Testing Hypotheses about Determinants of Protein Structure with High-Precision, High-Throughput Stability Measurements and Statistical Modeling†

Fang Yi,‡ Dorothy A. Sims,§ Gary J. Pielak,‡,|| and Marshall Hall Edgell*,‡,§

Departments of Biochemistry and Biophysics, Microbiology and Immunology, and Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599

Received January 13, 2003; Revised Manuscript Received April 1, 2003

ABSTRACT: Statistical modeling provides the mathematics to use data from large numbers of mutant proteins to generate information about hypotheses concerning protein structure not easily obtained from anecdotal studies on small numbers of mutants. Here we use the unfolding free energies of 303 unique eglin c mutant proteins obtained from high-precision, high-throughput chemical denaturation measurements to assess models concerning helix stability. A model with helix propensity as the sole determinant of stability accounts for 83% of the mutant-to-mutant variation in stability for 99% of the mutant proteins (three outliers). When position effects and side chain−side chain interactions are added to the model, the fraction of variation explained increases to 92%. The propensity parameters in this model are identical to helix propensity values derived from other approaches. Measurement error accounts for another 1% of the mutant-to-mutant variation in stability. While the data support terms for several of the expected stabilizing/destabilizing effects, it does not support terms for several others, including i, $i + 3$ effects in the center of the helix and helix−dipole effects. In addition, the model does better with terms for several stabilizing/destabilizing effects for which we cannot identify the physical basis. The precision of our unfolding stability measurements (±0.087 kcal/mol) allows us to conclude that the 7% of variation in stabilities of the mutant proteins not accounted for by the model or by measurement variation is both real and large with respect to the nonpropensity terms in the model. The analysis also shows that the common practice of using $C_m m_{av}$ instead of $C_m m_{mut}$ to calculate $\Delta G_{HOH,N-D}$ values for each mutant protein results in a loss of information. We see no correlation between the residuals derived from the full model and $m_{mut} - m_{wt}$, and hence it is unlikely our m_{mut} values reflect mutant-to-mutant differences in the denatured state.

One approach to deal with the "fold problem", that is, predicting tertiary structure from amino acid sequence, is to identify the various molecular forces that underlie the physical basis of conformational stability and then to determine the contribution to stability of each in the residue−residue interactions of a protein. Site-directed mutagenesis (*1*) has been used extensively to investigate the effects of individual amino acid mutations on the folded and unfolded state, the stability of proteins, and their binding activities (*2*−*7*). Effects of mutations on protein stability can be measured by the free energy change between the denatured state (or the state of the protein in high denaturant) and the native state, $\Delta G_{HOH,N-D}$. While site-directed mutagenesis is a powerful tool for decomposing stabilizing effects, the complexity of the interactions of single residues in proteins and the frequency of thermodynamic coupling means that an experimentally obtained $\Delta G_{HOH,N-D}$ value for a given amino acid change may not be derived solely from the energy effect of interest. This complexity implies that characterizing a few tens of mutants is at root anecdotal.

A common way to rise above anecdotal analyses and deal with complexity is to use a more formal approach for hypothesis testing, that is, statistical modeling or regression analysis. Statistical modeling decomposes the observable of interest, here the denaturation free energy, into component parts and uses data from large numbers of cases to determine model parameters that best fit the data. While this divide and conquer approach is attractive, Mark and van Gunsteren (*8*) made a compelling argument that decomposing the free energy into terms for specific interactions or specific groups may not correctly account for entropic effects. On the other hand, others (*9*, *10*) have reassessed the same data and asserted that free energy decomposition *is* possible. Whatever the resolution of this controversy, there seems to be a general agreement that the approach has, at minimum, empirical predictive utility (*8*−*10*).

Many studies have focused on developing empirical relationships to predict the denaturation free energy of a protein or the free energy consequences of mutation (*11*−*15*). With the exception of the recent work of Serrano and colleagues (*16*) involving ∼1000 mutants, those studies are based on fitting the behavior of modest numbers of mutants from modest numbers of model proteins. We extend this approach, focusing on data from large numbers of mutants from specially designed combinatorial libraries. We showed previously that helix propensities can be determined by using

† Funding for this work was provided by NIH Grant GM58665 and NSF Grant MCB0212939.

* To whom correspondence should be addressed. Phone: (919) 962-0147. Fax: (919) 962-8103. E-mail: marshall@med.unc.edu.

‡ Department of Biochemistry and Biophysics.

§ Department of Microbiology and Immunology.

|| Department of Chemistry.

10.1021/bi0340649 CCC: $25.00
Published on Web 05/30/2003

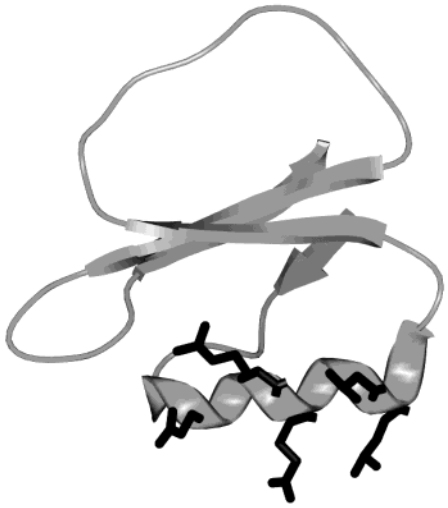

FIGURE 1: Ribbon diagram of wild-type eglin c. The diagram was generated by using SPOCK (*25*) and the crystal structure coordinates (Protein Data Bank entry 1CSE). The four solvent-exposed residues in the α-helix that are varied in the libraries (R22, E23, T26, L27) are shown in stick format as is D19.

a model for stability that contains only helix propensity terms or descriptors, parametrized with specific activity data from 455 isolates from three combinatorial mutant libraries (*17*). The best fit parameters correlated well (Pearson correlation coefficient, $R = 0.85$) with the helix propensity values derived from traditional approaches, but the model accounted for only 31% of the mutant-to-mutant variation in activity (*17*). This observation is not surprising because the model is about stability, but the metric was specific activity. Additionally, our specific activity data were relatively imprecise (coefficient of variation ~12%). Here we use a more structure-relevant metric, $\Delta G_{HOH,N-D}$, derived from the same set of mutants to reassess the analysis.

Any protein could be chosen as a model for these studies, but to obtain accurate stability data that can be interpreted in a straightforward way, monomeric proteins that denature by a reversible two-state reaction at equilibrium are preferred. To facilitate high-throughput determinations of $\Delta G_{HOH,N-D}$ by chemical denaturation we also need a protein that folds rapidly. Eglin c is a small (70 residue) monomeric protein from the potato inhibitor I family of serine proteinase inhibitors (*18*). The structure of eglin c is known from both NMR (*19*) and X-ray crystallography (*20*) studies. Eglin c consists of a four-stranded β-sheet flanked on one side by an α-helix and a binding loop on the other. The helical region of wild-type eglin c corresponds to valine 19 through tyrosine 29 in the intact protein. Its denaturation thermodynamics are well established as following a two-state mechanism (*21*). Its structural homologue, chymotrypsin inhibitor 2 (CI2), has been a useful model protein for folding and mutagenesis studies (*22*−*24*). To parametrize a helix propensity model, we constructed three combinatorial libraries in which four solvent-exposed sites (R22, E23, T26, L27) in the α-helix of eglin c (Figure 1) were mutated to any of seven possible amino acids. Six of the amino acids (E, K, Q, D, N, H) are hydrophilic and common to all three patterned libraries, while the seventh is P for one library (114 variants), G for another (154 variants), and A for the third (187 variants) (*17*). A F10W mutation was introduced into the synthetic eglin c gene prior to library construction to allow fluorescence-monitored chemical denaturation measurements. To facilitate high-precision, high-throughput measurements of $\Delta G_{HOH,N-D}$, we used robotics to prepare the proteins and solutions for chemical denaturation and a semiautomated dual channel titrating fluorometer to monitor denaturation (*26*).

