# RESEARCH PAPERS

*Acta Cryst.* (1994). D**50**, 237–249

## An Evaluation of the Use of Databases in Protein Structure Refinement

BY JIN-YU ZOU

*Department of Molecular Biology, Box* 590, *Biomedical Center, Uppsala University, S*-751 24 *Uppsala, Sweden*

AND SHERRY L. MOWBRAY*

*Department of Molecular Biology, Box* 590, *Biomedical Center, Swedish Agricultural University, S*-751 24 *Uppsala, Sweden*

(*Received* 13 *April* 1993; *accepted* 14 *September* 1993)

### Abstract

The speed of electron-density fitting during X-ray structure solution and refinement, and the quality of the protein model resulting, can both be enhanced by the use of databases of main- and side-chain conformations. Three structures are compared in this report, one refined at high resolution (1.7 Å), and two at lower resolutions using either the database method (2.4 Å resolution) or more traditional empirical electron-density fitting (1.9 Å resolution). An analysis of peptide orientation was used as an aid in finding unusual portions of main-chain structure. The fit of side chains to known rotamer conformations was used to help determine the accuracy of these atomic positions. In addition, the use of an objective measure of the fit of structures to electron-density maps was evaluated, both alone and in combination with side-chain conformational information.

### Introduction

Most X-ray structures probably contain some 'errors', largely because of the fact that a few portions of any protein will always be ill defined in the electron-density maps, but also because refinement is a somewhat tedious process, during which it is difficult to remain totally attentive to all aspects of protein chemistry and structure. Information about main-chain and side-chain conformations previously seen in highly refined structures can be used to help evaluate whether particular structural decisions are likely to be correct. This is in many ways similar to the long-accepted use of stereochemical information in applying constraints and restraints for positional refinement of X-ray coordinates (Sussman, Holbrook, Church & Kim, 1977; Hendrickson & Konnert, 1980). All of these types of information can also be used in the analysis of structures after refinement.

To determine the effects of including conformational information from main-chain and side-chain databases on the quality of the final models, three refined X-ray structures of the periplasmic glucose/galactose-binding protein (GBP) involved in chemotaxis and transport in Gram-negative bacteria were analyzed.

### Methodology

The tools of the program package *O* (Jones, Bergdoll & Kjeldgaard, 1990; Jones & Kjeldgaard, 1992) provided the basis for most of the work described here.

Information about main-chain conformations seen previously in highly refined X-ray structures was included during electron-density fitting and refitting using the *O* option lego_ca. A number of polyalanine fragments with $C^{\alpha}$ positions similar to those of a given model segment are located in this way from the *O* database of 33 structures refined at high resolution. The search algorithm first uses sets of intramolecular $C^{\alpha}$—$C^{\alpha}$ distances to find segments with similar structure (Rossmann & Liljas, 1974), then calculates the least-squares fit between the $C^{\alpha}$ atoms of the model and each of these segments after the appropriate transformations have been applied (Jones & Thirup, 1986). The user selects the segment that best fulfills the r.m.s. fit and any other important criteria; the coordinates for the first and last residues of the segment are not updated. In the analysis of refined models, the pep_flip (peptide orientation analysis) option (Jones, Zou, Cowan & Kjeldgaard, 1991) was used to help locate unusual (and possibly incorrect) peptide orientations. For

* Author to whom correspondence should be addressed.

each position $i$ in the sequence, up to 20 pentapeptides from the database with an r.m.s. fit better than 1.0 Å to the $C^{\alpha}$ atoms of the zone from $i-2$ to $i+2$ are located using the procedure described above. The r.m.s. deviation of the main-chain carbonyl O atom of residue $i$ of the model fragment from the equivalent O atoms of the database fragments is then calculated to give the pep_flip value for this residue.

The side chains observed in highly refined structures also show a great preference for certain conformations, those that would be predicted from energy considerations (Janin, Wodak, Levitt & Maigret, 1978). Tabulations of these most probable conformations (James & Sielecki, 1983; McGregor, Islam & Sternberg, 1987; Ponder & Richards, 1987) have made possible the convenient use of 'rotamer' information in structure building, refinement and analysis (Jones *et al.*, 1991). The normal *O* database includes only those rotamers found in at least 10% of the representative side chains analyzed by Ponder & Richards (1987). These common side-chain conformations are used in structure building and rebuilding with the *O* option lego_side_chain. Using the existing main-chain and $C^{\beta}$ coordinates for a given residue as guide points, the user can thus view the different rotamers at the graphics terminal and select that which best fits the electron density. It had been observed, however, in the Ponder & Richards study that the positions of the terminal atoms of some residue types (particularly lysine, arginine and glutamine) are not as well defined, and so the distributions of only the first two $\chi$ angles could be given with any confidence. All atoms of these residue types are included with each rotamer for use in building and refitting with *O*; the terminal atoms can be fitted further to the electron density with the torsion option (which allows manual adjustment of the side-chain $\chi$ angles).

The side-chain conformations in a particular model can be compared to those of the common rotamers using the rsc_fit option. Given the main-chain and $C^{\beta}$ coordinates for a residue (except glycine or alanine) the positions of the side-chain atoms are predicted assuming each rotamer in turn. The r.m.s. fit of these atomic positions to those of the model is then calculated. The rotamer with the smallest r.m.s. difference is listed, and its r.m.s. difference reported as the rotamer side-chain (RSC) value. The RSC values for glycine and alanine are, by definition, 0.0. For the present study, the analysis was altered to reflect the uncertainty of the terminal atoms in long residues noted above. For lysine and arginine, the modified side-chain analysis used the positions of only CB, CG and CD (*i.e.* $\chi_1$ and $\chi_2$ only) thus allowing the RSC values to reflect only those atoms for which well defined conformations were described.

The RSC calculation was also modified to correct an error in an earlier version of *O* (Jones *et al.*, 1991) by which functionally equivalent atoms (*e.g.* OD1 and OD2 of aspartate) were not treated as equivalent in the analysis. The amino-acid types affected were tyrosine, phenylalanine, aspartate and glutamate. Since there are two common histidine rotamers that are similarly shaped (differing by a 180° rotation in $\chi_2$) only one of which was included in the *O* database, histidine was also treated as symmetrical in the present analysis.

The real space fits of main- and side-chain atoms to the electron-density maps were obtained with the *O* option rs_fit (Jones *et al.*, 1991). In this method, three density functions are used. One corresponds to the experimental electron density, previously calculated on a suitable grid. A second density is calculated on the same grid using the coordinates of the protein. For each atom, the electron density $\rho$ at position $r$ is modelled by a Gaussian distribution of the form,

$$\rho(r) = (Z/A^3)\exp(-\pi r^2/A^2),$$

where $Z$ is the atomic number of the atom and $A$ is the atomic radius (Jones & Liljas, 1984, following Diamond, 1971); the density for the protein is the sum of that for the individual atoms contributing. The atomic radius is, in turn, generated from the atomic temperature factor using the relationship,

$$B = 4\pi(cA^2 - A_o^2),$$

where $B$ is an overall temperature factor, $c$ is a constant chosen to remove systematic differences between $F_{obs}$ and $F_{calc}$ and $A_o$ is the 'zero-temperature radius' (Deisenhofer & Steigemann, 1975). The third density is calculated as for the second one, but using only selected portions of a particular residue. The atoms to be used are defined by a dictionary and may be, for example, the main-chain atoms only (to evaluate main-chain connectivity), the side-chain atoms only (to evaluate the fit of individual side chains) or all of the atoms making up the residue. This third density acts as an envelope to allow the selection of points to be used in evaluating how well the chosen atoms fit the experimental density. In the original formulation (Jones *et al.*, 1991) this evaluation function was a grid sum $R$ factor, but in the current version of *O*, a correlation coefficient is used. Thus for every non-zero value in the third density function, the equivalent points in the first two maps are used in the calculation of the correlation function. This function has values between −1.0 (anticorrelated) and 1.0 (a perfect fit).

