Downloaded via Jos&eacute; &Aacute;ngel Mart&iacute;nez Gonz&aacute;lez on July 12, 2022 at 11:06:34 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Journal of
CHEMICAL INFORMATION
AND MODELING

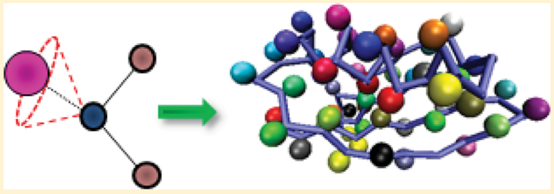
Article

pubs.acs.org/jcim

Design of a Rotamer Library for Coarse-Grained Models in Protein-Folding Simulations

María Larriva[†] and Antonio Rey*

Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, E-28040 Madrid, Spain



ABSTRACT: Rotamer libraries usually contain geometric information to trace an amino acid side chain, atom by atom, onto a protein backbone. These libraries have been widely used in protein design, structure refinement and prediction, homology modeling, and X-ray and NMR structure validation. However, they usually present too much information and are not always fully compatible with the coarse-grained models of the protein geometry that are frequently used to tackle the protein-folding problem through molecular simulation. In this work, we introduce a new backbone-dependent rotamer library for side chains compatible with low-resolution models in polypeptide chains. We have dispensed with an atomic description of proteins, representing each amino acid side chain by its geometric center (or centroid). The resulting rotamers have been estimated from a statistical analysis of a large structural database consisting of high-resolution X-ray protein structures. As additional information, each rotamer includes the frequency with which it has been found during the statistical analysis. More importantly, the library has been designed with a careful control to ensure that the vast majority of side chains in protein structures (at least 95% of residues) are properly represented. We have tested our library using an independent set of proteins, and our results support a good correlation between the reconstructed centroids from our rotamer library and those in the experimental structures. This new library can serve to improve the definition of side chain centroids in coarse-grained models, avoiding at the same time an excessive additional complexity in a geometric model for the polypeptide chain.

INTRODUCTION

One of the most important unresolved items in biophysical science is how a protein amino acid sequence determines its 3D native structure in solution. Many natural proteins fold into unique compact structures, in spite of the huge number of possible conformations a flexible chain molecule may adopt.[1]

Folding is not a chemical reaction in which a polypeptide chain reaches the native state, mainly because the process involves neither the breakage nor the formation of chemical bonds (apart from possible sulfur bridges). When a polypeptide chain folds, it suffers a conformational change. This change in the structure, from a rather flexible chain with a large conformational freedom to an essentially unique structure that remains stable in solution, allows many proteins to play a specific role in living organisms. Even the smallest protein is extremely complicated when considered together with its aqueous environment at atomic detail if one wants to calculate the free energy landscape of the system in a reasonable time using the computational resources currently available. In order to face these technical obstacles, it is useful to reduce the degrees of freedom inherent to proteins in aqueous solution. Reduced (or coarse-grained) models are really effective to study the folding process of proteins in a comprehensive way to avoid the computational limitations of more-detailed full-atom calculations. Levels of detail in the geometric representation of proteins vary therefore from the full atom representations to those in which each residue is represented by a single bead.[2−4]

Even in simple simulation models, the study of protein folding usually requires the definition of interactions depending on the amino acid sequence of the considered protein. To this end, different low-resolution mean field or knowledge-based potentials have been derived, which try at least to reproduce the partition between hydrophobic and hydrophilic residues in the protein core or surface, respectively, although they can also include the chemical nature of the interacting residues.[5−8] These potentials can use as interaction centers just the α-carbon positions, when very crude representations of proteins with a single bead per amino acid are used. The resulting potentials are usually too wide and shallow,[9,10] with a trend toward a too unspecific packing of the protein core that has to be avoided with additional contributions to the system energy.[6] Therefore, it is usually more convenient to use interaction centers that are not so distant in space as α-carbons are in compact conformations; this means to introduce in every residue some representation of the side chain, which is then used as the center for the sequence-dependent interactions.[11−13] In the last decades, different research groups have designed computer models for protein folding based on these premises. As an important part in these designs, different representations of side chains can be found in the literature.[14−18] However, and with few exceptions,[18−20] their derivations are not fully explained nor are their intrinsic values addressed because they are always related to the performance of the full simulation model and thus mainly to the force field employed. Here, we try to describe a careful derivation of a

Received: October 8, 2013
Published: December 20, 2013

rotamer library for side chains compatible with coarse-grained simulation models for protein folding and to assess its geometric quality independently of any interaction potential it may serve in the future.