EXPERIMENTAL PROCEDURES

Construction of the Combinatorial Library and Clone Isolation. The library construction and clone isolation have been described (*17*).

Full-Length DNA Sequencing. In our previous study of 455 eglin c mutants, codons 1 through 55 were sequenced for 211 isolates, and codons 10 through 33 were sequenced for the other 244 mutants (*17*). For this study, we sequenced the entire gene for all 455 mutants. Double-stranded DNA was prepared and sequenced at the University of North Carolina DNA Sequencing Facility. We were unable to obtain a full-length sequence from 51 of the clones due to low DNA yields. Of the remaining 404 mutants 32 have mutations at sites other than those at the four desired positions (R22, E23, T26, L27). These 83 mutants were not included in this study. In addition, 18 of the mutants were removed from the analysis because they were duplicates.

High-Throughput Mutant Protein Purification. His-tagged mutant proteins were purified from 12 mL of culture by using nickel affinity resin disks (Pierce) in 96-well microtiter plates as described in the companion paper (*26*). We were unable to obtain adequate protein for $\Delta G_{HOH,N-D}$ measurements from 47 of the variants, and hence they were excluded from the analysis.

High-Throughput Protein Stability Determination. Stabilities were determined by fluorescence-monitored denaturation induced by guanidine hydrochloride at 25 °C in 50 mM Tris and 100 mM NaCl, pH 8.5, as described in the companion paper (*26*). The sequences, the available measured stabilities, and the predicted stabilities are provided in the Supporting Information. Four of the proteins gave denaturation profiles that could not be fit to a two-state model and hence were excluded from the analysis.

Parametrizing Helix Propensity Determinants with Regression Analysis. Helix propensity represents the idiosyncratic position and context-independent component of helix stability contributed by each amino acid. Hence a single amino acid specific parameter for each amino acid is used to represent the helix propensity portion of helix stability. The change in stability caused by the mutant amino acids in each protein can be represented as the sum of the contributions for each of the nine amino acid types in the four mutable sites, that is, the number of each of the nine possible amino acids in the four sites times its amino acid specific helix propensity parameter:

$$\Delta(\Delta G_{HOH,N-D}) = \sum_i k_i n_i \qquad (1)$$

where k_i is the amino acid helix propensity parameter for the *i*th amino acid type, n_i is the specific number of the *i*th amino acid type in the four mutable sites and hence can range from 0 to 4, and the amino acid types, *i*, are A, D, E, G, H, K, N, P, and Q.

The helix propensity parameters are then determined by regression analysis to best fit the stability data. Regression

analysis models in which the sum of the descriptor values is fixed (in this case, the sum of the number of each of the amino acid types in a given mutant protein is always four) is called a "mixture" model (*27*) and must be fitted with no intercept. For linear regression analysis we used the JMP software package (version 5; SAS Institute, Cary, NC). The standard least-squares method was used to minimize the difference between predicted and measured $\Delta G_{HOH,N-D}$ values and to obtain best fit estimates for the parameters.

Descriptor Format. The models used for regression analysis have parameters or descriptors that represent observables or things in the mutant proteins that we can count. In the text and tables these parameters are described with a shorthand best defined by an example. The descriptor DE/K*i* (22, 23), DE/K$i + 4$ is incremented if the *i*th position in the mutant protein, which can be either 22 or 23, is a D or E and the $i + 4$ position is also a D or an E or if the *i*th position is a K and the $i + 4$ position is also a K. For any given mutant protein this particular descriptor can take on the values 0 (if none of the conditions are met), 1 (if the conditions are met for position 22 or 23), or 2 (if the conditions are met for both positions 22 and 23).

Assessing the Maximal Energetic Contributions of Effects Other than Propensity. In the construction of the full model, described below, it turned out that the data did not support several interactions that were expected to affect stability. Some of these effects had small numbers of cases in our library of mutant proteins, and hence an issue arises as to whether the effect is truly missing or simply not significant due to the small number of cases. We used an analysis of simulated data sets to assess the smallest parameter value for the effect that could have been detected with our particular set of mutants. The simulated data sets had exactly the same properties as the real mutant proteins with the exception that an energy term was added to the mutant proteins that would have the expected effect if it were present. In the simulated data sets the stability of each mutant protein was determined by the equation:

$$\Delta G_{N-D,simulated} = \Delta G_{N-D,predicted} + \Delta G_{random} + \Delta G_{effect\ to\ test} \quad (2)$$

where $\Delta G_{N-D,predicted}$ is the stability predicted by the full model (minus the effect to be tested), ΔG_{random} is a number drawn from a normal distribution with the mean and standard deviation of the residuals (measured denaturation free energy minus denaturation free energy predicted by the full model) for the entire mutant protein population, and $\Delta G_{effect\ to\ test}$ is the energetic contribution of the effect (effect contribution per instance times the number of instances in the mutant protein).

To test how strong the effect would need to be to be detected by the actual number of cases in our mutant protein library, we carried out regression analysis on simulated data sets with the effect having various magnitudes (0.4, 0.3, 0.2, 0.1, 0.075, 0.05, 0.025, and 0.01 kcal/mol per instance). The smallest parameter that had a *P* value (probability of observing an even greater *t*-statistic given the hypothesis that the parameter is zero) of less than 0.05 in the regression analysis was reported as the upper limit that the effect could have and still not show up in the regression analysis using the measured stability values instead of the simulated data.

RESULTS

GdnHCl-Induced Chemical Denaturation. On resequencing the DNA from the original library of 455 eglin c mutants, 32 were found to have second site mutations and 18 were duplicates. We were unable to get a good DNA sequence from 51 other mutants, usually due to low yields of plasmid DNA. Mutant proteins were prepared from the remaining 354 mutant-containing strains and 12 different isolates containing wild-type eglin c. Yields of protein ranged from 2 to 70 μg/mL of culture and are approximately proportionate to protein stability. Forty-seven of the strains gave too little protein for chemical denaturation analysis. GdnHCl denaturation data were acquired with a dual channel semiautomated titrating fluorometer (ATF105, Protein Solutions Inc.) from the remaining 307 mutant proteins. Since wild-type eglin c denatures via a two-state, equilibrium-reversible process (*21*), we analyzed all of the mutant proteins with a two-state model. Four of the mutant proteins gave denaturation profiles that could not be fit to a two-state model and were excluded from the analysis. Typical normalized GdnHCl-induced denaturation curves for the wild-type and several mutant proteins can be seen in the companion paper (Figure 5 in ref *26*). Fluorescence intensity data were fitted to a six-parameter model (*26*) to obtain values for $[GdnHCl]_{1/2}$, the midpoint of the denaturation curve, m_{N-D}, the dependence of the denaturation free energy on molar GdnHCl concentration, and the standard fitting errors of each. Using the linear extrapolation method (*28*), $\Delta G_{HOH,N-D}$ is the product of $[GdnHCl]_{1/2}$ and m_{N-D}.

