Protein Engineering vol. 16 no. 12 pp. 963–969, 2003
DOI: 10.1093/protein/gzg143

Advantages of fine-grained side chain conformer libraries

Reshma P.Shetty[1], Paul I.W.de Bakker[2,3], Mark A.DePristo and Tom L.Blundell

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

[1]Present address: Biological Engineering Division, Massachusetts Institute of Technology, Boston, MA 02139, USA

[2]Present address: Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA

[3]To whom correspondence should be addressed.
E-mail: debakker@molbio.mgh.harvard.edu

We compare the modelling accuracy of two common rotamer libraries, the Dunbrack–Cohen and the 'Penultimate' rotamer libraries, with that of a novel library of discrete side chain conformations extracted from the Protein Data Bank. These side chain conformer libraries are extracted automatically from high-quality protein structures using stringent filters and maintain crystallographic bond lengths and angles. This contrasts with traditional rotamer libraries defined in terms of χ angles under the assumption of idealized covalent geometry. We demonstrate that side chain modelling onto native and near-native main chain conformations is significantly more successful with the conformer libraries than with the rotamer libraries when solely considering excluded-volume interactions. The rotamer libraries are inadequate to model side chains without atomic clashes on over 20% of targets if the backbone is held fixed in the native conformation. An algorithm is described for simultaneously modelling both main chain and side chain atoms during discrete *ab initio* sampling. The resulting models have equivalent root mean square deviations from the experimentally determined protein loops as models from backbone-only ensembles, indicating that all-atom modelling does not detract from the accuracy of conformational sampling.

Keywords: all-atom modelling/conformational sampling/side chain conformations/RAPPER/rotamers

Introduction

Ab initio methods for protein structure modelling have been an active research area for many years. Most of these methods focus on the construction of the protein backbone, the N, Cα, C and O atoms, and rely on independent side chain modelling programs to model side chains at a later stage. This approach to all-atom modelling has the advantage of dividing the construction of a complete structure for a given amino acid sequence into two separate challenges of approximately equivalent complexity: backbone construction and side chain assignment (Levitt *et al*., 1997). By drawing a clear demarcation between the two steps, they are kept conceptually distinct and each can be addressed with appropriate tools. Methods that explicitly account for side chains during the modelling of the main chain generally include either the Cβ atom alone (the position of which is determined by the main chain coordinates) or use a low-resolution virtual-atom (centroid) representation of the side chain (Levitt, 1976; Kang *et al*., 1993; Keskin and Bahar, 1998; Gibbs *et al*., 2001). The common goal is to exploit structural information available from the side chains without incurring the extra computational cost of modelling more atoms.

Regardless of whether side chains are ignored or simplified in some way during backbone construction, a side chain assignment procedure must be used to obtain a complete model of a protein or section of a protein. This is generally considered to be a difficult problem because at each position in the protein, multiple conformations are possible, leading to a combinatorial explosion of side chain assignments that need to be explored. Ponder and Richards observed that side chains in native structures tend to cluster around certain χ torsional angles (a phenomenon referred to as rotamericity), thus enabling a discrete but fairly comprehensive representation of allowed side chain conformations (Ponder and Richards, 1987). Thus, the set of natively occurring side chains can be described by relatively few conformations and the number of possible side chain assignments to a protein backbone is dramatically reduced. Various algorithms and energy functions can be applied to find the set of side chain conformations that eliminate clashes and optimize electrostatic interactions (Jacobson *et al*., 2002).

Recent work has called the use of rotamer libraries in side chain conformational assignment into question in a logical extension of the argument originally posed by Schrauber *et al*. that there are systematic outliers in rotamericity (Schrauber *et al*., 1993). Xiang and Honig (2001) noted that the use of an extensive coordinate rotamer library (containing up to thousands of rotamers), generated by taking the Cartesian coordinates of side chain atoms in a database of protein structures, results in higher accuracy for side chain prediction onto native backbones than the Dunbrack and Cohen backbone-dependent rotamer library and other rotamer libraries (some at lower resolutions) derived from the same set of structures and defined in terms of dihedral angles. This is in agreement with the analysis by Tuffery *et al*. (1991), who observed that better accuracy can be achieved with their updated rotamer library (called RC3 with 214 rotamers) than the Ponder and Richards library (84 rotamers) (Ponder and Richards, 1987) or the rotamer library by Tuffery *et al.* (110 rotamers) (Tuffery *et al*., 1991). Similarly, the flexible rotamer model of Mendes *et al*. (1999), in which 'subrotamers' are generated by varying torsion and bond angles in their initial rotamer library, improves predictions with respect to a rigid rotamer model. These studies all suggest that traditional rotamer libraries are often incomplete and lack the necessary resolution for accurate prediction of experimental side chain conformations.