In order to prepare a rotamer library, one has to understand why folded proteins present geometrical preferences in the local conformations of the amino acid side chains. These libraries are based on the existence of some preferred regions for the side chain torsional angles (usually named as χ-angles in textbooks), as evidenced by a detailed study of spatial and orientational distributions of amino acid side chains. Dihedral angle preferences for side chains were already established by Ramachandran and co-workers a long time ago[21] and have been experimentally measured and tabulated. These distributions are not random,[22] and preferred conformations of amino acid side chains in protein structures can be well established through statistically obtained rotamer libraries. Therefore, a rotamer library is a collection of rotamers for each residue type. A rotamer, on the other hand, is described by a set of geometrical parameters that allows determining the location of the atoms in an amino acid side chain. Several rotamer libraries that store full atom information for the side chains have been established, providing highly useful applications for the scientific community.[23−27] These libraries are usually classified into backbone-independent and backbone-dependent ones.[28] Belonging to the first group are those in which there is no reference to the particular backbone conformation; they are calculated from all available side chains of a certain amino acid type. Backbone-dependent rotamer libraries, on the other hand, present side chain conformations and/or frequencies that are dependent on the local backbone conformation of the corresponding residue.

It is also possible to adopt an intermediate situation by developing secondary structure-dependent libraries, whose rotamers do not depend on the specific local backbone conformation but on the type of secondary structure element in which the residue is located, i.e., these libraries present different angles and/or population values for side chains located at α-helix, β-sheet or coil secondary structures. In ref 28, Dunbrack et al. provide a general overview about the published rotamer libraries from the 1970s to the publication date, beginning with the backbone-independent library developed from only three proteins whose structure had been resolved at that time[21] to 2000, when Lovell and co-workers developed their backbone-independent library from 240 structures using a strict filter to select the side chains included into the analysis.[29]

Typically, the parameters sorted into these libraries permit tracing a side chain atom by atom. However, in order to design, implement, and apply a reduced protein model, the most important choice is the level of detail intended for the representation of the polypeptide chain. Some of the coarse-grained or "toy" models used for proteins normally consist of a description using two beads per residue, the α-carbon atom (named C^α in the remaining of this manuscript) and the centroid of the side chain,[30,31] although many other representations are possible, using for example the β-carbon or the most distant atom[13,17,20,31] to compute the interactions centered at the side chains. According to this fact, it is desirable to adapt the rotamer resolution bearing in mind the complexity of the protein representation and/or the potential employed[20,32] because the interaction scheme among residues is strictly connected to the geometric description of the protein.

In this work, we describe in full detail the building of a new backbone-dependent library for side chain rotamers compatible with a low resolution protein model. The idea is that if one is interested in studying the protein-folding process, a vast amount of conformations for the model chain has to be sampled along the numerical calculations involved in molecular simulations. Therefore, the definition of the model, including the side chain rotamers, has to be very easy to compute in a fast manner, so that the total simulation time remains reasonable, even with relatively modest computational resources. Within this aim, the model that inspires this work contains two interaction centers per residue: one on the backbone, identified with the C^α, and the second one on the amino acid side chain, represented by the geometric center of its heavy atoms. With this idea, we have estimated the side chain orientational preferences in natural amino acids from a statistical analysis of a structural database that contains a significant number of structures elucidated by X-ray crystallography from the Protein Data Bank (PDB).[33]

The study summarized in this work is composed of three parts. First, we have analyzed the possible backbone dependence of the side chain orientation by using only the distances between C^α atoms separated by one residue, instead of the real backbone dihedral angles or any other angular parameter related to the full-atom polypeptide backbone, which will be only available in more complex representations of the protein. The next step involves a simultaneous analysis of the distance and orientational preferences of side chain positions with respect to the backbone, taking into account the intervals predefined for the protein backbone resulting from the first analysis, explained in detail in the Methods section. Finally, we have obtained for each amino acid type a variable number of discrete conformations not equally populated, represented by one distance (between the C^α atom and its corresponding side chain centroid) and two angles that set the orientation of the centroid with respect to a local reference system attached to the C^α. This is our rotamer library. To check it, a set of structures from the PDB, not included in the initial training database, have been chosen. We have selected the proteins for this blind test according to several structural and experimental characteristics to see whether any of these factors may influence the performance of the library. We have also compared the results computed using our rotamer library to rebuild the centroids of all the side chains in the residues of the training and test structures with the reconstructed centroids obtained using the Park and Levitt rotamer model[19] on the same native protein backbones. This model is a very simple one and very adequate for the type of simulations of protein folding we plan to set up in the near future. The main difference between the Park and Levitt rotamer model and ours is that we include a larger flexibility in the side chain conformations, which we try to implement with a very small computational cost.

The results presented in this manuscript support the good correlation between the reconstructed side chain centers from our rotamer library and those in experimentally determined PDB structures for each of the proteins tested.

■ METHODS

We have estimated the side chain orientational preferences in natural amino acids from the statistical analysis of a structural training database consisting of 1584 structures with less than 30% of sequence homology[34] elucidated by X-ray crystallography (whose resolution is better than 2.0 Å) from the PDB.[33]

In these structures, residues with more than one set of coordinates or an incomplete number of atoms, or those that are at the chain ends or at the boundaries of chain gaps (in structures with missed residues inside the sequence) have been eliminated from the analysis. The number of residues of each type included in our analysis are reported in the first column of Table 1.

