Local electrostatic optimization in proteins

Maureen Toner Oliva and John Moult

Center for Advanced Research in Biotechnology, University of Maryland
Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850, USA
1Present address: 8309 Country Oaks Station, West Chester, OH 45069, USA
2To whom correspondence should be addressed

A simple electrostatic model has been used to investigate the extent to which the structure of protein molecules is organized to optimize the internal electrostatic interactions. We find that the model provides a favorable total intra-protein electrostatic energy for almost all polar and charged groups of atoms, suggesting a high degree of structural optimization. By contrast, a significant fraction of individual group–group interactions are found to be unfavorable. An analysis as a function of the range of interactions included shows the electrostatic organization is generally relatively short range (up to 6 or 7 Å between group centers). Although the model is very simple, it is useful for assessing the overall quality of protein experimental structures, for pin-pointing some types of errors and as a guide to improving protein design.

Keywords: protein design/protein electrostatics/strain/structure quality

Introduction

It is a common observation that protein molecules are characterized by many favorable internal electrostatic interactions, particularly hydrogen bonds (Baker and Hubbard, 1984; Stickle et al., 1992; McDonald and Thornton, 1994; Bordo and Argos, 1994) involving both backbone and side chains, and salt bridges between oppositely charged groups (Bryant and Lawrence, 1991; Spassov and Atanasov, 1994). Attention has also been drawn to the role of non-hydrogen bond short range electrostatic interactions in stabilizing particular types of secondary structure (Avbelj and Moult, 1995; Maccallum et al., 1995a,b), and to the presence of specific types of unfavorable interactions (Avbelj and Moult, 1995). The extent to which the internal electrostatic interactions in proteins are favorable is not clear. If protein molecules have sufficient degrees of freedom, structures could be organized such that all electrostatic groups have a favorable energy. If not, there must be some local strain. An investigation of the steric strain associated with main chain torsion angles (Herzberg and Moult, 1991) has shown that strain is rare for this contribution to the energy, and that when it does occur it is usually associated with regions of the structure directly involved in function. We now ask whether there exists an analogous rule for strain in internal electrostatic interactions in proteins.

The rigorous treatment of electrostatic interactions in proteins remains a challenging problem in computer modeling. The complex charge distribution around the atoms, its modification by the molecular environment (polarization) and by charge transfer (Moult et al., 1985) and the often dominant role of solvent effects demand sophisticated treatment in order to obtain accurate quantitative results. Many approaches to accurately treating both the charge distributions effects (for example, Dudek and Ponder, 1995) and solvent (Davis and McCammon, 1990; Sharp and Honig, 1990; Warshel and Aqvist, 1991) are being aggressively pursued. For the present purposes, we wish to use the simplest possible representation of the electrostatics that will capture the properties we are interested in. We therefore include only intra-protein energy contributions using a conventional point charge model with no polarization, and neglect interactions with the solvent. Only short range interactions are considered, to focus on the local electrostatic organization. The energy calculated from the model thus represents only one contribution to the total electrostatic free energy. Throughout the paper, the word ‘energy’ should be taken to mean ‘local intra-protein electrostatic energy’. This partial electrostatic energy is the component that most directly captures the extent to which the local three-dimensional structure is organized to provide a favorable electrostatic environment, and therefore is appropriate for our purposes. In this respect, it is a generalization of the partial electrostatic energy used to define main chain hydrogen bonds in some secondary structure identification algorithms (Kabsch and Sander, 1983). In addition to providing insight into electrostatic organization, the model is shown to provide an effective and convenient diagnostic tool.

We will see later that there is a large degree of local optimization of electrostatic interactions in proteins. As in the case of rare occurrences of steric strain (Herzberg and Moult, 1991), probing the limits of this organization requires careful consideration of the role of artifacts in the experimental structures. We therefore concentrate on a thorough analysis of a few high resolution structures to establish the overall picture.