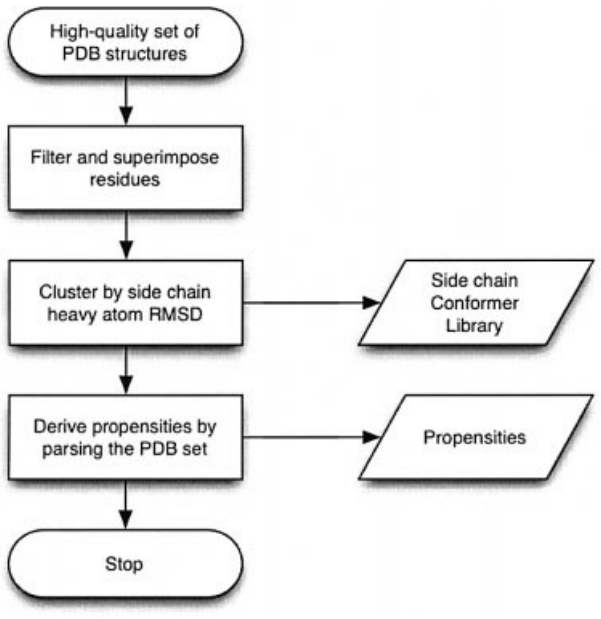

Fig. 1. Flow chart of the derivation of the side chain conformer libraries.

Table I. Statistics of the side chain conformer libraries

	SCL0.2		SCL0.5		SCL1.0		PRL	
	No.	%	No.	%	No.	%	No.	%
Arg	1240	20.7	415	29.1	135	30.3	34	22.4
Asn	248	4.1	48	3.4	15	3.4	7	4.6
Asp	204	3.4	35	2.5	11	2.5	5	3.3
Cys	14	0.2	4	0.3	3	0.7	3	2.0
Gln	667	11.1	148	10.4	43	9.6	9	5.9
Glu	582	9.7	108	7.6	30	6.7	8	5.3
His	344	5.7	62	4.3	21	4.7	8	5.3
Ile	79	1.3	18	1.3	6	1.3	7	4.6
Leu	207	3.5	36	2.5	16	3.6	5	3.3
Lys	670	11.2	195	13.7	46	10.3	27	17.8
Met	393	6.6	85	6.0	33	7.4	13	8.6
Phe	354	5.9	55	3.9	20	4.5	4	2.6
Pro	14	0.2	3	0.2	1	0.2	2	1.3
Ser	26	0.4	8	0.6	4	0.9	3	2.0
Thr	22	0.4	5	0.4	3	0.7	3	2.0
Trp	459	7.7	105	7.4	32	7.2	7	4.6
Tyr	435	7.3	88	6.2	23	5.2	4	2.6
Val	30	0.5	8	0.6	4	0.9	3	2.0
Total	5988	100.0	1426	100.0	446	100.0	152	100.0

Listed are the number of side chain conformers per amino acid type for a given side chain heavy atom r.m.s.d. cutoff (0.2, 0.5 and 1.0 Å). Also shown are details of the 'Penultimate' rotamer library (PRL) (Lovell *et al*., 2000).

Thus far, much of the work on explicit all-atom modelling has focused on the utility of side chain interactions in distinguishing native-like conformations from a decoy ensemble. Samudrala and Moult showed that by accounting for side chain interactions with surrounding atoms, their residue-specific all-atom conditional probability discriminatory functions (RAPDF) can select native-like conformations more accurately (Samudrala and Moult, 1998). They also show that a simplified (virtual atom) representation of their energy function (RVPDF) has worse discriminatory power than the all-atom RAPDF function. In fact, even a naive approach to side chain assignment (by taking the most frequently observed rotamer) can substantially improve discrimination by their all-atom function despite the likely presence of atomic clashes (Samudrala *et al*., 2000). In recent work, we demonstrated that if the side chain assignment program SCWRL (Bower *et al*., 1997) is used to assign side chains, the calculated residual clash energy, which indicates the severity of clashes present in the model, has no discriminatory power for selecting native-like conformations (de Bakker *et al*., 2003). Hence it appears that when side chains are ignored during backbone construction, it is possible to generate models that have low main chain root mean square deviations (r.m.s.d.s) from the native structure but that have clashing side chain conformations.

Our previous work demonstrates that *ab initio* sampling under simple constraints such as idealized geometry (Engh and Huber, 1991), residue-specific propensity-weighted φ/ψ state sets and hard-sphere repulsion can yield accurate backbone conformational ensembles for loops (de Bakker *et al*., 2003; DePristo *et al*., 2003a). There are two possible scenarios for the impact of explicit all-atom modelling on the conformational ensemble. First, the ensemble may on average be driven towards the experimentally determined structure, since the inclusion of more atoms may force the main chain to adopt a more native-like conformation. Secondly, the conformational ensemble may move away from well-packed conformations since atomic clashes can be avoided more readily by, for instance, projecting out into solvent. To address these issues, we describe here an extension to the conformational sampling program RAPPER (de Bakker *et al*., 2003; DePristo *et al*., 2003a,b) that permits all-atom modelling of protein fragments (the term fragment refers to any eight-amino acid sequence in a protein structure). We also derive an extensive side chain conformer library that consists of side chain atomic coordinates taken from a database of high-quality protein structures and filtered according to a set of quality criteria. We show that for side chain modelling onto a native or near-native backbone within a fixed environment, the Dunbrack and Cohen backbone-dependent rotamer library (DCRL) (Dunbrack and Karplus, 1993; Dunbrack and Cohen, 1997) and the Lovell *et al*. 'Penultimate' rotamer library (PRL) (Lovell *et al*., 2000) are insufficient and a fine-grained side chain library is called for.

