doi:10.1006/jmbi.2001.5009 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **312**, 885–896

JMB

Prediction of Functionally Important Residues Based Solely on the Computed Energetics of Protein Structure

Adrian H. Elcock

Department of Biochemistry University of Iowa, Iowa City IA 52242-1109, USA

Catalytic and other functionally important residues in proteins can often be mutated to yield more stable proteins. Many of these residues are charged residues that are located in electrostatically unfavorable environments. Here it is demonstrated that because continuum electrostatics methods can identify these destabilizing residues, the same methods can also be used to identify functionally important residues in otherwise uncharacterized proteins. To establish this point, detailed calculations are performed on six proteins for which good structural and mutational data are available from experiments. In all cases it is shown that functionally important residues known to be destabilizing experimentally are among the most destabilizing residues found in the calculations. A larger scale analysis performed on 216 different proteins demonstrates the existence of a general relationship between the calculated electrostatic energy of a charged residue and its degree of evolutionary conservation. This relationship becomes obscured when electrostatic energies are calculated using Coulomb's law instead of the more complete continuum electrostatics method. Finally, in a first predictive application of the method, calculations are performed on three proteins whose structures have recently been reported by a structural genomics consortium.

Keywords: structural genomics; continuum electrostatics; prediction

Introduction

Structural genomics projects, which aim to solve large numbers of protein structures in rapid and largely automated processes,[1] have the potential for dramatically altering the way that the structural and functional properties of proteins are investigated. Until recently, the biochemical function and characteristics of a protein were usually known many years in advance of its structure being solved. Structural genomics initiatives promise to reverse this order of events: in future, structures of proteins will routinely be obtained before their functions are unambiguously determined. As a result of this, there is now an increased demand for methods capable of identifying a protein's function (or its functionally important residues) from examination of its structure.

Of course, for proteins for which there are evolutionarily related (i.e. homologous) sequences already known, functionally or structurally important residues can be identified simply on the basis of their degree of conservation across the family of aligned sequences.[2,3] However, for structural genomics applications, this approach will often not be useful: many proteins identified as prime candidates for structure determination are chosen precisely because they bear little sequence similarity to known structures, and therefore have an increased likelihood of adopting a novel fold.[4] Now, if the structure of such a target ultimately proves to bear strong similarity to an already known fold, then alignment of the structures can be used to identify important residues by analogy to the structurally similar partner.[5] This strategy will not work, however, if the protein adopts a truly novel fold. For such cases, it is clear that if functional sites or residues are to be predicted at all then this must be done based on an analysis of the structure alone.

There have been a variety of previous efforts in this direction. Herzberg & Moult[6] have shown that residues with backbone dihedral angles in strained

E-mail address of the author: adrian-elcock@uiowa.edu

0022-2836/01/040885–12 $35.00/0

conformations are often functionally important, proposing for example that they can be responsible for correctly aligning residues in the active sites of enzymes. Others have shown that the location of active-site residues can be identified by searching for sizeable cavities or clefts in the protein structure (large enough for substrates to bind).[7,8] Similar types of structural analyses, coupled with measures of surface properties such as hydrophobicity have also been used to identify sites on the surfaces of proteins that are involved in protein-protein interactions.[9,10] As a final example, Zhu & Karlin[11] have shown that statistically significant clusters of charged residues can be identified and often shown to be of structural or functional importance.

Here, a different approach is adopted. The aim is to exploit an interesting feature of functionally important residues: namely, that they often destabilize proteins. The most explicit demonstrations of the destabilizing effects of functional residues have come from site-directed mutagenesis experiments on enzymes, where several elegant studies have shown that mutation of catalytic residues can result in more stable (albeit now inactive) proteins.[12–14] Other studies have shown that residues involved in forming interfaces with other proteins or ligands can also be replaced to produce more stable proteins.[15,16] These results provide the basis for the first part of a hypothesis explored here: functional residues destabilize proteins. The second part of the hypothesis is that destabilizing residues can be identified by a purely computational analysis of the energetics of a protein's structure. Several studies on small proteins provide the support for this suggestion. Using related but different approaches, computational techniques have been used to identify charged residues located in unfavorable electrostatic environments. Subsequent site-directed mutagenesis experiments demonstrated that the residues identified could be converted to neutral or oppositely charged residues to yield proteins of increased stability.[17–19] These studies, together with some additional examples, have been the subject of a review by Sancluz-Ruiz & Makhatadze.[20]

Here, the intention is therefore to investigate whether functional residues can be predicted by simply finding charged residues that are calculated to electrostatically destabilize the protein. To do this, we first carry out continuum electrostatics calculations (see Methods) on proteins for which experimental results have already shown that residues known to be functionally important can be mutated to increase stability. These proteins serve as a benchmark for assessing the potential strengths and weaknesses of the computational methodology, providing an indication of whether destabilizing residues can be identified with confidence and, if so, whether these destabilizing residues are, on average, likely to be functional. In an attempt to demonstrate the validity of the idea in a more general setting, we then conduct a much larger scale analysis of 216 protein structures, and investigate the relationship between the calculated electrostatic energies of charged residues and the extent to which they are evolutionarily conserved. Finally, the method is used in a truly predictive scenario relevant to structural genomics, by application to three protein structures recently solved by a structural genomics consortium.[21]

Results

Detailed analysis of six proteins

As a first example of the potential for predicting functional residues on the basis of electrostatic energetic calculations, the case of retinoic acid binding protein, CRABP, was considered. Site-directed mutagenesis studies have indicated that two residues (R111 and R131) that are crucial for binding retinoic acid, also destabilize the protein.[15] Figure 1 shows the calculated electrostatic energies of each amino acid side-chain in CRABP. The charged and polar residues generally have positive values of ΔG_{elec}, indicating that from a purely electrostatic perspective, the side-chains are destabilized in the folded protein relative to solution. This is a common result of continuum methods,[19,22] the usual interpretation of which is that the favorable energetic contributions to folding come from the burial of hydrophobic groups.[22] The more important point here, however, is that of the four most destabilizing residues, two are the crucial R111 and R131 (Figure 1; Table 1). Of the other two strongly destabilizing residues, K30 is identified only because it is in an unusual compact conformation that places its charged NH_3^+ group in a

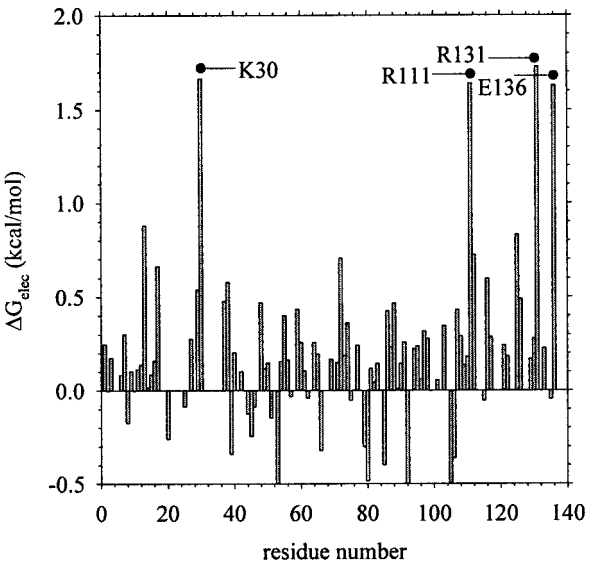

Figure 1. Calculated electrostatic free energies of folding for side-chains of residues in CRABP. Positive values indicate residues that electrostatically destabilize the protein.