The layout of the paper is as follows: in the next section, we describe the electrostatic model, discuss the avoidance of crystallographic artifacts and introduce the protein structures used in the analysis and tests. An analysis of the total energy of polar and charged groups and individual group–group energies in some high quality protein structures is then presented, leading to the main conclusions about electrostatic organization. Next, the model is used to compare the electrostatic energy distributions in pairs of correct and incorrect crystallographic structures, low and high resolution versions of the same structure, NMR and crystallographic versions of a structure and to identify and correct apparent errors in a protein design. Finally, we summarize the conclusions of the paper, and discuss their relationship to other work in the field.

Materials and methods

Point charges and electrostatic groups
The common approximation of partial atomic point charges at the atomic centers is used. As in many empirical force fields,
the charge values are restrained to provide a net charge of approximately zero or an integral number of electrons over a group of atoms (Hünenberger and van Gunsteren, 1997). For example, the dipole formed by the carbon and oxygen atoms of a backbone carbonyl group has a total charge of zero, and the charged guanidino group of arginine carries a net positive charge, with a magnitude of approximately one electron, distributed over its atoms under normal physiological conditions. The charge distribution is restricted to the more polar atoms, primarily oxygen and nitrogen, and the hydrogen atoms associated with these, as well as some neighboring aliphatic carbons, in the common ‘polar hydrogen’ approximation. All other atoms in the molecule, that is, most aliphatic and aromatic carbons and the associated hydrogens, are considered electrostatically neutral and make no contribution to this analysis. The charges are from a version of the VFF force field (Dauber-Osguthorpe et al., 1988). Group and charge definitions may be found at http://moult.carb.nist.gov/es_analysis. The analysis focuses on the environment and interactions of the dipole and charged groups so defined, referred to as ‘groups’ in the rest of the paper. The main chain peptide link, formed by the carbonyl and amide groups between two successive Cα atoms of the polypeptide backbone, is treated as a single group.

**Electrostatic model**

The partial electrostatic interaction energy between a pair of polar or charged groups i and j is calculated using Coulomb’s law with a dielectric constant of unity, in the usual manner:

\[ E_{ij} = K \sum q_i q_j / r_{ik} \]

where the sums are over all atoms k of group i and atoms l of group j, \( q_k \) and \( q_l \) are the partial atomic charges in electrons, and \( r_{ik} \) is the distance between atoms i and k in Ångströms. K is the scaling constant (332) converting the energies to kcal/mol. Interactions between a pair of groups are included if the centers of charge are less than a cut-off distance \( d_c \) apart. The center of charge of a group \( \mathbf{r}_c \) is defined as:

\[ \mathbf{r}_c = \sum q_i \mathbf{r}_i / \sum q_i \]

The effect of different cut-off distances, \( d_c \), is considered below. In most cases, atomic co-ordinates are taken from X-ray crystal structures. Intermolecular protein interactions in the crystal were not generally included, but the effect of these on abnormally unfavorable energies was investigated. Crystallographically bound ions are included. Their effect on electrostatics is discussed.

**Treatment of crystallographic co-ordinates**

**Addition of hydrogens to polar and charged groups**

With the exception of a few very high resolution and neutron studies, crystallography is not able to identify the positions of hydrogen atoms in proteins. Most of the required polar hydrogen positions were added using considerations of covalent geometry and steric factors in the standard manner, as follows. Hydrogens on the nitrogens of backbone amide groups, asparagine and glutamine side chains, the guanidino groups of arginine and the imidazole groups of histidine residues have positions determined by the covalent bond. For neutral histidine residues, a single hydrogen was added to either the ND1 or ND2 atom, according to which gave the most favorable interaction energy with the environment. Similarly, for a neutral carboxyl group, a hydrogen was added to the oxygen resulting in the more favorable total interaction energy. Hydrogens were built onto charged amino groups such that they were staggered with respect to the previous carbon atoms in the side chain. Hydrogens on the hydroxyl groups of threonine and serine cannot be positioned by covalent and steric factors alone, and were aligned to give the maximum favorable interaction energy with the local electric field, in a manner similar to that of Brünger and Karplus (1988). For tyrosine, the hydrogen of the hydroxyl group was positioned in the plane of the phenyl ring, on the side giving the most favorable interaction with the electric field. Hydroxyl hydrogen positions were adjusted iteratively to obtain a self-consistent set of positions.