The appropriate values of $A_o$ and $c$ will depend on the resolution of the map and other factors. The defaults given in *O* for these parameters are 0.9 Å

and 1.04, respectively, which seem to be acceptable for use with $2F_o - F_c$ maps at better than 2.0 Å resolution or multiple isomorphous replacement (MIR) maps. For other maps, they can be optimized to provide the best discrimination between residues with good and poor electron-density fit. Since $A_o$ and $c$ co-vary, a number of combinations of values will give rise to similar correlation coefficients for an individual residue. In practice, $A_o$ can thus be set to 0.9 and $c$ varied in the range of 0.6–1.2 in steps of 0.05 to choose the value of $c$ which results in the highest correlation coefficient (0.90–0.95) for one or more residues with good electron density (selected at the graphics terminal). This approach seems generally to give the largest difference between the correlation coefficients obtained for residues with good and poor electron density. It is not appropriate to simply refine the parameters to give correlation coefficients with the highest overall value (*i.e.* to minimize the difference between the observed and calculated maps for all residues, regardless of their fit to the electron density). Some combinations of $A_o$ and $c$ will give substantially higher correlation coefficients for residues with poor fit, and only slightly lower ones for residues with good fit, than with the optimal combination.

### Refinement of *Salmonella* GBP structures

The structure of GBP from *Salmonella typhimurium* in complex with β-D-glucose (GBP-S1), was solved at 3 Å resolution by the method of multiple isomorphous replacement (Mowbray & Petsko, 1983) and refined at 2.4 Å resolution using the program *X-PLOR* (Brünger, Kuriyan & Karplus, 1987; Brünger, 1988). The space group was $C2$, and the cell dimensions were $a = 119.59$, $b = 37.28$, $c = 80.23$ Å and $\beta = 123.37$ . This structure was fit/refit to the MIR and $2F_o - F_c$ maps using the main-chain and side-chain rotamer databases of *O*. The main chain was built using overlapping polyalanine fragments (usually five residues in length) chosen on the basis of fit to the electron density and any other constraints (such as hydrogen bonding or sequence dependence of conformation) and modified as required. Similarly, side-chain rotamer conformations were chosen on the basis of fit to the electron density as well as appropriate hydrogen bonding; the terminal torsion angles of larger residue types were modified if necessary. Where the electron density was poor, the most common rotamer compatible with the local structure was used. Side-chain conformations not found in the rotamer library of *O* were included only when required to satisfy the electron density and local structure. During refinement, residues with pep_flip values $\geq 2.5$ Å were compared individually with the similar main-chain conformations found in the database, and the structure was corrected if needed. The final model contains only 305 amino acids, as the first two and last two residues were too ill defined in the electron-density maps to be certain of their exact conformation. A total of 106 water molecules, the sugar and a Ca atom were included in addition to the protein. This coordinate set and the associated structure factors are available from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) as 3GBP; they are further described in Mowbray, Smith & Cole (1990).

The structure of the same GBP in complex with β-D-galactose (GBP-S2) was solved beginning with the coordinates of GBP-S1, then cyclically refitting and refining at 1.7 Å resolution using *X-PLOR*. The space group was $C2$, and the unit-cell dimensions were $a = 119.57$, $b = 37.39$, $c = 80.14$ Å and $\beta = 123.18$ , essentially identical to those of the previous form. The strategy followed for including the database information during refinement was the same as that used for the lower resolution structure. All 309 residues are present in the final model, as well as a total of 153 water molecules, the sugar and calcium. This coordinate set and associated structure factors are available from the Protein Data Bank as 1GCA; they are further described in Zou, Flocco & Mowbray (1993).

Some of the conventional refinement statistics for these two protein models are given in Table 1. The r.m.s. coordinate difference between GBP-S1 and GBP-S2 is 0.15 Å using the $C^\alpha$ atoms of residues 3–307. These differences are distributed randomly in the molecule; no structural changes are attributable to the different sugars bound. Essentially all of the solvent molecules located in GBP-S1 are found to be present in GBP-S2 as well. These final models are analyzed further below.

### Fit to electron density

Real-space correlation coefficients were used as a measure for the agreement of GBP-S1 and GBP-S2 to their respective $2F_o - F_c$ maps (Fig. 1). The average values for main- and side-chain atoms were 0.898 and 0.896 for GBP-S1, and 0.922 and 0.903 for GBP-S2, respectively. The uniformly high correlation coefficients observed for the main chain in each model give a fair representation of the clear and continuous main-chain electron density observed throughout both protein chains. Similarly, the relative values observed for the different side chains give a very good feel for the quality of their fit to the electron-density map. (For consistency, the calculation for side-chain values in the present case made use of only the atoms of the reduced library described above. The primary effect was an improvement in the correlation coefficients obtained for the

Table 1. *Statistics for the X-ray structures analyzed*

| Protein model | Resolution range (Å) | *R* factor* | R.m.s. deviations† Bonds (Å) | Angles (°) | Dihedral angles (°) | Improper angles (°) |
|---|---|---|---|---|---|---|
| GBP-S1 | 7.5–2.4 | 16.1 | 0.010 | 2.55 | 24.6 | 0.86 |
| GBP-S2 | 7.5–1.7 | 19.0 | 0.015 | 2.67 | 24.0 | 1.16 |
| GBP-E | 10.0–1.9 | 14.6 | 0.029 | 4.23 | 24.8 | 3.14 |
| | | | Bond distance (Å) | Angle distance (Å) | Fixed planar torsion $\omega$ (°) | |
| GBP-E‡ | 10.0–1.9 | 14.6 | 0.024 | 0.045 | 6.6 | |
| GBP-S1-*i*§ | 7.5–2.4 | 15.9 | 0.019 | 0.057 | 4.7 | |
| *f* | 7.5–2.4 | 14.2 | 0.014 | 0.030 | 3.5 | |
| GBP-S2-*i*§ | 7.5–1.7 | 20.0 | 0.019 | 0.052 | 4.5 | |
| *f* | 7.5–1.7 | 18.9 | 0.015 | 0.028 | 2.3 | |

* $R$ factor $= \Sigma| F_{obs} - F_{calc} |/\Sigma F_{obs}$, where $F$ is the structure-factor amplitude.

† As reported by the *X-PLOR* analysis routines.

‡ This protein was refined using the program *PROLSQ* (Konnert & Hendrickson, 1980); the *R* factor and statistics are as reported in coordinate set 2GBP.

§ Since these proteins were refined in *X-PLOR*, they were also introduced into *PROLSQ* to allow a fairer comparison with GBP-E. Values before refinement (*i*) and after six refinement cycles (*f*) are given, using the same set of reflections that had been used in *X-PLOR*.

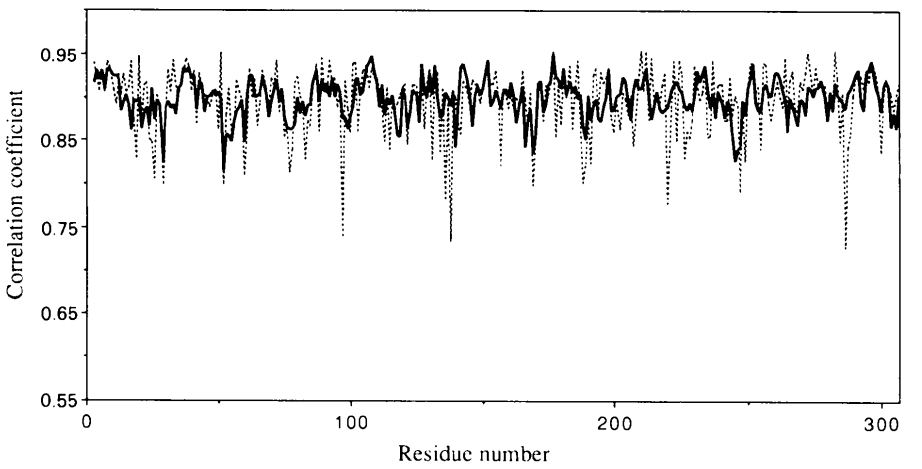
lysines that comprise almost half of the residues with poor electron density in these maps.)