Table 1. Destabilising residues in six characterized proteins

Protein	PDB code	No. of charged residues	Known destabilizing functional residues (rank)
CRABP	1cbi	40	R131 (1), R111 (3)
Barstar	1a19	24	E76 (1), E80 (2), D39 (9), D35 (15)
Barnase	1a2p	26	R87 (2), K27 (5), R59 (6), H102 (8)
RNase HI	1ril	45	D10 (1), D70 (4), D134 (8)
T4 lysozyme	2lzm	44	E11 (2), D20 (7)
Chicken lysozyme	1hel	27	D52* (2), E35* (4)

Calculated ranks of known destabilizing, functional residues in six proteins. Rank (given in parentheses) indicates the position of the residue when all residues are listed in order of decreasing ΔG_{elec} values. e.g. R111 (3) indicates that R111 is calculated to be the third most destabilizing residue of all residues in the protein. Asterisks (*) indicates that these residues are inferred to be destabilizing by analogy to corresponding residues in T4 lysozyme.

highly hydrophobic pocket formed by M27 and V31. This is almost certainly not a meaningful conformation: in the asymmetric unit of the crystal is a second monomer of CRABP in which K30 adopts a far more extended, solvent-exposed conformation that is calculated not to be destabilizing (data not shown). The fourth destabilizing residue, E136 is identified as unfavorable only because it is the final residue in the polypeptide chain and therefore interacts unfavorably with the adjacent carboxyl terminus. Both K30 and E136 are therefore false positives.

A second example of a protein stabilized by mutation of its ligand-binding residues is barstar, the intracellular inhibitor of the nuclease barnase. In the crystal structure of the barnase-barstar complex,[23] four negatively charged residues on barstar (D35, D39, E76 and E80) are present in the protein-protein interface. Charge-neutralizing mutations of all four residues have been shown to stabilize barstar, but particularly pronounced effects ($> \sim 1$ kcal/mol; 1 cal = 4.184 J) are only observed for mutations of E76 and E80.[16] Figure 2 shows the calculated electrostatic energetics of all side-chains in the crystal structure of uncomplexed barstar. E76 and E80 are calculated to be the two most destabilizing residues in the protein (Table 1). D39 is also found to be significantly destabilizing, although there are clearly several other residues that are more strongly destabilizing. With the possible exception of E32, none of these other residues is located close to the interface and so they are unlikely to be of functional importance. Interestingly, two uncharged residues are found to have strongly destabilizing effects. Q18 is involved in an unusual contact with E32 in which the hydrogen-bonding groups of the two residues are arranged in a face-to-face arrangement. Since this arrangement (which destabilizes both Q18 and E32) does not seem to be present in other crystal structures of barstar, Q18 appears to be a false positive identified not because of errors in the calculations, but because of idiosyncrasies of the structure itself. T63 forms the first part of a type I′ β-turn and is destabilized primarily because it is largely desolvated. Again, in calculations on another barstar structure (an NMR minimized structure; PDB code 1bta) this residue is not found to be particularly destabilizing.

Mutagenesis experiments similar to those performed on barstar have also been conducted on its target, barnase. Experimentally, three active-site residues, K27, R59 and H102, can be replaced by neutral residues to produce more stable proteins.[12] As in the above cases, when the crystal structure of uncomplexed barnase is analyzed, we find that these three residues are amongst the most electrostatically destabilizing residues in the protein (Figure 3; Table 1). A fourth residue involved in binding to barstar (R87) is also calculated to be one of the eight most destabilizing residues. Although the emphasis here is on identifying destabilizing residues, it is interesting to note in passing that two active-site residues (D54 and E73) that experimentally appear to stabilize the protein[12] are calculated to have favorable electrostatic energies (Figure 3). These two residues, together with E60, have all been shown to reduce the otherwise extremely fast rate of association of barnase with barstar,[24] and therefore from a functional perspec-

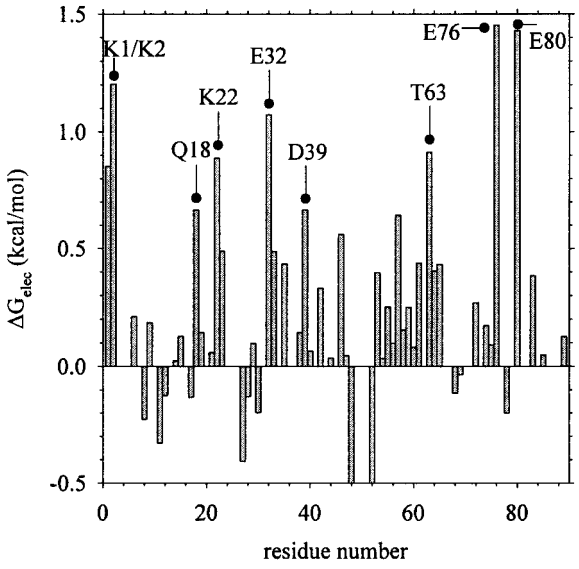

Figure 2. Calculated electrostatic free energies of folding for side-chains of residues in barstar.

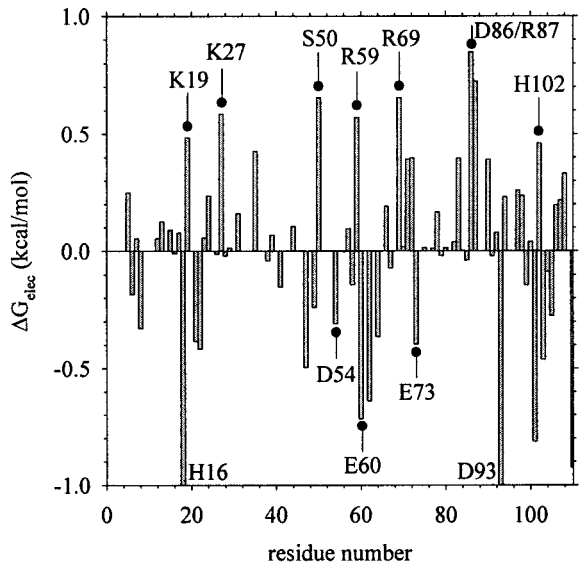

Figure 3. Calculated electrostatic free energies of folding for side-chains of residues in barnase.

tive might be considered detrimental. If one interprets this as suggesting that these residues must therefore be present for reasons of stability, it is encouraging to find that E60 is also calculated to have a favorable electrostatic energy.