**Re-orientation of the side chains of asparagine, glutamine and histidine residues**

Because of the small difference in the electron density of oxygen, nitrogen and carbon atoms, there are the two possible orientations of these side chains compatible with an electron density map. In the few cases where the electrostatic energies were found to be unfavorable, these groups were rotated 180°. These changes are noted in the results.

**Selection of protonation states for aspartic acid, glutamic acid and histidine side chains**

Charges were assigned to these residues according to the standard pK values and the reported pH of the crystals. Occasional uncertainties of charge, for aspartic and glutamic acid side chains in the pH range 4–5, and for histidine in the range 6–8, are noted.

**Avoidance of crystallographic artifacts**

Generally only high resolution (2 Å or better), well refined (R factor less than 20%) structures were examined. In these structures, significant errors are rare, and generally restricted to atoms with high temperature factors. Thus, when unfavorable group energies were encountered, the temperature factors of the groups concerned are checked.

**Protein structures considered**

**Streptomyces griseus protease A (SGPA)**

Streptomyces griseus protease A is a bacterial serine protease of the chymotrypsin family, with a single polypeptide chain of 181 residues. Co-ordinates were taken from the protein data bank file 2SGA (Sielecki et al., 1979; Moult et al., 1985). These crystals were grown at pH 4.1 from 1 M sodium phosphate. The resolution is 1.5 Å and the final R factor after Hendrickson–Konnert refinement (Hendrickson, 1985) was 13%. Thirty-one of the 265 polar and charged groups have one or more atoms with temperature factors higher than 25 Å². The side chains of Q122 and Q182 were rotated 180° about \( \chi_1 \) to obtain negative total energies for the corresponding polar groups. One side chain, that of R221, had not been assigned crystallographic co-ordinates because of disorder, and was built in a fully extended conformation. A sodium ion that forms an inter-protein molecule bridge in the crystals was included in the electrostatic calculations.

**Bovine trypsin, pH 8**

Trypsin is also a serine protease of the chymotrypsin family, remotely related in sequence to SGPA. Co-ordinates of the single polypeptide chain were taken from the PDB file 2PTN (Walter et al., 1982). These crystals were grown from 2.4 M ammonium sulphate at pH 8. There is a bound calcium ion associated with each protein molecule. The resolution is 1.55 Å, and the R factor 19%, following a version of Jack and Levitt (1978) refinement. Fifty-nine of the 340 polar and charged groups have one or more atoms with temperature factors.
higher than 25 Å². Four side chains were rotated 180° around the last χ angle: H91, N100, Q188 and N223. Thirty-seven of the 59 high temperature factor polar and charged groups were reported as not reliably observed in the experimental electron density map.

**Bovine trypsin, pH 5**

Co-ordinates were taken from the PDB file 1TPO (Marquat et al., 1883). These crystals, grown at pH 5, are isomorphous with those at pH 8 described above, and have identical cell dimensions. As with the higher pH form, there is a bound calcium ion associated with the protein molecule. The resolution is 1.7 Å, and the R factor 18%, following a version of Jack and Levitt (Jack and Levitt, 1978) refinement. Sixty-one of the 337 polar and charged groups have one or more atoms with temperature factors higher than 25 Å². Five side chains were rotated 180° around the last χ angle: Q64, N48, H57, N100 and Q210. Thirty-six polar and charged groups and one non-polar one were reported as unobserved in the experimental electron density map.

**Ribonuclease**

Bovine ribonuclease consists of a single polypeptide chain of 124 residues. Co-ordinates were taken from the PDB file 7RSA (Wlodawer et al., 1988). Crystals were grown in t-butyl alcohol at pH 5.3, and there is one ordered t-butyl alcohol molecule observed in the asymmetric unit. The Hendrickson–Konnert refinement led to an R factor of 15% at a resolution of 1.26 Å. Two side chains (residues K7 and K31) were poorly defined in the electron density map. Seventeen of the 199 polar and charged groups have temperature factors higher than 25 Å². Hydrogen positions (determined from neutron diffraction data) were taken from the PDB file. Those on hydroxyl groups were reoriented for best interaction with the local electric field, as with the other proteins. For the 12 side chains reported to have dual conformations, the highest occupancy conformation was used. In the cases of equal occupancy, the side chain conformation with the lowest average temperature factor was used. No side chains were rotated.