Since the side-chain fit results differed in some details from conclusions drawn at the graphics terminal, the side chains of GBP-S2 were individually classified as to whether they were well or poorly determined based on direct inspection of the electron-density map; representative examples of the two situations are shown in Fig. 2. The electron density of 25 side chains (8% of the total) was qualitatively considered to be poor in the $2F_o - F_c$ map; these residues are indicated in Fig. 1(*b*) for comparison with the real-space correlation coefficients. Side chains with clear electron density essentially always had correlation coefficients greater than 0.8, while those with poor electron density generally had correlation coefficients less than this value. Four surface residues (Lys26, Gln83, Lys113 and Lys300) had correlation coefficients between 0.8 and 0.9 with both the full and reduced libraries, although their side chains were very poorly defined in the electron-density map contoured at $1\sigma$. These residues do, however, show a single conformation with continuous electron density at a lower contour level ($0.7$–$0.8\sigma$) reflecting the fact that the correlation coefficient is relatively insensitive to the absolute scale of the electron density. Only one side chain (that of Ala279) was considered to have good electron density despite a correlation coefficient of 0.79; this is associated with slightly higher temperature factors for the main chain in this region, whereby the single overall temperature factor used in the real-space fit calculation becomes less appropriate. Using a simple cutoff of 0.8 for the real-space fit, 22 out of 309 residues (7%) would fall in the poorly defined category for GBP-S2. A similar analysis for GBP-S1 suggests that a cutoff of 0.8 would be appropriate in that case, as well.

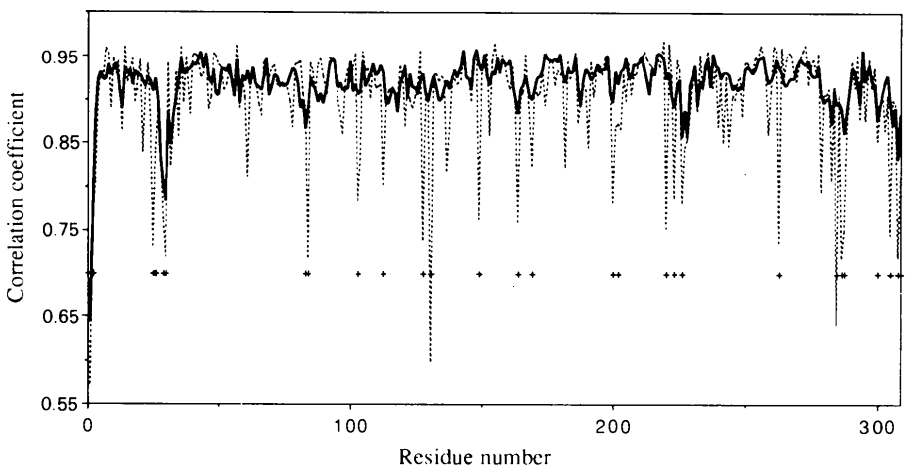
An earlier version of the real-space fit analysis employed a grid sum *R* factor (Jones *et al.*, 1991) instead of a correlation coefficient for the description of fit to electron density. A comparison of the results of these two procedures is shown for the side chains of GBP-S2 (with the reduced library) in Fig. 1(*c*). While the *R* factors do reflect the degree of clarity in the map, the currently implemented correlation-coefficient analysis gives the best match between the relative value obtained and the subjective assessment of the quality of the electron density.

## Main-chain analysis

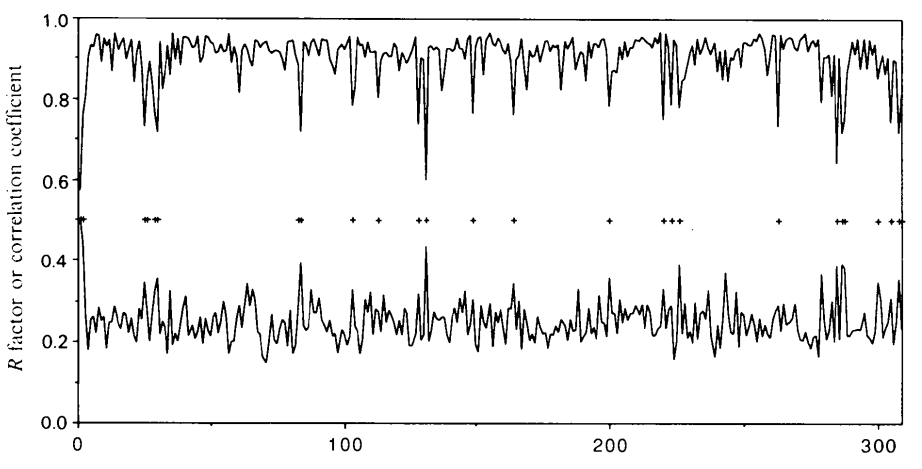
Not all of the pep_flip values greater than 2.5 Å seen during refinement were actually errors; some high values exist in both GBP-S1 and GBP-S2 (Table 2) even though the positions of all main-chain carbonyl O atoms are clear in both electron-density maps. The observed values for GBP-S1 agree very well with those obtained for GBP-S2, and so it was concluded that this structure contains no gross errors in the orientation of peptide planes. The largest pep_flip value found in both (3.6 Å for residue Asp236) represents a peptide orientation 180° away from that which would commonly occur with similar $C^\alpha$ positions. This particular conformation is forced by the need to provide a hydrogen bond between the main-chain amide N atom of residue 237 and a buried aspartic acid side chain nearby (Asp211), and is necessary for the proper structure in the hinge (Mowbray, 1992). With $\varphi = 146.40$, $\psi = -28.26°$, Asp236 is a Ramachandran violation (Ramakrishnan & Ramachandran, 1965), but it is the only




(*a*)




(*b*)




(*c*)

Fig. 1. Real-space correlation coefficients for the current $2F_o - F_c$ maps of models (*a*) GBP-S1 and (*b*) GBP-S2. The main-chain values are shown as solid lines, and the side-chain values (using the reduced side-chain library) as dashed lines. Values of $A_m = 0.90$ (Deisenhofer & Steigemann, 1975) and an overall temperature factor of 20.0 Å$^2$ (chosen close to the actual overall protein temperature factor) were used for both structures. The values for $c$ and the integration radius were 0.82 and 3, respectively, for the 2.4 Å map (sampled on a 0.8 Å grid), and 1.04 and 4 for the 1.7 Å map (sampled on a 0.4 Å grid). (*c*) A comparison of the side-chain correlation coefficients obtained for GBP-S2 (top line) with the grid-sum $R$-factor results using the same parameters (bottom line). Residues of GBP-S2 with poor side-chain electron density (based on inspection at the graphics display) are indicated as + in all plots.

residue with a pep_flip score above 2.0 Å that is. The only other Ramachandran violation in the GBP structure (residue 91) has a much lower pep_flip value (1.42 Å).

### Side-chain analysis

The rotamer side-chain analysis (RSC) was used to locate side-chain conformations that do not agree with the common rotamers found in the *O* database. The overall average of the RSC values for GBP-S2 is 0.540 Å if all residues are included, or 0.675 Å if alanine and glycine are excluded; all values greater than 2.0 Å are listed in Table 3. Obviously, even this 'final' high-resolution structure has some residues that deviate from the rotamer database using the RSC criterion.