Materials and methods

Side chain libraries

The side chain conformer library was built from the Top-500 database of high-resolution protein structures (download from http://kinemage.biochem.duke.edu/). Figure 1 contains a schematic flow chart of our derivation protocol. A test set of 100 structures was compiled from the Top-500 by including proteins from the Fiser *et al*. (2000) loop benchmark set and additional randomly selected proteins. All residues in the 400 remaining protein structures were collected that satisfy the criteria proposed by Lovell *et al*. (2000): *B*-factor <30.0; no alternate side chain conformations; no missing atoms; no van der Waals overlaps of >0.4 Å for any atom in a residue, as calculated by the contact analysis program PROBE (Word *et al*., 1999). These filtered residues are superimposed on their N, C and Cβ atoms with their Cα atoms at the origin and clustered according to the global r.m.s.d. computed over all side chain heavy atoms. Three side chain libraries were derived, referred to as the SCL0.2, SCL0.5 and SCL1.0, differing only by the r.m.s.d. cutoff used to generate the library.

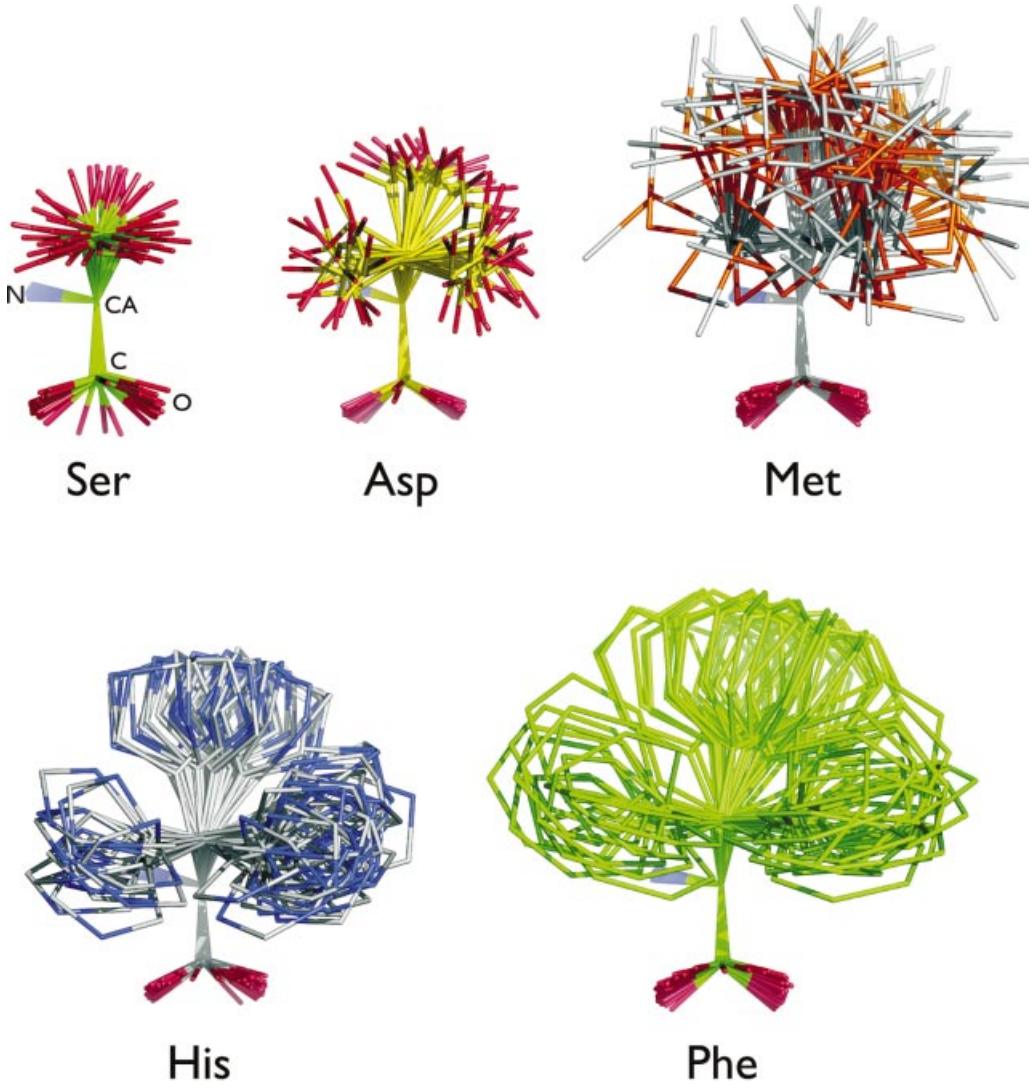

Fig. 2. Illustration of the conformations for amino acids Ser, Asp, Met, His and Phe contained in the fine-grained SCL0.5 library. This figure was made using PyMOL.

