Relation Between Protein Stability, Evolution and Structure, as Probed by Carboxylic Acid Mutations

Raquel Godoy-Ruiz, Raul Perez-Jimenez, Beatriz Ibarra-Molero and Jose M. Sanchez-Ruiz*

Native proteins are marginally stable. Low thermodynamic stability may actually be advantageous, although the accumulation of neutral, destabilizing mutations may have also contributed to it. In any case, once marginal stability has been reached, it appears plausible that mutations at non-constrained positions become fixed in the course of evolution (due to random drift) with frequencies that roughly reflect the mutation effects on stability ("pseudo-equilibrium hypothesis"). We have found that all glutamate → aspartate mutations in wild-type Escherichia coli thioredoxin are destabilizing, as well as most of the aspartate → glutamate mutations. Furthermore, the effect of these mutations on thioredoxin thermodynamic stability shows a robust correlation with the frequencies of occurrence of the involved residues in several-hundred sequence alignments derived from a BLAST search. These results provide direct and quantitative experimental evidence for the pseudo-equilibrium hypothesis and should have general consequences for the interpretation of mutation effects on protein stability, as they suggest that residue environments in proteins may be optimized for stabilizing interactions to a remarkable degree of specificity. We also provide evidence that such stabilizing interactions may be detected in sequence alignments, and briefly discuss the implications of this possibility for the derivation of structural information (on native and denatured states) from comparative sequence analyses.

Keywords: protein evolution; protein structure; protein stability; mutation effects

Thermodynamic stability of proteins is marginal,1–4 as indicated by low experimental values (of the order of a few tens of kJ/mol) for the denaturation free energy at physiological temperature. Low stability may have been selected during evolution due to its several potential advantages. Thus, low stability may be associated to a high degree of flexibility (likely required for function), facilitate protein degradation and help to ensure that the protein is not trapped in incorrectly folded states during folding.4 Also, destabilizing interactions, such as those involving buried polar groups,5,6 may be required for structural uniqueness. On the other hand, the fact that, probably, many (otherwise non-deleterious) mutations may be slightly destabilizing might have also contributed to the low stability of proteins. These neutral, but destabilizing, mutations could perhaps have accumulated during some early stage in evolution up to the point in which stability was marginal and further destabilization would have compromised the native states.