Table 1. Number of Residues in Our Analysis and Values for Side Chain Virtual Bond Lengths between the C$^\alpha$ Atom and Side Chain Centroid as a Function of Residue Type *m*

amino acid	no. cases	d_m (Å)[a]		d_m^{PL} (Å)[b]
Ala	31173	1.56	–	1.5
Arg	16226	4.02	4.66	4.1
Asn	15896	2.52	–	2.5
Asp	21427	2.51	–	2.5
Cys	4782	2.11	–	2.0
Gln	13278	2.73	3.38	3.1
Glu	22678	2.76	3.38	3.1
His	8439	3.19	–	3.1
Ile	20085	2.11	2.43	2.3
Leu	32324	2.65	–	2.6
Lys	19737	3.15	3.73	3.5
Met	4451	2.66	3.20	3.0
Phe	14685	3.45	–	3.4
Pro	16606	1.90	–	1.9
Ser	12304	1.94	–	1.9
Thr	20586	1.98	–	1.9
Trp	5528	3.85	–	3.9
Tyr	13091	3.82	–	3.8
Val	26089	1.98	–	2.0

[a] Distance values in our rotamer model. [b] Distance values corresponding to the Park and Levitt model.[19]

In order to build a simple coarse-grained model, the backbone in every residue has been represented by its C$^\alpha$ coordinates alone. In addition, to define the side chain centroids in our model, we have used the average coordinates of their side chain heavy atoms. Henceforth, this centroid will be represented as SC. According to this, every residue in our model has been represented by the coordinates of its C$^\alpha$ and a set of possibilities for the position of its SC.

Setting the Parameters: Definition of a Backbone-Coupled Reference System. Considering any given amino acid *i*, other than the two chain ends, we have set three axes forming an orthonormal basis on its C$^\alpha$ atom, as indicated in Figure 1(a). The vectors starting at this atom and leading toward the neighbor α-carbons are all that is needed to define these axes, as indicated in the figure. This means a very simple approach when compared to previous ones that use the N and C′ atoms in the backbone, although the geometrical building of the reference set is quite similar.[20] Once the reference system has been defined, three variables, the two angles θ_i, ϕ_i and the distance d_i between the C_i^α atom and the centroid SC$_i$, describe the localization of the side chain bead with respect to the backbone (Figure 1(b)). Because of the definition of the axes, which require the presence of neighbor residues, it is not possible to set an orthonormal basis on the end residues. Thus, they are excluded from our analysis. In the case of glycines, which do not possess a side chain, we directly locate its "centroid" on the C$^\alpha$ atom for the sake of completeness of the model geometry. This is equivalent to setting for this type of residues $d_i = 0$, and obviously, it is not necessary to calculate any kind of rotamers on them.

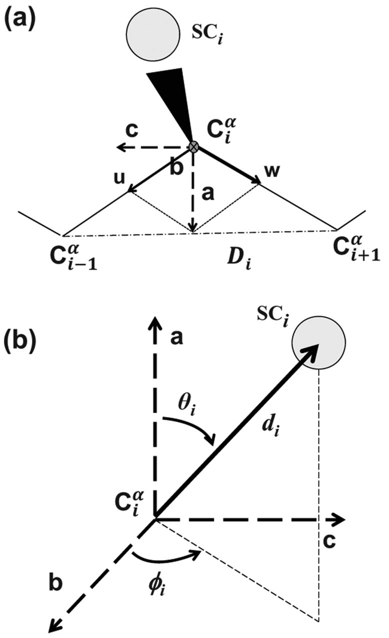


Figure 1. Schematic representation of a coarse-grained polypeptide chain. (a) The peptide backbone is represented by C$^\alpha$ atoms. Virtual bonds connecting consecutive α-carbons are taken from the PDB file of the considered protein. D_i represents the distance between C_{i-1}^α and C_{i+1}^α (and determines the backbone dependence of residue *i*). The side chain attached to the C_i^α is represented as SC$_i$. On every residue, two unit vectors (u and w) have been defined. The addition of these vectors generates vector a, in the same plane, while their cross product generates vector b, perpendicular to both. The cross product a × b generates vector c. With this vector, the definition of the reference system on a given residue *i* is completed. (b) In this reference system, the position and the orientation of the side chain centroids are fully located using one distance (d_i) and two angles (90° ≤ θ_i ≤ 180° and −180° ≤ ϕ_i ≤ 180°).

The distance D_i, computed between C_{i-1}^α and C_{i+1}^α and indicated in Figure 1(a), provides the backbone dependence in our library. To define it, most of the previous rotamer studies have included the local backbone structural preferences employing either the Ramachandran angles or the element of secondary structure in which the residue is sitting (see, for example, ref 28 and references therein). However, we have developed a different approach consistent with the coarse-grained characteristics of a simple description of the polypeptide chain. We have completed a rigorous statistical analysis on the selected PDB structures of our training set, computing all the possible D_i distances between C$^\alpha$ atoms that are second-neighbors along the sequence and obtaining their distribution function, which is shown in Figure 2(a). It clearly shows two mainly populated peaks. The highest one (which we have used to define an interval for D_i labeled $I^{(a)}$) is centered around 5.4 Å; it mainly includes residues located in α-helices and tight turns. The second one (which is used to define another interval for D_i, labeled as $I^{(c)}$), centered around 6.7 Å, corresponds to amino acids mainly located on β-sheets and other extended regions of the chain. In addition to these two intervals for D_i, we have decided to include an additional one, $I^{(b)}$, in an intermediate range of distances between 5.75 and 6.25 Å (Figure 2(a)). It corresponds to the region between both peaks, which still shows a significant population due to the