Backbone-dependent propensity values for collected side chain conformers are compiled from the same 400 protein structures as a function of φ and ψ in which each dihedral angle is divided into 40° bins. Side chain conformations with a χ_1 dihedral angle difference within 20° are treated as identical for the purpose of computing propensities. Thus, the SCL0.2, SCL0.5 and SCL1.0 represent side chain libraries of decreasing granularity. Table I contains the statistics for each of the SCL libraries and the rotamer library by Lovell *et al.* (see below). The SCL derived at 0.2 Å contains four times as many conformers than the SCL0.5, which in turn has almost three times as many conformers as the SCL1.0. Figure 2 shows the superimposed SCL0.5 conformers for Ser, Asp, Met, His and Phe amino acids.

We compare the SCLs to two side chain rotamer libraries, the Lovell *et al*. 'Penultimate' rotamer library (http://kinemage.biochem.duke.edu/) (PRL) (Lovell *et al*., 2000) and the July 2001 version of the Dunbrack and Cohen backbone-dependent rotamer library (http://dunbrack.fccc.edu/bbdep/) (DCRL) (Dunbrack and Karplus, 1993; Dunbrack and Cohen, 1997).

Excluded volume constraints

All atoms are represented in RAPPER as hard spheres with atomic van der Waals radii taken from PROBE (Word *et al*., 1999) and reduced by 20%. An atomic clash results if the distance between the centres of any two atoms is less than the sum of their van der Waals radii.

Side chain modelling onto a fixed backbone

Two strategies are used to model side chains onto a fixed main chain conformation. First, a pool of viable side chain conformations is computed for each position in the protein by eliminating all side chain conformations that clash with the main chain or surrounding protein structure. In the first, exhaustive method, every combination of side chain conformations is evaluated to identify all clash-free sets of side chains. This method is computationally tractable only for the two rotamer libraries (DCRL, PRL) over short stretches of amino acids of up to eight residues.

The second, stochastic method involves randomly assigning side chain conformers from the pool of possible conformers at each position in a sequential manner. A side chain conformer is

successfully fixed to a position only if its atoms do not clash with previously assigned side chains. When all side chain conformations from the pool result in a clash, an earlier side chain is reassigned in order to permit successful side chain modelling at the current position. If the side chain–side chain clash can be resolved, then assignment will continue, otherwise the process is restarted, but now moving along the sequence in the opposite direction. At most 1 000 000 passes of this algorithm are made to collect 100 unique models. If fewer than 100 models are obtained (as is sometimes the case with the rotamer libraries), the exhaustive algorithm is applied to collect additional models.

Ab initio polypeptide construction

The program RAPPER (de Bakker *et al*., 2003; DePristo *et al*., 2003a,b) generates an ensemble of self-consistent conformations of a specified segment of a protein structure within the context of a fixed (native) protein environment. A single conformation is generated by iteratively growing the polypeptide chain from the N to the C terminus. All heavy main chain atoms (N, Cα, C, O) of the backbone are modelled with idealized stereochemistry (Engh and Huber, 1991). The backbone φ/ψ and ω dihedral angles are sampled according to residue-specific propensities (Lovell *et al*., 2003).

A side chain group is assigned immediately following each extension of the polypeptide chain by a single amino acid. Given a main chain, a side chain is randomly selected from the pool of conformations, according to their propensities. If this side chain conformation does not clash with an atom in the surrounding protein or the partially constructed chain, the entire amino acid conformation is kept. Otherwise a new conformation is tried until all conformations in the library have been eliminated. When the library is exhausted, a side chain in an amino acid early in the constructed chain is remodelled, in an attempt to permit side chain placement at the current position. If no pair of side chains can be resolved without clashes, the extended amino acid is rejected. Note that this assignment method is exhaustive, as all side chain conformations are examined before a main chain conformation is rejected as invalid.

A gap closure restraint is applied to ensure that the generated chain re-attaches to the C-terminal anchor and is further subjected to a simplex minimization to improve the joint covalent geometry (DePristo *et al*., 2003a). No fragments are permitted with global main chain r.m.s.d.s of <0.2 Å.

Side chain modelling onto a flexible backbone

By permitting some backbone variation from the native conformation, the backbone may be able to adjust to facilitate side chain conformation assignment. When modelling with backbone flexibility, fragments are remodelled *ab initio* as described above with the additional restraint that the model Cα atoms must be placed within 1 Å from the corresponding Cα position of the crystal structure (DePristo *et al*., 2003b).