Accuracy of Stability Values. There are three types of reportable errors associated with $\Delta G_{HOH,N-D}$ determinations. One is the error calculated from fitting the data from a single measurement to the two-state model. This error is usually very small. In our case, the standard errors for $[GdnHCl]_{1/2}$ and m_{N-D} from curve fitting are ± 0.0026 M and ± 0.01 kcal mol^{-1} M^{-1}, respectively, resulting in a standard error for $\Delta G_{HOH,N-D}$ of ± 0.03 kcal/mol from error propagation analysis. The other two errors reflect repeatability. This can be obtained from multiple measurements from a single protein preparation or from measurements from multiple protein measurements. Our measurements for $\Delta G_{HOH,N-D}$ from a single preparation of wild-type eglin c is 6.14 ± 0.04 kcal/mol. Measurements from 19 different preparations of wild-type eglin c gave 6.12 ± 0.05 kcal/mol. The error for reproducibility from different protein preparations most directly represents one's capacity to know the $\Delta G_{HOH,N-D}$ values, and that is what we would like to report for our measurements. However, this approach is not practical for 303 mutant proteins. So, we determined reproducibility from four different protein preparations for each of 14 mutant proteins selected to represent the range of $\Delta G_{HOH,N-D}$ values in the full set. The standard errors for $\Delta G_{HOH,N-D}$ from these 56 protein preparations average ± 0.087 kcal/mol. If these values apply to all of the mutant proteins, these measurements will contribute only 1.3% to the variance of otherwise perfect models (*26*).

Regression Analysis To Parametrize the Helix Propensity Model. The propensity component of helix stability is the residue intrinsic, context-free, contribution, and hence there is a single parameter for each residue in the model. Regression analysis was used to find parameters that provide

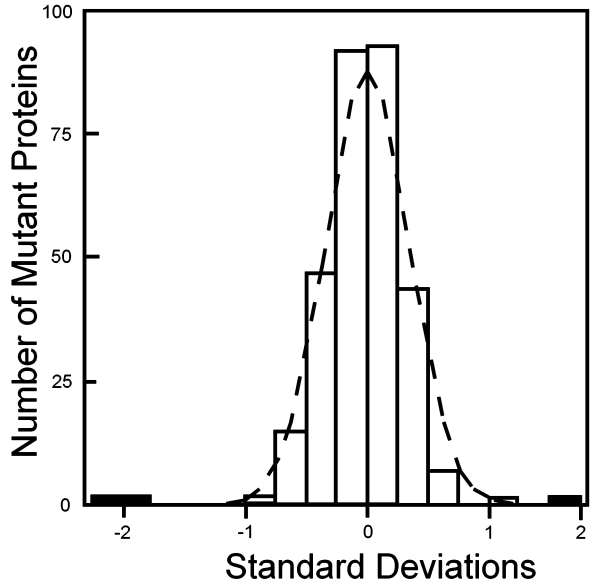

FIGURE 2: Distribution of $\Delta(\Delta G_{HOH,N-D})$ residuals from the simple propensity model. The *x*-axis is in units of standard deviations from zero residual. The three filled boxes represent single mutant proteins with residuals so large as to be unlikely to be from the same population as the bulk population and hence are defined as outliers. The dotted line represents a normal distribution fitted to the data.

the best fit to the measured $\Delta G_{HOH,N-D}$ values from the 303 mutant proteins. The fit for the entire set of 303 mutant proteins has an R^2 of 0.79 (R^2 is the square of the correlation between actual and predicted response and represents the proportion of the variation in response around the mean that can be attributed to terms in the model rather than random error). However, the distribution of residuals (measured $\Delta G_{HOH,N-D}$ minus predicted $\Delta G_{HOH,N-D}$) shows the presence of three outliers (Figure 2). It is not the presence of mutant proteins with large residuals that indicate the presence of outliers but rather the presence of mutant proteins whose residuals have such low probabilities that they would be unlikely to be present in a population size of 303. Three mutant proteins appear to belong to a different population of proteins based on the capacity of the propensity model to predict their stability. We can therefore pull the two populations apart and analyze each separately using the propensity model. The major population of 300 mutant proteins is predicted well by the propensity model. The parameters derived from this subset (Table 1) account for 83% of the measured variance in $\Delta G_{HOH,N-D}$ for 99% of the mutant proteins. As would be expected, fitting a nine-parameter model to the smaller population of three mutant proteins gives a very good fit ($R^2 = 0.93$), but the derived parameters have no correlation with those determined either from the bulk population of 300 mutant proteins or from previous determinations of helix propensities.

The use of regression analysis to parametrize any particular dissection of the free energy into contributions of specific groups or interactions is subject to statistical and experimental errors. One measure of adequate sampling of a combinatorial library is that the regression parameters do not change significantly with increasing sample size. The parameters for the helix propensity model become stable when the sample size becomes larger than 100 (Figure 3).

Interestingly, the parameters in this model become statistically significant before they become stable. That is, the propensity parameters derived from the smallest of the $\Delta G_{HOH,N-D}$ subsets have *P* values less than 0.05. This test for robustness is similar to the test for overfitting, where parameters are derived from a "training" set and the performance of those parameters is then tested on a completely unrelated set of samples. In a similar fashion we divided the 300 $\Delta G_{HOH,N-D}$ values into two subsets of 150 with equal numbers of A-, G-, and P-containing mutant proteins and derived the parameters from each. The parameters for each subset account for almost the same fraction of variance ($R^2 = 0.82$ and $R^2 = 0.83$). The parameters from the two subsets were essentially identical ($R = 0.98$).

There Is Useful Information in Individual m_{N-D} Values. The rate of the change of ΔG_{N-D} as a function of denaturant

Table 1: α-Helix Propensity Parameters Determined by Regression Analysis from the $\Delta(\Delta G_{HOH,N-D})$ Effects of Amino Acid Composition in Four Solvent-Exposed Helical Positions in Eglin c

amino acid	parameter estimate[a]	SE	*P* value[b]
K	0.955	0.022	<0.0011
A	0.801	0.038	0.0347
Q	0.735	0.019	<0.0001
E	0.578	0.019	<0.0001
H	0.472	0.020	<0.0001
N	0.414	0.024	<0.0001
D	0.348	0.021	<0.0001
G	0	0.021	<0.0001
P	−1.160	0.129	<0.0001

[a] The parameters for each of the nine amino acids were derived from the regression analysis of $\Delta G_{HOH,N-D}$ values for 300 eglin c variants using the JMP software package, version 5 (SAS Institute, Cary, NC). The parameters (kcal/mol per instance) were normalized to set the glycine parameter at zero by subtracting the glycine value (−1.037 kcal/mol) from all of the parameters. [b] The probability of getting an even greater *t*-statistic given the hypothesis that the parameter is zero.