In any case, once marginal stability has been reached, we may expect mutations at non-constrained positions to be fixed with probabilities related with their effects on stability (i.e. fixation of destabilizing mutations being more unlikely). Actually, it has been recently suggested,7,8 mostly on theoretical grounds, that, over evolutionary time, sequences are visited according to energy (i.e. high energies in the native state being unlikely), in a manner analogous to the Boltzmann
The proteins were pure as measured by the effect of mutation on denaturation temperature (right-axis) and denaturation free energy (right-axis). D → E and E → D mutations are shown with open and filled bars, respectively. The numbers alongside the bars indicate the mutated residue in WT thioredoxin. Experimental details on protein preparation and stability determination are as follows. Site-directed mutagenesis: Oligonucleotides used for mutagenesis were obtained from Genetek. Mutations in the codons corresponding to E or D positions in the amino acid sequence of thioredoxin were introduced by the QuikChange™ Site-Directed Mutagenesis method developed by Stratagene®. Briefly, the Quik-Change method is based on a PCR amplification using two complementary oligonucleotide primers containing the desired mutation. The parental non-mutated DNA is finally digested by an endonuclease. Mutation was verified by DNA sequence analysis. Protein expression and purification: Plasmid pTK100 encoding thioredoxin (a gift from Dr. Maria Luisa Tasayco, The City College of New York) was transformed into E. coli JF521 strain for protein overexpression. Cells were grown, starting from single colonies, at 37°C in Luria broth with 40 μg/ml of kanamycin to select for the plasmid-bearing cells. The final 750-fold dilution of the cell broth was allowed to grow during 12 hours after stationary phase was reached. After centrifugation, cell pellets were frozen at −20°C until purification. Protein was purified as described. The proteins were pure as measured by SDS-PAGE gel densitometry. The molecular weight of pure proteins was confirmed by mass spectrometry. Protein concentration was determined spectrophotometrically at 280 nm using a published value of the extinction coefficient. Stability determinations: Differential scanning calorimetry (DSC) experiments were carried out with a VP-DSC calorimeter from MicroCal (Northampton, MA) at a scan-rate of 1.5 K/minute. Protein solutions for the calorimetric experiments were prepared by exhaustive dialysis against the buffer (5 mM Hepes, pH 7.0). Additional experimental details of the calorimetric experiment have been described in detail. A protein concentration dependence for WT thioredoxin denaturation temperature has been reported in the literature and attributed to protein dimerization. Therefore, we carried out all the DSC experiments at comparatively low protein concentrations (about 0.5 mg/ml or below) and we checked that no protein concentration effects on denaturation occurred within the 0–0.5 mg/ml range. Denaturation of all thioredoxin variants studied was highly reversible and fittings of the two-state equilibrium model to the heat capacity profiles were excellent and similar to that we have previously described for WT thioredoxin. The effects of mutations on denaturation free energy were small and, therefore, we could use Schellman’s equation to calculate them without introducing significant errors. The Schellman equation can be written as: ΔΔG = ΔHm(variant) / Tm(variant) / Tm(WT), where ΔΔG is the perturbation free energy (i.e. ΔG(variant)−ΔG(WT)), Tm(variant) and ΔHm(variant) are the denaturation temperature and denaturation enthalpy change (at the denaturation temperature) for the unperturbed protein (i.e. the WT form), and ΔTm is the perturbation effect on denaturation temperature (i.e. Tm(variant)−Tm(WT)).
Figure 2. Correlation between effects of D → E mutations on thioredoxin stability and the frequency of occurrence of D and E residues (as measured by the Z scores) in sequence alignments derived from a BLAST search carried out using as query sequence that of E. coli thioredoxin (see further below in this legend). Stability changes are measured by the effect of mutation on denaturation temperature (left-axis) and denaturation free energy (right-axis). All the stability changes given describe the effect of D → E mutations. The open symbols correspond to the original D → E mutations in WT thioredoxin. The filled symbols are derived from the original E → D mutations, but the sign of the stability changes has been reverted, so that the values shown correspond to D → E effects (for instance, ΔTm = Tm(WT, E) − Tm(variant, D)). The small numbers alongside the symbols indicate the position mutated. At least three stability determinations were performed for each variant; average values of mutation effects on stability and associated standard errors are shown. The large numbers in the main Figure and the insets indicate the sequence similarity range used in the calculation of ∑E and ∑D (see below for definitions). The BLAST search and the calculation of Z scores were performed as follows. Sequence alignments were obtained using WU BLAST 2 (Gish, W. (1996–2003) http://blast.wustl.edu) at the European Bioinformatics Institute (http://www.ebi.ac.uk/blast2/). The default options of the search and the SWALL database were used. The 500 alignments were used as provided by the MVIEW tool. No attempt was made to expurgate them from possible false positives; we assumed that, since sequence similarities with the query sequence were about 0.25 or higher, most of the proteins found in the search share the thioredoxin fold. Also, we did not try to expurgate the alignments from redundancies, although our choice of weights (see below) will give zero weight to any sequence that is identical with the query one. We quantified the frequency of occurrence of a given residue type (X, for instance, D or E) at a given position (p) in the set of sequences using the following score: 