**Results**

The total intra-protein electrostatic energies of groups are almost always favorable

Figure 1a shows the distribution of total intra-protein electrostatic energies for the polar and charged groups in SGPA, using a cut-off of 6 Å, calculated as described above. Almost all are negative, that is, favorable. There are only four exceptions, and two of these, for the hydroxyl groups of S40 and T213, have positive energies of less than 0.1 kcal/mol. The other two are for the peptide links after Y120A and A114, with energies less than +2 kcal/mol. The Y120A peptide has unfavorable energy contributions from the two peptide links immediately following in the chain. These four residues, 120A–120D, are an insertion in the sequence relative to the other serine proteases, and form a short hairpin loop.

The unfavorable energy of the A114 peptide link arises from its interaction with the positive charge on the N-terminus of the polypeptide chain and the charged imidazole group of H108. There is no obvious reason for this exception. Neither the Y120A peptide link nor that of A114 have a positive energy when a shorter cut-off of 5 Å is used, but both do with a longer 7.5 Å cut-off. Omission of the intermolecular sodium ion from the calculations leads to an additional unfavorable energy of the C-terminal carboxyl group, suggesting the observed position of this group is affected by the crystal environment.

A similar distribution of total group energies is found for the other high resolution, well refined protein structures examined (Table I), although not quite as highly optimized as for the highly refined SGPA case. In trypsin, the influence of a number of unobserved charged side chains that were built in arbitrary conformations increases the number of positive energy groups substantially; in the pH 5 structure, out of a total of 22, nine involve these groups. We discuss only the 12 groups with energies greater than 3 kcal/mol. Five of these (side chains of K222, N223 and N25 and the peptide links before T125 and C220) result from the unobserved side chains. A further four can be attributed to the long range influence of the doubly charged bound calcium ion: peptide links (before residues D71, N74 and V75), have strong unfavorable interactions with the calcium ion, but the energies improve dramatically when a longer cut-off of 7.5 Å is used. The fourth group, that of the carboxyl of D71, has an energy of less than 3 kcal/mol at the longer cut-off, because of the inclusion of...
a favorable interaction with the calcium ion. These changes with cut-off suggest that doubly charged ions are accommodated by longer range organization. Two further unfavorable group energies are for the peptide before residue S93 and the imidazole group of H91, and are caused by an unfavorable interaction between them. It was assumed that this histidine is positively charged at pH 5. At pH 8, where the histidine was treated as neutral, these two groups do have negative energies. Finally, the peptide link preceding residue 243 has an unfavorable interaction with the C-terminal carboxyl group. At pH 5, the carboxyl may be neutral, rather than charged, as we assumed. However, at pH 8, there is also a positive energy, although less than 3 kcal/mol.

The pH 8 trypsin structure has 15 unfavorable total group energies. We discuss only seven which are greater than 3 kcal/mol. Two of these are associated with disordered side chains, as at pH 5. The same four groups as at pH 5 are affected by the calcium ion. The remaining group is not unfavorable in the pH 5 structure. This is the peptide link before Y29, which interacts unfavorably with a number of other peptide links.

For ribonuclease, there are a total of 16 groups with positive energy, eight of which have energies greater than +3 kcal/mol. Of these eight, two are groups involved in catalysis, H12 and K41. These groups are both charged in the calculation, and have a strong, unfavorable interaction with each other. A third group, the peptide before T45, is also unfavorable because of an interaction with the H12 ring. At the pH of the crystals (5.3) one would expect the histidine to be charged, and both ring nitrogens are protonated, according to the neutron diffraction structure. The temperature factor of the charged group of K41 is rather high (31 Å²). For these three groups, a longer cut-off and the inclusion of intermolecular interactions do not affect the picture. Four other well ordered positively charged amino groups, on K61, K91 and K98 and the N-terminus, have unfavorable energies. The conformation of K61 is stabilized in the crystal by a strong salt bridge to a carboxyl group on a neighboring molecule. K98 is very exposed to solvent. Use of a 7 Å cut-off results in a net favorable energy for both of these. The N-terminus interacts unfavorably with the first peptide link in the chain, a relationship that is essentially fixed by the covalent geometry. A 7 Å cut-off worsens the energy, as does the inclusion of symmetry related molecules.