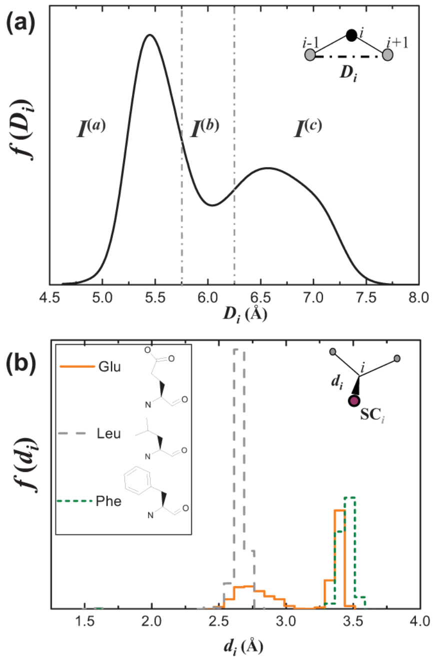


Figure 2. (a) Distribution function for the distance D_i between $C^{\alpha}_{i-1} - C^{\alpha}_{i+1}$ atoms calculated using all the proteins in the training set. (b) Distribution function of the distance d_i between C^{α}_i and SC_i for three residues. Leu and Phe present only one peak. Glu is an example of a typical distribution for amino acids with more than one peak in this curve, not all them equally populated.

presence of many amino acids in sections without a defined secondary structure along the native conformation of the corresponding proteins.

Computing the Library: Coarse Graining the Spatial Preferences of Side Chains in Proteins. The next step in our procedure consists of a simultaneous analysis of the distances d_i between the C^{α} and the SC in a given residue i and the orientational preferences of the side chain centroid positions with respect to the backbone, taking the three intervals predefined for the protein backbone mentioned above into consideration.

First, we have analyzed for every type of nonglycine amino acid the distribution function for its possible d_i values. An example of the type of distance distributions we have obtained in our library is shown in Figure 2(b) for a few representative residues. The complete list of resulting average d_m values for the single or double peaks in these distributions for every type of residue m is shown in Table 1. A very similar set of average distances for the different amino acids had been already observed by Keskin and Bahar,[31] although the distance values between the two beads per residue in the protein model described by these authors (l^s_i) slightly differ from the distances presented in this work (d_i) due to the different criteria employed to select the position of the bead representing the side chains.[22]

For many residue types (Table 1), only one peak is found in the distribution curve. For example, we have observed this behavior in Leu (dashed line in Figure 2(b)), whose peak appears centered at 2.65 Å, and in Phe (dotted line), whose single peak appears at 3.45 Å. For all these residues, the values are almost identical to those previously obtained in the rotamer library of Park and Levitt,[19] shown in the last column of the table. The very minor differences are probably due to the larger database of native proteins that we have been able to use. In Arg, Gln, Glu (the latter also shown as an example in Figure 2(b)), Ile, Lys, and Met, two peaks are observed in our analysis. In these cases, we use every single peak to compute an independent average for d_m, resulting in the two different values collected in Table 1 for these residues. Now, the results from Park and Levitt[19] usually stay between our peaks. However, including all the possibilities into a single C^{α}−SC distance can result in an average value that corresponds to a negligible population of centroid positions. Therefore, the possibility to include the side chain flexibility, even at the basic level of the distance between a side chain centroid and its α-carbon, represents an improvement of the resulting library at a rather modest cost.

Therefore, every peak resulting from our statistical analysis is treated independently from each other in the subsequent orientational analysis of the side chain centroids in native proteins. This new stage of the study, which will be detailed below, involves the screening of the set of orientational parameters (θ_i and ϕ_i) for every distance d_i found for each type of residue under one of the backbone situations described above.

Thus, we have evaluated the preferred spatial orientation of the centroids for every nonglycine amino acid with respect to the backbone. This analytical process is detailed, using Asp as an example, in Figure 3. First, we must obtain for every value of d_i and every backbone distance interval $I^{(x)}$ (with $x = a,b,c$ as defined in Figure 2(a)) a scattered graph as shown in the left column graphs of Figure 3. A spot in these graphs represents

