P(a_1)}\ \overline{P(a_2)} \quad (11)$$

This implies:

$$\Delta q = C''\log\left(\frac{M_{a_1a_2}M_{a_2a_1}}{(P(a_1)P(a_2))^K} - 2\log\kappa\right) \quad (12)$$

where $C'' = \frac{C^2}{C'^2}$, and,

$$K = \frac{2(1+\lambda^2)}{1+2\lambda^2} \quad (13)$$

if λ is defined as: $\lambda = \frac{\gamma^2}{\Gamma^2}$. As λ can vary from 0 to ∞, this implies K can vary from 1 to 2.

Thus, if the Gaussian distribution for q_{opt} is broader (as reflected in a larger value of γ) than that of q, the above equations predict a K value near 1. On the other hand, if q has a broader distribution, then one would expect a K value of 2.

3 Methods and Results

As discussed above, we have used in this study optimal structure dependent mutation matrices, which were derived in earlier work [20]. These matrices were derived using a data set of 82 proteins, using phylogenetic trees, and employing a Bayesian methodology to optimize the probability that a given matrix fit all the mutations seen within a given data set. Some of the data sets used were also secondary structure and surface accessibility specific, allowing us to generate optimal mutation matrices for all 15 separate cases: 8 matrices, one for each secondary structure/surface accessibility class; 4 matrices, one for each secondary structure regardless of surface accessibility; 2, one for exposed and one for buried residues, and 1 for all residues. We have correlated these 15 matrices and the Dayhoff PAM250 matrix [21], normalized to a similar evolutionary distance, with various physical-chemical parameters: ΔG of transfer from water to octanol and cyclohexane [2], alpha helical and beta sheet propensities [3], size, and charge. In order to correlate these quantities, we considered the absolute difference of the values of a given parameter for two amino acids *vs.* the value of $\log(M_{ab}M_{ba}/(P(a_1)P(a_2))^K)$ with K ranging from 0 to 10, as described in the theory section.

In addition to the correlation coefficients, for each of the cases listed above, we also calculated the probability that a random, uncorrelated sample would give that correlation coefficient or higher. As the number of data points differs for each case, it is this probability which is actually the more important value for determining which correlations are significant. (*i.e.* for a large data set, a smaller correlation coefficient is often more unlikely than a substantially higher correlation coefficient for a smaller data set.)

The optimal K values (k_{opt}), correlation coefficients (r), and probabilities of a random distribution matching or exceeding that correlation coefficient (P_r), are shown for some cases in Table 1. Figures 1a through 1f shows the actual distribution of data points for various parameters correlated against our matrix for all residues. For each plot, the best fit line is plotted.

4 Discussion

Using structure-dependent mutation matrices, the best correlation coefficients found against the structure-dependent matrices are higher than those against the Dayhoff matrix in all cases, usually by a factor of 1.5 or more. Clearly, simply gathering statistics on pairs of highly similar sequences does not encompass the greater complexities seen in evolution.