Since the RSC procedure measures an r.m.s. difference between the coordinates found and the ones expected with the most similar rotamer, the results may differ from those obtained with an analysis based on deviations of side-chain $\chi$ angles (as was actually used in the Ponder & Richards paper from which the *O* database arises). A manual evaluation of the side chains using the same criteria as Ponder & Richards was therefore undertaken to allow the two methods to be compared. Side chains were considered to match a particular rotamer conformation if their torsion angles were within 2.5 standard deviations of the means, using the values reported (which varied with the residue type and angle measured). For consistency, only those angles defined by the authors for their rotamers were analyzed (*e.g.* $\chi_1$ and $\chi_2$ only for lysine and arginine). Side chains of GBP-S2 could thus classified as to whether they were common rotamers (greater than 10% frequency in that study and, therefore, present in the database of *O*), less common rotamers (frequency less than 10%, and so not present in the *O* database) or non-rotamer conformations (not described by Ponder & Richards). Glycine, alanine and proline were considered to be common rotamers for the purposes of the analysis. These results have been correlated with the RSC values and with qualitative judgments made about the electron density for GBP-S2 (as described above) in Tables 3 and 4.

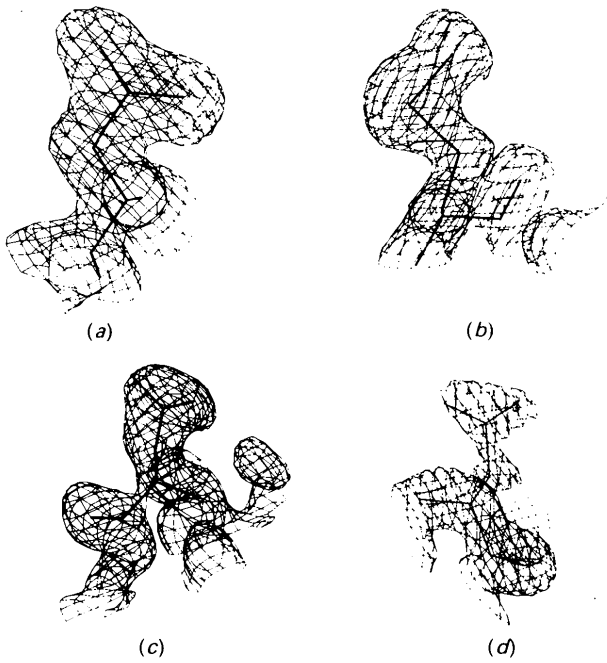



Fig. 2. Examples of well and poorly defined residues in the 1.7 Å $2F_o - F_c$ map for GBP-S2 (contoured at $1\sigma$ with a grid size of 0.425 Å). Three well defined residues, Asp212 (*a*), Met192 (*b*) and Asn259 (*c*), are shown. The last is located in a region with locally high temperature factors, but its conformation is still considered to be clear. A poorly defined residue, Asn84 (*d*), seems to have at least two alternate conformations. Side chains for which no electron density was observed were considered to fall into this category as well.

Table 2. *Pep_flip values greater than* 2 Å

| GBP-S1 (2.4 Å) | | GBP-S2 (1.7 Å) | | GBP-E (1.9 Å) | |
|---|---|---|---|---|---|
| Residue | R.m.s. | Residue | R.m.s. | Residue | R.m.s. |
| 236 | 3.58 | 236 | 3.58 | 236 | 3.62 |
| 138 | 3.04 | 138 | 3.02 | 276* | 3.20 |
| 41 | 2.44 | 278 | 2.50 | 278 | 3.12 |
| 278 | 2.42 | 235 | 2.43 | 138 | 2.63 |
| 235 | 2.28 | 41 | 2.41 | 41 | 2.56 |
| 182 | 2.27 | 182 | 2.26 | 235 | 2.27 |
| 256 | 2.22 | 280 | 2.10 | 182 | 2.25 |
| 42 | 2.10 | 42 | 2.08 | 109 | 2.06 |
| 181 | 2.03 | 181 | 2.05 | 280 | 2.03 |
| 148 | 2.01 | (256) | (1.96) | 42 | 2.03 |

* This represents a difference in peptide orientation between GBP-E and GBP-S (see text).

A number of conclusions could be drawn from this analysis. First, the coordinate r.m.s.- and angle-derived rotamer assignments agree very well. Essentially all side chains with RSC values less than 2.0 Å did represent the rotamer reported by *O*. (The situation will be slightly different during refinement, as discussed below.) Second, roughly half of the RSC values greater than 2.0 Å are due to the legitimate presence of less common rotamers. About 3% of all residues in GBP-S2 with clear electron density fall into this category. Third, some residues of the non-rotamer class have good electron density and conformations that ought to be reasonable, that is their $\chi$ angles fall within the distributions centered near $-60$, $+60$ or 180 (or $-90$, 0, $+90$ for angles involving an aromatic group) (*e.g.* Janin *et al.*, 1978). Specifically, the leucine, tryptophan, methionine and glutamine rotamers of Ponder & Richards do not cover all conformations in GBP-S2 that are clearly substantiated by electron density. (Tryptophan and

Table 3. *Comparison of the RSC analysis of the final structures with a rotamer analysis based on side-chain $\chi$-angle deviations*

For GBP-S2, all RSC values above 2.0 Å are given, as well as those for any less common or non-rotamer side chains (as defined in the text) which fall below that value. For GBP-S1 and GBP-E, values are given for all residues shown for GBP-S2 (sorted to be on the same line), as well as for any others greater than 2.0 Å in these structures themselves. RSC values less than this cutoff are also given for all residues that were designated as possible problems in these structures as described in the text; only these residues which are different are coded according to their rotamer status.

| GBP-S2 (1.7 Å) | | GBP-E (1.9 Å) | | GBP-S1 (2.4 Å) | |
|---|---|---|---|---|---|
| Residue | R.m.s. | Residue | R.m.s. | Residue | R.m.s. |
| Trp195§ | 2.96 | Trp195 | 2.72 | Trp195 | 2.74 |
| Gln45‡ | 2.81 | Gln45 | 2.71 | Gln45 | 2.77 |
| Tyr12‡ | 2.69 | Tyr12 | 2.34 | Tyr12 | 2.59 |
| Trp133§ | 2.68 | Trp133 | 2.43 | Trp133 | 2.69 |
| Glu165‡ | 2.60 | Glu165 | 2.81 | Glu165 | 2.56 |
| Glu240‡ | 2.47 | Glu240§ | 1.63 | Glu240 | 2.50 |
| Thr180‡ | 2.35 | Thr180 | 2.32 | Thr180† | 0.06 |
| Val162‡ | 2.31 | Val162 | 1.62 | Val162† | 0.95 |
| Thr253‡ | 2.21 | Thr253 | 2.24 | Thr253§ | 1.31 |
| Met192§ | 2.10 | Met192 | 2.23 | Met192 | 2.03 |
| Asn288*§ | 2.08 | Asn288 | 0.85 | Asn288 | 2.03 |
| Thr307‡ | 2.06 | Ser307§ | 1.33 | Thr307† | 0.73 |
| Gln142§ | 2.03 | Gln142 | 1.92 | Gln142 | 2.10 |
| Met250† | 2.03 | Leu250§ | 1.25 | Met250 | 1.86 |
| Leu268§ | 1.92 | Leu268 | 1.95 | Leu268† | 0.73 |
| Lys285*§ | 1.84 | Lys285 | 1.60 | Lys285 | 1.94 |
| Asp2*§ | 1.73 | Asp2 | 1.23 | Asp2 | — |
| Leu178‡ | 1.69 | Leu178 | 1.36 | Leu178 | 1.85 |
| Leu145‡ | 1.55 | Leu145 | 1.76 | Leu145§ | 1.72 |
| | | Val87‡ | 2.27 | Thr3‡ | 2.28 |
| | | Gln26 | 2.22 | Lys26* | 2.09 |
| | | Leu55§ | 2.08 | Val19§ | 2.06 |
| | | Gln83 | 2.03 | Lys191‡ | 1.61 |
| | | Leu37§ | 2.00 | Lys47† | 1.15 |
| | | Asp280§ | 1.95 | Arg21† | 0.69 |
| | | Leu36§ | 1.81 | Ser229† | 0.51 |
| | | Leu99§ | 1.55 | Ser95† | 0.37 |
| | | Ser247† | 0.68 | | |
| | | Ile230† | 0.23 | | |

* Residues for which the electron density was poor.
† Side-chain conformations close to a common rotamer.
‡ Side-chain conformations close to a less common rotamer.
§ Side-chain conformations that did not match any rotamer (non-rotamer).

