Protein Engineering vol.10 no.4 pp.361–372, 1997

Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space

P.Tufféry[1], C.Etchebest[2] and S.Hazout

U155 INSERM, Centre de Bioinformatique, Université Paris 7, Tour 53, 1er étage, 2 place Jussieu, 75251 Paris cedex 05 and [2]URA 77 CNRS, Institut de Biologie Physico-Chimique, 13 rue P. et M.Curie, 75005 Paris, France

[1]To whom correspondence should be addressed

We have studied the effects of backbone inaccuracy on the efficiency of protein side chain conformation prediction using rotamer libraries. The backbones were generated by randomly perturbing the crystallographic conformation of 12 proteins and exhibit C_α r.m.s.d.s of up to 2 Å. Our results show that, even for a perturbation of the backbone fully compatible with the temperature factors of the proteins, the predicted side chain conformations of approximately 10% of the buried side chains remain variable. This fraction increases further for larger backbone deviations. However, for backbone deviations of up to 2 Å r.m.s.d., the predicted side chain r.m.s.d. varies only in a ratio of $<$1.4. Moreover, a possible strategy for obtaining side chain conformations close to the experimental ones consists of extracting the consensus conformations of the side chains from a series of backbone conformations. Such a procedure allows the computation of the side chain conformations with no loss of accuracy for backbones exhibiting r.m.s.d.s of up to 1 Å from the crystallographic coordinates. For larger backbone deviations (up to 2 Å r.m.s.d.) the r.m.s.d. of the buried side chains increases from 1.33 up to 1.60 Å. We also discuss the influence of the size of the rotamer library on the quality of the prediction.

Keywords: modelling/protein/rotamers/side chains

Introduction

When building protein structural models, once a frame for the backbone has been defined, one of the problems is the calculation of the side chain conformations. To achieve this complex goal, different strategies have been proposed. A commonly accepted hypothesis consists of assuming that, for homologous proteins, the side chain conformations remain close (Summers *et al.*, 1987; Summers and Karplus, 1989; Eisenmenger *et al.*, 1993; Wilson *et al.*, 1993; Laughton, 1994). Thus, one can derive side chain conformations from those of a sufficiently homologous protein. However, when modelling proteins that exhibit a low homology to any experimentally determined structures, this hypothesis becomes less and less acceptable as the number of substituted side chains increases and as the conformation of the backbone differs more and more. Several algorithms, designed for determining the optimal side chain conformations associated with a backbone conformation and only based upon an energy criterion, have been described. Different conceptual search strategies have been employed, such as simulated annealing (Holm and Sander, 1991, 1992; Lee and Subbiah, 1991; Niugles and Brunner, 1991), genetic algorithms (Tuffery *et al.*, 1991, 1993), heuristic search (Reid and Thornton, 1989), combinatorial search (Tuffery *et al.*, 1991, 1993; Wilson *et al.*, 1993), dead-end elimination (Desmet *et al.*, 1992) and neural networks (Hwang and Liao, 1995). In order to overcome the problem of exploring the complex energy hypersurface associated with the side chain conformations and despite the fact that it was suggested that rotamers may not correspond to any side chain conformational reality (Schrauber *et al.*, 1993), many approaches make use of rotamer libraries to limit the search to a small number of conformations for each type of side chain (Ponder and Richards, 1987; Reid and Thornton, 1989; Holm and Sander, 1991; Tuffery *et al.*, 1991, 1993; Desmet *et al.*, 1992). The use of such libraries has allowed the design of search methods that are fast enough to compute side chain conformations in a few seconds or minutes even for proteins having more than 200 residues. Their efficiency has been assessed by comparing the predicted side chain conformations obtained to those of the crystal structures.

However, little information is available about the robustness and sensitivity of such methods when applied in cases where the backbone is not 'perfect', in particular when building a model. Holm and Sander (1992) observed a decrease in the prediction accuracy for side chains positioned on the backbones rebuilt from the C_α data, compared with those obtained for backbones defined by their crystallographic coordinates. In a few cases, Wilson *et al.* (1993) studied the influence of the sequence homology on the prediction, by permutating the sequences of homologous proteins of known structure. A recent study (Chung and Subbiah, 1996) used a similar approach to tackle the problem of the relevance of side chain packing methods as a function of the homology to a template more systematically.

In the present study, we analyse the side chain prediction accuracy in the rotamer space as a function of the backbone error. To achieve this, we have simulated a series of backbone deformations for a set of 12 proteins and studied the stability of the side chain conformations as a function of the backbone deviation to its crystallographic conformation. In addition, several catalogues were used to assess the influence of the rotamer library choice.

Materials and methods

Selection of a collection of proteins

We chose to study a total of 12 proteins of known crystallographic structures (resolved to better than 2.4 Å and having *R* values $<$0.22). The structures were selected from a catalogue of non-redundant structures (Hobohm and Sander, 1994) from the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). They were also chosen in order to exhibit the different folds (α, β or α/β) according to Orengo *et al.* (1993, 1994) and to have sizes large enough so that each protein should have a well-defined internal core, since it is commonly recognized that side chain prediction is less efficient for residues located at

the surface of the proteins. Their protein codes (associated number of amino acids, fold type and fraction of buried residues) are 1bfg (126, β, 0.46), 1lz1 (130, α + β, 0.47), 1lis (131, α, 0.35), 1aak (150, α + β, 0.43), 1bgc (158, α, 0.47), 4gcr (174, β, 0.47), 1lmb (179, α, 0.44), 1gky (186, α/β, 0.44), 1sacA (204, β, 0.53), 1ahc (246, α + β, 0.52), 5timA (249, α/β, 0.56) and 1nbaA (253, α/β, 0.49), where A denotes the A chain for multimeric proteins built of repeated identical monomers. The fertilization protein 1lis exhibits a particularly low fraction of buried residues (0.35) while all other proteins exhibits ratios between 0.43 and 0.56. A visual inspection of this protein shows that its shape is more flat than globular. In fact, it consists of a three-helix bundle, with one helix crossing the others. Only a few residues have no contact with the solvent. Thus, we can consider that the core of this protein is an intermediate between a real 'hydrophobic core' and an exposed residue set.

These proteins exhibit a total of 2186 residues and the amino acid distribution is 187 (120) Ala, 139 (20) Arg, 104 (31) Asn, 113 (30) Asp, 32 (27) Cys, 97 (27) Gln, 125 (27) Glu, 150 (70) Gly, 32 (14) His, 127 (107) Ile, 197 (159) Leu, 120 (11) Lys, 42 (27) Met, 87 (65) Phe, 103 (32) Pro, 153 (52) Ser, 100 (43) Thr, 34 (26) Trp, 98 (60) Tyr and 146 (105) Val, where the numbers in parentheses correspond to the buried residues.

Determination of the exposed/buried residues

The solvent accessible surfaces were calculated using the method described by Richmond (1984), using a sphere of radius 1.4 Å. All residues having less than 20% of the accessible surface of the same residue in an Ala-X-Ala fragment with an α-helical conformation were classified as buried. The internal and external residues were detected according to their solvent accessible surfaces computed from the PDB files. These assignments were maintained whatever the backbone deformation was.

Energy computations

The energies were computed using the 'Flex' all-atom force field (Lavery *et al.*, 1986a,b). This force field is suited to internal coordinates and includes the standard van der Waals, torsion angle, electrostatic and hydrogen bond energy contributions. In our calculations, a sigmoidal dielectric function $\varepsilon(R)$ was used as a model for the dielectric damping of the electrostatic interactions between two charges in a polar solvent.

Generation of a collection of backbone conformations

For each of the 12 proteins, five sets of 50 different backbone conformations were generated within different root mean square deviation (r.m.s.d.) ranges. In one set (referred to as the set sTF), the backbone conformations were constrained so as to have deviations compatible with the temperature factors of the C_αs of the proteins. Other sets were built by selecting conformations having backbone r.m.s.d.s of between 0.25 and 0.5 Å (s05), 0.5 and 1.0 Å (s10), 1.0 and 1.5 Å (s15) and 1.5 and 2.0 Å (s20). Care was taken to select backbone conformations covering the whole range of r.m.s.d.s of each set. The mean C_α r.m.s.d.s of the sets, compared to the crystallographic structures, are 0.18, 0.37, 0.75, 1.25 and 1.75 Å for sets sTF, s05, s10, s15 and s20 respectively.

The conformations were generated using the following procedure.

(i) Randomly select a variable describing the backbone (only the ϕ or ψ dihedrals were considered and prolines were excluded).
(ii) Select a random dihedral perturbation to apply to the variable. The range of acceptable values usually employed was close to $\pm 1°$. By using such small steps, it is expected that the structures are not likely to exhibit large deviations from the crystallographic conformations.
(iii) Go to step (i) and repeat the process 1000 times. This large number of modifications ensures that a large number of residue conformations will be affected.
(iv) Superimpose the generated and the crystallographic conformations and check that the C_α r.m.s.d. is acceptable (see below).
(v) Check that the variation in the backbone–backbone energy relative to that of the crystallographic structure is $<10\%$. This was done to ensure that the generated backbone conformations, even if not optimal, are not unrealistic (i.e. no conformation exhibiting crossing backbones or backbone–backbone steric conflicts can be selected). It also avoids the bias due to the influence of non-acceptable backbone conformations in the side chain positioning. The energies were computed including all the atoms of the backbone.