A second nuclease that has been the subject of stability/activity studies is RNase HI. At least three active-site residues (D10, D70 and D134) involved in binding Mg^{2+} have been shown to destabilize the protein relative to charge-neutralizing mutants.[14] Again, these three residues are amongst the eight most electrostatically destabilizing residues identified (out of a total of 45 charged residues in the protein), and one of them, D10, is the most destabilizing residue found (Table 1). Since the crystal structure used for the calculations is of the Mg^{2+}-free form of the enzyme, this suggests that these residues are pre-organized, ready to bind the divalent cation, and that the free energy to pay for the pre-organization must be provided by the folding of other residues in the protein. Such an interpretation is consistent with Warshel's arguments concerning preorganization of enzyme active sites.[25,26]

Two final demonstrations of the potential for identifying functionally important residues are provided by phage T4 lysozyme and hen egg-white lysozyme. T4 lysozyme is one of the classic examples for which active-site mutations have been shown to stabilize the protein:[13] two residues implicated in catalysis (E11 and D20) have been shown to be replaceable by more stabilizing neutral residues. The calculations identify these two residues as being amongst the most electrostatically destabilizing residues: in this case, E11 is calculated to be the second most destabilizing residue out of a total of 44 charged residues in the protein (Table 1). Similar calculations performed on hen egg-white lysozyme also identify the corresponding residues (E35 and D52) as being destabilizing (Table 1). Experimentally, E35 is known to be destabilizing because its pK_a is elevated from the usual value of ~4.4 to 6.1 (discussed by Yang & Honig[27]). It is well known that shifting a pK_a value toward the physiological range of around 7 allows catalytic residues to participate more easily in acid-base reactions (for a review of factors affecting pK_a values, see Antosiewicz *et al.*[8]), but for the present context, it is worth noting that such shifts are not without cost to the overall stability of the protein.

A large-scale study

In some respects, the above examples provide only anecdotal evidence that functionally important residues can be identified by continuum electrostatics calculations. To establish the validity of the idea in a more general way requires a large-scale analysis of many structures. Such a study is complicated by the fact that it is difficult to develop an automated method for extracting information on which residues are functionally important: there is little agreement on how to denote such residues in the literature. In lieu of a more direct indicator, it was decided simply to interpret the degree of evolutionary conservation of a residue as a measure of its functional importance. This property has the advantage that it is relatively easy to quantify (see Methods) and requires only a suitable alignment of related protein sequences, which can be obtained in an automated fashion. Clearly, implicit in this approach is the assumption that there is a direct relationship between functional importance and evolutionary conservation. It should be remembered that there are prominent examples where this relationship fails: for example, the crucial, enzyme-binding residues of the family of proteins known as Kazal inhibitors, are actually the least conserved residues in the entire protein.[29]

For 216 protein structures, electrostatic energies of all the charged residues were calculated with the method used above. For the same residues, the degree of evolutionary conservation has been calculated in terms of its sequence entropy (see Methods): residues with lower sequence entropies are more highly conserved. For each protein, the charged residues were separately ranked in order of increasing entropy and decreasing electrostatic energy. Having obtained ranks for all residues in the protein, a histogram was constructed whose elements *i,j* contain the number of charged residues found with an electrostatic energy of rank *i*, and entropy of rank *j*. Summing results from all 216 proteins results in a histogram in which a total of 13,474 residues are represented. The degree to which the energy and entropy ranks are correlated in this histogram allows us to assess the general validity of our hypothesis.

Prior to conducting the analysis, it was hoped it would be found that highly conserved residues would have electrostatic energies that are either highly favorable or highly unfavorable. Obviously, those with favorable energies would be present for reasons of stability; those with unfavorable energies are hypothesized here to be responsible for function. In fact, such a relationship is indeed observed. Figure 4 shows the distribution of electrostatic energies of charged residues that rank amongst the top 10% in terms of conservation. Results are reported separately for proteins containing 25-49, 50-74, 75-99 and 100+ charged residues respectively: the close correspondence between the four distributions indicates that the relationship applies equally to proteins of any size. In all four cases, there is a clear tendency for the top 10% conserved residues to be either the most stabilizing or (especially) the most destabilizing residues in the protein.

More importantly in the present context, a biased distribution is also observed when the distribution of sequence entropies is plotted for the top 10% most destabilizing residues (Figure 5). According to our hypothesis, highly destabilizing residues, being functionally important, should be more likely to be conserved than to be non-conserved. Figure 5 shows that this is true: it therefore provides a clear demonstration that the hypothesis investigated here has validity on a broad scale. Having said that, it is also apparent that the extent of the effect is not as dramatic as we might have hoped: strongly destabilizing residues are only twice as likely to be found in the top 10% of conserved residues than in the bottom 10%. It is therefore worth questioning whether the signal that we observe is strong enough to be useful in a practical setting. This is difficult to demonstrate at this stage, but an indication that this is likely to be true is suggested by the fact that in 158 of the 216 protein structures (i.e. in 73%), at least one of the residues that ranked amongst the top 10% most destabilizing residues also ranked amongst the top 10% most conserved residues.

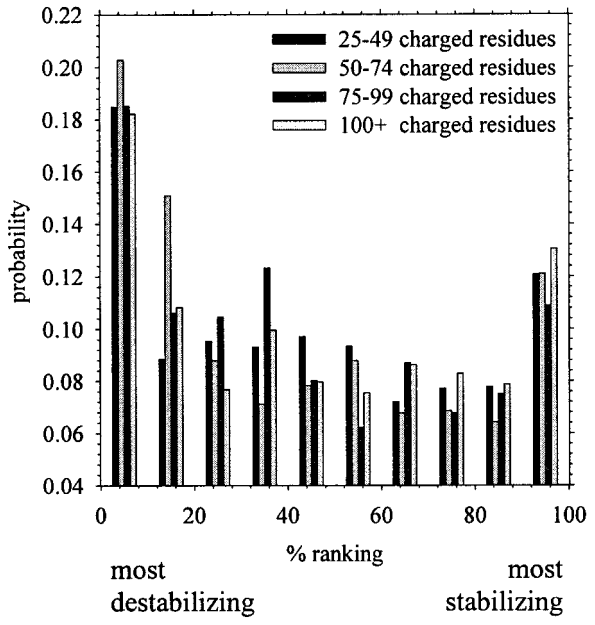

Figure 4. Histogram showing the distribution of energy ranks for the top 10% most conserved charged residues in proteins of varying sizes. Residues ranked in the 0-10% range are the most destabilizing residues, residues ranked in the 90-100% range are the most stabilizing.