Benchmark sets

Three sets of eight-residue targets are used to examine the performance of the side chain libraries. The first and second sets are composed of α-helical and β-strand fragments, respectively, derived from the subset of 100 structures. The third set comprises eight-residue loops from the loop benchmark set developed by Fiser *et al*. (2000).

Table II. The number and percentage of eight-residue α-helical, β-strand and loop targets that each of the five side chain libraries fail to assign when the fragment backbone is either held fixed in the native conformation ('Fixed backbone') or permitted flexibility ('Flexible backbone')

	α-Helices		β-Strands		Loops	
	No.	%	No.	%	No.	%
No. of targets	24		93		33	
Fixed backbone						
SCL0.2	0	0	0	0	0	0
SCL0.5	0	0	8	9	1	3
SCL1.0	1	4	14	15	4	12
DCRL	5	21	33	35	9	27
PRL	5	21	40	43	9	27
Flexible backbone						
SCL0.2	0	0	0	0	0	0
SCL0.5	0	0	1	1	0	0
SCL1.0	1	4	1	1	0	0
DCRL	1	4	2	2	0	0
PRL	1	4	3	3	0	0

'No. of targets' refers to the total number of targets in the benchmark set for each type of fragment. During side chain modelling onto native backbones, the stochastic method was used for the SCLs and the exhaustive method was used for the DCRL and PRL.

Assessment of the conformational ensemble

The accuracy of a model is evaluated by its r.m.s.d., without superposition, to the experimentally determined structure. In an ensemble of models, the model with the lowest r.m.s.d. is the upper bound on accuracy, and the mean or average r.m.s.d. of the ensemble reflects the expected accuracy of a model selected at random.

The r.m.s.d. gives more weight to residues with larger side chains since they contribute more atoms to the calculation. Side chain modelling accuracy is also assessed by the percentage of residues whose side chain dihedral angles (χ_1, χ_2, χ_3 and χ_4) are within 40° of the corresponding angle in the experimentally determined structure. The χ_{1+2} accuracy is the percentage of residues with both χ_1 and χ_2 angles within 40° of the crystal structure.

Results

Side chain modelling onto a fixed, native backbone

Atomic clashes with the native backbone, the rest of the protein structure and other side chain atoms limit the number of conformations that a side chain can assume at each position. Since all side chain conformations are enumerated for the DCRL and the PRL, we can conclusively determine whether the library can assign side chains to each target without any atomic clashes. For α-helices, β-strands and loops, the DCRL cannot assign 21, 35 and 27% and the PRL fails on 21, 43 and 27% of the targets, respectively (see Table II). In contrast, the SCL0.2 succeeds in modelling side chains on each of the targets in the benchmark set via the stochastic method. The number of target failures increases with decreasing granularity of the side chain conformer library. The SCL0.5 fails on 0% of the α-helices, 9% of the β-strands and 3% of the loops whereas the SCL1.0 fails on 4, 15 and 12% of the same target sets. This trend reflects the connection between how detailed the side chain library is and the percentage of failed targets. Although the DCRL and PRL are both backbone-dependent libraries and arguably optimized for assignment of secondary structure, they

Table III. Accuracy of conformational ensembles generated by RAPPER for all eight-residue loop targets using *ab initio* loop modelling

	R.m.s.d. (Å)			
	Main chain		All-atom	
	Lowest	Average	Lowest	Average
None	1.58	4.72	–	–
SCL0.2	1.58	4.81	2.27	6.00
SCL0.5	1.65	4.75	2.38	5.97
SCL1.0	1.61	4.82	2.35	6.02
DCRL	1.63	4.88	2.35	6.06
PRL	1.61	4.95	2.45	6.13

'Lowest' refers to the average of the lowest r.m.s.d. found in each ensemble for a given loop length. 'Average' refers to the average of the ensemble-averaged r.m.s.d. of all models over all eight-residue loop targets. The r.m.s.d. values are relative to the native structure and include either the N, Cα, C and O atoms ('Main chain') or all heavy atoms ('All-atom'). 'None' refers to *ab initio* modelling using main chain atoms only.

fail on a surprisingly large number of protein fragments with regular secondary structure.

Given the high level of detail of the SCL0.2, it is interesting that the cumulative execution time (over all targets) for the SCL is almost 150 times faster than the DCRL and 15 times faster than the PRL. This perhaps unexpected speed-up can be attributed to the greater likelihood of successful assignment in fewer passes (using the stochastic method) using our fine-grained conformer library, whereas the coarse-grained rotamer libraries often result in an exhaustive search (see Materials and methods).

Side chain modelling onto a flexible backbone

The percentage of failed targets drops dramatically to 4, 2 and 0% with the DCRL and 4, 3 and 0% for the PRL for α-helices, β-strands and loops, respectively (see Table II). Self-consistent models can be found for all targets using the SCL0.2. The SCL0.5 fails on only 1% of β-strands and the SCL1.0 fails on 4% of α-helices and 1% β-strands. As loops are often found on the surface of the protein and are therefore less constrained by excluded-volume interactions with the rest of the protein, it is understandable that the side chain libraries can assign all loop targets by introducing backbone flexibility. α-Helices and β-strands, however, are more likely to be in the highly constrained core of the protein, where even backbone flexibility may be unable to compensate for the coarse nature of the side chain library.