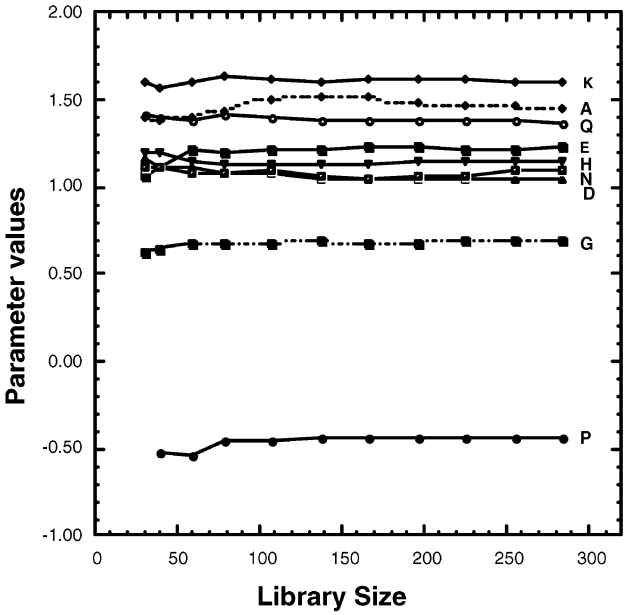

FIGURE 3: Parameter values versus library size. Regression parameters were calculated from subsets of mutant proteins. The subsets contained an equal number of $\Delta G_{HOH,N-D}$ values from the three libraries (6 hydrophilics plus A, or G, or P) until a subset size of 132, at which point the 44 mutant proteins with $\Delta G_{HOH,N-D}$ values in the proline library were exhausted; after that point, only variants from the alanine- and glycine-containing libraries were sampled.

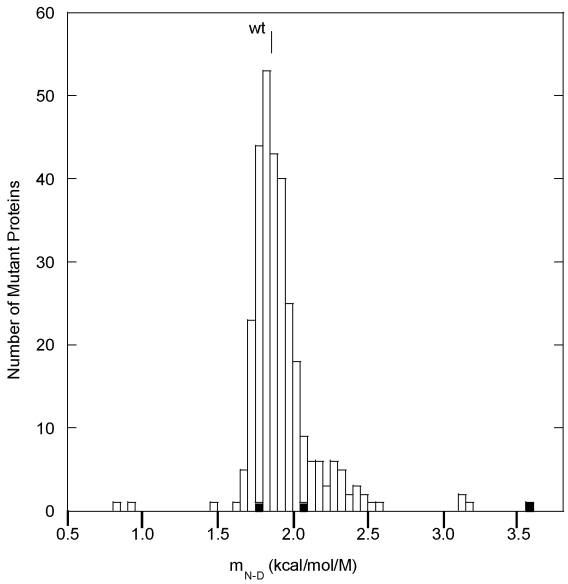

FIGURE 4: Distribution of m_{N-D} values for 303 eglin c mutant proteins. This distribution has a median of 1.86 kcal mol^{-1} M^{-1} and an average of 1.92 kcal mol^{-1} M^{-1}. The wild-type m_{N-D} value, indicated by the bar in the figure, is 1.86 ± 0.03 kcal mol^{-1} M^{-1} (average and standard deviation from different protein preparations). The three mutant protein outliers excluded from the propensity analysis model are indicated as black filled bars.

concentration, m_{N-D}, varies significantly from the wild-type value for many mutant proteins in many protein systems (*29*−*33*). However, the determination of m_{N-D} is less precise than the determination of C_m, the midpoint of the transition. Consequently, in many studies, an average m_{N-D} value derived from wild type or all of the mutant proteins in the study has been used for calculating $\Delta G_{HOH,N-D}$. The assumption is that most of the variation in measured m_{N-D} values is from measurement imprecision. The m_{N-D} values for the 303 mutant eglin c proteins in our data set vary over an approximately 3-fold range (Figure 4), similar to what is seen in staphylococcal nuclease mutants (*33*). The standard deviation for the m_{N-D} value distribution (±0.20 kcal mol^{-1} M^{-1}) for the entire population is four times larger than the experimental error (±0.05 kcal mol^{-1} M^{-1}) for any particular m_{N-D} value, suggesting that the observed mutant-to-mutant variation in our m_{N-D} values contains information about the mutant proteins. When individually determined m_{N-D} values are used to calculate $\Delta G_{HOH,N-D}$, the propensity model accounts for 83% of the mutant-to-mutant variation in $\Delta G_{HOH,N-D}$. When the mean m_{N-D} is used to calculate the $\Delta G_{HOH,N-D}$ values for all of the mutant proteins, the propensity model accounts for only 77% of the data variance, and parameters for each amino acid in the propensity model are smaller than those calculated using the m_{N-D} values from each individual mutant protein.

Building a More Complete Model for Helix Stability. For a model to be parametrized by regression analysis, we seek variables, terms, or descriptors that represent things we can count in each mutant protein [e.g., the sum of prolines in the four mutable sites (0−4), the number of alanines at position 27 (0 or 1), the number of glutamic acids at position 22 when there is a lysine at position 26 (0 or 1)]. To build a more complete model, we used a two-pass approach. We first tested every candidate descriptor as a single term added to the propensity model. All of the descriptors that had individual *P* values better than 0.05 when tested by regression analysis against the stability values of the 300 mutant proteins were then added to a candidate full model. That candidate model was used in regression analysis, and all of the descriptors with *P* values greater than 0.05 were removed from the model. Each of the rejected descriptors was then tested one at a time as an alternative to an existing descriptor (e.g., a stabilizing position effect for alanine at position 22 instead of the destabilizing effect for alanine at position 27). This was done by removing the descriptor that seemed the likely alternative from the full model, adding the alternative descriptor, and assessing its fit and the *P* value in this alternative "full" model. If the *P* value was greater than 0.05, the alternative was rejected. If the alternative replaced more than one other descriptor and passed all of the other tests, then it was chosen because a model with fewer descriptors is better than one with more. If there was a physical basis to select a descriptor over its alternative, it was chosen; if not, the descriptor that gave the best fit to the data (largest R^2) was chosen, but this is a weak basis for a choice. Ultimately, one needs to identify a physical basis for the choices.

Position Effects. Amino acid helix propensities are defined as the position-independent helix-stabilizing effects of individual amino acids. The simple propensity model had a separate descriptor for each amino acid, representing the number of that amino acid in the four mutable sites. However, there are several known position-dependent effects in α-helices, i.e., capping (*34*−*36*), helix−dipole (*37*, *38*), and side chain−side chain effects (*39*, *40*). To assess the utility of a position-dependent descriptor for a given amino acid, above and beyond its propensity, we added a single term for that amino acid at the position of interest to the simple propensity model and used regression analysis of the 300 stability values to parametrize context effect descriptors. This procedure was repeated for all of the nine amino acids involved in this study, one by one, for each of the four sites subject to mutation. Table 2 lists the position effects with *P* values greater than 0.05.

Histidine and proline have no significant position parameters. There were too few proline-containing mutant proteins with stability data to assess position effects. Most of the proline-containing mutant proteins gave insufficient yields due to proteolysis in *Escherichia coli*. Only glutamine has a single, significant, nonzero position effect parameter. The remainder of the residues tested have multiple position effect parameters that pass the $P < 0.05$ test. When all of the position effect descriptors that pass the *P* value test are added together to the candidate full model, only aspartic acid has multiple descriptors that again pass the *P* value test (D at 22 and D at 23). For the remaining cases of multiple position effect descriptors for a single amino acid, regression analysis by itself does not indicate whether it is the stabilizing or the destabilizing effect that should go into the full model. So, we are left with eight position effect descriptors, most as alternatives: a destabilizing effect of Q at 23; a stabilizing effect of K at 22 or a destabilizing effect at 23 or 26; a destabilizing effect of D at 22; a stabilizing effect of D at 23; a destabilizing effect of E at 22 or a stabilizing effect at 27; a destabilizing effect of N at 22 or a stabilizing effect at 27; a stabilizing effect of G at 22 and 26 or a destabilizing effect at 23 and 27; and a stabilizing effect of A at 22 or 26 or a destabilizing effect at 27.