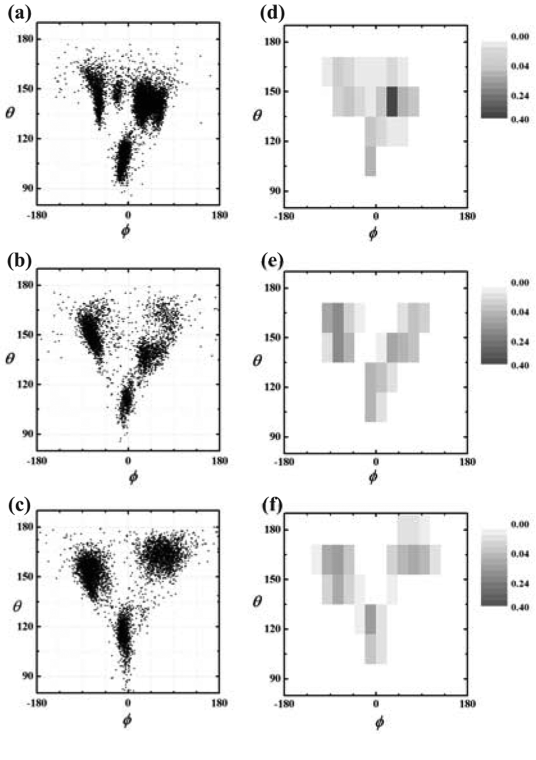


Figure 3. Definition of the angular dependence for the rotamers in the library. Left column graphs: Every point on the scattered plots represents the angular coordinates for each Asp residue found in the backbone intervals, (a) $I^{(a)}$, (b) $I^{(b)}$, (c) $I^{(c)}$, in the proteins of the training database. Right column graphs: Corresponding 2D-histograms obtained for Asp to evaluate the average angles and the probability p of each rotamer, including at least 95% of all the cases analyzed.

the angular coordinates θ and ϕ, defined in our reference system, for the centroid of one of the residues found in the training database. Values of θ below 90° have not been significantly found, neither in Asp nor in any of the other amino acids, due to the chirality of the C^{α} in natural amino acids comprising the polypeptide chains. Using these graphs, we can readily assess the regions with a higher density of spots in a qualitative way, establishing a preliminary estimation of the number of rotamers for the evaluated residue in the considered intervals. To quantify the information shown in these graphs, we have calculated standardized bidimensional histograms (the right column graphs in Figure 3). The bin width in these histograms is set to 10° and 15° for the axes θ and ϕ, respectively. This fact allows us to use the values of θ and ϕ registered for each spot, integrating the corresponding region on the 2D-histogram to compute average values for these two angles, which define a particular rotamer. Moreover, the probability p for the different regions can also be evaluated from their distinct populations as an important additional information, which is also included in the rotamer library.

The histogram itself has been carefully computed and analyzed in order to include at least 95% of the spots appearing in the left column graphs. As a matter of fact, in some cases obtaining the rotamers is not a trivial process, mainly because of the complexity of the computed 2D-histograms. An example of this situation is shown in Figure 4, where we see one of the histograms for the amino acid Arg. As it can be appreciated, there is an almost continuous filling of the available space, which makes it very difficult to establish the boundaries between rotamers or even the number of them. In cases like this one, we have included an additional intermediate step to ensure the validity of the defined discrete rotamers, according to the quality standards we want to enforce. After computing a very small preliminary set of rotamers from the 2D-histogram, we have evaluated for each of the resulting rotamers j the width of the distribution that yields to its average values. To do that, we have established a maximal divergence criterion, i.e., we do not allow the width of these distributions to exceed half of the distance between the side chain centroid and its C^{α}, d_m, where m corresponds again to the amino acid type. Distance values of d_m corresponding to the Park and Levitt model (Table 1) are used in this evaluation stage. This way we use an external reference in this part of the calculations. This cutoff permits a maximal divergence of approximately 30° in an angular scale between the experimental and the rebuilt centroid positions. To implement this requirement, we have rebuilt each of the centroids (SC_k) for every residue k whose side chain centroid is located in the region that we have selected to define rotamer j. Then, we have evaluated the distance d^{kj} between the original centroid SC_k (computed from the PDB coordinates) and the reconstructed centroid obtained on C_k^{α} by using the coordinates θ_j and ϕ_j of rotamer j that is being tested. A wide distribution in the values of d^{kj} that goes beyond the allowed limit mentioned above, or a multipeaked curve in the distribution of these distances, indicate a poor selection of bins to calculate rotamer j, bringing as a consequence the splitting of the original rotamer into two or more independent ones. This refinement process is repeated for each rotamer definition as much as needed in order to obtain narrow and unimodal distributions of d^{kj}. The number of rotamers may become large in cases as the one shown in Figure 4. However, as we shall see in the Results and Discussion section, the quality of the final results proves that this procedure is definitely contributing to the excellence of the resulting rotamer library.

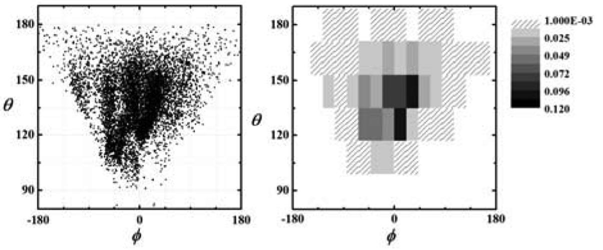


Figure 4. Example of the definition of the angular dependence for difficult rotamers in the library. Left: Individual angular coordinates for Arg residues in the backbone interval $I^{(a)}$ and in the peak with average d_i = 4.02 Å. Right: Corresponding 2D-histogram.