The r.m.s. fit was performed using the procedure described by Sippl and Stegbuchner (1991). The selection criterion was to ensure that either the fit between the two structures was within a given range of the C_α r.m.s.d. or that the locations of the protein C_αs do not deviate more than the maximal deviation U associated with the temperature factor B of the crystallographic structure, using the relation $B = 8\times\pi^2\times U^2$.

If the r.m.s.d. or backbone–backbone energy criteria were not met, the whole procedure was restarted.

This approach was based upon the generation of random backbone conformations and was preferred to extracting conformations from trajectories of molecular dynamics (MD), since it is considerably faster: large r.m.s.d.s could only be obtained in MD either at high temperatures or with long simulation times. Nevertheless, the maximal variations in the backbone energy are quite compatible with that obtained by MD. In addition the procedure that we employed for positioning the side chains is able to build side chain conformations adapted to each backbone.

Determination of the side chain conformations

The calculation of the side chain conformations, given the backbone coordinates, was performed using the algorithm SMD (Tuffery *et al.*, 1991). This algorithm performs a conformational search in the rotameric space, based on an energy criterion. For the protein sizes considered in this study, it was shown to exhibit a good search efficiency (Tuffery *et al.*, 1993), so that the influence of the search on the results is expected to be negligible. The conformations resulting from the SMD algorithm are usually refined using a fast quasi-Newton minimization procedure (QNMP). It corresponds to a local refinement once the global conformational search has been performed. Since we want to compare the conformers obtained starting from different backbone conformations, this procedure was not used in this study.

Estimation of the side chain conformation deviations

The side chain conformation deviations were measured as the deviations from the side chain crystallographic conformations which were taken as reference, since one does not know the actual conformation that the side chains will adopt for the

different backbone conformations and since one aim of the present study was to assess the robustness of side chain positioning when it is performed starting from erroneous backbone conformations.

The r.m.s.d. for every side chain, performed in local orthonormal references built from the N, C_α and C atoms of each amino acid, is a measure independent of the variations between the backbone coordinates. This calculation is equivalent to superimposing the N, C_α, C and C_β atoms of the amino acids. Thus, only a measure of the difference in the side chain conformations is taken into account. The r.m.s.d.s were computed on the heavy atoms of the side chains.

We have also considered the differences in terms of χ agreement: two χ values were considered as 'similar' when their difference was $<\pm 40°$ (according to Hwang and Liao, 1995).

These two measures are somewhat different, since the χ_1 agreement does not warrant that the r.m.s.d.s are low for long side chains. In addition, a χ_1 and χ_2 agreement may still result in large r.m.s. values for bulky side chains. As an indication of the correspondence between these measures, rotating the χ_1 (χ_2 and $\chi_1 + \chi_2$ respectively) of the side chains of the 12 proteins by 40° leads to a mean r.m.s.d. of 1.43 (0.69 and 1.60 Å respectively).

Definition of the rotamer libraries

Three rotamer libraries were considered.

(i) The catalogue (RC1) defined by Ponder and Richards (1987) exhibits a total of 84 rotamers to describe the 20 amino acids.
(ii) The catalogue (RC2) defined by Tuffery *et al.* (1991) exhibits a total of 110 rotamers for the 20 amino acids.
(iii) An extended catalogue (RC3) including 214 rotamers was built from a survey of the buried residues of 200 non-redundant structures taken from the catalogue proposed by Hobohm and Sander (1994).

The method of determining the rotamers of RC3 is based on a combination of data segmentation and dynamic clustering and is identical to that used to determine those of RC2. Compared to RC2, the changes mostly concern residues having more than two χs: Glu was described by 12 rotamers, Lys by 49, Met by 17, Gln by 19 and Arg by 39. This improvement corresponds mainly to the description of χ_3–χ_5, for which we observed a tendency for the standard 60, −60 and 180° values. Compared to RC2, some supplementary rotamers were also introduced for residues having two χs: Leu (–84, 75; −167, −82; 64, 160), Ile (–79, 87; 68, 98), Asp (–168, 80; 57, 103), Asn (–167, −128; −169, 62; −170, −46; 61, 64; 58, −75; −75, 75; 68, 179), His (–66, 176; −162, 173; 60, −161), Trp (–173, 21), Tyr [(RC2 −64, 102) changed into (–67, 82) and (–61, 137); −68, −29]. For Asn, the rotamers were defined so that they sample the angular space, since no well-defined cluster is observed for values of χ_1 close to 60 and 180°. Instead, it seems that χ_2 can adopt any angular value. Overall, all the conformations described in RC1 or RC2 are represented in RC3, even if small differences are observed between the angular values of the corresponding rotamers (generally $<20°$). RC3 contains, in addition, some new conformations, particularly for the side chains having more than two χs.

Assessment of side chain conformational variability by an 'equivalent rotamer number'

To quantify the variability of the residue conformations observed for a series of the backbone of a given protein, we have used the Shannon entropy,

$$H = \Sigma_{i=1}^{nr} (-p_i \ln p_i)$$

where *nr* corresponds to the total number of rotamers describing a given amino acid and p_i denotes the frequency of occurrence of rotamer *i* observed in the 50 solutions. From this entropy, an equivalent rotamer number N_{equ} can be derived as

$$N_{equ} = e^H = \Pi_{i=1}^{nr} p_i^{-p_i})$$

This measure varies between 1 and *nr*. N_{equ} is strictly equal to 1 when only one rotamer is observed (i.e. the side chain does not move). N_{equ} takes the value *nr* when all the rotamers are equally observed. A residue described by three rotamers observed with frequencies 0.9, 0.05 and 0.05 exhibits a value of 1.48 for N_{equ}, while observed frequencies of 0.65, 0.34 and 0.01 exhibit a N_{equ} of 1.99. In this study, two thresholds were employed: 1.5 and 1.9. A value of 1.5 for N_{equ} can be considered as indicative of the fact that the side chain mainly adopts one given conformation but others are marginally observed. Such sites will be denoted as variable since the side chain conformation adopted for a series of different backbone conformations is not constant. A value >1.9 indicates that the side chain adopts at least two rotameric states with non-negligible frequencies. Such sites will be denoted as highly variable since the most frequently selected conformation is not selected in a large number of cases. Note that the variability measured here is different from the conformational instability of a side chain observed during an MD simulation. It is simply a measure of the side chain conformation stability when predicted from different backbone conformations. Thus, it measures how sensitive this side chain is to the perturbation of the backbone conformation.

Results

The side chain conformations were predicted for each of the 50 generated backbones describing sets sTF to s20 and using the three rotamer catalogues.

Side chain calculation for the crystallographic PDB backbone conformations

Table I reports the overall r.m.s.d.s calculated for the 12 proteins and for the three rotamer libraries. The conformations of the side chains issued either from a rotamer best fit (i.e. obtained by taking the rotamer that exhibits the lower r.m.s.d. to the crystallographic conformation for each side chain) or from the SMD algorithm are compared to their crystallographic conformations. The 'best fit' (BF) r.m.s.d. represents a quantification of how the rotamer library can describe the crystallographic side chain conformations, omitting any energy consideration. As expected, the narrower the discretization of the conformational space of the side chains the better the fits are. RC3 also exhibits the smallest difference between the BF_a (all residues) and BF_b (buried residues only) which suggests a better approximation for the external residues: the number of rotamers describing side chains preferentially located at the surface of the proteins was much greater in this library.

Considering the predicted conformations (SMD values), the r.m.s.d.s exhibit large differences compared to those of the BF values. As illustrated in a previous study (Tuffery *et al.*, 1993), this does not correspond to a failure of the search algorithm, but rather illustrates the fact that the BF conformations are not necessarily associated with the energy minimum within the rotamer space. This might also be partly due to the fact that we do not consider the crystal packing forces that might

Table I. Side chain r.m.s.d.s (Å): three rotamer sets (RC1, RC2 and RC3) are considered

Protein	RC1				RC2				RC3			
	BF[a]	BF[b]	SMD[a]	SMD[b]	BF[a]	BF[b]	SMD[a]	SMD[b]	BF[a]	BF[b]	SMD[a]	SMD[b]
1bfg	0.98	0.81	1.94	1.08	0.85	0.61	1.72	0.96	0.65	0.50	1.82	1.06
1lz1	0.96	0.63	1.65	0.96	0.97	0.62	1.60	1.14	0.66	0.50	1.68	1.09
1lis	0.89	0.70	2.26	2.07	0.78	0.59	2.28	1.82	0.61	0.52	2.49	1.85
1aak	1.06	0.84	2.01	1.88	1.01	0.70	1.92	1.73	0.81	0.64	2.13	1.66
1bgc	0.85	0.73	1.72	1.27	0.72	0.54	2.06	1.64	0.61	0.50	1.85	1.25
4gcr	1.11	0.95	2.03	1.95	0.87	0.73	1.73	1.32	0.71	0.68	1.83	1.07
1lmb	0.87	0.62	1.63	0.94	0.78	0.55	1.86	1.07	0.64	0.52	1.91	1.02
1gky	0.76	0.72	1.83	1.42	0.68	0.60	1.82	1.73	0.58	0.49	1.56	1.19
1sac	0.85	0.80	2.07	1.91	0.77	0.68	1.95	1.77	0.65	0.61	1.83	1.51
1ahc	0.91	0.69	2.00	1.67	0.87	0.62	2.02	1.63	0.61	0.51	1.87	1.28
5tim	0.91	0.82	1.84	1.53	0.83	0.74	1.76	1.33	0.69	0.64	1.60	1.31
1nba	0.88	0.74	1.86	1.73	0.75	0.68	1.75	1.45	0.57	0.48	1.87	1.69
Mean	0.91	0.75	1.90	1.53	0.82	0.63	1.87	1.46	0.64	0.54	1.87	1.33