$$Z_X(p) = \sum_s w_s(p,s)$$

where w_s(p,s) is a weight associated to residue type X in position p of a given sequence, s. This weight is given a value of zero if the residue present at position p in the sequence s is not of the X type. If the residue present at p in s is of the X type, then we consider two possibilities: (1) The residue present at position p in the query sequence (that of E. coli thioredoxin) is not of type X. Then, the weight is set equal to the sequence similarity between s and the query sequence (i.e. the number of matches in the sequence alignment of s to the query sequence divided by the number of residues in the latter). (2) The residue at p in the query is of type X. Then the weight is set equal to 1 minus the sequence similarity. It should be clear that our score basically counts the number of occurrences of a residue type at a given position, but that the occurrences are weighted according to sequence similarity with the query sequence. That is, our choice of weights is an attempt to take into account the fact that coincidences between high-similarity sequences most likely reflect that the corresponding proteins diverged recently in evolution. We note, however, that using unweighted numbers of occurrences as Z scores yields very similar ΔΔG_{D→E} versus ΔΔG_{E→D} correlations (results not shown). The sum in the equation defining the Z score extends to sequences within a specified range of sequence similarities with the query sequence.
pseudo-equilibrium hypothesis holds to some significant extent, we may expect the experimental values of the mutation effect on $\Delta G$ (i.e. the experimental $\Delta \Delta G_{D=E}$ values for several positions in the protein) to correlate with $\Gamma_{D=E}$. In fact, the correlation (including the 13 variants studied here) is rather good, as shown in Figure 2. It is also a robust correlation; that is, a good correlation is obtained when using a wide range of sequence similarity for $Z$ score calculation (for instance, the 1–0.3 range, which comprises 309 sequences; see Figure 2), but also when using a narrow range (for instance, the 1–0.6 range, which comprises only 23 sequences; see Figure 2). Finally, it is worth noting that the value for the slope of $\Delta \Delta G_{D=E}$ versus $\Gamma_{D=E}$ (approximately 0.35) is reasonably close to unity and that the value of the intercept is close to zero (see Figure 2).

The above results indicate that, as a first approximation, molecular evolution places D and E residues at the studied positions in a random-like manner, but with frequencies that roughly reflect the corresponding effects on protein stability. At this point, it is important to note that the contribution of a carboxylic acid residue to stability is position-dependent, as demonstrated by the different values of $\Delta \Delta G_{D=E}$ (and $\Gamma_{D=E}$) shown in Figure 2. This position dependence most likely reflects the interactions with the different environments of the D/E residues in the native structure of the protein (although, certainly, denatured-state effects cannot be ruled out: see Concluding remarks). However, the environment of a given group is not “static”, but will also change during the course of molecular evolution. Thus, under the pseudo-equilibrium hypothesis, we may expect the simultaneous presence of interacting groups to be favored in the course of evolution according to the stabilizing character of the interaction. In an attempt to detect such interactions in the sequence alignments, we have carried out chi-squared association tests as described in the legend to Figure 3. As an illustrative example, we show in Figure 3 the $\chi^2$ values for the association between position 9 (D in WT E. coli thioredoxin) and all other positions in the thioredoxin sequence. Many of the calculated $\chi^2$ values are higher than the value (3.84) corresponding to a 5% significance level. We assume that most of these statistically significant associations have no structural meaning and that they only reflect sequence similarity (i.e. two residues present in E. coli thioredoxin are also likely to be present in sequences showing significant similarity with that of thioredoxin). However, the two higher $\chi^2$ values (61 and 133) do correspond to residues (N63 and T66) that are very similar to D9 takes part. The set of sequences used for these $\chi^2$ calculations is defined by the 1–0.3 range of similarity with the query sequence of the BLAST search (that of E. coli thioredoxin).
Table 1. Potential residue–residue interactions involving glutamate and aspartate groups in E. coli thioredoxin suggested by the chi-squared association analysis of sequence alignments