Ab initio loop modelling

As shown in Table III, the explicit modelling of side chain atoms during conformational sampling of polypeptides results in ensembles of comparable similarity to native as those generated when only the main chain atoms are modelled. During all-atom modelling, the lowest and average main chain r.m.s.d. value of the generated ensemble from the native structure are 1.58 and 4.81 Å, respectively, as averaged over all loop targets using SCL0.2. This compares with 1.58 and 4.72 Å, respectively, for backbone-only models. Despite forcing the backbone conformation to accommodate side chain conformations assigned solely on the basis of backbone-dependent propensities and excluded-volume constraints, it is still possible to obtain conformational ensembles similar to the experimentally determined structure (as measured by main chain r.m.s.d.).

Table IV. Percentage of χ_1 and χ_{1+2} angles within 40° of the experimentally determined structure for all eight-residue loops using *ab initio* simultaneous all-atom loop modelling

	χ_1 (±40°)		χ_{1+2} (±40°)	
	Highest	Average	Highest	Average
SCL0.2	93	49	69	18
SCL0.5	91	48	67	18
SCL1.0	87	46	63	19
DCRL	93	44	64	19
PRL	91	49	67	24

The overall performances of the DCRL and PRL are slightly worse than the SCL0.2 (see Table III). These results are expected on the basis of all-atom modelling with backbone flexibility: small deviations in backbone conformation are sufficient to permit side chain placement without atomic clashes using coarse-grained side chain libraries. Similarly, when comparing the best and average model all-atom r.m.s.d. to the native fragment, we find that the three SCLs are similar in accuracy to both rotamer DCRL and PRL libraries.

If side chain placement is assessed in terms of dihedral angle deviation from native, the χ_1 angle is predicted correctly in 93% of residues for the best model and achieves a mean accuracy of 49% for all models as averaged over all loop targets when using the SCL0.2 (see Table IV) with only slightly worse accuracies for the SCL0.5 and SCL1.0. The corresponding values for the DCRL are 93% (highest) and 44% (mean) and for the PRL 91 and 49%, respectively. If both the χ_1 and the χ_2 dihedral angles are considered, the best model has 69% of residues predicted correctly and the ensemble has on average 18% of residues correct when using the SCL0.2. The most detailed conformer library is equivalent to or only slightly better than the rotamer libraries in predicting the χ_1 dihedral angle. Although the SCL0.2 is slightly preferable to the other libraries in terms of the best model generated on average if both the χ_1 and χ_2 angles are taken into account, the PRL leads to the most accurate ensemble average.

Discussion

Side chain modelling onto native and flexible backbones

A detailed side chain library is necessary to model side chains over a diverse set of fragments in which the side chain conformations are constrained by van der Waals interactions with nearby atoms and the backbone is in a native or near-native conformation. Conventional rotamer libraries tend to seek optimal coverage of the side chain conformational space with a minimal number of side chains (usually one per rotamer state). This objective has previously been justified by the reduction in the number of combinations of side chain conformations that need to be examined. However, in the restraint-based modelling situation considered here, we have shown that the coarse-grained representation of side chain conformational space can be a serious obstacle to successful all-atom modelling. When modelling side chains onto fragments within a fixed structural environment, a coarse-grained library does not provide sufficient flexibility to avoid severe atomic clashes with other atoms in >20% of targets, regardless of secondary structure type. As progressively more detailed

conformer libraries are used, the number of target failures drops to zero, reiterating the notion that there is a direct correspondence between the granularity of the side chain library and ability to model side chains onto native backbones in a self-consistent manner.

It is stressed that the SCLs are non-redundant libraries owing to clustering, effectively making them much smaller, and more manageable, than the libraries from Xiang and Honig (2001). Following the work of the 'penultimate' rotamer library (Lovell *et al*., 2000), another advantage of the SCLs is the use of strict criteria for data inclusion, the importance of which was highlighted by Dunbrack in a recent review (Dunbrack, 2002).

The inability of the rotamer libraries to model side chains without causing severe van der Waals overlaps can be ameliorated by permitting backbone flexibility. A significant observation, however, is that there remain some α-helix and β-strand targets that still cannot be assigned with the rotameric libraries.

There are some targets for which only few valid models can be generated using the rotamer libraries. This implies that, in all-atom modelling, the ensemble of conformations will be forced away from the native structure to accommodate non-clashing side chains. Our results demonstrate that a more extensive side chain library is crucial when modelling side chains on regular secondary structures in a protein core, as backbone flexibility cannot always mitigate the coarseness of the rotamer libraries.