When comparing our results to those of the study by Pielak *et al.* we find

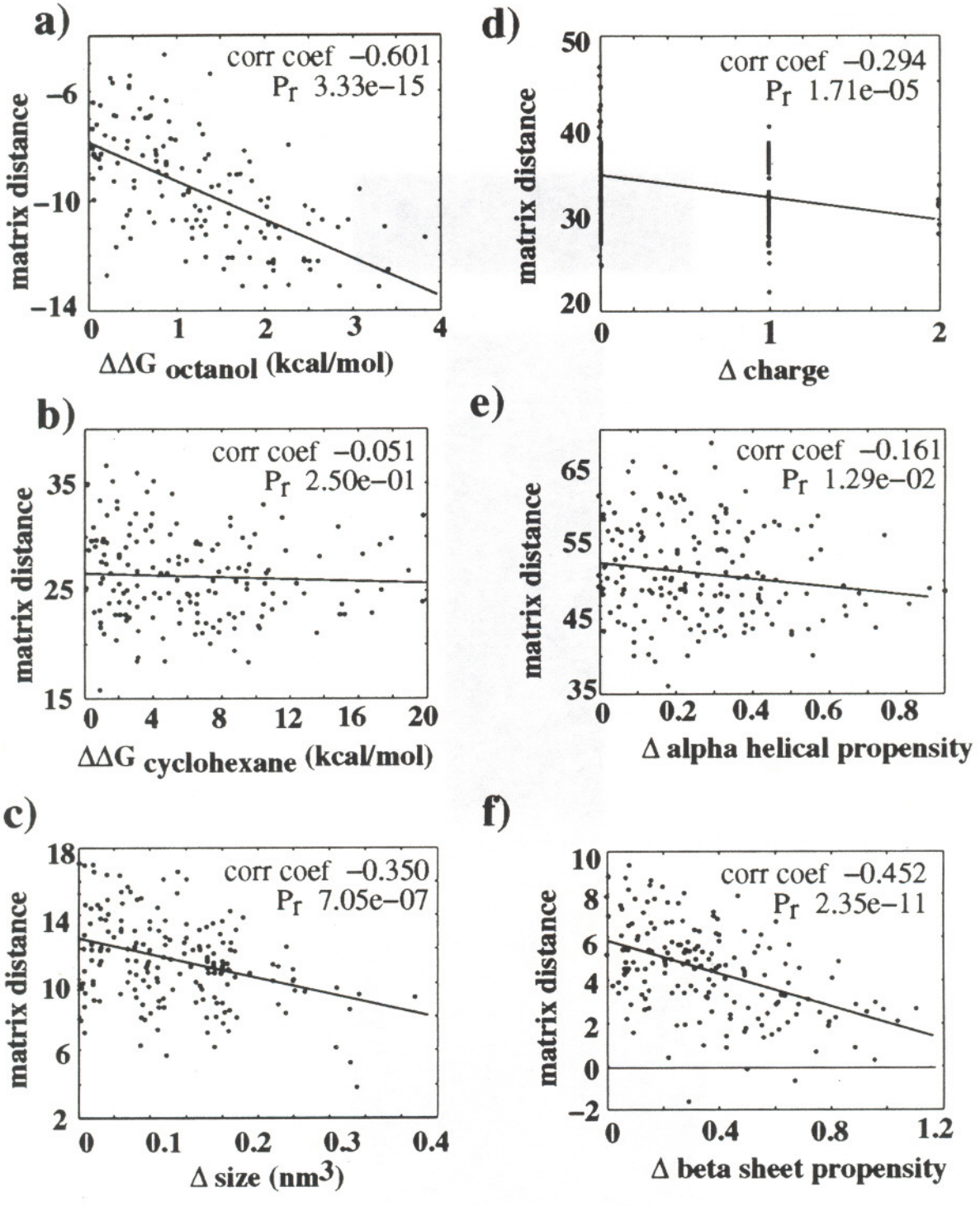

Figure 1: a: Scatter plot of the $\Delta\Delta G$ of transfer to octanol[2] *vs.* the matrix for all residues, with the best fit line drawn. Matrix distances are in arbitrary units. The correlation coefficient, r, and probability that r arises from uncorrelated data, P_r, are also given. b: same plot for the $\Delta\Delta G$ of transfer to cyclohexane[2]. c: for Δ size. d: for Δ charge. e: for Δ alpha helical propensity[3]. f: for Δ beta sheet propensity[3].