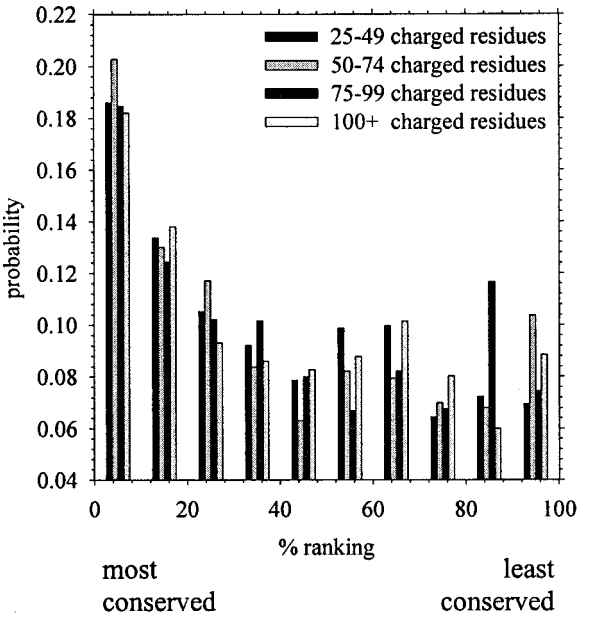

Figure 5. Histogram showing the distribution of sequence entropy ranks for the top 10% most destabilizing charged residues in proteins of varying sizes. Residues ranked in the 0-10% range are the most conserved residues, residues ranked in the 90-100% range are the least conserved.

Predictions for three new structures

A recent report from a structural genomics consortium has described the solution of the first ten structures of a set of 424 proteins selected from a thermophilic archaeon.[21] Seven of the structures could be assigned a function either on the basis of structural similarity to known proteins, or because bound cofactors such as NADH were identified. Three, however, remained with only very loose functional assignments; in order to extend our study into the academically riskier realm of actual prediction, each of these three proteins is examined in turn.

In the structure of protein MTH1184, four positively charged residues stand out as being strongly destabilizing (Table 2). Two of these (R11 and R22) map to a region of the protein surface that also

Table 2. Destabilizing residues in three uncharacterized proteins

Protein	PDB code	No. of charged residues	Most destabilizing residues (ΔG_{elec})
MTH1184	1gh9	27	R22 (1.8), K33 (1.5), R11 (1.3), R28 (1.2), R40 (0.7), E46 (0.7), R35 (0.7)
MTH1175	1eo1	29	K88 (1.3), K67 (1.0), N100 (0.9), D11 (0.9), E98 (0.8), K2 (0.8), R22 (0.7)
MTH538	1eiw	33	E91 (2.5), D106 (2.5), D102 (1.4), D109 (1.4), E105 (1.2), D40 (1.1), D16 (0.9)

List of residues calculated to be the most destabilizing in three uncharacterized proteins. Residues are listed in order of decreasing ΔG_{elec} values, given in parentheses in units of kcal/mol.

contains a high concentration of exposed hydrophobic side-chains: the combination of hydrophobic and destabilizing charged residues may well identify this region as a binding site for some other (unidentified) molecule (Figure 6(a)). That this region is highly likely to be functionally important is supported by the fact that it is also immediately adjacent to a multiple Cys motif (residues C7, C9, C24 and C26) that has already been suggested to be a metal-binding site (Figure 6(a)[21]).

A similar situation is also seen in the second of the three proteins, MTH1175, for which the two most destabilizing residues, K67 and K88 (Table 2), also map to a region occupied by several exposed hydrophobic residues (Figure 6(b)). Again, a betting scientist might suggest that this is a potential binding site for another molecule. Interestingly, MTH1175 also contains an unstructured arginine-rich C terminus that has been suggested to be a potential RNA-binding domain.[21] The five arginine residues in this tail (residues 119-124) do not, however, register as being particularly destabilizing in the calculations (Table 2). From an energetic perspective, this is actually a reasonable result: the interactions between the residues of an unstructured region will be more or less identical in both the folded and unfolded states, so unless they form different interactions with the remainder of the protein when folded or unfolded, they will not make any difference to the protein's stability. Nevertheless, the arginine-rich nature of the tail means that it may well be of functional importance. If it is, then it is not likely to be detectable on the basis of the destabilization criterion investigated here.

A highly charged tail is also found with the third protein, MTH538, a 111 residue protein that bears structural similarity to both the flavodoxin family of proteins and response regulator proteins of two-component bacterial signaling pathways such as CheY. Despite these similarities, the protein does not appear to bind flavins and lacks an aspartate residue near the structural position occupied by the phosphate-accepting D52 of CheY.[30] The single most destabilizing residue identified is D91 (Table 2), which interestingly is located adjacent to residues known to undergo chemical shift changes on binding of Mg^{2+} (residues 92, 95-97;[29]). D91 is destabilized primarily because of unfavorable electrostatic interactions with negatively charged residues concentrated in the C-terminal helix (disordered from residue 106 onwards), residues which themselves are destabilized (Table 2; Figure 6(c)). The destabilizing behavior of the helix/tail in these calculations contrasts with that observed for the arginine-rich tail of MTH1175, and probably reflects the fact that aspartate and glutamate residues have considerably shorter side-chains than arginine and are therefore less able to assume conformations in which unfavorable electrostatic interactions are avoided. Again however, similar unfavorable electrostatic interactions will probably also be present in the unfolded state of the protein, so it is unlikely that the tail residues themselves will strongly destabilize the protein. Because of this, it is probably fair to say that the tail residues stand out in the calculations only because of a shortcoming of the computational methodology; namely, in its idealized description of the unfolded state of the protein. It is, of course, tempting to suggest that the highly charged tail might be of some functional importance, but if it is, then its identification through the current computational methodology must be considered fortuitous.

Discussion

There are two key results presented here. First, from detailed calculations on six well-studied proteins, it has been shown that electrostatic energy calculations can be used to identify functionally important charged residues on the basis of their destabilizing effects. Second, in a broader study of 216 proteins, a clear signal was found that residues calculated to be destabilizing are also more likely to be evolutionarily conserved. This provides a novel, theoretically based extension of the conflicting relationship, previously identified in anecdotal experimental studies, between the requirements for stability and function in proteins.[12–16] The implication of both results is that calculations can be used to predict functionally important residues in otherwise uncharacterized structures. Of course, the calculations do not tell us what the specific functional roles of these residues are: they may be involved in binding or catalysis (or, less likely, some other function), but they do allow us to identify potentially interesting residues for further experimental study.