Thus, most of the exceptions to the rule of low or negative total group energy are associated with features such as undetermined side chain positions, strong charge effects and perhaps catalytic mechanisms. The effect of unidentified ions and errors in position for some poorly ordered side chains may also be significant.

**Extent of favorable and unfavorable group–group interactions in a refined protein structure**

Figure 1b shows the distribution of group–group electrostatic energies for all polar and charged group interactions in the SGPA structure. Approximately 38% of intergroup interactions are found to be positive. Contrary to the impression given by analyses that focus on hydrogen bonds, a large number of short range interactions are in fact unfavorable. Thus, although protein molecules are organized such that total group energies are nearly always favorable, they are far from being organized so that every contributing individual group–group interaction is favorable. This picture holds for the other high resolution structures examined (Table I).

**Sensitivity of the analysis to the range of interactions included**

We next examine the sensitivity of the findings to the range of interactions included. Table II shows that at a very low cut-off value, 4 Å, a high fraction of groups have a total energy greater than zero, so that this cut-off is clearly too short to include the relevant structural organization. The fraction falls to a minimum at 6 Å, and then climbs again with increasing cut-off. These data suggest that the optimization of electrostatic organization in proteins is short range. The analysis of the number of group–group contributions that are unfavorable shows a rather different picture. The fraction is lowest at short distances, and climbs steadily beyond a cut-off of 5 Å. In particular, between 5 and 6 Å, although the number of groups with a total unfavorable energy falls, the fraction of unfavorable contributions increases from about 1/4 to about 1/3. At cut-offs of 7.5 Å or larger, the fraction of unfavorable group–group interactions approaches the random value of 50%, indicating no significant long range optimization.

As noted earlier, an exception to the general finding that a 6 Å cut-off is near the optimum for total group energy are the situations involving doubly charged groups. There are two such instances in the proteins examined—the bound calcium ion in the trypsin structures, and the iron in the iron sulphur clusters in the ferredoxin (discussed below).

At a cut-off distance of 5 Å, SGPA has a total of nine groups with unfavorable total energies, four of which involve contributions from atoms with high (>25 Å²) temperature factors. Two of the other groups, the peptides following H57 and before D102 have unfavorable interaction energies with

<table>
<thead>
<tr>
<th>Protein</th>
<th>Number of groups</th>
<th>Unfavorable groups</th>
<th>% Unfavorable groups</th>
<th>% Unfavorable contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2sga</td>
<td>265</td>
<td>4</td>
<td>1.5</td>
<td>37.8</td>
</tr>
<tr>
<td>7rsa</td>
<td>199</td>
<td>16</td>
<td>8.0</td>
<td>40.4</td>
</tr>
<tr>
<td>2ptn</td>
<td>339</td>
<td>15</td>
<td>4.4</td>
<td>38.3</td>
</tr>
<tr>
<td>1tpo</td>
<td>337</td>
<td>22</td>
<td>6.8</td>
<td>38.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cut-off (Å)</th>
<th>Unfavorable groups</th>
<th>% Unfavorable groups</th>
<th>% Unfavorable contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>39</td>
<td>15.7</td>
<td>26.9</td>
</tr>
<tr>
<td>5.0</td>
<td>9</td>
<td>3.5</td>
<td>25.0</td>
</tr>
<tr>
<td>6.0</td>
<td>4</td>
<td>1.5</td>
<td>37.8</td>
</tr>
<tr>
<td>7.5</td>
<td>10</td>
<td>3.7</td>
<td>46.4</td>
</tr>
<tr>
<td>10.0</td>
<td>8</td>
<td>3.0</td>
<td>48.6</td>
</tr>
</tbody>
</table>

Data are for SGPA. The columns show the number of groups with total electrostatic energy greater than zero, percentage of total number of groups and percentage of all group–group energies greater than zero, as a function of interaction distance cut-off.
Table III. Unfavorable total electrostatic energies for peptide links in main chain only structures