	ΔΔG octanol			ΔΔG cyclohexane			Δ size			Δ charge		
matrix	K opt	r	P_r	K opt	r	P_r	K opt	r	P_r	K opt	r	P_r
exposed alpha helix	0.6	-0.497	2.84e-10	4.8	-0.221	1.76e-03	1.3	-0.185	5.09e-03	4.5	-0.486	4.28e-13
beta sheet	0.5	-0.466	4.14e-09	4.5	-0.183	7.85e-03	2.2	-0.240	4.06e-04	5.8	-0.449	3.13e-11
turn	0.1	-0.588	1.71e-14	3.9	-0.205	3.35e-03	3.8	-0.295	1.63e-05	5.5	-0.511	1.77e-14
coil	0.1	-0.536	6.53e-12	5.8	-0.192	5.68e-03	2.4	-0.254	1.92e-04	5.9	-0.401	4.21e-09
buried alpha helix	0.3	-0.496	2.96e-10	9.4	0.085	1.32e-01	1.1	-0.110	6.48e-02	9.9	0.211	1.63e-03
beta sheet	0.0	-0.479	1.43e-09	9.9	0.228	1.26e-03	0.7	-0.247	2.80e-04	9.9	0.341	6.52e-07
turn	0.0	-0.480	1.23e-09	9.9	0.092	1.13e-01	3.4	-0.309	6.37e-06	0.0	-0.113	5.94e-02
coil	0.0	-0.522	2.70e-11	9.9	0.110	7.45e-02	1.9	-0.168	9.85e-03	9.9	0.231	6.22e-04
alpha helix	0.9	-0.574	8.94e-14	3.1	-0.108	7.87e-02	2.2	-0.193	3.63e-03	6.0	-0.266	9.83e-05
beta sheet	0.0	-0.552	1.10e-12	9.9	0.119	6.01e-02	2.3	-0.281	3.98e-05	0.0	-0.084	1.23e-01
turn	0.0	-0.563	3.28e-13	2.8	-0.169	1.32e-02	3.8	-0.377	3.61e-08	7.7	-0.447	3.89e-11
coil	0.0	-0.583	3.34e-14	4.3	-0.092	1.14e-01	2.9	-0.286	2.87e-05	4.9	-0.238	4.32e-04
exposed	0.2	-0.628	7.92e-17	4.1	-0.213	2.45e-03	2.5	-0.303	9.44e-06	5.3	-0.523	3.78e-15
buried	0.0	-0.536	5.93e-12	3.1	-0.129	4.53e-02	3.3	-0.350	3.29e-07	4.9	-0.393	8.94e-09
all residues	0.0	-0.601	3.33e-15	5.8	-0.051	2.50e-01	3.4	-0.340	7.05e-07	7.0	-0.294	1.71e-05
Dayhoff PAM	0.0	-0.404	4.61e-07	9.7	-0.043	2.85e-01	9.9	-0.230	6.78e-04	9.9	-0.227	7.64e-04

Table 1: Optimal K values, correlation coefficients(r), and probability that a similar distribution could arise from uncorrelated data (P_r) are given for the $\Delta\Delta G$ of transfer to octanol and cyclohexane, Δ size, and Δ charge *vs.* the various mutation matrices listed above.

several similarities. Pielak *et al.* found the $\Delta\Delta G$ of transfer to octanol to be a good indication of changes in stability for most amino acid substitutions in the protein core [10], and correspondingly we find it has a high correlation against our matrix for buried residues. The most likely explanations for this high correlation is that the ΔG of octanol serves as a good model for moving residues from the aqueous environment to the hydrophobic core during folding. This interpretation is supported by Pielak's observation that mutation matrices are highly correlated with changes in stability for these substitutions [10].

One surprising result was that the high correlation observed between the ΔG of octanol and the mutation matrix for exposed residues. As the environment of these residues remains very similar during folding, this correlation can not be explained by the need to stabilize the folded conformation. If we look at the "reverse hydrophobic effect", however, we can find a likely explanation for this correlation [22–24]. The ΔG of octanol correlates well with the exposed matrix because it is not the stability, but the foldability of the protein that is being optimized. One of the major factors in efficient folding of the protein is the destabilization of incorrect conformations; the polar nature of surface residues prevents stabilization of states in which these residues are buried. Thus, it is the patterns of hydrophobicity which are most important in protein folding [25,26], and the correlation of the ΔG of octanol with our exposed matrix simply reflects this conservation of hydrophobicity.

Another surprising fact was that we did not observe any significant correlation between the ΔG of cyclohexane and any of our matrices, even for buried residues. The result of Pielak *et al.*would suggest that the ΔG of cyclohexane should have been a good model for the protein core, but the correlations we observe do not support that conclusion. We even tried correlating the ΔG of

transfer to cyclohexane with *only* the hydrocarbon amino acids, as suggested by the results of Pielak *et al.*, but found no significant correlations with any of our matrices. Specifically, none of the 16 correlations against our matrices and the Dayhoff matrix had a P_r value of less then 0.06, and only 2 were under 0.1. This result suggests that the ΔG of cyclohexane, which can not form hydrogen-bonds, might be a good model only for artificial site mutations, where nature is less able to adjust the surrounding structure to fit the mutation. In evolution, mutations are better modeled by the ΔG of hydrogen-bond forming octanol, because naturally occurring mutations are likely to be those which can take advantage of possible hydrogen-bond partners.