Table 2: Position Effects

descriptor[a]	cases[b]	in full model	parameter value[c]	incremental R^2[d]	P value[e]
A at 22	16		0.236	0.004	0.0099
A at 26	14		0.222	0.004	0.0140
A at 27	20	yes	−0.431	0.016	<0.0001
K at 22	43	yes	0.335	0.023	<0.0001
K at 23	37		−0.197	0.007	0.0007
K at 26	32		−0.165	0.004	0.0059
Q at 23	50		−0.103	0.002	0.0417
E at 22	60		−0.104	0.003	0.0278
E at 23	46		0.117	0.003	0.0336
E at 26	47		−0.105	0.002	0.0499
E at 27	39		0.145	0.004	0.0107
N at 22	34	yes	−0.266	0.015	<0.0001
N at 27	41		0.171	0.005	0.0031
D at 22	43		−0.183	0.007	0.0006
D at 23	46	yes	0.266	0.013	<0.0001
G at 22	34		0.158	0.003	0.0149
G at 23	37		−0.219	0.005	0.0036
G at 26	27		0.202	0.004	0.0089
G at 27	16		−0.249	0.005	0.0043
D/E (22)	103	yes	−0.183	0.012	<0.0001
G (22, 26)	61	yes	0.300	0.012	<0.0001
G (23, 27)	43		−0.300	0.012	<0.0001
K*i* (23), DE*i* + 3	8		−0.205	0.002	0.0523
K*i* (23), DHK*i* + 3	19		−0.143	0.002	0.0529
N*i* (22, 23), ND*i* + 4	20	yes	−0.156	0.002	0.0404
D*i* (22, 23), NDE*i* + 4	39		−0.130	0.003	0.0140
DE/K*i* (22, 23), DE/K*i* + 4	65	yes	−0.154	0.006	0.0013
DE/K*i* (22, 23), K/DE*i* + 4	53	yes	0.215	0.010	<0.0001
AGADIR	300		0.129	0.001	0.1924
FOLD-X	300		−0.119	0.013	<0.0001

[a] See Experimental Procedures for the descriptor format. [b] Number of cases of the descriptor in the 300 mutant proteins. [c] The parameter value (kcal/mol per instance at pH 8.5 and 25 °C in the absence of denaturant) where positive values represent stabilizing effects in a model comprising the propensity descriptors plus the single descriptor in column 1. [d] The increase in R^2 attained by adding the descriptor in column 1 to the propensity model which has an R^2 of 0.828. [e] The P value (probability of observing an even greater t-statistic given the hypothesis that the parameter is zero) for the model comprising the propensity descriptors plus the single column 1 descriptor.

Side Chain−Side Chain Interactions. We chose to evaluate two types of sequence-dependent side chain interactions, steric clashes and charge−charge interactions. In eglin c, the sequence of the helical region (residues 19−29) is DQAREYFTLHY, where the four underlined residues (R22, E23, T26, L27) are mutated to any of the six amino acids (Q, E, K, D, N, H) plus P or G or A. Side chains containing residues of opposite charges at (*i*, *i* + 3) and (*i*, *i* + 4) positions so that they face each other on the surface of the helix can contribute to helical stability by forming ion pairs (*38*). Destabilization can arise from residues at these positions with the same charge or via steric clashes. Interactions of this nature between mutated residues with residues at nonmutable positions would appear in our analysis as position effects for that residue in that position. For example, the position effects of K, D, and E at mutable position 22 (Table 3) are presumably due to side chain interactions with the nonmutated aspartic acid at position 19 (Figure 1). Hence, to derive side chain−side chain interaction parameters involving only the mutable positions, we looked at *i*, *i* + 4 interactions (22 with 26 and 23 with 27) and *i*, *i* + 3 interaction (23 with 26). We enumerate for each mutant the number of interactions that meet the criteria and then evaluate these as potential descriptors in the full model as described in the previous section. Six side chain−side chain interaction descriptors survived the P value test when tested as a single addition to the propensity model (Table 2). Descriptors that were tested but did not survive the P value test when added to the helix propensity model test are addressed in the Discussion.

Table 3: Full Model Descriptors and Contributions to α-Helix Stability

descriptor[a]	presumed physical basis	parameter[b]	fraction response explained[c]	P value[d]
G at 22 or 26	unknown	0.289	0.0097	<0.0001
K at 22	charge−charge with D at 19	0.286	0.0127	<0.0001
DE/K*i* (22, 23), K/DE*i* + 4	charge−charge	0.176	0.0041	<0.0001
D at 23	unknown	0.136	0.0041	0.0010
A at 22, 23, 26, 27	propensity	0.064	0.0011	0.0641
K at 22, 23, 26, 27	propensity	−0.067	0.0016	0.0064
DE/K*i* (22, 23), DE/K*i* + 4	charge−charge	−0.068	0.0012	0.0847
Q at 22, 23, 26, 27	propensity	−0.104	0.0081	<0.0001
N*i* (22, 23), ND*i* + 4	steric clash *i* with *i*+4	−0.114	0.0010	0.0431
Q*i* (22, 23), Q*i* + 4	steric clash *i* with *i*+4	−0.121	0.0013	0.0355
DE at 22	charge−charge with D at 19	−0.127	0.0040	0.0003
E at 22, 23, 26, 27	propensity	−0.250	0.0319	<0.0001
N at 22	unknown	−0.265	0.0093	<0.0001
N at 22, 23, 26, 27	propensity	−0.379	0.0756	<0.0001
A at 27	unknown	−0.396	0.0121	<0.0001
H at 22, 23, 26, 27	propensity	−0.423	0.2273	<0.0001
D at 22, 23, 26, 27	propensity	−0.541	0.1243	<0.0001
G at 22, 23, 26, 27	propensity	−1.037	0.3423	<0.0001
P at 22, 23, 26, 27	propensity	−1.974	0.1292	<0.0001

[a] See Experimental Procedures for the descriptor format. [b] In kcal/mol per descriptor at pH 8.5 and 25 °C in the absence of denaturant. Positive terms are stabilizing. [c] Fraction of the variability accounted for by the full model that is accounted for by this descriptor. [d] Probability of getting an even greater t-statistic given the hypothesis that the parameter is zero.

The Full Model. Of the six side chain−side chain descriptors (bottom of Table 2) that were significant in the simple model (propensity descriptors plus the single descriptor to be tested), only three (Table 3) were also significant in the full model (propensity descriptors plus all of the single descriptors significant in the simple model). Among these were several alternative descriptors for side chain effects. The fractions of variation in stability of the mutant proteins accounted for by the various alternative stabilizing *i*, *i* + 4 models were approximately the same so we chose the model with fewer parameters. For the same reason we chose a single descriptor for the destabilizing *i*, *i* + 4 effects. The resulting 19-parameter model (Table 3) accounts for 92% of the variation in stabilities of the 300 mutant proteins (Figure 5). This is 9% more than the model with parameters for helix propensities only. It is worth noting that while the propensity terms account for a large fraction of the predictive capacity of the model (Table 3, column 4), there are notable exceptions. The fraction of the variability accounted for in the full model that is accounted for by both the alanine and lysine descriptors is smaller than the position effects involving alanine and lysine.