<table>
<thead>
<tr>
<th>Protein</th>
<th>Number of peptides</th>
<th>Unfavorable peptides</th>
<th>% Unfavorable peptides</th>
<th>% Unfavorable contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2sga</td>
<td>173</td>
<td>14</td>
<td>7.7</td>
<td>37.4</td>
</tr>
<tr>
<td>7rsu</td>
<td>125</td>
<td>9</td>
<td>7.2</td>
<td>40.2</td>
</tr>
<tr>
<td>2ptn</td>
<td>212</td>
<td>21</td>
<td>9.9</td>
<td>38.0</td>
</tr>
<tr>
<td>1tpo</td>
<td>210</td>
<td>22</td>
<td>10.4</td>
<td>38.4</td>
</tr>
</tbody>
</table>

As elsewhere, a cut-off of 6 Å between group centers was used. ‘Number of peptides’ lists the number of inter-residue peptide links included. ‘Unfavorable peptides’ are those links with a positive total electrostatic energy and ‘unfavorable contributions’ are peptide–peptide interactions with positive energies. Trends here are similar to those seen when all electrostatic groups are included (Table I).

the adjacent charged side chains of the catalytic triad, H57 and D102, respectively. Thus, slightly longer range organization than usual is involved in stabilizing these functional components. A similar though less pronounced strain is present in the peptide link following H57 in trypsin at pH 5, but not at the higher pH of 8, where the histidine is uncharged. There is additional angular strain in the χ1 angle of H57 in all three of these structures, with a value close to 90°. At a longer cut-off of 7.5 Å, SGPA has 10 unfavorable total energies, four of which are contributed by unfavorable relatively long range charge–charge interactions, suggesting that at longer distances, solvent screening, not represented in our model, may be more significant.

Extent of optimization of intra-main chain electrostatics

Backbone peptide units provide the majority of electrostatic groups in proteins. We next ask to what extent main chain conformations are such that interactions between only these groups are favorable. Table III shows the same type of data as in Table I, but this time including only interactions between backbone groups (including the N- and C-terminal charges). Although the fraction of unfavorable total group energies has increased by a few per cent for three of the four proteins, the general behavior is very similar.

Use of the electrostatic model for detecting errors in structures

We can utilize the short range optimization of electrostatic interactions in proteins to make sensitive tests for incorrect features in experimental, modeled and designed structures. We illustrate this with a number of examples.

Initial and final structures of the 7Fe ferredoxin from Azotobacter

The initial crystal structure of this protein deposited in the PDB (Ghosh et al., 1982) was incorrect. The mistake did not come to light until the structure was solved again by a different group (Stout et al., 1988) when it became apparent there had been an incorrect choice of enantiomorph for the crystal. Figure 2 shows the distribution of group total electrostatic energies for the initial structure, compared with that for the final one, calculated as described above. There are 41 groups with total energies greater than 3 kcal/mol in the initial structure, compared with 12 such groups in the final one. Thus errors of this magnitude can be very easily detected by electrostatic analysis.

Initial and final structures of gene5 protein

As described above, the incorrect choice of enantiomorph in the initial structure determination of ferredoxin (Ghosh et al., 1982) results in errors throughout that structure. The case of gene5 protein provides a test of the ability of the electrostatic analysis to reveal more localized errors. The initial structure of gene5 protein (Brayer and McPherson, 1987), PDB code 2gn5, has many amide groups with positive energies, particularly in the N-terminal half of the sequence. The later structure (1vqb) shows a distribution of energies similar to that of other high quality structures, with almost all of the energies less than zero. Comparison of the two structures shows a difference of sequence alignment by two residues in the first 40 positions of the sequence and by one residue in the region 70–75.
Although not completely diagnostic, most of the high energy groups are in the misaligned N-terminal portion of the sequence.