The optimal K values for most matrices for octanol were low, many of them zero. This suggests that the simple linear model might be the best fit for octanol. An interesting exception to this, however, is found in the matrices for alpha helices. The alpha helix matrix had an optimal K value of $K_{opt} = 0.9$, very close to the lower limit of 1 set by the Gaussian model. The optimal K value for exposed alpha helices was also higher than the rest ($K_{opt} = 0.6$). Thus, for alpha helices, in which factors such as patterns of hydrophobicity are important, the simple linear model is not sufficient, and more complicated models such as the the Gaussian model may be more appropriate.

In addition to the ΔG of transfer to octanol and cyclohexane, we also examined correlations of our matrices against alpha helical and beta sheet propensity. Pielak *et al.* found little correlation of alpha helical propensity with changes in stability, and in a similar fashion, we find little correlation of alpha helical propensity with the various mutation matrices. (The best P_r values for correlations with Δ alpha helical propensity were on the order of 10^{-3}). This tells us that helical propensity is not an important factor in deciding what mutations are allowed. The negligible correlation of the mutation matrices with alpha helical propensity agrees with the results of Chakrabartty *et al.* for instance, that show only alanine is a helix former - leucine and arginine are indifferent, and all others are helix-breakers[27]. It has also been found that hydrophobic interactions are more important for changes in stability than alpha helical propensity[25,28,29], and that patterns of hydrophobicity are sufficient to induce helix formation[26]. Thus, these results suggest that it is really factors such as patterns of hydrophobicity which matter in the formation of alpha helices.

Beta sheet propensity showed a higher correlation, with the highest correlation against the matrix for beta-sheet residues (r = -0.488, P_r = 3.44e-13). This higher correlation is not simply a dependence on physical-chemical properties such as size or hydrophobicity, as we found no correlation between beta sheet propensity and these characteristics. We also noted that buried beta

sheets had a higher correlation than exposed beta sheets (r = -0.386 *vs.* r = -0.306). This fact suggests that patterns of hydrophobicity are less important in sheets than in other secondary structures, possibly because beta sheets are more likely to be buried.

The correlations of size with our matrices showed some very specific results. We found the best correlation was with the turn matrix, and the second best with the buried matrix. Those correlations, along with the correlation with the buried turn matrix, were all significantly higher than the average over all matrices. This is not surprising, as size is a very important factor in turns, where steric clashes are the major concern, and in buried positions, where internal packing plays an important role. In the case of charge, the highest correlation was with the exposed matrix, and the second highest with the exposed turn matrix. The correlation with the exposed matrix is not surprising, as charged amino acids are less likely to be found in the interior of proteins, but the strong correlation with turns was not expected. As turns often separate secondary structure elements, perhaps charged residues play more of a role in stabilizing the ends of secondary structures than is thought.

For the physical-chemical parameters, other than transfer free energies, the simple linear model does not fit the data well. In the case of size and charge, the optimal K values for all significant correlations were above 2. For size, all significant correlations fell between 2 and 4, and for charge, optimal K values were between 4 and 6. These optimal K values suggest that the Gaussian model is more nearly correct - *i.e.* the optimal parameter value is more nearly a central q_{opt}, than an extreme q_{opt} value, or that the fitness varies over residue positions, or some combination of both. Clearly the Gaussian model has some limitations, as theory predicts a maximum value of 2 for K, but this model is likely more relevant than a simple linear fitness function. One conclusion we can draw from the large values of K is that the fitness function for size, $f(q)$, probably has a wide peak. This is not surprising, as one would not expect size to have a very sharply peaked fitness function - previous studies have shown that structure can adapt to a range of sizes [6,19,25,30].

5 Conclusion

There are several major conclusions which our data brings to light. One of these is that the ΔG of transfer from water to octanol is a very good model for evolutionarily constrained mutations. With the octanol model, we also found that while stability is probably the parameter being optimized for buried residues, it is likely the foldability which is being optimized in exposed positions. When using the ΔG of transfer from water to cyclohexane, however, we found it to

be a poor model even for only the hydrocarbon amino acids, a surprising result considering the conclusions of Pielak *et al.* A second conclusion is that alpha helical propensity seems to have little bearing on evolutionary constraints on mutation, while beta sheet propensity is slightly more important. Finally, size and charge are important, but their significance varies with local structure.