BF: r.m.s.d.s obtained by approximating each side chain by its best fit rotamer.
SMD: r.m.s.d.s obtained by positioning the side chains using SMD.
[a]All side chains.
[b]Buried side chains only.

affect the side chain conformations. The influence of using RC1, RC2 or RC3 appears to be nil when we consider all the residues together (SMD_a), but a dependence appears for buried residues (SMD_b): the r.m.s.d. decreases down to a mean value of 1.33 Å for RC3. As a reference, the mean r.m.s.d. obtained by randomly assigning rotamers of RC2 to the 12 proteins is 2.9 Å.

Considering individual proteins, the situation appears more complex. The r.m.s.d.s vary from 1.60 to 2.49, while the r.m.s.d.s of the buried residues vary from 0.94 to 2.07. The effect of the rotamer library seems dissimilar. When considering all the side chains, some proteins present a smaller r.m.s.d. with a smaller number of rotamers: two-thirds of the proteins exhibit larger r.m.s.d.s with RC3 than RC2 or RC1. On the other hand, for buried residues, the situation is the opposite: two-thirds of the proteins are better predicted with RC3. The worst prediction of the study was obtained for the fertilization protein 1lis. In fact, the poor score obtained for its buried residues is mainly due to the poor prediction of Tyr111 and Tyr130. During the search, a conformation that flushes these side chains on the outside of the protein is selected. These two side chains thus become *de facto* exposed in their predicted conformations. This suggests that one must be careful with the assignment of the buried residues. However, for the remainder of the study, we kept the assignments of the buried residues as defined above.

In terms of the χ agreement (the fraction of the side chains having their χ exhibiting a deviation $<\pm 40°$ from the value observed within the crystallographic conformation), the values vary at around 72% for χ_1 and 60% for χ_2 when considering all the side chains together (buried and external). The agreement is better for buried residues, with mean values close to 82% for χ_1 and 72% for χ_2. The best values are obtained for RC3 where 85% of χ_1 are correctly predicted for buried residues and 73% for χ_2.

Side chain calculation for the different sets of perturbed backbones

Table II shows the mean r.m.s.d.s obtained for each protein and for the different backbone sets. A comparison of the r.m.s.d.s obtained when positioning the side chains on the PDB backbone with those obtained for set sTF shows that the mean values are similar overall. In some cases, notably for granulocyte colony-stimulating factor (1bgc) with RC2, the mean r.m.s.d. is lower than the r.m.s.d. obtained with the crystallographic backbone. When considering the standard deviations we observe that some of them are as high as 0.26 for the fertilization protein 1lis with RC2. In that case, the r.m.s.d. obtained for one backbone can differ by as much as ± 0.5 Å from the mean value: the side chain r.m.s.d. can be as high (as low) as 2.6 Å (1.5 Å) for the buried residues. Thus, a very small backbone perturbation can induce large conformational changes in the side chains. However, the mean standard deviations remain generally close to 0.1 Å, suggesting that for set sTF, the r.m.s.d. variations are restricted to ± 0.2 Å.

For increasing values of backbone error, the r.m.s.d.s tend to increase. Figure 1A shows the variation in the mean r.m.s.d.s associated with sets sTF to s20 for the buried residues and for the different rotamer libraries. The error bars correspond to the confidence intervals of the means. From sets s10 to s20, the means deviate significantly from set sTF. From set s05, the behaviour is approximately linear and the best results are obtained for RC3. Note that the slopes of the curves reflect a variation in the side chain r.m.s.d. which is much smaller than that of the backbone.

In terms of χ agreement, Figure 1B shows that the fraction of χ_1 predicted within $\pm 40°$ decreases for the three rotamer libraries from values of between 80 and 85% for set sTF down to values close to 70% for set s20. Again, for sets s10 to s20, the mean values are significantly smaller than that of set sTF. The same tendency is observed for χ_2, with values decreasing from close to 73% down to 60–65%. We also observe that RC3 gives the best results for χ_1 but the worst for χ_2.

Figure 2A shows the mean r.m.s.d.s for set sTF obtained for each residue type (on all the proteins, for buried residues). The average r.m.s.d. is 0.90 Å. For most residue types, small r.m.s.d. variations are observed between the profiles obtained for the different rotamer libraries, except for His which is predicted worse with RC1, Trp for which the prediction is improved from RC1 to RC3 and Arg for which the best results are obtained with RC2. For most of the residue types, the

Table II. Overall buried side chain r.m.s.d.s computed upon 12 proteins, considering different rotamer catalogues (RC1, RC2 or RC3) and different backbone sets (sTF, s05, s10, s15 and s20)

	Protein	sTF		s05		s10		s15		s20	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RC1	1bfg	0.98	0.09	1.09	0.19	1.48	0.32	1.56	0.26	1.74	0.30
	1lz1	1.08	0.09	1.18	0.21	1.28	0.24	1.78	0.49	2.09	0.38
	1lis	2.04	0.10	2.00	0.17	2.02	0.34	2.27	0.31	2.35	0.31
	1aak	1.89	0.05	1.82	0.12	1.83	0.16	1.97	0.21	2.07	0.24
	1bgc	1.37	0.11	1.46	0.14	1.69	0.16	1.91	0.21	2.01	0.21
	4gcr	1.92	0.16	1.98	0.15	2.10	0.17	2.18	0.17	2.22	0.19
	1lmb	0.94	0.09	1.00	0.11	1.11	0.19	1.30	0.22	1.40	0.20
	1gky	1.56	0.16	1.61	0.17	1.67	0.20	1.82	0.21	1.94	0.23
	1sac	1.89	0.17	2.03	0.22	2.05	0.28	2.19	0.27	2.34	0.26
	1ahc	1.70	0.08	1.68	0.06	1.75	0.12	1.79	0.16	1.88	0.21
	5tim	1.51	0.04	1.56	0.10	1.66	0.14	1.76	0.18	1.83	0.14
	1nba	1.75	0.11	1.80	0.15	1.91	0.17	1.93	0.18	1.98	0.15
RC2	1bfg	0.99	0.09	1.05	0.18	1.43	0.35	1.48	0.30	1.58	0.28
	1lz1	1.19	0.18	1.20	0.16	1.33	0.23	1.82	0.40	2.07	0.41
	1lis	2.17	0.26	2.02	0.29	2.07	0.31	2.23	0.33	2.37	0.33
	1aak	1.75	0.10	1.76	0.12	1.83	0.16	1.97	0.18	2.09	0.23
	1bgc	1.35	0.09	1.36	0.10	1.56	0.21	1.88	0.20	1.95	0.22
	4gcr	1.43	0.17	1.62	0.18	1.75	0.18	1.95	0.20	1.93	0.21
	1lmb	1.02	0.04	1.10	0.11	1.20	0.18	1.36	0.22	1.48	0.24
	1gky	1.50	0.16	1.50	0.14	1.56	0.20	1.65	0.25	1.88	0.20
	1sac	1.80	0.16	1.86	0.19	1.92	0.26	2.05	0.26	2.27	0.30
	1ahc	1.59	0.07	1.61	0.08	1.62	0.10	1.64	0.18	1.81	0.22
	5tim	1.35	0.07	1.42	0.12	1.57	0.15	1.69	0.17	1.80	0.17
	1nba	1.40	0.13	1.51	0.20	1.63	0.22	1.80	0.22	1.87	0.19
RC3	1bfg	0.83	0.11	0.93	0.22	1.30	0.30	1.42	0.26	1.53	0.34
	1lz1	1.05	0.05	1.08	0.06	1.20	0.21	1.68	0.44	1.93	0.38
	1lis	1.88	0.16	1.80	0.28	1.88	0.35	2.15	0.39	2.32	0.36
	1aak	1.68	0.05	1.68	0.12	1.85	0.20	1.99	0.23	2.12	0.27
	1bgc	1.31	0.13	1.37	0.20	1.68	0.23	1.96	0.25	2.02	0.21
	4gcr	1.34	0.20	1.33	0.20	1.46	0.22	1.63	0.20	1.69	0.24
	1lmb	0.98	0.08	1.02	0.10	1.15	0.17	1.28	0.20	1.47	0.22
	1gky	1.41	0.14	1.46	0.16	1.54	0.20	1.75	0.25	1.93	0.21
	1sac	1.59	0.14	1.71	0.16	1.73	0.22	1.82	0.27	2.06	0.23
	1ahc	1.48	0.12	1.48	0.12	1.54	0.16	1.63	0.21	1.77	0.26
	5tim	1.37	0.05	1.43	0.11	1.53	0.14	1.68	0.18	1.75	0.16
	1nba	1.68	0.08	1.76	0.09	1.88	0.14	1.94	0.12	1.98	0.18

The mean r.m.s.d.s and associated standard deviations obtained from 50 backbone conformations.

r.m.s.d. values are close to the average value. Above this threshold, we find essentially the aromatic amino acids (Trp and Tyr), the longest chains (Arg and Lys), Gln and Glu.