During this fine-tuning procedure, that we have applied to the 2D-histograms obtained for all the amino acid types, some problems with prolines (Pro) in cis conformation have been detected. Reconstructed cis-Pro side-chains significantly depart from the PDB cis-Pro coordinates, with deviations well above the tolerance limit fixed for this amino acid no matter how many rotamers (up to a sensible number) are defined. This is in part due to the fact that cis-Pro residues produce very short D_i backbone distances. Bearing in mind that the abundance of cis-Pro in natural proteins[35] does not exceed 7% and, as we have mentioned before, we have excluded from our 2D-histograms a maximum 5% of cases, being the amount of Pro in the cis configuration very close to this amount, we have dispensed with this type of prolines for the other stages of our study. This situation has also been observed in previous coarse-grained rotamer libraries[20] where cis-Pro residues have also been excluded from the calculations.

Validation of the Library. As a first test of our library, we have taken the C^{α} coordinates of the proteins in our training data set to reconstruct every single side chain centroid in two different ways:

(1) Using as a very simple reference the geometric model described by Park and Levitt,[19] the centroid coordinates of an i^{th} residue can be reconstructed from the C^{α} coordinates of the residues $i - 1$ and $i + 1$ by the relation

$$\mathbf{r}_{\text{SC}}^{\text{PL}} = l\cos\theta\mathbf{x} + l\sin\theta\mathbf{y} \qquad (1)$$

where l is the distance from a side chain centroid to its C^{α} atom, which depends on the residue type (registered as d_m^{PL} in Table 1), and θ is the out of plane angle used, fixed at 37.5° for all the residues. Unit vectors x and y are obtained from the C^{α} coordinates of the contiguous residues.[19] Equation 1 constitutes one very simple and probably the most computationally efficient way of implementing a single bead side chain representation in a coarse-grained model for a polypeptide chain.

(2) Using the geometrical and statistical information included in the rotamer library described in this work is the other way to reconstruct a single side chain centroid.

In this first test, we want to check whether the number of derived rotamers in our library, together with the averaging procedures leading to their definitions, result in an adequate representation of the protein side chains.

In addition, we have used another set of 46 proteins, different from those in the training set, which have also been analyzed in the way previously described. To complete a significant test, we have chosen this blind data set from the PDB taking into account the experimental technique by which proteins were resolved (NMR or X-ray methods) and some structural features as, for example, the number of residues and the main type of secondary structure present.

RESULTS AND DISCUSSION

We have developed a new backbone-dependent coarse-grained rotamer library. We have prepared it as a single large file that merges all the geometrical features mentioned in the previous section for every rotamer: backbone interval $I^{(x)}$, distance d from the side chain center to the backbone, and angles θ and ϕ, together with statistical information (probability p for every rotamer). This is all the information needed to reconstruct any side chain centroid SC_i of an amino acid i from the C^α coordinates of residues $i - 1$, i, and $i + 1$. As mentioned above glycines, cis-prolines, and terminal residues are excluded from the procedure.

As a summary of our rotamer library, we show in Figure 5 the populations of the four most populated rotamers for each amino acid type. In order to properly organize this statistical information, we have represented the values of p using bar charts with an independent graph for every backbone interval $I^{(x)}$ considered in our procedure. A first look at the three graphs in this figure gives us the first relevant result. In many of the cases, the most probable rotamer does not have a probability above 50%, i.e., choosing the rotamer with the highest value of p is less likely than choosing any of the remaining options. This fact demonstrates the usefulness of this rotamer library because an exceedingly simple geometrical reconstruction from backbone coordinates,[19] which reduces all the options into a single one or gives an exceedingly important role to the most probable rotamer,[20] may be clearly insufficient in many occasions. On the other hand, even in a coarse-grained representation, it seems necessary to take into account the different geometries an amino acid side chain can adopt in a polypeptide chain, based not only on the chemical nature of the residue but also on the local configuration of the backbone.

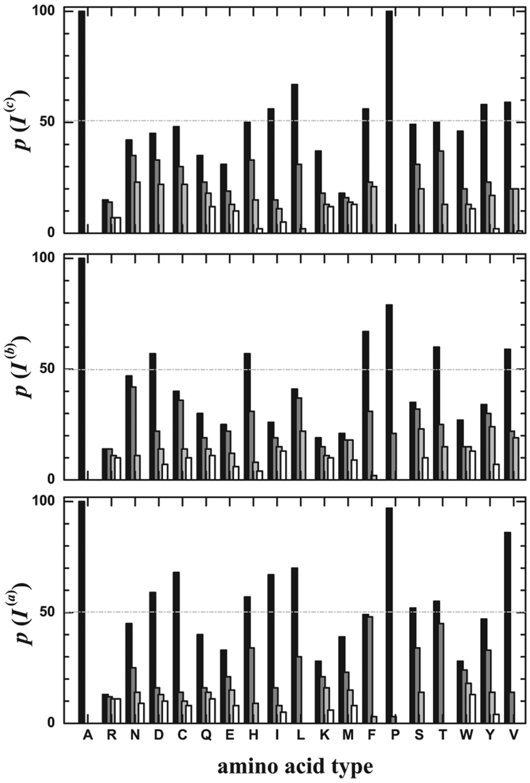


Figure 5. Library at a glance. Probability, for each backbone distance interval $I^{(x)}$, of different rotamers found for every amino acid type. Only the most populated rotamers are shown in the bar graphs, up to 4 per amino acid and per backbone interval.