Table 4. *Correlation of side-chain conformation with electron density and structural agreement*

| | Common rotamers | Less common rotamers | Non-rotamers |
|---|---|---|---|
| GBP-S2* | | | |
| Well determined | 269 | 9 | 6 |
| Poorly determined | 22 | 0 | 3 |
| Total | 291 | 9 | 9 |
| GBP-S1† | | | |
| Same | 261 | 4 | 5 |
| Different | 8 | 2 | 3 |
| Total | 269 | 6 | 8 |
| GBP-E† | | | |
| Same | 246 | 7 | 6 |
| Different (probably correct) | 13 | 0 | 0 |
| Different (check further) | 3 | 1 | 8 |
| Total | 262 | 8 | 14 |

* The well and poorly determined designations refer to the visual assessment of side-chain electron density described in the text, and as illustrated in Fig. 2.

† The analysis of GBP-S1 and GBP-E included only those side chains where the electron density of GBP-S2 was well determined. The side chains of each were categorized as 'same' or 'different' based on whether they fell within the expected angular ranges as described in the text. Conformational differences in GBP-E which involved sequence changes, or that could be altered as a result of different crystal packing were generally considered to be correct, unless a consideration of other factors indicated they were likely to be ill defined in the electron-density maps.

methionine side chains were sampled at a low frequency in that study, and so some of their rotamers may not have been observed often enough to be defined as a separate type. Glutamine possesses a chameleon-like ability to change with its environment, giving rise to a large number of documented conformations; GBP contains yet another.) This category represents about 2% of all side chains for which the electron density was clear. In contrast, one additional residue with good electron density (Tyr12 at the sugar-binding site) seems to be a distortion of a common rotamer, since its conformation ($\chi_1$ $-100$ , $\chi_2$ 40 ) was outside the expected angular distributions. Lastly, a few other residues have non-rotamer conformations with one or more $\chi$ angles that do not fall near the usual energy minima. The electron-density map in each case suggests a mixed population of two or more common rotamer conformations; the conformation found in the structure is midway between the rotamer conformations suggested by the map. Since the refitting strategy was to use of only common rotamers where the electron density was poor, these non-rotamer conformations must result from coordinate adjustments made by the refinement program.

The average RSC value for GBP-S1 was 0.584 where all residues were included, and 0.730 where alanine and glycine were excluded, slightly higher than the equivalent values for GBP-S2. All RSC values greater than 2.0 Å are given in Table 3, together with some observations about differences from GBP-S2 and the results of a side-chain $\chi$-angle analysis. A number of distinct side-chain conformations in GBP-S1 could be located quickly using differences between its RSC values and those of GBP-S2, as shown in Fig. 3(*a*). This type of analysis was useful in locating cases where one side chain is close to a rotamer and the other is not, but missed instances where both are close to, or distant from a rotamer (not necessarily the same one). In Table 4, the relationship between the frequency of rotamer conformations and differences in the two structures is also detailed.

A significant fraction (4 out of 13) of the differences between GBP-S1 and GBP-S2 involve residues

that appear in the former structure with common rotamer conformations, but as less common rotamers in the latter one (see Table 3 and Fig. 4). It was pointed out by Jones *et al.* (1991) that the inclusion of any but the common rotamers is likely to be inappropriate in an initial model. For the same reasons, it was felt that the less common rotamers should be included with caution in a lower resolution model. In a few cases in the GBP-S1 refinement, less common or non-rotamer conformations were in fact included, only to be picked out later as possible errors by analogy with GBP-S2. 'Different' was not, however, necessarily synonymous with 'incorrect', since some residues have conformations in GBP-S1 that did fit within the electron density of the 2.4 Å map as well as, or better than, the conformation found in GBP-S2 (for example, see Fig. 4). Of the side chains with distinct conformations in the two structures (indicated in Table 3), only two have side-chain correlation coefficients less than 0.80 in GBP-S1; the average is 0.865, only slightly less than the overall mean for the side chains. As an additional test, GBP-S2 was refined against the same 2.4 Å data used for GBP-S1 (starting $R$ factor 27%, final $R$ factor 17% with 100 cycles of Powell minimization only). Of the 13 residues in question, four had correlation coefficients which were significantly better (5–10% larger) in the original comparison between GBP-S1 and its map than are found for the newly refined structure *versus* its map, implying that