Other Predictors. To test the adequacy of our descriptors for helix stability, we tested whether adding existing models

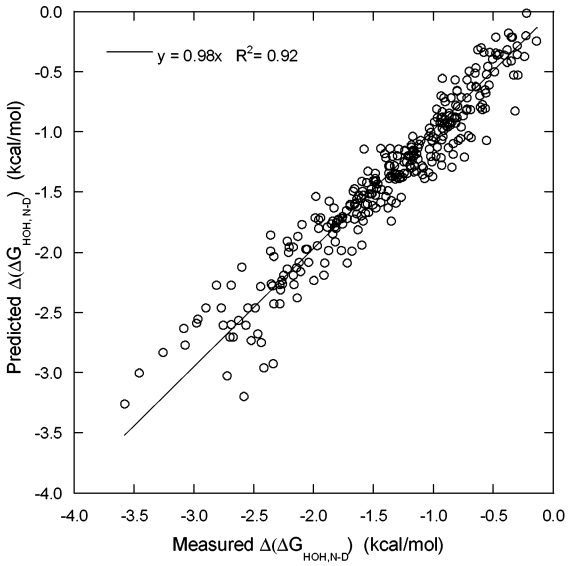

FIGURE 5: Measured versus predicted stability values for 300 mutant eglin c proteins. Predicted values were calculated using the 19-parameter full model from Table 3.

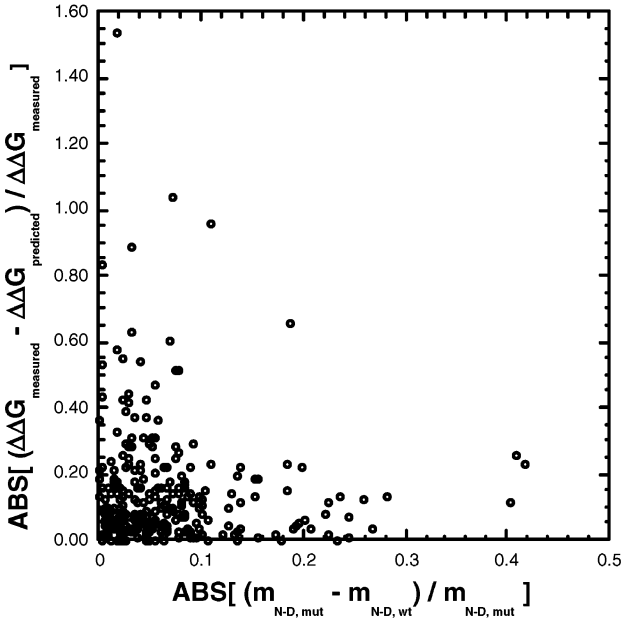

FIGURE 6: Normalized residuals versus normalized denaturation slopes. The maximum value of the *x*-axis has been truncated slightly, eliminating a few values to show more clearly the lack of relationship between the residuals and the denaturation slope, m_{N-D}.

for protein stability or helicity to the full model increases the fraction of mutant-to-mutant variation in stability predicted by the model. FOLD-X is a computer algorithm (*16*) that estimates the energetic contribution of mutations to proteins and complexes. FOLD-X considers van der Waals interactions, hydrogen bonds, side chain−side chain interactions, solvation energy, water stabilization, and charge−charge interactions. Although we have included terms for position effects and side chain−side chain interactions, we expect that if FOLD-X includes neglected or underestimated interactions, inclusion of the FOLD-X predicted term should increase the fraction of stability variation explained by the enhanced model. FOLD-X adds 1.3% to the R^2 over and above the propensity descriptors (Table 2) and adds 0.1% to the full model, indicating that there are effects in FOLD-X that are relevant to our mutant proteins that are not already in our full model. When we used FOLD-X alone to predict the $\Delta(\Delta G_{N-D})$ values for our mutant protein population, it accounts for only 40% of the variation in stabilities of the mutant proteins as compared to 92% for our full model.

We also tested the Web assessable version of AGADIR (*13*) to assess whether that program brings to bear stabilizing effects not present in our full model. The peptide helicity calculated by AGADIR adds 0.1% to the R^2 over and above the propensity descriptors, but the resulting parameter has a very high *P* value (Table 2), indicating that there are no stabilizing effects relevant to our mutant proteins in the AGADIR helicity numbers that are absent from our full model. When we used AGADIR helicity values alone to predict the $\Delta(\Delta G_{N-D})$ values for our mutant proteins, it accounts for only 6% of the variation in stabilities of the mutant population as compared to 92% for our full model.

Reference State and m_{N-D} Values. The denaturation slope, m_{N-D}, reflects the sensitivity of the protein unfolding free energy to denaturant concentration. Formally, the m_{N-D} value is simply one of the parameters in the equation used to describe the two-state denaturation reaction. One assertion for its physical correlate is the amount of protein surface exposed to solvent upon unfolding (*2*, *42*). Hence one might expect that mutant proteins with m_{N-D} values significantly different from that of the wild-type protein would have an altered reference state that could account for some of the change in free energy of unfolding. If the mutant proteins with m_{N-D} values most different from wild type have some of their ΔG_{N-D} difference from wild type due to an altered reference state, then they would be the mutant proteins least well explained by the model since it is based solely on native state considerations. The m_{N-D} values for the 300 eglin c variants vary over an approximately 3-fold range (Figure 4), indicating a potentially significant impact on the reference state for some of the mutant proteins. However, there is no correlation (Figure 6) between the normalized change in m_{N-D} values of the mutant proteins from wild type and their residuals (the difference between predicted and measured stabilities). If anything, there is a larger spread of residuals for the mutant proteins with more wild-type m_{N-D} values.

DISCUSSION

Comparing Helix Propensity Values from Regression Analysis to Those from Other Studies. The parameters for each of the nine amino acids varied in the experiment were scaled to set glycine to zero (by subtracting the glycine parameter derived from regression analysis from all nine parameters) to facilitate the comparison of our parameters with the helix propensities obtained from other studies (*43*−*46*). The propensity values derived from this study agree well with these other scales of helix propensities (Table 4), which include those based on $\Delta\Delta G_{HOH,N-D}$ values measured in a synthetic coiled coil peptide (*43*), at residue 44 in T4 lysozyme (*44*), at residue 32 in barnase (*45*), and at residue 21 in ribonuclease T1 (*46*).