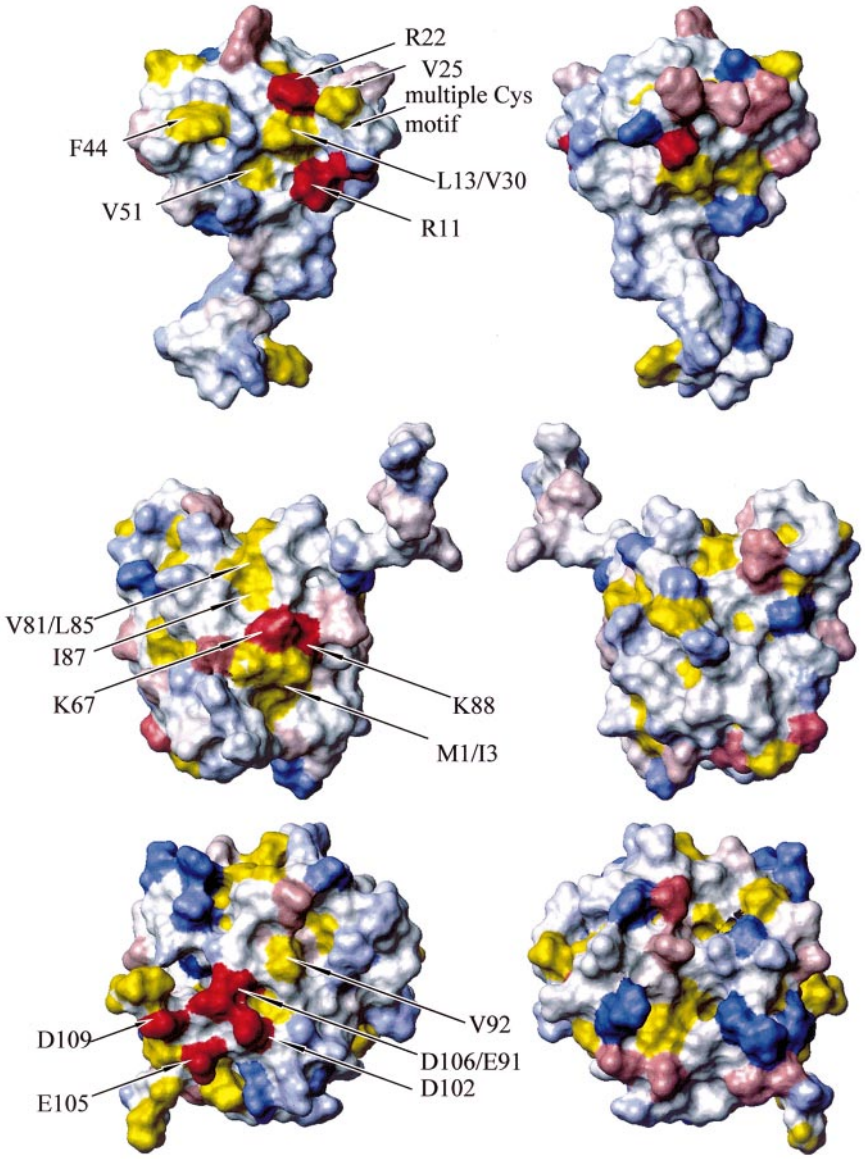

Figure 6. Structures of (top) MTH1184, (center) MTH1175, and (bottom) MTH538 with the protein surface coloured according to the ΔG_{elec} values of the residue side-chains. Red indicates strongly destabilizing residues, blue strongly stabilizing, and white residues with near-zero effect. Yellow indicates the side-chains of hydrophobic residues Ala, Ile, Leu, Met, Phe and Val. Two views of the proteins are presented: a 180° rotation around the vertical axis relates the two views.

The success of this work is crucially dependent on the method used to describe the electrostatic energetics of side-chains in proteins (see Methods). There are many aspects of the use of continuum electrostatics methods that remain under discussion, not least of which is the meaning and magnitude of the term ''dielectric constant'' as it applies to proteins.[31–33] Since the absolute values of ΔG_{elec} can be extremely sensitive to the choice of dielectric constant (e.g. see Hendsch & Tidor[34]), it is worth stressing the following point. In the current application, very accurate results are not actually required: all that is needed is to know the relative order of stabilities of the residues. In order not to confuse the main issue here, we have deliberately avoided trying to make quantitative comparisons between our calculated numbers and experimental effects of mutations on stability: in fact, there is no straightforward way of doing this, since our calculations do not directly correspond to ''real world'' mutations (for a discussion, see Spector *et al.*[19]). However, such a comparison would in any case miss the point presented here, which is simply to identify those residues in a structure that (in solely electrostatic terms) make the largest unfavorable (or least favorable) contributions to stability. Intuitively we expect that the relative ordering of residue stabilities would be much less sensitive to the details of the calculations than the absolute values of ΔG_{elec}.

In our opinion, the success of the continuum electrostatics method here can be attributed to the fact that it provides at least a qualitative description of the two factors that most affect the stability of charged residues in proteins. The first of these is the screened Coulombic interaction with other charged or polar residues, which may be unfavorable or favorable depending on the sign and proximity of other (partial) charges. The second is the unfavorable desolvation effect that inevitably results when a charged residue is removed from the aqueous environment and placed in the protein interior. It turns out that both contributions can be important for correctly identifying destabilizing residues, and because of this, a simpler electrostatic model such as Coulomb's law may not be as useful as the present methodology for identifying destabilized residues, even though it will probably work well in cases where desolvation effects are unimportant (e.g. see Grimsley *et al.*[17]). We base this assertion on two results. First, for several of the six proteins studied in detail here, desolvation effects are far from negligible: in fact, calculations using Coulomb's law conspicuously fail to identify the key residues of T4 lysozyme, hen egg-white lysozyme and CRABP (data not shown). Second, when the large-scale analysis of 216 structures is repeated using Coulomb's law, the relatively strong signal seen in Figure 5 is destroyed: the most destabilizing residues do not now show a tendency to be more conserved (see Figure 7). That said, it may well prove possible to usefully employ Coulombic calculations if they are combined with some other means of describing desolvation effects, such as measuring (or perhaps visually assessing) the degree of burial of a residue. The advantage of the present methodology of course is that it provides a unified framework for estimating both factors in a single set of calculations.