For larger backbone deviations, the residue prediction is affected as shown in Figure 2B. It shows the r.m.s.d. differences with respect to set sTF for the different backbone sets. It reports only the results obtained with RC3, but similar profiles are observed with RC1 and RC2. For all the rotamer libraries and all the residue types, the divergence relative to set sTF increases with the backbone deviation. Note that the divergence values remain much smaller than the reference values of set sTF. Between sets s05 and s20, the mean r.m.s.d. varies at a ratio of 4.7 (i.e. 0.37 to 1.75) for the backbone, but at a ratio <2 for the different types of side chains. Beside this general tendency, the relative behaviour of each amino acid can be strongly different. Ser, for instance, is the least sensitive to the backbone variations whereas large differences are observed for His. Globally, the largest differences are observed for the same residues as mentioned above, the biggest and the longest side chains.

Discussion

Comparison of the behaviour of exposed versus buried residues

First, we briefly discuss our observations concerning the exposed versus buried residues relative to the crystallographic backbones. Obviously, increasing the number of rotamers mostly seems to improve the prediction of the buried residues, despite the fact that for RC2 and RC3 it is mainly for amino acids preferentially located on the surface that the number of rotamers has been increased.

In fact, among the pseudo-optimal rotamer conformations that exhibit energies no more than 10% greater than that of the optimum found by the SMD algorithm, we observe that the side chains for which different rotamers are selected correspond mostly to the external side chains, in a ratio close to three times that for the buried side chains. This fact simply reflects that the surface residues are more labile than the buried ones for a smaller energy cost and, hence, suggests a weak power of discrimination for the exposed residues. This situation could be changed if the solvent effect could be taken into

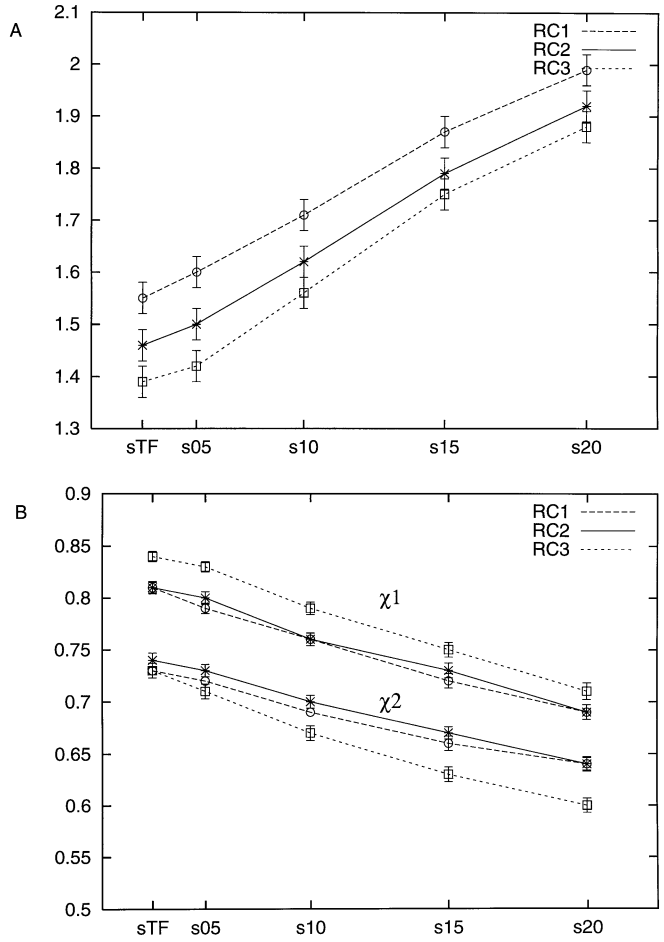

Fig. 1. (**A**) Overall mean r.m.s.d.s for the predicted conformations of the buried side chains of 12 proteins, for three rotamer libraries (RC1, RC2 and RC3). The profiles are a function of the mean backbone r.m.s.d. to the crystallographic conformation. The error bars correspond to the confidence intervals of the means. (**B**) Corresponding overall fraction of χ agreement. Top: fraction of χ_1 agreement. Bottom: fraction of χ_2 agreement.

account. In addition, it has to be noted that, due to their larger lability, the side chain conformations of the exposed residues are less precisely determined by crystallography (in some cases, they might simply be constructed using standard rotameric conformations).

Due to our definition of the buried/exposed residues, the question remains of evaluating how much the poor external side chain prediction can affect the prediction of the buried side chains. To evaluate this, we have performed a search with the external side chains frozen in their crystallographic conformations. The new mean r.m.s.d.s for the buried residues are 1.47, 1.34 and 1.37 Å for RC1, RC2 and RC3 respectively. Thus, a maximal difference of 0.12 Å is obtained for RC2. Clearly, considering the average accuracy ranges observed in this study, the influence of the surface amino acid on the buried residue conformations is weak.

Side chain conformational variability for non-optimal backbone

We now investigate the mechanisms underlying the decrease in the prediction efficiency. First, we analyse the variability of the side chain conformations observed for the different backbone sets. Table III shows the fractions of side chains for which the equivalent rotamer number N_{equ} is >1.5 or 1.9. When considering all the side chains, we observe that, for the three rotamer libraries, at least 21% of the side chains can be labelled as variable (i.e. they appear to be affected by the backbone modification) for set sTF. This fraction evolves up to a maximal value of 71% for set s20 using RC3. For buried residues alone, the fraction of the side chains affected is systematically higher (from 23 to 79%). In addition, we observed an increasing difference between the buried and all residues: while the difference is close to 0 for set sTF, we observed differences of 3–4% (3–6, 6–9 and 8–11%) for set s05 (sets s10, s15 and s20 respectively). Concerning the highly variable side chains, their proportion is at least 12% for set sTF and increases further up to 67% for set s20. Again, a slight increasing difference (0–10% from sets sTF to s20) is observed between the proportion obtained for all and the buried side chains. Moreover, we also noted a larger fraction of highly variable side chains for RC2 and RC3.

These results suggest that the loss of prediction efficiency can be attributed more to an increasing number of mispositioned side chains rather than to larger deviations for a small number of residues. Concerning the core of the proteins, the results suggest that the stronger steric constraints lead to an enhanced conformational variability in terms of the rotamers. As a corollary, increasing the number of rotameric states describing each type of side chain facilitates the adaptability of the side chain conformations, but leads to larger fractions of highly variable side chains.

We have investigated the location of the side chains that appear highly variable to analyse whether the fluctuating side chains are randomly distributed within the structures. Figure 3 shows the location of the highly variable buried side chains for 5timA, for sets sTF to s15. As can be seen, for set sTF the highly variable side chains are not uniformly distributed within the structure. For larger backbone deviations, the variability seems to evolve by diffusing from ‘seeds’ corresponding to the side chains highly variable for set sTF (here three main clusters). To evaluate whether this scheme corresponds to any general tendency, we have studied the fraction of the highly variable side chains that are close to each other. The spatial proximity of the side chains was simply detected using an interatomic distance criterion, supplemented by an angular condition upon the $C_{\alpha_i} - C_{\alpha_j}$ vector and the two $C_{\alpha_i} - B_i$ $C_{\alpha_j} - B_j$ vectors, where B_i and B_j are the centres of geometry of the atoms of the side chains i and j. This last condition selects the side chains that face each other. Table IV shows that even for set sTF, nearly 75% of the buried residues that are highly variable (representing close to 12% of the buried residues; see Table III) are close to another highly variable residue, while the non-highly variable sites have only ~45% highly variable neighbours. Considering all the residues, this fraction (see Figure 3B–D for 5timA) increases, on average for the three rotamer libraries, from 65% for set sTF to 92% for set s20. It is larger for buried residues. This suggests that one part of the variability is conditioned by the side chain dependence in the core of the protein. It seems to ‘diffuse’ from the seeds observed for set sTF. However, the distribution of the seed side chains within the structures is not uniform.