As is known and evident from Figure 5, alanine (residue symbol A) side chain lacks any conformational freedom, whatever its local backbone geometry is. In this amino acid, the computed "rotamer" corresponds to the β-carbon of its side chain, as expected. More interesting is what happens with the other types of amino acids. They present a variable number of rotamers. The larger and more flexible a side chain is, the higher becomes the number of available rotamers in the library, a usual situation found in Arg (R), Gln (Q), Glu (E), Lys (K), Met (M), Asn (N), and Trp (W) in any of the backbone intervals. The different rotamers for a given amino acid frequently present rather different probabilities in our library, as we have mentioned above. In some cases, there is not a predominant rotamer with a value of p greater than 50%, although this fact can be conditioned by the backbone geometry. For example, this is observed in Cys (residue symbol C). For the backbone interval $I^{(a)}$ (bottom graph, Figure 5), one rotamer has a probability that clearly overrides the value of p for the others. This situation does not hold either in the $I^{(b)}$ or $I^{(c)}$ intervals (middle and top graphs, respectively). Moreover, in the backbone interval $I^{(c)}$, only three rotamers with a significant population appear for this residue, while four rotamers are found in the other two backbone intervals. A similar situation can be described for Asp (D), His (H), Ile (I), Leu (L), Phe (F), Ser (S), Thr (T), and Tyr (Y). On the other hand, Pro (P, with a very rigid side chain given its cyclic character) and Val (V) present a clearly predominant rotamer whatever the backbone interval is, although the number of populated rotamers also varies from one interval to other. These facts confirm the existence of remarkable differences among the three backbone intervals described in this work, both in the number of rotamers and in their probabilities. Hence, the relative populations among different rotamers seem not to be preserved across the different backbone intervals. The differences in interval $I^{(b)}$ with respect to $I^{(a)}$ and $I^{(c)}$ also support our choice of three different intervals for the local backbone dependence of our rotamer library, a fact which could have been initially thought as arbitrary from the results of Figure 2.

The main goal of developing our rotamer library is to design and implement a new tool in order to reconstruct a simple representation of the amino acid side chains in proteins without losing sight of the distinctive features of coarse-grained models, i.e., using only the essential information preserved in the C^α coordinates from the protein backbone. Therefore, it is important to be sure of the quality of the set of rotamers included in our library for every amino acid. One of the key issues for us is to be able to find among all these rotamers one that allows us to reconstruct the side chain centroid as near as possible to the experimentally determined native centroid, whose coordinates are taken from a PDB file. It does not need to be the most populated one, which as we have just discussed may not always be a good representative of the available conformational space for a given side chain. To assess how correctly we are able to reconstruct the centroids, we have carried out a first test on the 1584 proteins of the structural database used as training set to check the consequences of the different clustering and averaging procedures included in our definition of the different rotamers, as defined in the Methods section. We have reconstructed all the centroids for the internal residues in these proteins using our rotamer library and compared them with the corresponding PDB centroids for each nonglycine residue i using the general expression

$$r(i) = |\mathbf{r}_{\mathrm{SC}}(i) - \mathbf{r}_{\mathrm{SC}}^{\mathrm{PDB}}(i)| \qquad (2)$$

where $r(i)$ represents the distance between the experimental centroid used as reference (computed from the PDB coordinates), whose position is given by $\mathrm{r}_{\mathrm{SC}}^{\mathrm{PDB}}(i)$, and the centroid reconstructed from our rotamer library, with coordinates $\mathrm{r}_{\mathrm{SC}}(i)$. We have followed three different methods to compute this distance, each of them representing an independent test:

TEST 1: We have used the geometrical definition given by Park and Levitt[19] in order to set the centroid positions $\mathrm{r}_{\mathrm{SC}}(i)$ using eq 1. As we have already mentioned, this is a simple yet rather rigid way to have an estimation of the side chains.

TEST 2: We have used our library choosing the rotamer with the highest probability p among all the possible rotamers for a given amino acid, according to its backbone local geometry. This way we can continue with the evaluation of a possible choice of this rotamer alone in a very reduced version of the library.

TEST 3: We have used our library to extract among all the available rotamers for every amino acid in the corresponding backbone configuration the nearest one to the centroid in the native structure of the protein, i.e., the one which makes $r(i)$ minimum. This test tries to check the flexibility of our rotamer library to accommodate the different situations found in the studied protein conformations.

In order to present all this information in a reasonably compact manner, we have calculated with each of these methods an average $\overline{r}$ over the values $r(i)$ of the residues belonging to every protein in the training database

$$\overline{r} = \frac{\sum_{i=2}^{N-1} r(i)}{N'} \qquad (3)$$

where N is the number of residues in that particular protein. Neither the possible glycines nor the residues at the ends of the polypeptide chain are considered in this calculation. Cis-prolines have also been excluded, so the number of residues N' included in the calculation for every protein is $N' < N$.

The results of this analysis are shown in Figure 6, where the averages obtained in the three tests for every protein are plotted against the protein size (actually, the number N' of rebuilt side chains in every protein). For Test 1 (Park and Levitt centroids, black symbols) and Test 2 (centroids from rotamers in our library with the highest value of p, in red), the results are roughly comparable, with the mean values slightly above 1 Å in both tests, although the deviations between the reconstructed and experimental centroids are consistently smaller with our rotamers. This implies that the rotamer library from Park and Levitt, in spite of its simplicity, may result in a convenient representation of the side chain centroids if the internal flexibility of the side chains is not important in the considered model. We have not appreciated any relevant connection between the size of the protein and the values of $\overline{r}$ calculated.