For the six proteins that we have studied in detail, it is apparent that in addition to the known functional residues, several other residues are identified as being destabilizing. Since the potential for predicting truly functional residues depends on the method's ability to adequately separate out signal from noise, it is clearly important to ask whether these additional residues represent false positives. The answer may be partly a matter of definition. If the calculations identify a residue that in reality (i.e. in experiments) turns out to stabilize the protein, then this is an unambiguous false positive. Many of the residues identified here as being destabilizing have not been subjected to mutagenesis studies. We can, however, suggest that a significant proportion will almost certainly turn out to be false positives, and that this can arise because of inadequacies in the computational methodology and because of idiosyncrasies of the structure used in the calculations. In terms of methodology, a clear shortcoming is suggested by the results obtained for the admittedly unusual case of the charged tail of MTH538. The present methodology assumes that side-chains do not interact with each other in the unfolded state. In future, more elaborate treatments of the unfolded state might prove to be useful for dealing with unstructured or highly charged regions.[35] In terms of structural idiosyncrasies, it is clear even from the few structures that we investigate in detail, that not all side-chains in protein structures are correctly positioned; this is in line with large-scale assessments of the quality of protein structures.[36] One way to decrease the sensitivity of results would be to use multiple structures in the analysis. These are routinely provided for proteins solved by NMR techniques (and are utilized here in the calculations on the structural genomics proteins) but can also be obtained for other structures by using molecular dynamics (MD) simulations to sample a range of potential configurations.

In addition to potentially labeling a stabilizing residue as being destabilizing, the calculations may also identify a residue that is truly destabilizing, but which is not functionally important. In such a case, the calculations are successful in one sense (identifying destabilizing residues), but unsuccessful in another (identifying functional residues). One way of providing an independent check on the likely functional importance of a residue in the absence of direct experimental evidence is to examine whether it is subject to a high degree of evolutionary conservation. When we use this criterion in a large-scale analysis of proteins we obtain evi-

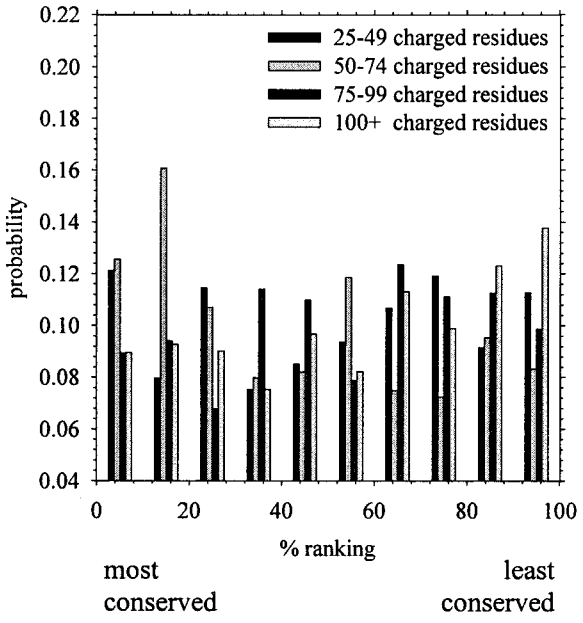

Figure 7. Histogram showing the distribution of sequence entropy ranks for the top 10% most destabilizing charged residues in proteins of varying sizes, but with electrostatic energies calculated using Coulomb's law. Residues ranked in the 0-10% range are the most conserved residues, residues ranked in the 90-100% range are the least conserved.

dence that destabilizing but non-functional charged residues may actually be quite common: although there is a clear preference for highly destabilizing residues to be more conserved, the magnitude of the effect is not dramatic (Figure 5). This is not an unreasonable result: most proteins will be able to tolerate several destabilizing residues without adverse consequences (i.e. unfolding). One interesting direction for the future will be to investigate whether tolerated-but-unimportant destabilizing residues can be more easily identified (and the signal in Figure 5 amplified) by examining hyperthermophilic proteins, which are likely to be much less able to accept such residues. In support of this idea, one recent study has already identified two destabilizing residues that are present in a mesophilic protein but absent from its hyperthermophilic cousin[37] and evidence has been presented that electrostatic interactions in hyperthermophilic proteins in general appear to be more favorable than those in mesophiles.[38–40]

Finally, it is worth noting that energetically unfavorable residues include not only those that are destabilized electrostatically, but also residues in strained conformations and (often, but not necessarily) residues with exposed hydrophobic groups. The potential importance of the latter type of residues has been repeatedly demonstrated, as they are often implicated as binding sites for other molecules.[9–10] The mapping of the electrostatically destabilizing residues to the protein surface, where their proximity to hydrophobic groups can be assessed (Figure 6), can be considered a first step toward combining the two types of information. It is also worth noting that an alternative method of describing residue energetics has been developed by Freire and co-workers. In their model, the energetics are equated (primarily) with the degree to which the residue is buried. When combined with an intelligent and computationally intensive sampling procedure, their method also appears useful for identifying stabilizing and destabilizing residues, and by implication, binding sites in proteins.[41] An approach that combines the detailed electrostatics of the present approach with the surface area-based energetics of the above, may ultimately prove to be of even better predictive ability.

Methods

Throughout this work, electrostatic energies of charged and polar residues are calculated using continuum electrostatics methods (for a review, see Honig & Nicholls[42]). These methods, which are based on the Poisson-Boltzmann formalism, are commonly used to investigate questions of protein stability.[19,34,39,40,43] The approach that is followed here closely mirrors these studies and in particular is more or less identical with that used by Tidor and co-workers, to whose work the reader is referred for further details.[19,34] Briefly, calculations are performed with the aim of quantifying the change in electrostatic free energy ΔG_{elec}, that results when a given amino acid side-chain is transferred from aqueous solution (in which state it is assumed not to interact with other side-chains) into the fully folded protein. Standard expressions are used for calculating the electrostatic free energies in the two environments. Subtracting the two results gives a value for ΔG_{elec}, relative to that obtained in a reference state in which the residue that has all its partial charges set to zero (see Spector *et al.*[19] and Elcock *et al.*[33] for further discussion of this subject).

Structures of all the proteins investigated here were obtained from the Protein Data Bank†. Missing side-chains and hydrogen atoms were added to all structures using the WhatIf web interface[44]‡: the correct positioning of hydrogen atoms can be crucial for reasonable results with continuum methods.[45] All aspartate, glutamate, lysine and arginine residues were assumed to be in their charged forms, whilst the protonation states of histidine residues were assigned on the basis of their potential for hydrogen bonding interactions. Electrostatics calculations were performed with the finite difference PB program UHBD,[46] with partial charges and radii for all atoms being assigned from the PARSE parameter set.[47] Note that since this parameter set assigns zero charges to all atoms of hydrophobic side-chains, the electrostatic folding free energies of all such side-chains are zero. The ionic strength was set to 100 mM, and solvent and protein dielectric constants were set to 78.40 and 12.0, respectively. The latter value is in line with the results of recent analyses of the effects of placing charged amino acid side-chains in highly desolvated environments.[48] Depending on the size of the protein, calculations were performed using an initial grid of dimensions 50 Å×50 Å×50 Å×4-6 Å, followed by four successive ''focusing'' steps to a final grid resolution of 0.25 Å. For proteins solved by NMR techniques, the analysis was carried out on all structures and the calculated energies for each residue in turn were Boltzmann-weighted. Since the calculated ΔG_{elec} values can often be sensitive to idiosyncrasies of a solved structure (see Results), this weighting ensures that the results are not overly affected by the presence in the sample of a single structure in which a residue finds itself in a highly energetically unfavorable environment.