Ab initio loop modelling

We show here that it is possible to model simultaneously all atoms for the construction of self-consistent, representative conformational ensembles of polypeptides within a fixed protein structural environment. When generating all-atom models of protein fragments, one common approach is initially to generate an ensemble of backbone conformations and then use a separate side chain assignment program to add side chain heavy atoms. Despite the use of various algorithms to explore efficiently side chain conformational space, separately modelling the main chain and side chain generally leads to models with severe van der Waals overlaps, as the side chains are forced onto pre-generated main chain conformations. We find here that clash-free all-atom models can be efficiently obtained by simultaneously modelling both main chain and side chain using a detailed side chain conformer library. Side chain placement consists of the selection of the highest propensity conformation that satisfies excluded volume constraints. The coupling of main chain and side chain construction simplifies the development of complete, internally consistent models. The need to explore exhaustively all combinations of side chain conformations to find a valid side chain assignment to a fixed backbone is eliminated. Interestingly, we find that use of the same simple approach to side chain modelling on backbones that are constructed with only main chain atoms requires an increased number of samples to collect the same number of all-atom models satisfying all constraints (data not shown). Hence simultaneous all-atom modelling has an advantage over more traditional methods in that it permits a pairwise interaction between the main chain and side chain conformations, facilitating the generation of complete models free of atomic clashes.

An unexpected finding of this work is that inclusion of side chain atoms during *ab initio* fragment modelling does not significantly alter the main chain r.m.s.d. of either the model most similar to native or the ensemble average. The difference of 0.0 Å in the best model and 0.09 Å (see Table III) of the backbone-generated and the SCL0.2 ensemble translates to negligible differences in atom positions. As discussed in the Introduction, two extreme scenarios could result when modelling all atoms versus modelling only main chain atoms: the ensemble could be driven towards the native structure because of the added constraints imposed by the presence of explicit side chain atoms or the ensemble could be driven away from the well-packed native structure because it permits rapid side chain placement with fewer potential atomic clashes. If the effect of the presence of side chains on the main chain r.m.s.d. of the ensemble is examined on a per target basis (data not shown), we find that in a few cases the lowest and average main chain r.m.s.d. is improved significantly by the explicit inclusion of side chain atoms, whereas in others, the main chain r.m.s.d. is adversely affected by the presence of side chain atoms. For most targets, however, there is little change in the main chain r.m.s.d. of the ensemble. Hence it would appear that the data presented in Table III represent a balancing of the two situations such that the average main chain r.m.s.d. over all loop targets (whether considering the lowest or the mean r.m.s.d.) is not significantly affected by the presence of all atoms. In the light of the essentially equivalent average main chain r.m.s.d. of the ensemble developed using simultaneous all-atom modelling and the ensemble developed using main chain-only modelling, the former ensemble is clearly superior owing to the presence of all atoms and the absence of atomic clashes.

A couple of trends emerge in comparing the performances of the side chain libraries during *ab initio* all-atom modelling of loops. First, the most fine-grained conformer library (SCL0.2) does only slightly better than the two rotamer libraries in terms of all-atom r.m.s.d. (see Table III). The relatively small differences in accuracy between this conformer library and the rotamer libraries is due to the effect observed in near-native assignment, namely that small conformational changes in the backbone can mitigate the effects of coarse-grained representations of side chains. Furthermore, the three libraries (SCL0.2, DCRL and PRL) produce models with almost equivalent χ_1 accuracies; the only noteworthy difference is that the SCL0.2 generates models with an average χ_1 accuracy of 5% better than the DCRL (as averaged over all loop targets) (see Table IV). These results are consistent with the evaluation in terms of all-atom r.m.s.d. The χ_{1+2} accuracy is more unusual in that the SCL0.2 tends to generate the best model when considering the average χ_{1+2} accuracy of the best model, whereas the rotameric libraries are superior according to the ensemble average χ_{1+2} accuracy. This suggests that the conformer library contains the side chain conformation closest to native but the coarser rotameric libraries perform better on average. Nevertheless, as with the evaluations using all-atom r.m.s.d. and χ_1 accuracy, there does not appear to be a clearly superior library for use in *ab initio* all-atom loop modelling. At first glance, this result may appear contradictory to our earlier observation that the more fine-grained SCL is better able to model side chains onto native and near-native backbones than the DCRL and PRL. The rotamer libraries simply lack the resolution to accommodate the constrained (native or near-native) backbone without incurring atomic clashes. In *ab initio* loop modelling, however, the main chain has sufficient conformational freedom such that the backbone can accom-

modate the coarser DCRL and PRL without penalizing the all-atom r.m.s.d. relative to the SCLs. All of these observations lead us to conclude that the errors in main chain modelling far outweigh any differences in the side chain libraries. However, since main chain-only *ab initio* modelling leads to conformations incompatible with clash-free all-atom models, it is imperative to model both the main chain and side chain simultaneously.