Finally, since the proteins belong to different structural classes, we have analysed the differences observed between the classes to check whether the different structural constraints associated with the different classes influence the side chain conformational variability. For example, for set sTF and RC1,

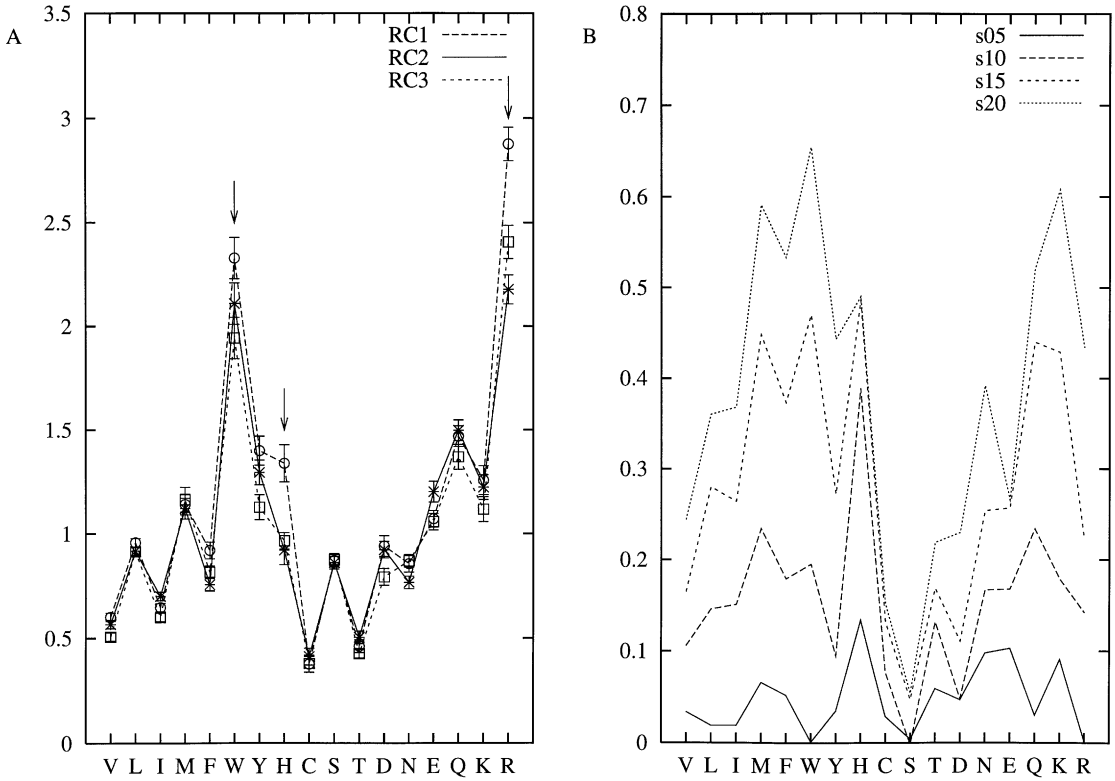

Fig. 2. (**A**) Mean r.m.s.d.s for the predicted conformations of the different types of buried side chains of 12 proteins having their backbone perturbations compatible with the crystallographic temperature factors. The arrows indicate residues exhibiting large differences for the different rotamer libraries. The error bars correspond to the confidence intervals of the means. (**B**) Differences between the mean r.m.s.d.s observed for set sTF and those obtained for larger backbone perturbations, for the rotamer library RC3.

Table III. Fraction (%) of side chains having an equivalent number of conformations N_{equ} >1.5 or 1.9. RC1, RC2 and RC3 correspond to different rotamer catalogues. All or only the buried side chains are considered. Side chains that cannot be described by more than one rotamer (Gly, Ala and Pro) were not included

N_{equ}		sTF		s05		s10		s15		s20	
		All	Buried	All	Buried	All	Buried	All	Buried	All	Buried
1.5	RC1	21	23	35	39	49	57	60	69	64	75
	RC2	24	24	38	41	53	59	62	71	67	77
	RC3	31	30	43	43	59	62	67	73	71	79
1.9	RC1	11	12	21	26	34	40	44	54	50	60
	RC2	12	12	24	26	37	43	48	56	52	62
	RC3	19	16	28	27	44	46	54	59	60	67

the fraction of highly variable side chains appears the highest for α proteins (14%) and the lowest for α/β (9%). Looking at the ratio of the 'fraction of highly variable side chains having at least one highly variable neighbour' on the 'fraction of non-highly variable side chains having at least one highly variable neighbour' gives values of 2 for β proteins, 1.8 for α + β proteins, 1.6 for α/β proteins and 1.0 for α proteins. This would indicate a stronger dependence between side chains for β than for α proteins. This indication remains however to be confirmed upon a larger set of proteins to obtain reliable statistics.

Side chain positioning for a non-optimal backbone: consensus conformations

The results discussed above are based on the average properties computed from 50 different simulations in given backbone deviation ranges. Even if small, the standard deviations associated (Table II) clearly show that, for a given backbone deviation, different possibilities exist for positioning of the side chains. Therefore, one can ask how much information on the side chain crystallographic conformations can be retrieved from the different conformations adopted by the side chains for the different backbones. We have considered the consensus conformations, namely the most probable rotamer combination deduced from the series of conformations obtained by simulation. Table V shows the r.m.s.d.s obtained for the buried residues by using such a consensus deduced from the series of 50 backbones. The prediction appears better than that obtained previously (Table II). In all cases, the analysis of the differences between the consensus and the mean data (using a signed rank test) shows that the consensus significantly improves the prediction. For set sTF, the consensus brings only a few improvements to the prediction. However, for sets

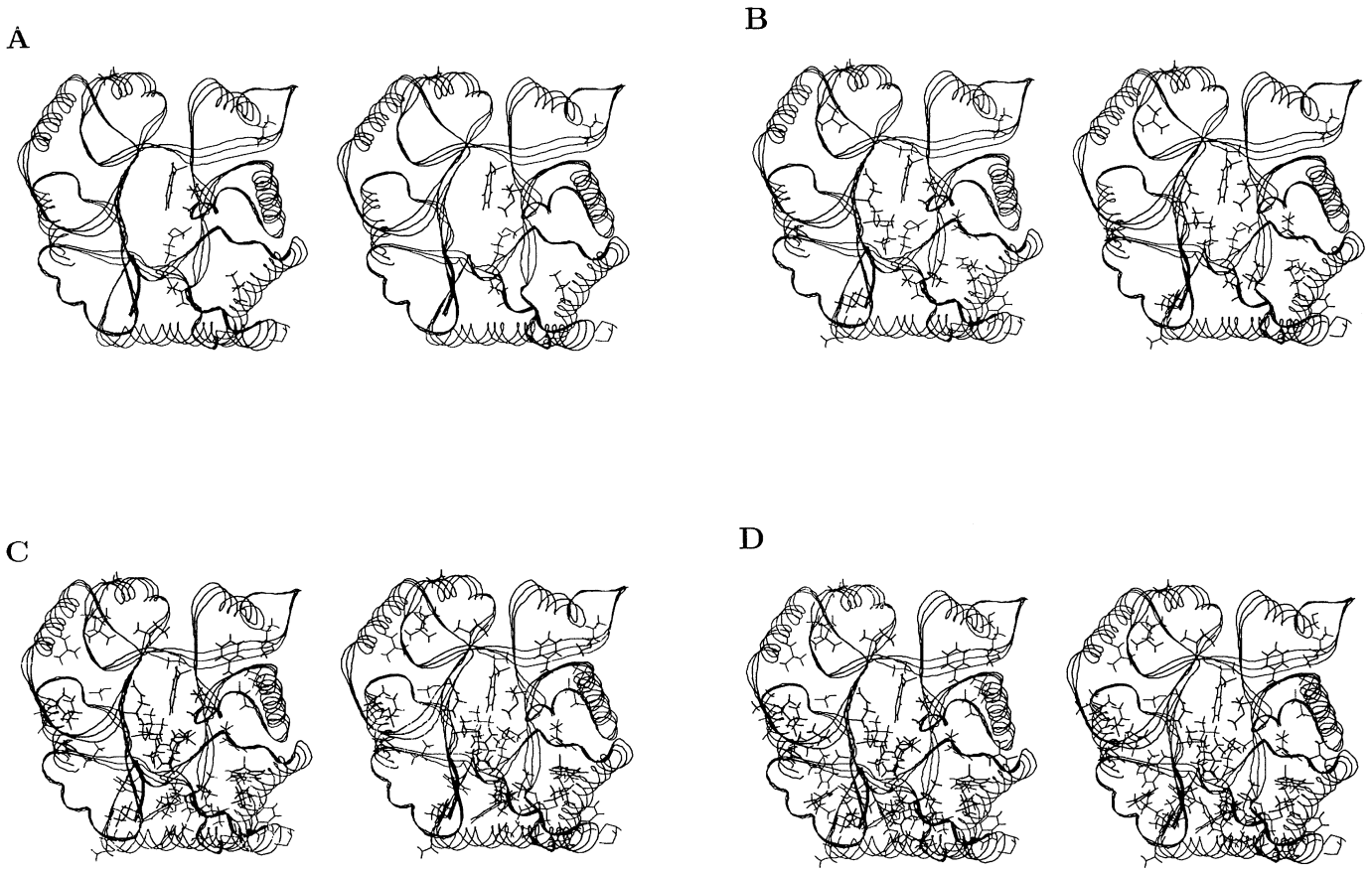

Fig. 3. Location of the highly variable residues of 5tim for (**A**) set sTF, (**B**) set s05, (**C**) set s10 and (**D**) set s15. Stereo views obtained with XmMol (Tuffery, 1995).