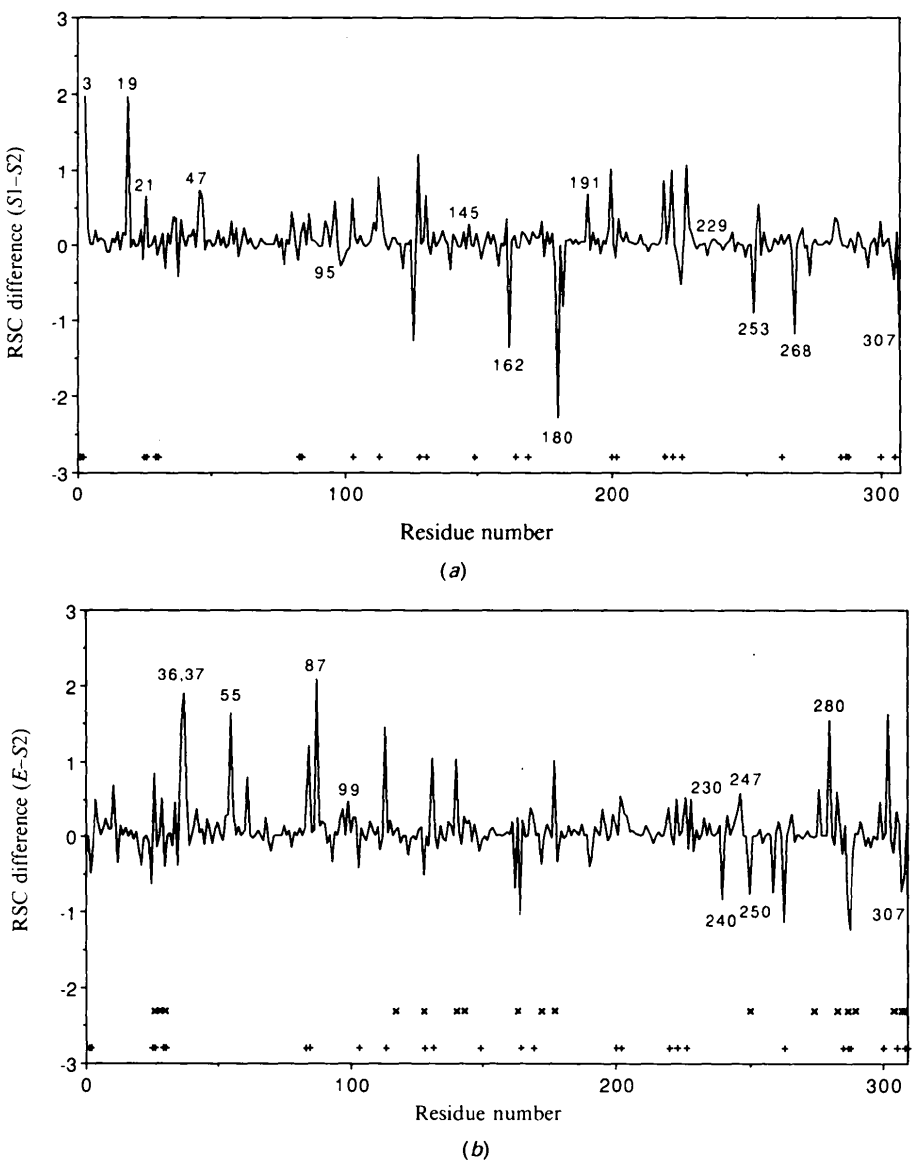



Fig. 3. Differences in rotamer side-chain fit for the different coordinate sets, using the reduced library: (*a*) GBP-S1 – GBP-S2, and (*b*) GBP-E – GBP-S2. Residues are marked with + where the side-chain electron density was poor in the GBP-S2 map, and × where there were sequence changes between GBP-S and GBP-E. Residues that were identified as having possible errors in side-chain conformation (see Tables 3 and 4) are indicated by their sequence numbers.

the original S1 conformation may be correct for that structure (residues Thr3, Lys26, Lys191, Ser229). In the remainder of cases, the correlation coefficients for the new structure/map were the same or slightly larger, implying that for some side chains the S2 conformation was correct for S1 as well. Slight differences in crystallization conditions may result in some distinct side-chain conformations.

### Two-dimensional RSC/real-space fit plots

It was observed during refinement that side chains with high RSC scores often had low correlation coefficients in the real-space fit analysis. This relationship suggested that a combined analysis of the two properties, as illustrated by the two-dimensional plots in Fig. 5, could be a powerful one.

A well refined structure such as GBP-S2 (Fig. 5*a*) should fit both the electron density and the rotamer database well, that is most residues should lie toward the top left-hand corner of the plot. The structure may have some side chains that are not common rotamers, but they should be well substantiated by electron density, as are those found toward the top right-hand corner of this plot. By the same principle, it should have few or no values at the lower right corner, that is, there should be few residues that fit neither the rotamer database nor the electron density. Residues that do not fit the electron density should probably be built as common rotamer conformations (lower left-hand corner of the plot).

A similar plot is shown in Fig. 5(*b*) for an early model in the 2.4 Å refinement of GBP (conventional *R* factor 28.2%). No solvent had been added at this stage, but $C^{\alpha}$ atoms had been correctly placed for all residues of the final model and all side-chain atoms were present. In this plot the distribution is somewhat more uniform. Fewer residues fall toward the upper left corner of this plot; ones falling outside this region provide clear targets for improvement of the protein model. While these side chains would be located with either of the real-space fit or RSC analyses separately, this two-dimensional plot allows the user to check the overall behaviour of the model, to select appropriate cutoffs for correcting the worst residues first and to track the progress of the refinement. As such, it would have been a useful aid. A plot for a partially refined structure which was built without the use of rotamers would be expected to have many more values falling in the low real-space fit/high RSC value (lower right) region.

A plot is shown for the final model GBP-S1 in Fig. 5(*c*). It shows the same overall characteristics of the plot shown for GBP-S2, and is a noticeable improvement over that for the partially refined model. It seems likely that only two of the possible errors in GBP-S1 would have been located on inspection of this type of plot. This is probably a reflection of both the lower resolution of the electron-density map (and the occasional uncertainty that results when trying to choose between similarly shaped rotamers based on electron density) and the fact that the GBP-S1 structure does fit both its map and the rotamer conformations very well.

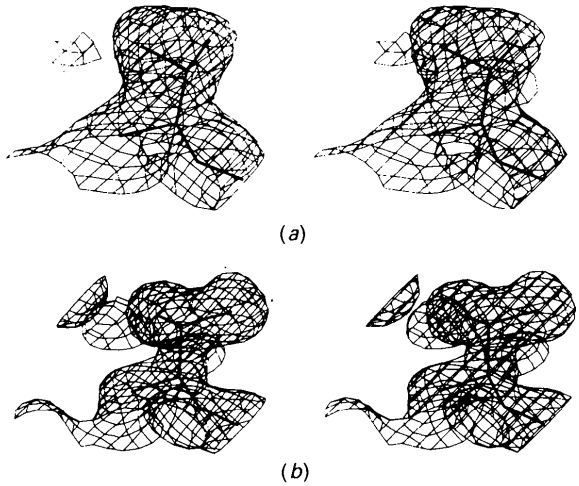



Fig. 4. Residue Val162 of GBP-S1 was fit to the 2.4 Å map as a common rotamer (*a*), but should probably be a less common one by reference to GBP-S2 and its 1.7 Å map (*b*). Both maps are contoured at a level of 1$\sigma$ with a grid of 0.45 Å.

### Conformational analysis of *E. coli* GBP

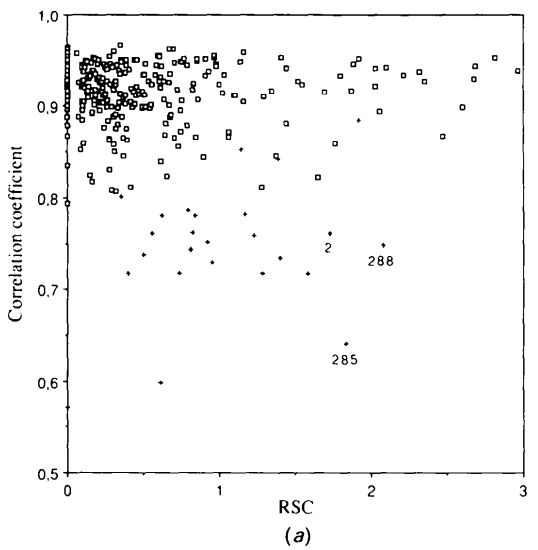
It is desirable to know how the results obtained when refinement utilizes database information compare with those obtained where the electron density is fit 'by eye'. Therefore, a glucose/galactose-binding protein of *E. coli* (GBP-E) which has 94% sequence identity with the *Salmonella* proteins was also analyzed for fit to the main- and side-chain databases.

The structure of GBP-E complexed with $\beta$-D-glucose (Vyas, Vyas & Quiocho, 1987, 1988) was independently solved at 3.0 Å resolution by the method of multiple isomorphous replacement (Vyas, Vyas & Quiocho, 1983) and refined at 1.9 Å resolution using the program *PROLSQ* (Konnert & Hendrickson, 1980) with alternating cycles of fitting/refitting to the multiple isomorphous replacement and $2F_o - F_c$ maps with the graphics program *FRODO* (Jones, 1982). The space group was $P2_1$, and the unit-cell dimensions were $a = 66.00$, $b = 37.05$, $c = 61.57$ Å and $\beta = 106.80°$. Atomic coordinates for this model were obtained as 2GBP from the Protein Data Bank. A total of 214 water molecules are included, as well as all residues of the protein sequence, the sugar and a calcium ion. In

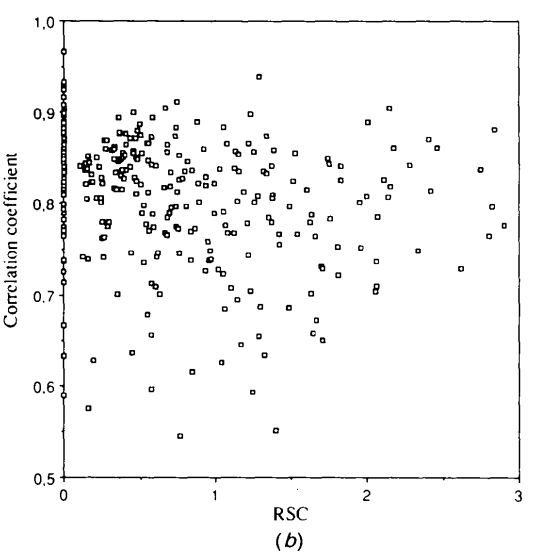
order to compare the reported statistics for GBP-E more directly with those of the *Salmonella* structures, the latter were refined for several cycles with *PROLSQ*. The angular deviations reported by that program for these two proteins both before and after that refinement are shown in Table 1.