Conclusion

We have derived fine-grained side chain conformer libraries using stringent quality criteria and compared these with traditional side chain rotamer libraries in terms of side chain modelling on eight-residue targets within a fixed native protein structure. Two common side chain rotamer libraries often do not produce a clash-free side chain assignment when the main chain is held fixed in its native conformation. These inadequacies are partially overcome by introducing backbone flexibility, as minor movements in the backbone can counterbalance the effect of a coarse-grained representation of side chain conformational space. Even so, a detailed side chain conformer library outperforms rotamer libraries when assigning side chains onto native and near-native backbones.

Simultaneous modelling of main chain and side chain atoms during *ab initio* conformational sampling permits the generation of conformational ensembles in which all atoms are explicitly represented and free of clashes. These ensembles are similar to backbone-only conformational ensembles in terms of main chain r.m.s.d. from the experimentally determined structure. Both the conformer and rotamer side chain libraries perform equivalently well in loop modelling, as errors in main chain placement are likely to dwarf side chain assignment errors and accommodate side chain placement irrespective of the coarseness of the library. Nonetheless, the primary advantage of an integrated approach for backbone generation and side chain assignment is that high-energy conformations can be directly detected and eliminated, making these all-atom ensembles superior for further analysis and selection.

The side chain conformer libraries (also those derived on the basis of the entire Top-500) are available for download at http://www-cryst.bioc.cam.ac.uk/rapper/.

Acknowledgements

The authors thank Simon Lovell for valuable discussions and for providing the 'Penultimate' rotamer library. The authors are grateful to the Barry M.Goldwater Foundation and Cambridge–MIT Institute (to R.P.S.), the Cambridge European Trust, Isaac Newton Trust and BBSRC (to P.I.W.D.B.) and the Cambridge Overseas Trust, the Marshall Aid Commemoration Commission and the US National Science Foundation (to M.A.D.P.) for financial support.

References

Bower,M.J., Cohen,F.E. and Dunbrack,R.L.,Jr (1997) *J. Mol. Biol.*, **267**, 1268–1282.

de Bakker,P.I.W., DePristo,M.A., Burke,D.F. and Blundell,T.L. (2003) *Proteins*, **51**, 21–40.

DePristo,M.A., de Bakker,P.I.W., Lovell,S.C. and Blundell,T.L. (2003a) *Proteins*, **51**, 41–55.

DePristo,M.A., de Bakker,P.I.W., Shetty,R.P. and Blundell,T.L. (2003b) *Protein Sci.*, **12**, 2032–2046.

Dunbrack,R.L.,Jr (2002) *Curr. Opin. Struct. Biol.*, **12**, 431–440.

Dunbrack,R.L.,Jr and Cohen,F.E. (1997) *Protein Sci.*, **6**, 1661–1681.

Dunbrack,R.L.,Jr and Karplus,M. (1993) *J. Mol. Biol.*, **230**, 543–574.

Engh,R.A. and Huber,R. (1991) *Acta Crystallogr. A*, **47**, 392–400.

Fiser,A., Do,R.K. and Sali,A. (2000) *Protein Sci.*, **9**, 1753–1773.

Gibbs,N., Clarke,A.R. and Sessions,R.B. (2001) *Proteins*, **43**, 186–202.

Jacobson,M.P., Friesner,R.A., Xiang,Z. and Honig,B. (2002) *J. Mol. Biol.*, **320**, 597–608.

Kang,H.S., Kurochkina,N.A. and Lee,B. (1993) *J. Mol. Biol.*, **229**, 448–460.

Keskin,O. and Bahar,I. (1998) *Fold. Des.*, **3**, 469–479.

Levitt,M. (1976) *J. Mol. Biol.*, **104**, 59–107.

Levitt,M., Gerstein,M., Huang,E., Subbiah,S. and Tsai,J. (1997) *Annu. Rev. Biochem.*, **66**, 549–579.

Lovell,S.C., Word,J.M., Richardson,J.S. and Richardson,D.C. (2000) *Proteins*, **40**, 389–408.

Lovell,S.C., Davis,I.W., Arendall,W.B.,III, de Bakker,P.I.W., Word,J.M., Prisant,M.G., Richardson,J.S. and Richardson,D.C. (2003) *Proteins*, **50**, 437–450.

Mendes,J., Baptista,A.M., Carrondo,M.A. and Soares,C.M. (1999) *Proteins*, **37**, 530–543.

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

Samudrala,R. and Moult,J. (1998) *J. Mol. Biol.*, **275**, 895–916.

Samudrala,R., Huang,E.S., Koehl,P. and Levitt,M. (2000) *Protein Eng.*, **13**, 453–457.

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

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

Word,J.M., Lovell,S.C., LaBean,T.H., Taylor,H.C., Zalis,M.E., Presley,B.K., Richardson,J.S. and Richardson,D.C. (1999) *J. Mol. Biol.*, **285**, 1711–1733.

Xiang,Z. and Honig,B. (2001) *J. Mol. Biol.*, **311**, 421–430.

Received April 30, 2003; revised October 24, 2003; accepted October 30, 2003