To test the likely validity of the hypothesis that electrostatically destabilizing charged residues are important for function, a large-scale analysis of 216 proteins was conducted. The structures chosen for analysis are a subset of 969 non-redundant structures identified by Nussinov's group§.[49] Because of maddening inconsistencies in the way that residues are numbered in PDB files, it is often difficult to correctly match residues with their entries in a sequence file. Accordingly, structures for which discrepancies arose were eliminated (this usually involved structures for which residues are missing in the PDB file). Also eliminated from consideration were structures with very few similar sequences (see below), and structures with fewer than 25 charged residues. The remaining 216 structures and chain identifiers are listed in Table 3. For each protein, similar sequences were identified by a BLAST search[50] of the non-redundant database.[51] Following this search, multiple sequence alignments for each protein were constructed using CLUSTALW.[52] Finally, the sequence entropy, s, at each position in the sequence was then calculated as described:[53]

† Available online at http://www.rcsb.org

‡ Available online at http://cmbi1.cmbi.kun.nl:1100/WIWWWI

§ Available online at http://protein3d.ncifcrf.gov/tsai

Table 3. Proteins investigated in large-scale analysis

135l:	1bdm:B	1div:	1gof:	1lxa:	1pfk:A	1svb:	2abl:
1aa6:	1bkf:	1dkx:A	1gpl:	1lya:B	1pfx:L	1tbd:	2ayh:
1aab:	1bmt:A	1dmo:	1gri:A	1mco:L	1php:	1tbr:R	2aza:A
1ab3:	1bp1:	1doi:	1gtp:A	1mhl:C	1pii:	1tco:B	2baa:
1abr:B	1bst:	1dor:A	1hdj:	1mka:A	1plq:	1tcr:A	2bmh:A
1ac5:	1bv1:	1dos:A	1hge:A	1mng:A	1poc:	1tf4:A	2eng:
1aca:	1bvd:	1dpe:	1hjr:A	1mse:C	1pox:A	1thv:	2kau:A
1acp:	1cau:B	1dpg:A	1hmy:	1mut:	1prc:M	1tmc:A	2lbp:
1add:	1cdb:	1eaf:	1hrd:A	1nah:	1pyt:A	1tnw:	2mhr:
1adn:	1cel:A	1ede:	1hsa:A	1ncx:	1qrd:A	1tpf:A	2min:B
1ads:	1cew:I	1edg:	1htp:	1nhk:R	1quk:	1tsy:	2nll:B
1afw:B	1cfb:	1eft:	1hul:A	1nsy:A	1rcb:	1u9a:A	2pia:
1aih:A	1cfc:	1efu:B	1ice:A	1ntr:	1rcf:	1ubs:A	2pii:
1ajs:A	1cfp:A	1etp:A	1ige:A	1nzy:A	1req:A	1urk:	2pol:A
1akz:	1chm:A	1fba:A	1ill:G	1obw:A	1rip:	1vap:A	2rmc:A
1alk:A	1cid:	1fd2:	1iow:	1occ:B	1rmv:A	1vmo:A	2stt:A
1alo:	1cmf:	1fli:A	1irp:	1occ:E	1roe:	1vps:B	2tct:
1amm:	1cpc:B	1fmk:	1jbc:	1ofg:A	1rpa:	1whi:	2tmd:A
1anu:	1crk:A	1fnf:	1jcv:	1ois:	1rsy:	1wht:A	2tmn:E
1aor:A	1cse:E	1fyc:	1jet:A	1one:A	1rtc:	1wio:A	2trx:A
1aov:	1csh:	1gca:	1klo:	1opg:H	1rvv:1	1xgs:A	2vik:
1aoz:A	1ctt:	1gd1:O	1ktq:	1opr:	1ryt:	1yge:	3cla:
1aps:	1cyu:	1gdt:A	1kva:	1ord:A	1sac:A	1yna:	3pmg:A
1apy:B	1dad:	1ghr:	1lam:	1ort:A	1scu:A	1ytb:A	4aah:A
1ast:	1dea:A	1gky:	1lcf:	1osa:	1shc:A	1znb:B	4rhn:
1asu:	1def:	1gln:	1luc:A	1pam:A	1sly:	1zxq:	5p21:
1axn:	1dim:	1glq:A	1lut:	1pbn:	1smd:	2abk:	8ruc:I

List of 216 proteins investigated in a large-scale analysis. Protein Data Bank codes are followed by the chain identifier (where applicable) of the chain for which calculations were conducted.

$$s = \sum p(i)\ \ln(p(i)) \qquad (1)$$

where $p(i)$ is the probability that the position in the sequence is occupied by a residue of type i. In order to distinguish conservative mutations (e.g. the exchange of isoleucine by valine) from non-conservative changes (e.g. the replacement of isoleucine by arginine), we divide the 20 amino acids into the following six groups: (1) Arg, Lys; (2) Asp, Glu; (3) His, Phe, Trp, Tyr; (4) Asn, Gln, Ser, Thr; (5) Ala, Ile, Leu, Met, Val, Cys; (6) Gly, Pro. This is the grouping scheme used by Mirny & Shakhnovich.[54] The summation in equation (1) is therefore conducted over six residue groups, not over 20 residue types.

Acknowledgments

The author is grateful for financial support from the Irene Wells Medical Research Fund at the University of Iowa, and thanks an anonymous reviewer for insightful comments.