The propensity values from the full model (Table 3) correlate better with helix propensity values determined from other systems than propensity values derived from a model with only propensity terms (Table 1). In the model with only

Table 4: Comparison of Regression Parameters with Helix Propensity Values from Other Systems

model system	intercept[a]	slope[a]	correlation coeff[a]
T4 lysozyme[b]	0.06	0.97	0.92
barnase[c]	−0.11	0.96	0.88
coiled coil peptide[d]	−0.15	0.87	0.86
RNase T[e]	0.06	0.73	0.95

[a] The intercept (kcal/mol per instance), slope, and correlation coefficient in comparisons of the regression parameters derived in this study and the α-helix propensities from biophysical studies of the indicated model systems. [b] Reference *44*. [c] Reference *45*. [d] Reference *43*. [e] Reference *46*.

propensity parameters the parameter for lysine (0.955 kcal/mol per instance) was larger than that for alanine (0.801 kcal/mol per instance) while the model that includes position effects (Table 3) gives a parameter for lysine that is smaller (0.970 kcal/mol per instance) than that for alanine (1.101 kcal/mol per instance). This discrepancy is due to covariation between the descriptor for the position-independent lysine descriptor and the descriptor representing the extra stabilizing effect of lysine at position 22 from its interaction with the aspartic acid at position 19. That is, in a model with propensity descriptors alone the parameter for lysine includes the stabilizing effects of propensity and the effects of the interaction with aspartic acid 19.

Physical Correlates with Descriptors. Our position analysis shows that there are alternative effects supported by the data. Regression analysis by itself does not provide a mechanism to decide between various alternative descriptors that deal with position effects. However, aspartic acid 19 provides a sensible physical correlate, charge−charge interactions, for descriptors involving charged residues at position 22, and hence we chose *i*, *i* + 3 descriptors involving position 22 rather than the alternatives involving position 27. The parameters for those descriptors are consistent with a stabilizing contribution from an *i*, *i* + 3 charge−charge interaction between the aspartic acid at position 19 and the lysine at position 22 and a destabilizing contribution from interactions with either glutamic or aspartic acid at position 22 with the aspartic acid at position 19. The descriptors involving *i*, *i* + 4 effects are consistent with expected interaction effects between charged residues.

We expected some effects that were not supported by the data. We see no position effects associated with *i*, *i* + 4 charge−charge effects involving positions 19 and 23. We see no *i*, *i* + 3 charge−charge effects involving positions 23 and 26. We see no position effects due to the interaction of charged residues with the helix dipole which should show up as position effects for charged residues in position 27 (Table 2). In addition, we see effects without an obvious physical correlate: a stabilizing effect of asparagine at position 22 (or destabilizing at position 27), a stabilizing effect of aspartic acid at position 23, a stabilizing effect of glycine at positions 22 or 26 (or destabilizing at positions 23 or 27), and a destabilizing effect of alanine at position 27.

Descriptors That Do Not Show Significant Effects on Stability. Many of the descriptors tested have parameters that are not significantly different from zero. For some descriptors this is because the number of cases for the descriptor is small.

Table 5: Upper Limit on Contributions of Effects Not Supported by the Data

descriptor[a]	cases[b]	upper limit on parameter value[c]
K*i* (23), D*i* + 3	5	0.3
K*i* (23), E*i* + 3	3	0.4
K*i* (23), DE*i* + 3	8	0.2
K*i* (23), K*i* + 3	6	0.3
K*i* (23), HK*i* + 3	8	0.2
K*i* (23), DHK*i* + 3	19	0.2
D*i* (23), K*i* + 3	3	0.3
D*i* (23), HK*i* + 3	12	0.3
D*i* (23), E*i* + 3	6	0.2
D*i* (23), D*i* + 3	5	0.3
D*i* (23), DE*i* + 3	11	0.1
DE*i* (23), DE*i* + 3	28	0.05
DE*i* (23), K*i* + 3	7	0.2
DE/K*i* (23),DE/K*i* + 3	34	0.03
DE/K*i* (23), K/DE*i* + 3	15	0.1
E*i* (23), K*i* + 3	4	0.3
E*i* (23), HK*i* + 3	12	0.1
E*i* (23), E*i* + 3	10	0.2
E*i* (23), D*i* + 3	7	0.1
E*i* (23), DE*i* + 3	17	0.2
H*i* (23), E*i* + 3	7	0.1
H*i* (23), HK*i* + 3	14	0.1
Q*i* (23), HK*i* + 3	19	0.2
Q*i* (23),Q*i* + 3	53	0.1
K*i* (22, 23), HK*i* + 4	15	0.3
H*i* (22, 23), DQE*i* + 4	30	0.1
H*i* (22, 23), HK*i* + 4	25	0.1
QN*i* (23), QN*i* + 4	18	0.2
Q*i* (22, 23), NDEHK*i* + 4	55	0.1
QN (22, 23), QND*i* + 4	69	0.1

[a] See Experimental Procedure section for the descriptor format. [b] Number of cases for the descriptor in the 300 mutant proteins. [c] Upper limit on the kcal/mol/instance of the effect. See Experimental Procedure section for how these values are calculated.

The numbers of cases with a proline at any position, with alanine at position 23, and with lysine at position 23 associated with a negative charge at position 26 are poorly represented in the library, and hence the absence of position effects involving these descriptors is expected. We were surprised to find that our data does not support the presence of *i*, *i* + 3 charge−charge effects between residues 23 and 26. Since the number of these cases in our library is relatively small, we determined how small the effects could be and still be detected with the small number of instances (Table 5). The main *i*, *i* + 3 effects in the eglin c helix [DE/K*i* (23), DE/K*i* + 3 and DE/K*i* (23), K/DE*i* + 3] must be smaller than 0.1 kcal/mol per instance to not be seen in our data set. The *i*, *i* + 3 effects supported by the stability data involving the D at 19 and K, D, or E at 22 have effects of 0.29, −0.13, and −0.13 kcal/mol, respectively (Table 3).

Applicability to Other Proteins. There is a risk in parametrizing models with data from a single protein, that the parameters will be idiosyncratic to that protein. Clearly, the propensity parameters derived here are applicable to other proteins, but it is not clear that this is true for the other effects in our model. On the other hand, all of the effects in our full model are expressed as descriptors that would be expected to be applicable in other proteins, and at least the signs of the parameters are all what are expected for all proteins. We see in eglin c strong *i*, *i* + 3 charge−charge effects involving residues 19 and 22, but we observe no such effects involving positions 23 and 26. We see a strong

destabilizing effect of asparagine at position 22 relative to position 27. We see a strong destabilizing effect of glycine at positions 23 and 27 relative to positions 22 and 26. We see a strong destabilizing effect of alanine at position 27 relative to positions 22 and 26. However, such effects cannot yet be said to represent portable descriptors until the effects can be rationalized in terms of some structurally and chemically sensible numerable features and the parameters verified in other proteins.

CONCLUSIONS

Statistical modeling or regression analysis increases the formalism of mutational studies of proteins at the expense of the need for characterizing large numbers of mutant proteins. Our implementation of a high-throughput procedure to process large numbers of mutant proteins has a downside in that low stability mutant proteins tend to be underrepresented. An upside is that the stability measurements are unusually precise (±0.087 kcal/mol). What does one get from the increased formalism? We found a satisfying level of robustness in the approach in the sense that we can extract helix propensity values from both relatively imprecise specific activity data (*17*) and precise stability measurements. Regression analysis generates model-independent parameters only when the descriptors in the model are independent of each other, and this requirement could easily limit its utility for obtaining parameters applicable to many proteins. Hence it is satisfying that while we see this issue at work, in that the calculation of the helix propensity for lysine is model dependent due to stabilizing charge−charge effects, the covariance effect is small.