Effect of crystallographic resolution and refinement on electrostatics

The four crystal structures used for characterizing the extent of electrostatic optimization in proteins are all at relatively high resolution and well refined. Do these factors affect the picture of electrostatics that emerges? To answer this question, we have analyzed an earlier version of the structure of one of the proteins, SGPA, using the obsolete PDB file 1sga (Brayer et al., 1978; http://pddbbs.sdsc.edu/PDBBobs.cgi). The resolution is 2.8 Å, as opposed to the final one of 1.5, and the structure determination was done before refinement methods became established. Figure 4 shows the distribution of total group energies in this structure, compared with that in the refined structure. It is apparent that there are many more positive energies: a total of 40 out of 256, as opposed to only four in the refined structure. The refinement of 2sga was performed with the Hendrickson–Konnert procedure (Hendrickson, 1985) which does not contain any electrostatic biases from a force field, so the change in the distribution in the final structure must be due to an improved model.

Comparison of structures derived by crystallography and NMR

NMR structures are usually believed to be less accurate than those obtained by crystallography (Billeter, 1992; Wagner et al., 1992; Gallagher et al., 1994). Tests based on statistical methods (Lüthy et al., 1992; Sippl, 1993) support this view. However, these tests use statistics derived from crystal structures, so arguably are unfairly biased in favor of that method. The electrostatic analysis has no statistical component, so provides a more objective measure. We have compared the distribution of total group energies in the crystal structure and average NMR structure of a small protein, the 56 residue B1 domain of protein G. The crystal structure (Gallagher et al., 1994), PDB file 1PGB, is at 1.92 Å resolution, with an R factor of 20%, so is a high quality structure by the criteria used in this paper. The NMR structure (Gronenborn et al., 1991), PDB code 2GB1, is a restrained, minimized average structure based on 60 independent conformations, all consistent with the NMR data. The structure is described as exceptionally well defined, with an r.m.s. spread over the 60 structures of 0.27 Å on backbone atoms and 0.39 Å for all atoms excluding disordered side chains. Figure 5 shows the two distributions. That for the X-ray structure is similar to those seen for the other high quality structures examined, with a slightly higher proportion of positive energies than usual (seven out of 92).

The distribution for the NMR structure is shifted towards more positive values with a higher proportion greater than zero (14 out of 88), in a manner similar to that seen for the preliminary X-ray structure of SGPA. Thus, at least in this case, the X-ray structure is the more electrostatically optimized, suggesting substantially higher accuracy.

Analysis and modification of a protein design

The strong tendency of electrostatic groups to have negative energies may also be used as an aid in protein design. Possible structural errors of a localized nature may be pinpointed, modified and re-examined. We illustrate this possibility on a designed four helix protein, referred to as ‘bundle’. Bundle was designed at an EMBL Protein Design workshop (Moult et al., 1987) and subsequently modified. Unlike several other four helix bundle designs (Bryson et al., 1995), it has never been expressed and characterized, so this is a rather abstract case. Nonetheless, it serves to demonstrate the potential of the method.

The original bundle design consists of a single polypeptide chain 112 residues long. Core packing was developed from the arrangement seen in the ROP dimer (PDB code 1ROP). Two symmetry related loops at one end of the bundle were designed to incorporate a calcium binding site. Ligands for the metal binding were based on the calcium site in parvalbumin (PDB 4cpv). Three GLUs, an ASP and GLN were sited on loops to fit the liganding observed in parvalbumin.

An electrostatic analysis was performed as in the other tests. Fourteen groups were found to have energies greater than +3 kcal/mol, and therefore were considered to be unlikely to occur in a correct protein structure. Some of the changes needed to correct these were cosmetic, in the sense that they involved rotations of side chains (residues T25, E27, D61, N66, D86 and Q89). However, other positive total energies suggested significant and specific design flaws. Table IV lists these high energy groups, the primary interaction responsible for the positive value, and the changes made to the model to eliminate them. Co-ordinates of the original
result is in keeping with the well known optimization of established. However, the finding here is more general. The specific short range interactions (Richards, 1993) is well and Argos, 1994; McDonald and Thornton, 1994) and other bonds (Baker and Hubbard, 1984; Stickle static energy of a group is favorable. Optimization of hydrogen environment around electrostatic groups in proteins is highly The major finding of this work is that the short range Short range electrostatics are highly optimized