Table IV. Fraction (%) of side chains having at least one highly variable neighbour

N_{equ}		sTF		s05		s10		s15		s20	
		All	Buried	All	Buried	All	Buried	All	Buried	All	Buried
Highly variable (⩾1.9)	RC1	56	65	77	81	85	94	90	97	91	98
	RC2	68	71	77	88	87	94	92	98	92	99
	RC3	71	84	81	89	88	95	93	99	94	99
Non-highly variable (<1.9)	RC1	28	41	47	69	63	86	66	89	65	93
	RC2	31	45	50	69	60	85	68	91	68	93
	RC3	42	56	53	74	68	91	69	93	74	97

For non-highly variable side chains, side chains that cannot be described by more than one rotamer (Gly, Ala and Pro) were not considered.

s05 to s20, a much larger improvement (range 0.08–0.30 Å) is observed. We also observed that the mean r.m.s.d.s obtained from the consensus appear less sensitive to the rotamer library choice than the means reported in Table II. The total results show, within the rotamer space, a decrease in the quality of the prediction of the buried residues from 1.3 Å for a backbone uncertainty compatible with the crystallographic temperature factors up to 1.6 Å for backbones having their C_α r.m.s.d.s between 1.5 and 2.0 Å. Comparing these results with those obtained from homology data (Chung and Subbiah, 1996), but using a non-discrete (rotameric) space for the search, we observed similar tendencies. However, the present results show a much more linear evolution of the side chain r.m.s.d.s for backbone r.m.s.d.s up to 2 Å. Indeed, the consensus values remain below 1.7 Å (the average values remaining below 2 Å), while for target/template main chain r.m.s.d.s of between 1.5 and 2.0 Å, the buried side chain r.m.s.d. error deduced from the homology data is between 1.8 and 3.5 Å. We must however remark that the number of structures simulated in the present study is much higher (600 different set s20 backbones considered) and that, for some set s20 individual simulations, we found buried side chain r.m.s.d.s close to 3 Å.

To analyse when the consensus differs from that of set sTF and because of the weak number of observations (12 proteins), we have used a signed rank test. It indicates that the series can be considered non-distinguishable from set sTF up to set s10 for RC1 and set s15 for RC2 and RC3. Thus, even if we observe an influence of the rotamer library, we can estimate that the limit amongst which the consensus differs is close to set s15, i.e. close to 1.5 Å r.m.s.d. We have analysed whether

Table V. R.m.s.d.s obtained for the consensus rotamer of the buried side chains of the proteins, RC1, RC2 and RC3 correspond to different rotamer catalogues. sTF, s05, s10, s15 and s20 correspond to different backbone sets

	Protein	sTF		s05		s10		s15		s20	
		cns	Δ_{m-cns}	*cns*	Δ_{m-cns}	*cns*	Δ_{m-cns}	*cns*	Δ_{m-cns}	*cns*	Δ_{m-cns}
RC1	1bfg	0.88	0.10	1.05	0.04	1.07	0.41	1.28	0.00	1.51	0.23
	1lz1	1.02	0.06	1.02	0.16	1.15	0.13	1.04	0.74	1.14	0.95
	1lis	2.07	–0.03	2.01	–0.01	2.12	–0.10	2.12	0.15	2.26	0.09
	1aak	1.88	0.01	1.88	0.06	1.65	0.18	1.67	0.30	1.97	0.10
	1bgc	1.26	0.11	1.23	0.23	1.56	0.13	1.74	0.17	1.77	0.24
	4gcr	1.75	0.17	1.86	0.12	1.94	0.16	2.17	0.01	2.02	0.20
	1lmb	0.94	0.00	0.92	0.08	0.92	0.19	0.88	0.42	0.93	0.47
	1gky	1.46	0.10	1.51	0.10	1.48	0.19	1.67	0.15	1.72	0.22
	1sac	1.68	0.21	1.68	0.35	1.70	0.35	1.71	0.48	1.80	0.54
	1ahc	1.60	0.10	1.60	0.08	1.62	0.13	1.63	0.16	1.63	0.25
	5tim	1.52	–0.01	1.44	0.12	1.52	0.14	1.53	0.23	1.52	0.31
	1nba	1.75	0.00	1.75	0.05	1.78	0.13	1.81	0.12	2.03	–0.05
m		**1.48**	**0.06**	**1.49**	**0.11**	**1.54**	**0.17**	**1.60**	**0.24**	**1.69**	**0.30**
RC2	1bfg	0.97	0.02	0.97	0.08	1.03	0.40	1.04	0.44	1.47	0.11
	1lz1	1.00	0.19	0.99	0.21	1.02	0.31	0.97	0.85	1.07	1.00
	1lis	2.21	–0.04	2.01	0.01	2.13	–0.06	2.08	0.15	2.11	0.26
	1aak	1.70	0.05	1.70	0.06	1.72	0.11	1.64	0.33	2.01	0.08
	1bgc	1.37	–0.02	1.35	0.01	1.24	0.32	1.76	0.12	1.64	0.31
	4gcr	1.27	0.16	1.52	0.10	1.54	0.21	1.66	0.29	1.67	0.26
	1lmb	0.98	0.04	0.98	0.12	0.99	0.21	0.98	0.38	1.09	0.39
	1gky	1.34	0.16	1.31	0.19	1.34	0.22	1.31	0.34	1.79	0.11
	1sac	1.74	0.06	1.65	0.21	1.78	0.14	1.91	0.14	1.99	0.28
	1ahc	1.56	0.03	1.56	0.05	1.52	0.10	1.54	0.10	1.66	0.15
	5tim	1.34	0.01	1.24	0.18	1.20	0.37	1.34	0.35	1.48	0.32
	1nba	1.38	0.02	1.47	0.04	1.51	0.12	1.73	0.07	1.61	0.26
m		**1.40**	**0.05**	**1.39**	**0.10**	**1.41**	**0.20**	**1.49**	**0.30**	**1.63**	**0.29**
RC3	1bfg	0.76	0.07	0.79	0.14	0.88	0.42	0.88	0.56	1.52	0.01
	1lz1	1.03	0.02	0.96	0.12	1.05	0.15	1.02	0.66	1.10	0.83
	1lis	2.04	–0.16	1.89	–0.09	2.04	–0.16	2.04	0.13	1.91	0.41
	1aak	1.68	0.00	1.67	0.01	1.63	0.22	1.69	0.30	2.12	0.00
	1bgc	1.25	0.06	1.17	0.20	1.38	0.30	1.70	0.26	1.77	0.25
	4gcr	1.23	0.11	1.22	0.11	1.14	0.32	1.48	0.15	1.48	0.21
	1lmb	0.89	0.09	0.89	0.13	0.96	0.19	0.86	0.42	1.04	0.43
	1gky	1.38	0.03	1.38	0.08	1.37	0.17	1.40	0.35	1.91	0.02
	1sac	1.37	0.22	1.61	0.10	1.54	0.19	1.61	0.21	1.53	0.53
	lahc	1.40	0.08	1.36	0.12	1.49	0.05	1.46	0.17	1.55	0.22
	5tim	1.35	0.02	1.43	0.00	1.45	0.08	1.39	0.29	1.46	0.29
	1nba	1.69	–0.01	1.73	0.03	1.72	0.16	1.73	0.21	1.82	0.16
m		**1.33**	**0.04**	**1.34**	**0.08**	**1.38**	**0.17**	**1.43**	**0.31**	**1.60**	**0.28**

cns: r.m.s.d. obtained from the consensus file. *m*: means of all proteins.
Δ_{m-cns}: improvement brought about by using the consensus rotamer compared to the mean r.m.s.d.s obtained for the series of 50 backbones of the different sets.

the consensus gives results comparable to those obtained for the crystallographic conformations. For RC1, no significant difference was found up to set s20. For RC2 and RC3, set s20 appears significantly different. We must however remark on the weak discrimination of the statistics. The profiles of Figure 4 suggest that the limit is somewhat closer to 1.0 Å. We have investigated whether the decrease in the efficiency of the consensus can be attributed to the fact that for increasing backbone deviations, the consensus is less informative or whether some mechanism inherent to the packing of the side chains drives their conformations towards another rotamer combination or both. To check this, we have compared the mean frequencies of occurrence of the rotamers defining the consensus for each side chain, for the different backbone sets. For RC3, the corresponding means are 0.84 (sTF), 0.81 (s05), 0.73 (s10), 0.68 (s15) and 0.64 (s20). A similar decrease is obtained for RC1 and RC2. This suggests that the main cause of the decrease of the consensus efficiency is because the consensus becomes statistically less representative when the backbone deviation increases.