<table>
<thead>
<tr>
<th>Residues</th>
<th>Distance between closest atoms (Å)</th>
<th>Closest atoms</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2–S1</td>
<td>1.30</td>
<td>N·–C</td>
<td>90.9</td>
</tr>
<tr>
<td>E48–K96</td>
<td>2.54</td>
<td>OE1·–NZ</td>
<td>60.5</td>
</tr>
<tr>
<td>D9–T66</td>
<td>2.58</td>
<td>OD2·–OG1</td>
<td>132.6</td>
</tr>
<tr>
<td>D104–Y49</td>
<td>2.87</td>
<td>OD1·–OH</td>
<td>67.3</td>
</tr>
<tr>
<td>D9–63N</td>
<td>2.98</td>
<td>N·–OD1</td>
<td>61.0</td>
</tr>
<tr>
<td>E44–K96</td>
<td>3.17</td>
<td>OE1·–NZ</td>
<td>56.7</td>
</tr>
<tr>
<td>E30–I60</td>
<td>6.84</td>
<td>N·–CG1</td>
<td>51.8</td>
</tr>
<tr>
<td>D61–P34</td>
<td>12.53</td>
<td>OD2·–CD</td>
<td>71.1</td>
</tr>
<tr>
<td>D15–I60</td>
<td>14.34</td>
<td>OD1·–N</td>
<td>59.7</td>
</tr>
<tr>
<td>E48–I75</td>
<td>21.82</td>
<td>OE2·–C</td>
<td>64.5</td>
</tr>
</tbody>
</table>

Interactions shown have $\chi^2$ values higher than 50 and are ranked according to the distance between the closest atoms in the native thioredoxin structure. See legend to Figure 3 for details on $\chi^2$ calculation.

Concluding remarks

The results we report here (Figures 1 and 2) provide clear evidence that, at positions not constrained for structural or functional reasons, mutations are accepted with probabilities related to their effects on stability. Besides clarifying the (poorly understood) mechanisms of fixation of mutations during evolution, these results should have immediate consequences for the interpretation of mutation effects on protein stability, since they strongly suggest that the environments of many groups in proteins may be optimized for stabilizing interactions to a remarkable degree of specificity (aspartate versus glutamate, for instance). We have also shown (Figure 3 and Table 1) that such stabilizing interactions may be detected in sequence alignments, a result which, as we elaborate below, may have implications for the derivation of structural information from comparative sequence analyses.

Only about half of the sequences in known genomes are homologous to proteins of known structure,20 prediction of structure from sequence is still a major issue in structural biology21 and the possibility of obtaining structural information (residue contacts) from the analysis of correlated mutations in sequence alignments has been suggested and explored. However, the results obtained so far using this approach do not seem to be overly encouraging. For instance, in a recent study22 using neural networks and comparatively large training and test data sets, an accuracy of only about 20% in assigning protein contacts was obtained, while an accuracy of, at least, 50% is thought to be required to recover the 3D structure.22 Our finding (Figure 3 and Table 1) that interactions responsible for contributions to stability of the order of a few kJ/mol may be detected in sequence alignments could perhaps lead to new strategies for the derivation of native-state structural information from comparative sequence analyses (note that the accuracy of the predictions in Table 1 is 60–70%) and, in addition, it does suggest the following intriguing possibility: that denatured-state interactions contributing significantly to stability may also be detected in sequence alignments. This possibility is admittedly speculative, but deserves attention since recent work has shown that stability effects associated with the denatured state can be quite dramatic. For instance, the pH-induced formation of hydrophobic clusters in the denatured state of a thermophilic ribonuclease H has been shown16 to cause a break (sudden change of slope) in the plot of denaturation temperature versus pH, indicating cooperative uptake of several protons in the denatured state. Also, single mutations have been shown17 to be able to disrupt the hydrophobic clustering in the denatured state of the same protein and, as result, to change drastically its stability curve (the plot of denaturation free energy versus temperature). In fact, a role for the denatured-state residual structure in tuning protein stability has been proposed recently.17 Clearly, it is not unreasonable that some correlations detected in analyses of sequence alignments be actually related with the denatured state. This could explain in part the comparatively low level of success in attempts to derive native-state structural information from analyses of sequence alignments and suggests that such analyses could perhaps yield useful information regarding the energetics and residual structure of denatured states.

Acknowledgements

This work was supported by Grant BIO2000-1437 from the Spanish Ministry of Science and Technology.
References


Edited by F. E. Cohen

(Received 24 September 2003; received in revised form 12 December 2003; accepted 12 December 2003)