Comparisons of the overall structures of the *Salmonella* and *E. coli* proteins have been reported in some detail in Mowbray (1992) and Zou *et al.* (1993), so those results will merely be summarized here. Each GBP is composed of two similar domains; the three hinge strands connecting them allow relative movements of the domains necessary for the protein's function (see discussion in Zou *et al.*, 1993, and references cited therein). Sequence changes are confined (with one exception) to surface residues, and appear to result in the different space groups obtained. Slight differences in the relative orientation of the two domains of GBP-E and GBP-S (2–3°) are apparently caused by different crystal packing. After accounting for the altered inter-domain angle and some smaller conformational changes associated with it, the agreement between the *Salmonella* and *E. coli* protein coordinates is of the same order as that between GBP-S1 and GBP-S2, *i.e.* 0.16 Å. Most of the solvent molecules of the GBP-S models are found in equivalent places in GBP-E. The fit of GBP-E to its electron-density map could not be evaluated, since structure factors were not available.

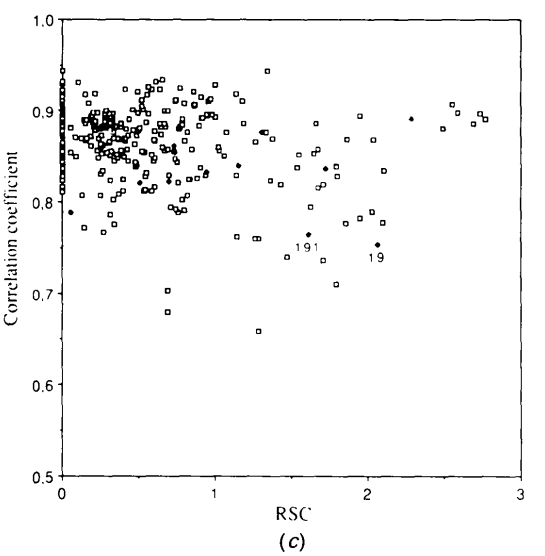
The pep_flip results obtained for GBP-E agree well with those of GBP-S1 and GBP-S2, with one exception. Residue Lys276 has a high value that is not seen in either of the *Salmonella* proteins; the peptide in question has the opposite orientation in GBP-E compared with its equivalent in the other two structures (Fig. 6). This portion of GBP-E has no hydrogen bonding to help define the main-chain conformation, either from the protein itself or through crystal contacts. The temperature factors are also locally higher than those found in the surrounding residues (18.5, 26.7, 31.1, 33.3, 29.3 and 23.9 Å$^2$




(*a*)




(*b*)




(*c*)

Fig. 5. (*a*) A two-dimensional plot of real-space fit correlation coefficients *versus* RSC values for GBP-S2. Residues that were considered to be well defined by the electron density are shown as □, while ones with poor electron density are shown as +. Residues previously mentioned as having poor fit to both rotamers and the electron density (see text and Table 3) are labelled by number. (*b*) Similar plot for an early model from the 2.4 Å refinement of GBP. (*c*) Similar plot for GBP-S1. Side chains that were determined to be the same in GBP-S2 (as indicated in Table 3) are shown as □, while ones that are different are shown as ●. Two residues that would most easily have been located as possible errors by this two-dimensional analysis are indicated by residue number. A value of $c$ = 1.04 was used for the real-space fit analysis in all three plots.

for the $C^{\alpha}$ atoms of residues 273–278). The peptide N atom of residue 276 of GBP-S is involved in a hydrogen bond with the side chain of Asp280; the resulting structural stabilization seems to give rise to a smaller variation in the main-chain $B$ factors (15.3, 18.3, 19.7, 18.6, 19.7, 19.7 Å$^2$ for the $C^{\alpha}$ atoms of residues 273–278 of GBP-S2). Residue 276 and those near it are in no way unusual with respect to their location in a Ramachandran plot (Ramakrishnan & Ramachandran, 1965) of any of these proteins. The $\varphi$ and $\psi$ values for residue 276 are $-92.2$ and $-9.5°$ ($\alpha$-helical region) for GBP-E, and $-86.0$ and $175.0°$ ($\beta$-strand region) for GBP-S2, respectively.

The deviations of side-chain $\chi$ angles in GBP-E from those of the rotamers are in general greater than observed for GBP-S1 and GBP-S2, although most do fit within the usual angular distributions. The average RSC values (0.615 where all residues were included, and 0.775 where glycine and alanine were excluded) are also slightly higher. These observations are not surprising, given both that database information was apparently not used in fitting GBP-E, and that somewhat looser stereochemistry was apparently allowed in the refinement of this structure (as shown in Table 1). As for GBP-S1, discussion of differences in side-chain conformation is restricted to those residues with clear electron density in GBP-S2. An RSC difference plot such as that discussed above for GBP-S1 is shown in Fig. 3(*b*). Some residues of GBP-E have different side-chain conformations, but there is no reason to suspect that they are not correct (such as when an altered sequence or crystal contact was involved). Some of the residues that differed fell into the non-rotamer class, and so deserve further inspection. Noteworthy in this category are five leucine residues (36, 37, 55, 99 and 250), four of which have $\chi_2$ values differing by 180° from the common rotamer conformations that are well supported by electron density in GBP-S2 (see Fig. 3*b*). The situation is exemplified by the case of Leu36 in Fig. 7. The RSC values for these leucine side chains of GBP-E fall between 1.5 and 2.1 Å. Some other side chains had non-rotamer conformations combined with high temperature factors, indicating that they were likely to be ill defined in the electron-density maps. An example of this is the conformation of Asp280, which was associated with the difference in the main chain of the models described above (see Fig. 6). The common rotamer observed for this residue in GBP-S results in improved hydrogen bonding in the region.

## Discussion

The utility of the database method was clearly demonstrated, both in refinement and in the final analysis of protein structures. Even a structure that is well refined at quite high resolution (as with the GBP-E structure mentioned here) can probably be improved in some details by reference to database information. The results suggest that a medium-resolution structure refined with the use of the databases can be as good as a higher resolution structure refined in the more traditional way. That this result can be obtained in a far less tedious fashion is obviously a bonus.

Some objective measure of the local fit of a structure to its electron-density map is desirable both for the location of errors and for highlighting portions that are ill defined for other reasons. The main advantages of using real-space fit values rather than the more common method of monitoring temperature factors lie in the facts that (*a*) they can be used with the original experimental map, early in refinement or at lower resolution when model temperature factors are not available, and (*b*) they encourage the user to concentrate on regions in which the fit to the electron density is poor, rather

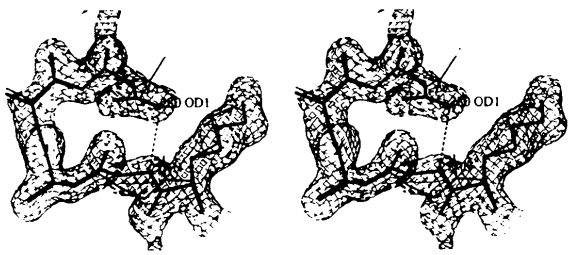



Fig. 6. The region surrounding Lys276 in GBP-S2 (thick lines) and GBP-E (thin lines). The hydrogen bonding between the main chain and side chain in the former protein is indicated by labelling the appropriate atoms. The corresponding region of the 1.7 Å resolution map for GBP-S2 is shown (dashed lines).