An attractive feature of statistical modeling is its capacity to quantify the completeness of the models tested. Here we show that the propensity component of the model accounts for 83% of the mutant-to-mutant variation in stability for 99% of the mutant proteins. Position and side chain−side chain effects bring the fraction of variation in stability accounted for by the model up to 92%. Measurement error accounts for another 1.3% of the variation in stability. This degree of precision means that the 7% of variation in stability values of the mutant proteins not explained by the full model and measurement variation is both real and large in that all of the nonpropensity effects in the full model account for only 9% of the mutant-to-mutant variation in stabilities. At this point we have no idea if the effects responsible for the missing 7% are uninteresting (i.e., arising from experimental methods) or interesting (i.e., arising from structural effects).

This method has allowed us to partition the effects from our population of 300 mutant proteins into four classes: expected effects supported by the data, expected effects not supported by the data, effects of unknown physical basis supported by the data, and effects missing from the model necessary to account for 7% of the variation in stability values. In addition, there is a small class of outliers that is badly predicted by the model. This is a surprisingly rich array of classes given that the determinants of α-helix stability are the most well-known.

ACKNOWLEDGMENT

We thank Dr. Charlie Carter and members of the Edgell and Pielak laboratories for useful discussions and Dr. Brenda Temple, Director of the UNC Structural Bioinformatics Facility, for help with graphics.

SUPPORTING INFORMATION AVAILABLE

One table of stability data for the 303 mutant proteins described and one table of predicted stabilities for 51 mutant proteins recovered at low yield or that did not have chemical denaturation transitions that could be fit to a two-state model. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Hutchison, C. A., III, Philips, S., Edgell, M. H., Gillams, S., Janke, P., and Smith, M. (1978) *J. Biol. Chem. 253*, 6551−6560.
2. Shortle, D., and Meeker, A. K. (1986) *Proteins: Struct., Funct., Genet. 1*, 81−89.
3. Matthews, B. W. (1987) *Biochemistry 26*, 6885−6888.
4. Fersht, A. R., Shi, J. P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Cater, P., Waye, M. M., and Winter, G. (1985) *Nature 314*, 235−238.
5. Jackson, S. E., and Fersht, A. R. (1994) *Biochemistry 33*, 13880−13887.
6. Pielak, G. J., Auld, D. S., Beasley, J. R., Betz, S. F., Cohen, D. S., Doyle, D. F., Finger, S. A., Fredricks, Z. L., Hilgen-Willis, S., Suanders, A. J., and Trojak, S. K. (1995) *Biochemistry 34*, 3268−3276.
7. Knecht, W., Sandrini, M. P. B., Johansson, K., Eklund, H., Munch-Petersen, B., and Piskur, J. (2002) *EMBO J. 21*, 1873−1880.
8. Mark, A. E., and van Gunsteren, W. F. (1994) *J. Mol. Biol. 240*, 167−176.
9. Brady, G. P., and Sharp, K. A. (1995) *J. Mol. Biol. 254*, 77−85.
10. Boresch, S., and Karplus, M. (1995) *J. Mol. Biol. 254*, 801−807.
11. Munoz, V., and Serrano, L. (1995) *J. Mol. Biol. 245*, 275−296.
12. Luque, I., Mayorga, O. L., and Freire, E. (1996) *Biochemistry 35*, 13681−13688.
13. Munoz, V., and Serrano, L. (1994) *Nat. Struct. Biol. 1*, 399−409.
14. Munoz, V., and Serrano, L. (1995) *J. Mol. Biol. 245*, 297−308.
15. Carter, C. W., Jr., LeFebvre, B. C., Cammer, S. A., Tropsha, A., and Edgell, M. H. (2001) *J. Mol. Biol. 311*, 625−638.
16. Guerois, R., Nielsen, J. E., and Serrano, L. (2002) *J. Mol. Biol. 320*, 369−387.
17. Lahr, S. J., Broadwater, A., Carter, C. W., Jr., Collier, M. L., Hensley, L., Waldner, J. C., Pielak, G. J., and Edgell, M. H. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 14860−14865.
18. Laskowski, M., Jr., and Kato, I. (1980) *Annu. Rev. Biochem. 49*, 593−626.
19. Hyberts, S. G., Goldberg, M. S., Havel, T. F., and Wagner, G. (1992) *Protein Sci. 1*, 736−751.
20. Hipler, K., Priest, J. P., Rahuel, J., and Grutter, M. G. (1992) *FEBS Lett. 39*, 139−145.
21. Bae, S. J., and Sturtevant, J. M. (1995) *Biophys. Chem. 55*, 247−252.
22. Jackson, S. E., and Fersht, A. R. (1991) *Biochemistry 30*, 10428−10435.
23. Jackson, S. E., Moraccim, M., elMasry, N., Johnson, C. M., and Fersht, A. R. (1993) *Biochemistry 32*, 11259−11269.
24. Jackson, S. E., elMasry, N., and Fersht, A. R. (1993) *Biochemistry 32*, 11270−11278.
25. Christopher, J. A., Swanson, R., and Baldwin, T. O. (1996) *Comput. Chem. 20*, 339−345.
26. Edgell, M. H., Sims, D. A., Pielak, G. J., and Yi, F. (2003) *Biochemistry 42*, 7587−7593.
27. Marquardt, D. W., and Snee, R. D. (1974) *Technometrics 16*, 533−537.
28. Santoro, M. M., and Bolen, D. W. (1988) *Biochemistry 27*, 8063−8068.
29. Elwell, M. L., and Schellman, J. A. (1979) *Biochim. Biophys. Acta 580*, 327−338.
30. Shirley, B. A., Stanssens, P., Steyaert, J., and Pace, C. N. (1989) *J. Biol. Chem. 264*, 11621−11625.
31. Betz, S. F., and Pielak, G. J. (1992) *Biochemistry 31*, 12337−12344.

32. Bowler, B. E., May, K., Zaragoza, T., York, P., Dong, A., and Caughey, W. S. (1993) *Biochemistry 32*, 183−190.
33. Shortle, D., Meeker, A. K., and Gerring, S. L. (1989) *Arch. Biochem. Biophys. 272*, 103−113.
34. Aurora, R., and Rose, G. D. (1998) *Protein Sci. 7*, 21−38.
35. Richardson, J. S., and Richardson, D. C. (1988) *Science 240*, 1648−1652.
36. Petukhov, M., Uegaki, K., Yumoto, N., and Serrano, L. (2002) *Protein Sci. 11*, 766−777.
37. Wada, A. (1976) *Adv. Biophys. 9*, 1−63.
38. Hol, W. G., van Duijnen, P. T., and Berendsen, H. J. (1978) *Nature 273*, 443−446.
39. Lyu, P. C., Liff, M. I., Marky, L. A., and Kallenbach, N. R. (1990) *Science 250*, 669−673.
40. Serrano, L., Sancho, J., Hirshberg, M., and Fersht, A. R. (1992) *J. Mol. Biol. 227*, 544−559.
41. Schellman, J. A. (1978) *Biopolymers 17*, 1305−1322.
42. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1995) *Protein Sci. 4*, 2138−2148.
43. O'Neil, K. T., and Degrado, W. F. (1990) *Science 250*, 646−651.
44. Blaber, M., Zhang, X. J., and Matthews, B. W. (1993) *Science 260*, 1637−1640.
45. Horovitz, A., Matthews, J. M., and Fersht, A. R. (1992) *J. Mol. Biol. 227*, 560−568.
46. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1997) *Biochemistry 36*, 10923−10929.

BI0340649