Conclusions

Short range electrostatics are highly optimized

The major finding of this work is that the short range environment around electrostatic groups in proteins is highly optimized. In almost all cases, the total intra-molecular electrostatic energy of a group is favorable. Optimization of hydrogen bonds (Baker and Hubbard, 1984; Stickle et al., 1992; Bordo and Argos, 1994; McDonald and Thornton, 1994) and other specific short range interactions (Richards, 1993) is well established. However, the finding here is more general. The result is in keeping with the well known optimization of dihedral angle values in proteins (for example: Ramakrishnan and Ramachandran, 1965; Herzberg and Moul, 1991; Dunbrack and Cohen, 1997). Taken together, these two observations tell us something about the nature of protein molecules. Although low energy conformational choices are restricted by steric factors and the partial double bond character of the peptide link, it is still possible to find arrangements that allow almost every group to sit in an energetically favorable environment. Further, the restrictions of an overall three-dimensional fold can be satisfied without introducing significant electrostatic strain. That is, it is not necessary to introduce significant local strain to obtain an overall low energy structure.

The limits of electrostatic optimization

Although the total intra-molecular electrostatic energy of most groups is favorable, a large proportion of contributing individual interactions are not (Figure 1b and Table I). One probable reason for this is that the steric and covalent restrictions placed on the polypeptide chain in the crowded folded state simply make it impossible for all interactions to be favorable. For example, in an α-helix, the dipoles of adjacent peptide groups are parallel, resulting in inescapably unfavorable interactions (Avbelj and Moult, 1995). An alternative explanation is that although from a functional point of view a protein molecule must be stable with respect to the unfolded state, it need not necessarily be very stable. In the unfolded state, protein electrostatic groups interact with surrounding water molecules, and these interactions will on average be favorable. However, their strength is limited by the internal interactions in the solvent, and by the disruption caused by the kinetic energy of the solvent molecules, so they are themselves only partly optimized. Some of this partial lack of optimization may be reflected in the same property in the folded state.

Electrostatic optimization is short range

Table II shows that the fraction of individual interactions that are unfavorable begins to rise when inter-group distances larger than 5 Å are considered. The fraction of total unfavorable intramolecular energies falls to a minimum around 6 Å, and rises slightly beyond that. These data reflect an increasingly random organization of group interactions at the longer distances: individual group pairs are equally likely to interact favorably or unfavorably, and, with many groups included, the total intra-protein electrostatic energy contribution approaches zero.
A simple electrostatic model is sufficient

As noted earlier, the applicability of this method is limited by the model’s lack of a solvation contribution and lack of other more subtle intra-protein contributions. We would expect that the optimization of electrostatic environments would be even more dramatic if an appropriate pairwise solvation model were available. The role of solvent screening was investigated, using a very simple model based on the Kirkwood-Tanford continuum theory (Tanford and Kirkwood, 1957), with a depth below the protein surface assigned to each atomic partial charge (for details, see Moult et al., 1985). Inclusion of this solvent screening does tend to reduce the size of the outliers in energy, both positive and negative, but leads to a slight increase in the total number of groups with positive energies. In the interest of simplicity, it was therefore not used for the main analysis. The Generalized Born model (Qiu et al., 1997; Luo et al., 1998) might be more suitable for this type of analysis. Other factors, such as possible interactions involving aromatic rings (Mitchell et al., 1994) and the effect of polarization and sensitivity to the charge set used have also not been investigated. Obviously, quantitative results, such as the pK of ionizable groups, cannot be obtained with this simple and unrefined method. Nevertheless, the conclusions reached in this paper are sufficiently robust that they do not need the support of a more sophisticated model.

Significance of the outliers

Analysis of outliers in main chain dihedral angles in proteins has shown that they are often associated with function (Herzberg and Moult, 1991). The implication there is that has shown that they are often associated with function and that function can be tested. Nevertheless, for the serine (Herzberg and Moult, 1991). The implication there is that

property is that the analysis has no ‘knowledge bias’, so that, for example, differences between NMR and X-ray structures and between high and low resolution X-ray structures cannot be attributed to the databases used to construct the evaluation functions.

Acknowledgements

This work was supported by NIH grant GM41034.

References


Received August 31, 1998; revised September 8, 1998; accepted June 4, 1999