Acknowledgements

We would like to thank Kurt Hillig for computational assistance. Also, Michael Thompson and Gary Pielak deserve thanks for their thoughtful insights. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, and NIH Grants GM08270 and R29 LM0577.

1. J. Fauchere and V. Pliska. *European Journal of Medical Chemistry*, 18:369–375, 1983.
2. A. Radzicka and R. Wolfenden. *Biochemistry*, 27:1664–1670, 1988.
3. P. Y. Chou. In G. D. Fasman, editor, *Prediction of Protein Structure and the Principles of Protein Conformation*, pages 549–586. Plenum Press, New York, 1989.
4. T. P. Creamer and G. D. Rose. *Proteins*, 19:85–97, 1994.
5. D. L. Minor and P. S. Kim. *Nature*, 367:660–663, 1994.
6. B. W. Matthews. *Biochemistry*, 26:6885–6888, 1987.
7. B. K. Shoichet, W. A. Baase, R. Kuroki, and B. W. Matthews. *Proc. Nat. Acad. Sci., U.S.A.*, 92:452–456, 1995.
8. M. Kimura. *Nature (London)*, 217:624–626, 1968.
9. J. L. King and T. H. Jukes. *Science*, 164:788–798, 1969.
10. G. J. Pielak and D. S. Auld and J. R. Beasley and S. F. Betz and D. S. Cohen and D. F. Doyle and S. A. Finger and Z. L. Fredericks and S. Hilgen-Willis and A. J. Saunders and S. K. Trojak. *Biochemistry*, 34:3268–3276, 1995.
11. H. Nicholson, W. J. Becktel, and B. W. Matthews. *Nature (London)*, 336:651–656, 1988.
12. W. A. Lim, D. C. Farruggio, and R. T. Sauer. *Biochemistry*, 31:4324–4333, 1992.
13. J. H. Hurley, W. A. Baase, and B. W. Matthews. *J. Mol. Biol.*, 224:1143–1159, 1992.
14. H. W. Hellinga, R. Wynn, and F. M. Richards. *Biochemistry*, 31:11203–11209, 1992.
15. X.-J. Zhang, W. A. Baase, and B. W. Matthews. *Protein Science*, 1:761–776, 1992.

16. A. E. Eriksson, W. A. Baase, and B. W. Matthews. *J. Mol. Biol.*, 229:747–769, 1993.
17. W. A. Lim, A. Hodel, R. T. Sauer, and F. M. Richards. *Proc. Nat. Acad. Sci., U.S.A.*, 91:423–427, 1994.
18. Jr. J. T. Kellis, K. Nyberg, and A. R. Fersht. *Biochemistry*, 28:4914–4922, 1989.
19. W. A. Lim and R. T. Sauer. *J. Mol. Biol.*, 219:359–376, 1991.
20. J. M. Koshi and R. A. Goldstein. *Protein Engineering*, 8:641–645, 1995.
21. M. O. Dayhoff, R. M Schwartz, and B. C. Orcutt. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, suppl. 3, page 345. National Biomedical Research Foundation, Washington, D.C., 1978.
22. A. A. Pakula and R. T. Sauer. *Nature (London)*, 344:363–364, 1990.
23. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. *Proc. Nat. Acad. Sci., U.S.A.*, 89:9029–9033, 1992.
24. B. E. Bowler, K. May, T. Zaragoza, P. York, A. Dong, and W. S. Caughey. *Biochemistry*, 32:183–190, 1993.
25. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. *Protein Science*, 4:561–602, 1995.
26. H. Xiong, B. L. Buckwalter, H. Shieh, and M. H. Hecht. *Proc. Nat. Acad. Sci., U.S.A.*, 92:6349–6353, 1995.
27. A. Chakrabartty, T. Kortemme, and R. L. Baldwin. *Protein Science*, 3:843–852, 1994.
28. R. J. Pinker, L. Lin, G. D. Rose, and N. R. Kallenback. *Protein Science*, 2:1099–1105, 1993.
29. M. Blaber, X. J. Zhang, J. D. Lindstrom, S. D. Pepiot, W. A. Baase, and B. W. Matthews. *J. Mol. Biol.*, 235:600–624, 1994.
30. B. W. Matthews. *Curr. Op. Struct. Biol.*, 3:589–593, 1993.