Consequently, using the consensus appears a means to obtain an enhanced side chain positioning when the backbone conformation is not optimal, such as when building a structural model. What is the optimal number of backbone conformations from which the consensus should be built? We have studied the evolution of the mean prediction r.m.s.d. obtained for a consensus built from a variable number of combinations (here 3, 5, 10 or 25), randomly selected amongst the 50 backbone conformations of each protein, for each of sets sTF to s20. As a reference, the prediction obtained for all 50 backbones is reported. Figure 4 shows for RC3 that increasing the size of the consensus leads in all cases to better-predicted conformations. Similar profiles are observed for RC1 and RC2. For sizes larger than 10 backbones, the improvement appears mainly located in the sets having the largest backbone deviations. For sizes larger than 25 backbones, the improvement becomes

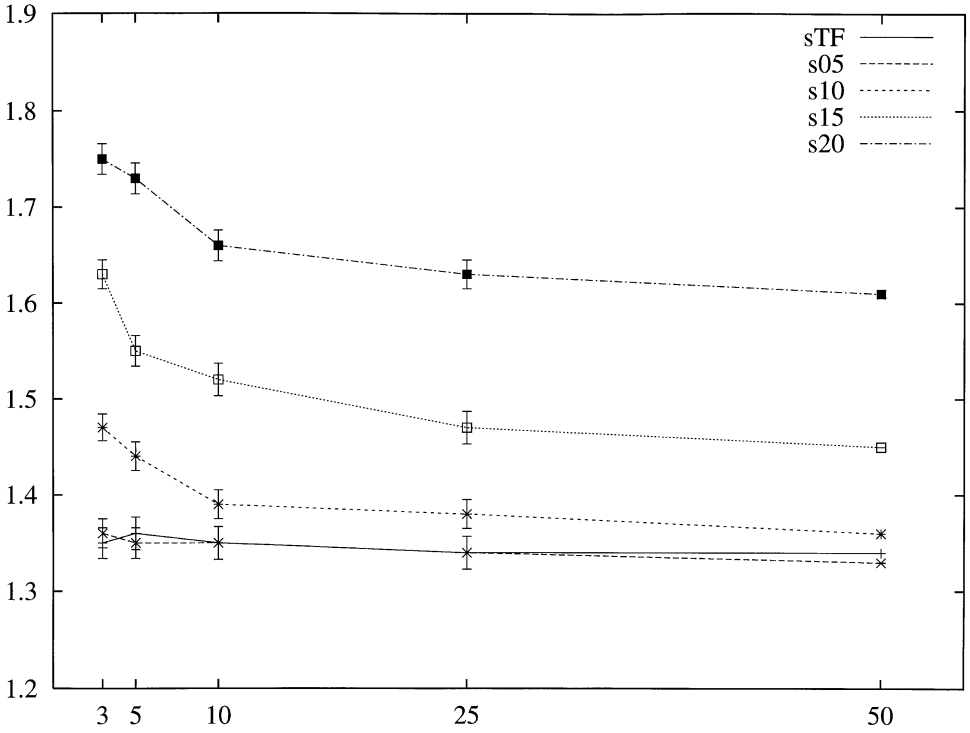

Fig. 4. Overall mean r.m.s.d.s for the predicted conformations of the buried residues of 12 proteins, using as the predicted conformation the consensus rotamer obtained with RC3 from a backbone number varying from 3 to 50. The error bars correspond to the standard deviations. The different lines correspond to the different ranges of perturbation of the backbones.

weak in all cases. In addition, we observed that the overall best prediction is obtained for set s05. Thus, it seems that perturbing a backbone by deviations from 0.25 to 0.5 Å corresponds to the flexibility necessary to allow the best repositioning of the side chains. This suggests a simple procedure to increase the robustness of the side chain conformation prediction for the structural models.

Influence of the rotamer library

We first discuss the relative prediction accuracy for the different rotamer libraries. As shown by Table II and Figure 1A, we observe a tendency towards better results for larger libraries. The best results are obtained for RC3 and, for buried residues, RC1 exhibits significantly higher means than RC2 and RC3, while RC3 appears better than RC2 (the error bars are associated with 2 standard deviations of the mean). For a given backbone set, the mean r.m.s.d.s tend to decrease from RC1 to RC3, while the variabilities remain similar. This indicates that the use of a more detailed rotamer library increases the accuracy of the prediction. This improvement tends, however, to be diminished when the backbone perturbation increases.

Looking separately at the proteins, the 'better prediction when using the largest library' rule is not met in some cases: for the lyzozyme (1lz1), the fertilization protein (1lis) and the lambda repressor (1lmb) where RC1 gives a lower r.m.s.d. than RC2 and for the *N*-carbamoylsarcosine amidohydrolase (1nba) where RC2 gives a lower r.m.s.d. than RC3. The case of 1lis has already been discussed above. For 1nbaA, an examination of the predicted conformations obtained, starting from the crystallographic backbone, shows that it is mostly due to a mismatch of the conformations of two residues in contact (Phe17 and Tyr206) when using RC3. Between RC2 and RC3, the Phe rotamers differ only slightly ($<10°$), while the rotamers describing Tyr were modified in RC3: one rotamer was split in two and another was added. In fact, the conformation selected by RC2 is included in RC3 but differs by 20° on χ_2. This deviation seems to trigger the selection of another pair of conformations for the two residues. Including the rotamers of RC2 within RC3 leads back to a correct prediction, as observed for RC2. Thus, a difference of only 20° on χ_2 for one rotamer of Tyr can induce a change in the combination of the rotamer selected. Furthermore, this change is preserved amongst the series of perturbed backbones, from sets sTF to s20. This poses the question of how sensitive the search is to the space discretization imposed by the rotamers.

Looking at the effects per residue type, for some residues we observed a significant improvement in the quality of the prediction depending on the rotamer library, notably for Trp, Tyr and Gln (Figure 2A). However, the gains remains small (0.5 Å for Trp). For Arg, we noted that, despite a much larger number of rotamers, the prediction in RC3 is somewhat worse than within RC2. However, the number of buried Arg residues is low (20).

To check whether increasing the number of rotamers can lead to non-used states, we considered the frequency of selection of each rotamer of each residue type. In all cases, the number of equivalent conformations was close to the real rotamer number, indicating that all the rotamers are evenly selected. Moreover, the influence of the backbone perturbation on this number is weak. In fact, the whole conformations of the rotamer libraries are used at least once in the prediction. Thus, the poor scores obtained for some residues cannot be attributed to a systematic non-selection of some rotamers.

In fact, Figure 1B suggests that the main effect of increasing the number of rotamers appears to be a better prediction of χ_1, while the discrimination is weakened for χ_2. If RC1 and

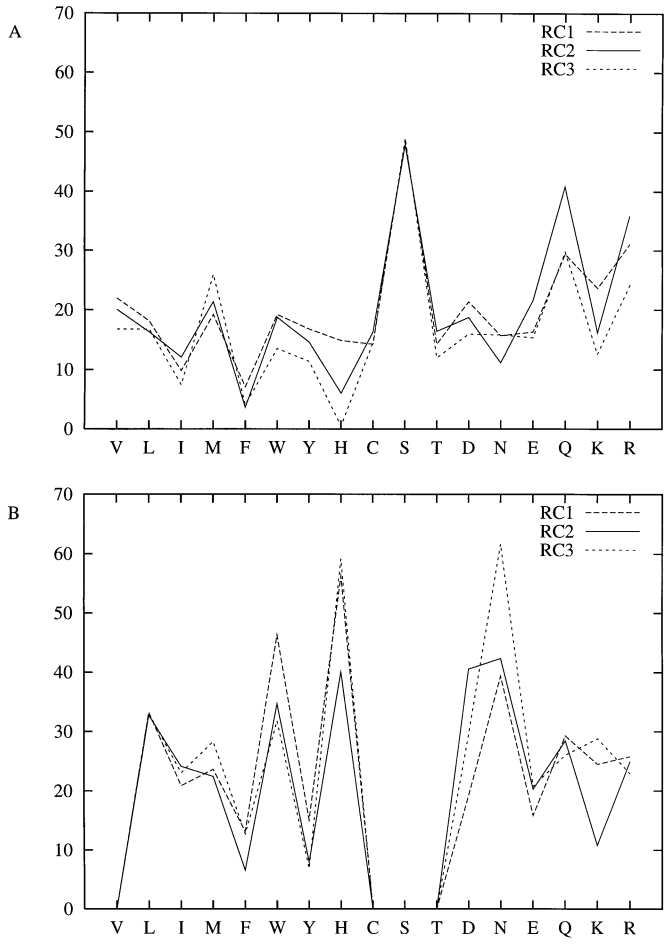

Fig. 5. Fraction (in percent) of (**A**) χ_1 and (**B**) χ_2 not predicted closer than $\pm 40°$ to the crystallographic value (backbone set sTF).