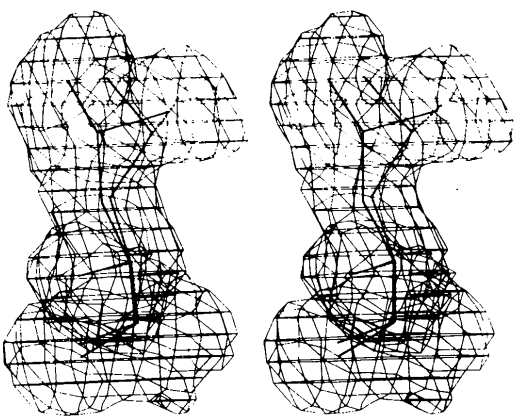


Fig. 7. Stereo plot showing the superimposed structures of GBP-S2 (solid line) and GBP-E (dashed line) at Leu36. The electron density of the current 1.7 Å resolution map for GBP-S2 is also shown.

than those with higher temperature factors (as found in loops, for example) but good density fit (see Mowbray & Cole, 1992, for example).

The pep_flip value represents a criterion independent of the Ramachandran plot for main-chain structural analysis. Pep_flip values greater than 2.0 Å (corresponding to an average difference greater than 65° from the peptide orientations normally found, given similar $C^{\alpha}$ positions) can be used to help locate unusual main-chain conformations. The largest value likely to be seen is approximately 3.6 Å, which represents a peptide orientation 180° away from the common ones. While not all high pep_flip values indicate errors in the structure, they do highlight places where closer inspection is warranted, and where features of structural interest occur. Most proteins will probably have some residues that 'disagree' with the current main-chain database, just as many have a few residues that are Ramachandran violations. This type of difference seems generally to arise from factors external to the segment of main chain under inspection, *i.e.* from the tertiary structure of the protein. Such situations often arise from special functions of the protein, such as those found at the sugar- and calcium-binding sites and hinge of GBP (residues 138, 181–182, 235–236 and 256; see Table 2). In other instances, larger pep_flip values are due to the existence of two strong structural clusters in the database (often with glycine or a bulky residue such as lysine, tryptophan or tyrosine at position 4 determining a subpopulation of conformations among the pentapeptides located from the database).

RSC values greater than 2.0 Å in a well refined high-resolution structure should generally be due to the legitimate presence of less common rotamers or non-rotamer conformations. The values lower down the list (in the range of 1.5–2.0 Å r.m.s. deviation) are often the real problem areas during refinement. This is largely due to the fact that refinement programs are rarely able to change from an incorrect rotamer to the correct one automatically, but the distortions resulting from poor side-chain packing will often show up as increases in the RSC value; the problem is often also associated with poor fit to the electron density. This correlation between the RSC and real-space fit values suggests that two-dimensional plots of these two properties, such as those shown in Fig. 5, should be a useful tool for tracking the progress of a refinement. RSC values in the 1.5–2.0 Å range for symmetrically branched side chains (leucine, valine and threonine) should be studied with particular attention, since they may represent differences of 180° in the terminal side-chain angle from the actual rotamer conformation, and will not necessarily be located by poor fit to the electron density.

About 97% of the residues of GBP-S2 have pep_flip values less than 2.0 Å, and 95% of the residues with clear electron density were accounted for by the common rotamers found in the *O* database. The cautious inclusion of less frequent main-chain or side-chain conformations should be possible based on sufficient structural and chemical evidence. GBP-S contains several instances where unusual main-chain or side-chain conformations are supported by good electron density, and have well defined structural origins. It should be noted, however, that the inclusion of less common conformations was highly correlated with possible errors in the lower resolution structures described here. A high pep_flip value found in GBP-E but not the GBP-S structures may represent a problem in the peptide orientation at this position of the former model. Of the non-rotamer conformations identified in GBP-S1 and GBP-E, 38 and 57%, respectively, were considered to be suspect or worthy of further evaluation. Closer attention to these differences from the database might have prevented some minor errors in both of these structures. It was concluded that the range of both main-chain and side-chain conformations incorporated into the *O* database is an effective one; it is correct often enough to be used on a near-automatic basis, and makes the user consider carefully before including less common conformations.

This work was supported by a grant from the Swedish Natural Science Research Council to SM (K-KU 9991-304) and funding to JZ from Uppsala University. The authors thank T. Alwyn Jones for many helpful comments on the manuscript and for support at many stages.

### References

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. T. JR, BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

BRÜNGER, A. (1988). *J. Mol. Biol.* **203**, 803–816.

BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). *Science*, **235**, 458–460.

DEISENHOFER, J. & STEIGEMANN, W. (1975). *Acta Cryst.* **B31**, 238–250.

DIAMOND, R. (1971). *Acta Cryst.* A**27**, 436–452.

HENDRICKSON, W. A. & KONNERT, J. H. (1980). *Indian Acad. Sci.* pp. 1–25.

JAMES, M. N. G. & SIELECKI, A. (1983). *J. Mol. Biol.* **163**, 299–361.

JANIN, J., WODAK, S., LEVITT, M. & MAIGRET, B. (1978). *J. Mol. Biol.* **125**, 357–386.

JONES, T. A. (1982). In *Computational Crystallography*, pp. 303–317. Oxford: Clarendon Press.

JONES, T. A., BERGDOLL, M. & KJELDGAARD, M. (1990). In *Crystallographic and Modeling Methods in Molecular Design*, edited by C. E. BUGG & S. E. EALICK, pp. 189–195. New York: Springer-Verlag.

JONES, T. A. & KJELDGAARD, M. O. (1992). *O: The Manual.* Uppsala, Sweden.
JONES, T. A. & LILJAS, L. (1984). *Acta Cryst.* A**40**, 50–57.
JONES, T. A. & THIRUP, S. (1986). *EMBO J.* **5**, 819–822.
JONES, T. A., ZOU, J.-Y., COWAN, S. W. & KJELDGAARD, M. (1991). *Acta Cryst.* A**47**, 110–119.
KONNERT, J. H. & HENDRICKSON, W. A. (1980). *Acta Cryst.* A**36**, 344–350.
MCGREGOR, M. J., ISLAM, S. A. & STERNBERG, M. J. E. (1987). *J. Mol. Biol.* **198**, 295–310.
MOWBRAY, S. L. (1992). *J. Mol Biol.* **227**, 418–440.
MOWBRAY, S. L. & COLE, L. B. (1992). *J. Mol. Biol.* **225**, 155–175.
MOWBRAY, S. L. & PETSKO, G. A. (1983). *J. Biol. Chem.* **258**, 7991–7997.
MOWBRAY, S. L., SMITH, R. D. & COLE, L. B. (1990). *Receptor*, **1**, 41–54.
PONDER, J. W. & RICHARDS, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.
RAMAKRISHNAN, C. & RAMACHANDRAN, G. N. (1965). *Biophys. J.* **5**, 909–933.
ROSSMANN, M. G. & LILJAS, L. (1974). *J. Mol. Biol.* **85**, 177–181.
SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S.-H. (1977). *Acta Cryst.* A**33**, 800–804.
VYAS, N. K., VYAS, M. N. & QUIOCHO, F. A. (1983). *Proc. Natl Acad. Sci. USA*, **80**, 1792–1796.
VYAS, N. K., VYAS, M. N. & QUIOCHO, F. A. (1987). *Nature (London)*, **327**, 635–638.
VYAS, N. K., VYAS, M. N. & QUIOCHO, F. A. (1988). *Science*, **242**, 1290–1295.
ZOU, J.-Y., FLOCCO, M. M. & MOWBRAY, S. L. (1993). *J. Mol. Biol.* **233**, 739–752.