References

1. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T. *et al*. (1999). Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23**, 151-157.
2. Volz, K. (1999). A test case for structure-based functional assignment: the 1.2 Å crystal structure of yjgF gene product from *Escherichia coli*. *Protein Sci.* **8**, 2428-2437.
3. Zvelebil, M. J. J. M. & Sternberg, M. J. E. (1988). Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* **2**, 127-138.
4. Brenner, S. E. & Levitt, M. (2000). Expectations from structural genomics. *Protein Sci.* **9**, 197-200.
5. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374-382.
6. Herzberg, O. & Moult, J. (1991). Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins: Struct. Funct. Genet.* **11**, 223-229.
7. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438-2452.
8. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884-1897.
9. Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.
10. Jones, S. & Thornton, J. M. (1997). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133-143.
11. Zhu, Z.-Y. & Karlin, S. (1996). Clusters of charged residues in protein three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **93**, 8350-8355.
12. Meiering, E. M., Serrano, L. & Fersht, A. R. (1992). Effect of active-site residues in barnase on activity and stability. *J. Mol. Biol.* **225**, 585-589.
13. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA,* **92**, 452-456.
14. Kanaya, S., Oobatake, M. & Liu, Y. (1996). Thermal stability of *Escherichia coli* ribonuclease HI and its active site mutants in the presence and absence of the Mg^{2+} ion - proposal of a novel catalytic role for Glu(48). *J. Biol. Chem.* **271**, 32729-32736.
15. Zhang, J. H., Liu, Z.-P., Jones, T. A., Gierasch, L. M. & Sambrook, J. F. (1992). Mutating the charged resi-

dues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability. *Proteins: Struct. Funct. Genet.* **13**, 87-99.
16. Schreiber, G., Buckle, A. M. & Fersht, A. R. (1994). Stability and function: two constraints in the evolution of barstar and other proteins. *Structure,* **2**, 945-951.
17. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M. P., Thurlkill, R. L. *et al*. (1999). Increasing protein stability by altering long-range coulombic interactions. *Protein Sci.* **8**, 1843-1849.
18. Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M. & Makhatadze, G. I. (1999). Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface. *Biochemistry,* **38**, 16419-16423.
19. Spector, S., Wang, M., Carp, S. A., Robblee, J., Hendsch, Z. S., Fairman, R. *et al*. (2000). Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry,* **39**, 872-879.
20. Sanchez-Ruiz, J. M. & Makhatadze, G. I. (2001). To charge or not to charge? *Trends Biotechnol.* **19**, 132-135.
21. Christendat, D., Yee, A., Dharamis, A., Kluger, Y., Savchenko, A., Cort, J. R. *et al*. (2000). Structural proteomics of an archaeon. *Nature Struct. Biol.* **7**, 903-909.
22. Yang, A.-S. & Honig, B. (1995). Free-energy determinants of secondary structure formation. I. Alpha helices. *J. Mol. Biol.* **252**, 351-365.
23. Buckle, A. M., Schreiber, G. & Fersht, A. R. (1994). Protein-protein recognition - crystal structural analysis of a barnase barstar complex at 2.0 Å resolution. *Biochemistry,* **33**, 8878-8889.
24. Schreiber, G. & Fersht, A. R. (1996). Rapid, electrostatically assisted association of proteins. *Nature Struct. Biol.* **3**, 427-431.
25. Warshel, A. (1978). The energetics of enzymatic reactions. *Proc. Natl Acad. Sci. USA,* **75**, 5250-5254.
26. Warshel, A. (1998). Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **273**, 27035-27038.
27. Yang, A.-S. & Honig, B. (1993). On the pH dependence of protein stability. *J. Mol. Biol.* **231**, 459-474.
28. Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1996). The determinants of p*K*(a)s in proteins. *Biochemistry,* **35**, 7819-7833.
29. Apostol, I., Giletto, A., Komiyama, T., Zhang, W. L. & Laskowski, M. (1993). Amino acid sequences of ovomucoid 3rd domains from 27 additional species of birds. *J. Protein Chem.* **12**, 419-423.
30. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H. & Kennedy, M. A. (2000). Structure-based functional classification of hypothetical protein MTH538 from *Methanobacterium thermoautotrophicum*. *J. Mol. Biol.* **302**, 189-203.
31. Sham, Y. Y., Muegge, I. & Warshel, A. (1998). The effect of protein relaxation on charge-charge interactions and dielectric constants of proteins. *Biophys. J.* **74**, 1744-1753.
32. Simonson, T. (2001). Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **11**, 243-252.
33. Elcock, A. H., Sept, D. & McCammon, J. A. (2001). Computer simulation of protein-protein interactions. *J. Phys. Chem. ser. B,* **105**, 1504-1508.
34. Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins - a continuum electrostatic analysis. *Protein Sci.* **3**, 212-226.
35. Elcock, A. H. (1999). Realistic modeling of the denatured states of proteins allows accurate calculations of the pH dependence of protein stability. *J. Mol. Biol.* **294**, 1051-1062.
36. Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). Errors in protein structures. *Nature,* **381**, 272.
37. Perl, D., Mueller, U., Heinemann, U. & Schmid, F. X. (2000). Two exposed amino acid residues confer thermostability on a cold shock protein. *Nature Struct. Biol.* **7**, 380-383.
38. Elcock, A. H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermostable proteins. *J. Mol. Biol.* **284**, 489-502.
39. Xiao, L. & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.* **289**, 1435-1444.
40. Kumar, S., Ma, B. Y., Tsai, C. J. & Nussinov, R. (2000). Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins: Struct. Funct. Genet.* **38**, 368-383.
41. Luque, I. & Freire, E. (2000). Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins: Struct. Funct. Genet.* **Suppl. 4**, 63-71.
42. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science,* **268**, 1144-1149.
43. Lounnas, V. & Wade, R. C. (1997). Exceptionally stable salt bridges in cytochrome P450cam have functional roles. *Biochemistry,* **36**, 5402-5417.
44. Rodriguez, R., Chinea, G., Lopez, N., Pons, T. & Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics,* **14**, 523-528.
45. Nielsen, J. E., Andersen, K. V., Honig, B., Hooft, R. W. W., Klebe, G., Vriend, G. & Wade, R. C. (1999). Improving macromolecular electrostatics calculations. *Protein Eng.* **12**, 657-662.
46. Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E., Luty, B. A., Ilin, A. *et al*. (1995). Electrostatics and diffusion of molecules in solution - simulations with the University of Houston Brownian dynamics program. *Comput. Phys. Commun.* **91**, 57-95.
47. Sitkoff, D., Sharp, K. A. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978-1988.
48. Dwyer, J. J., Gittis, A. G., Karp, D. A., Lattman, E. E., Spencer, D. S., Stites, W. E. *et al*. (2000). High apparent dielectric constants in the interior of a protein reflect water penetration. *Biophys. J.* **79**, 1610-1620.
49. Tsai, C. J., Lin, S. L., Wolfson, H. & Nussinov, R. (1997). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **260**, 604-620.
50. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
51. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2000). GenBank. *Nucl. Acids Res.* **28**, 15-18.
52. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of

progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.
53. Elcock, A. H. & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl Acad. Sci. USA,* **98**, 2990-2994.
54. Mirny, L. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.

Edited by B. Honig

(Received 18 June 2001; received in revised form 27 July 2001; accepted 27 July 2001)