RC2 show similar results for χ_2, those obtained with RC3 appear significantly lower. We have investigated whether this tendency could result from specific residue types. Figure 5 represents the fraction of χ not predicted closer than $\pm 40°$ to the crystallographic conformations for each residue type and upon the sTF sets of the 12 proteins. Again, we observed a better prediction for χ_1 (range 20 $\pm$ 6%) than for χ_2 (range 30 $\pm$ 8%). For χ_1, we noted a poor score for Ser, Glu, Gln and Arg for all the rotamer libraries. Despite the presence of three well-defined rotamers for Ser, the prediction is not efficient. In addition, we noted that for Thr no such problem appears. In fact, an analysis of the results shows that 21 of the 52 buried Ser residues are systematically well predicted, while the others exhibit varying conformations. Few of the incorrectly positioned Ser residues are involved in a hydrogen bond (for example, in the 12 proteins, only four buried Ser residues implicated in a hydrogen bond within the PDB conformations are incorrectly positioned within the RC2 consensus for set sTF). Instead, if we impose the conformations of the mispositioned Ser residues to their best fit rotamer (i.e. closest to the PDB conformation), we observe only slight energy differences between the combinations, in favour of the mispositioned combination. Thus, it appears that the small size of this side chain allows several quasi-equivalent conformations when it is not implicated in a hydrogen bond. Concerning χ_2, poor scores are obtained for Trp, His, Asp and Agn. We have analysed these large differences between Phe and Trp and Tyr and His. In fact, the rings of Trp and His exhibit a large tendency to adopt the conformation obtained by rotating χ_2 by 180°. Indeed, the fraction of χ_2 misalignment decreases to 15% for Trp and 20% for His if one accepts rotating the rings by 180°. Thus, it seems that for these residues the search is mainly able to discriminate the plane of the ring well.

We now examine the variations observed for the residue types, dependent on the rotamer library. Again, it is surprising that phenylalanine χ_2 is predicted worse with RC3 and RC1 than RC2 (Figures 1 and 5) since the three rotamer libraries exhibit the same number of rotamers. In total, the largest angular deviation between the rotamers describing Phe in RC1, RC2 or RC3 is only 16°. The same situation is found for His where the largest difference between the rotamers of RC2 and RC3 is 14°, while the corresponding value for RC1 and RC3 is 30°. However, only four residues (of 14 buried) appear to be mainly responsible for this poor score. As estimating the protonation state of one His is somewhat difficult, we have considered both states for these His residues, without any modification of the results. For Asn, as previously mentioned, χ_2 can in fact adopt any value. Again, RC3 contains in each case one rotamer close to those of RC1 and RC2. The maximal deviations for χ_2 are 28° and 7° for RC1 and RC2 respectively. The analysis of the differences between the rotamer libraries suggests that, for Phe, Trp, Tyr and His, duplicating the rotamer conformations to introduce a χ_2 variation close to 20° could improve the quality of the prediction. For the charged residues, this situation does not seem realistic, since an enhanced number of rotamers cannot be correlated to a systematic improvement. Furthermore, the large increase in the number of residues describing Lys and Arg in RC3, combined with the observation that all of them have been used within the simulations, suggest that, at this level of conformational discretization, the agreement between the force field and the conformations might be the limiting factor.

To test explicitly whether it is worth multiplying the number of conformations, we have built a rotamer library that includes new conformations obtained by moving the χ_2 values by $\pm 20°$. For His, Phe, Trp, Tyr and His the conformations where built starting from RC3. For Arg, Asn, Lys and Met only the rotamers of RC2 were considered to avoid a huge library. For Asp, the rotamers of RC1, RC2 and RC3 that are different were all included. Series of 10 conformations of set s05 for each protein were considered. Using this library, the mean r.m.s.d. of the consensus decreased to 1.28 Å. Compared to what was obtained with the other libraries, there is a gain in Phe, Trp, Tyr and His and also in Met, Cys and Ser (the values obtained for Glu, Lys and Arg cannot be simply analysed since their rotamers were those of RC2). Finally, looking at the χ agreement shows a gain for the χ_2 of Phe and His, no change for Trp and a loss for Tyr. For χ_1, the results show a deterioration tendency. These observations suggest that, even if the results can be improved by increasing the number of rotamers, we appear to reach a limit due to the balance between the complex energy hypersurface associated with the search and the conformational goal. Due to the small χ value differences considered ($<20°$), this would also suggest that algorithms that work in a non-discrete space could face the same barrier. Finally, it must be considered that the gain introduced by increasing the size of the libraries is obtained at a higher computational cost.

Conclusions

In this study, we have examined the sensitivity of side chain conformations predicted by a method using rotamers to the accuracy of the backbone conformation. We have also studied the opportunity of increasing the size of the rotamer libraries used.

Concerning the size of the rotamer library, our results show that an increase leads to an improvement in the prediction. However, the search algorithms appear confronted by the limit of the adequacy between the conformational sampling and the energetic search. Indeed, a difference of only 20° for the χ_2 of one residue can trigger the selection of wrong side chain conformations. This suggests that, for a thin space discretization, the differences between the energies associated with each state are not discriminating enough to distinguish between correct conformations rotated by a few degrees and another pseudo-optimal rotamer combination. In summary, the main improvement obtained by increasing the size of the rotamer library remains for χ_1, but at a higher computational cost.

Concerning the quality of the prediction for inaccurate backbones, our results indicate that the sensitivity is large in terms of the number of side chains that change their conformations when the backbones differ weakly. Indeed, even when the backbone coordinates differ by r.m.s.d.s of <0.2 Å, as many as 10% of the buried side chains exhibit varying conformations. It also appears that for deviations of the backbone >0.5 Å, the quality of the prediction of the conformations of the buried residues is affected. However, it is to be remarked that the loss of accuracy within the prediction of the side chains remains weak compared to the backbone deviations: a 2 Å r.m.s.d. for the backbones only leads to a difference of 0.5 Å r.m.s.d. for the predicted side chains.

Moreover, a means of diminishing the perturbation introduced within the positioning of the side chains by the backbone inaccuracy is to consider consensus conformations obtained through series of predictions performed for different backbones. Facing a concrete model construction, a plausible strategy consists of extracting the consensus conformation of the side chains from 10 backbones generated within 0.5 Å C_α r.m.s.d. Larger backbone sets can lead to better prediction but at a much higher computational cost. Using such an approach, one can expect to obtain results close to that obtained with the crystallographic backbone when starting from backbones having C_α r.m.s.d.s close to 1 Å. Overall, our best results show that the mean buried side chain r.m.s.d. increases from 1.33 Å for series of backbones compatible with the temperature factors of the crystallographic structures up to 1.6 Å (only a 20% loss) for backbone deviations up to 2 Å.

Finally, our results also show that when the backbone deviation increases, the number of side chains that exhibit varying conformations appears to increase by diffusion from seeds. One can wonder if the fluctuations observed correspond to adjustments due to the poor adequacy of the rotameric states that create local steric hindrances or whether the fluctuations can be interpreted as a measure of the cohesion of the structures.

References

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,H.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Chung,S.Y. and Subbiah,S. (1996) In Hunter,L. and Klein,T.E. (eds), *Pacific Symposium Biocomputing*, pp. 126–141. World Scientific, Hawaii, USA.

Desmet,J., Maeyer,B.M.D., Hazes,B. and Lasters,I. (1992) *Nature*, **356**, 539–542.

Eisenmenger,F., Argos,P. and Abagyan,R. (1993) *J. Mol. Biol.*, **231**, 849–860.

Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.

Holm,L. and Sander,C. (1991) *J. Mol. Biol.*, **218**, 183–194.

Holm,L. and Sander,C. (1992) *Proteins*, **14**, 213–223.

Hwang,J.-K. and Liao,W.-F. (1995) *Protein Engng*, **8**, 363–370.

Laughton,C.A. (1994) *J. Mol. Biol.*, **235**, 1088–1097.

Lavery,R., Sklenar,H., Zakrzewska,K. and Pullman,B. (1986a) *J. Biomol. Struct. Dynam.*, **3**, 989–1014.

Lavery,R., Parker,I. and Kendrickn,J. (1986b) *J. Biomol. Struct. Dyn.*, **4**, 443–462.

Lee,C. and Subbiah,S. (1991) *J. Mol. Biol.*, **217**, 373–388.

Niugles,M. and Brunner,A.T. (1991) *Protein Engng*, **4**, 649–650.

Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.

Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) *Nature*, **372**, 631–634.

Ponder,J.W. and Richards,F.M. (1987) *J. Mol. Biol.*, **193**, 775–791.

Reid,L.S. and Thornton,J.M. (1989) *Proteins*, **5**, 170–182.

Richmond,T. (1984) *J. Mol. Biol.*, **178**, 63–69.

Schrauber,H., Eisenhaber,F. and Argos,P. (1993) *J. Mol. Biol.*, **230**, 592–612.

Sippl,M.J. and Stegbuchner,H. (1991) *Comput. Chem.*, **15**, 73–78.

Summers,N.L. and Karplus,M. (1989) *J. Mol. Biol.*, **210**, 785–811.

Summers,N.L., Carlson,W.D. and Karplus,M. (1987) *J. Mol. Biol.*, **196**, 175–198.

Tuffery,P. (1995) *J. Mol. Graphics*, **13**, 67–72.

Tuffery,P., Etchebest,C., Hazout,S. and Lavery,R. (1991) *J. Biomol. Struct. Dyn.*, **8**, 1267–1289.

Tuffery,P., Etchebest,C., Hazout,S. and Lavery,R. (1993) *J. Comput. Chem.*, **14**, 790–798.

Wilson,C., Gregoy,L.M. and Agard,D.A. (1993) *J. Mol. Biol.*, **229**, 996–1006.

Received June 4, 1996; revised November 12, 1996; accepted December 9, 1996