On the other hand, the results from Test 3 (closest possible rotamer in the library, in green) indicate that most of the side chain centroids have been recalculated less than 0.5 Å away from the native side chain centroid taken from the PDB, with an average value of 0.4 Å over the full set. There is not any dependence on the number of residues in this case either. These results support once more the fact that in many cases the most probable rotamer is not the most suitable one to reconstruct a side chain centroid in protein structures and that a significant improvement is achieved by including a certain flexibility (a larger number of rotamers) in the library. Moreover, the results show that our rotamer library is complete enough to guarantee that a rotamer can be found for every residue that accurately reproduces the experimental one, at least at the level of resolution implemented in our model.

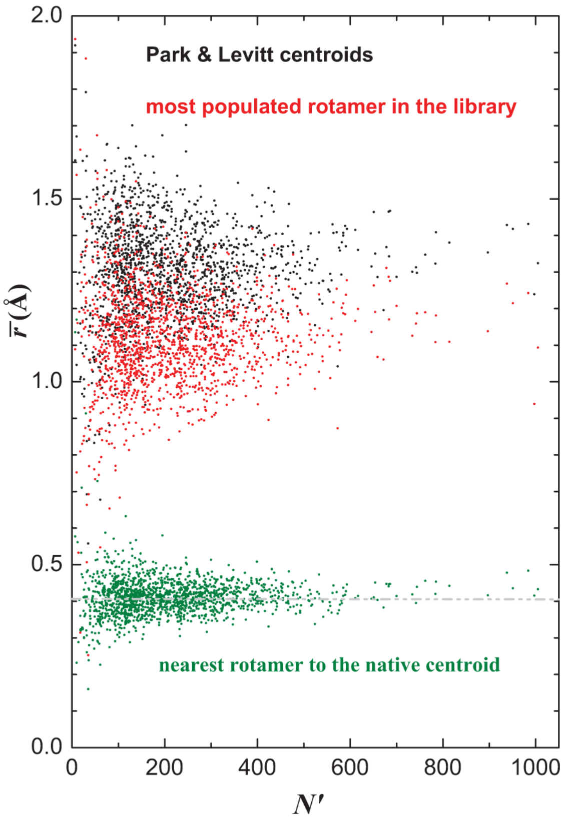


Figure 6. Average values (for each protein in the training set) of the distances obtained from eq 3, calculated following three different methods in order to test the reconstructed side chain centroids. Values of $\overline{r}$ calculated from Park and Levitt centroids[19] are colored in black. Values from the most probable side chain centroid of our library are colored in red. Values taken from the side chain centroid in our library that lies the closest to the experimental one are colored in green.

In order to blindly evaluate the performance of the library, we have chosen, as mentioned in the Methods section, an independent set of 46 protein structures not included in the training database from which the library has been calculated. These structures have been classified into five different groups. We have selected proteins for this set according to two different criteria: their main secondary structures and the experimental technique used to solve them. According to the secondary structure, we have established four groups: mainly α proteins (Table 2), mainly β proteins (Table 3), proteins with α and β segments (Table 4), and an additional group formed by proteins with a high fraction of coil segments (Table 5). All of these structures have been resolved by X-ray diffraction. Information about the size of the protein (N), number of

Table 2. Results for Differences between the Experimental and Closest Rotamer in Our Library for Mainly α Proteins in the Test Set[a]

PDB	N	N'	$\overline{r}_3 \pm \Delta\overline{r}_3$ (Å)	$\overline{z}_3 \pm \Delta\overline{z}_3$
1a7w	68	60	0.31 ± 0.03	0.23 ± 0.02
1af3	145	131	0.42 ± 0.03	0.31 ± 0.02
1ail	70	62	0.33 ± 0.03	0.23 ± 0.02
1cc5	83	65	0.51 ± 0.05	0.42 ± 0.03
1enh	54	50	0.36 ± 0.03	0.24 ± 0.02
1f1f	88	76	0.37 ± 0.04	0.29 ± 0.03
1gak	137	121	0.44 ± 0.03	0.28 ± 0.02
1p7n	176	150	0.38 ± 0.02	0.30 ± 0.01
1rop	56	54	0.33 ± 0.04	0.24 ± 0.03
2ig3	127	111	0.38 ± 0.03	0.25 ± 0.02

[a] Information about the protein size and mean values obtained in Test 3 is gathered in the table (see text for more details).

Table 3. Results for Differences between the Experimental and Closest Rotamer in Our Library for Mainly β Proteins in the Test Set

PDB	N	N'	$\overline{r}_3 \pm \Delta\overline{r}_3$ (Å)	$\overline{z}_3 \pm \Delta\overline{z}_3$
15c8	216	181	0.41 ± 0.02	0.33 ± 0.02
1a1x	106	92	0.39 ± 0.02	0.28 ± 0.01
1a3k	137	117	0.38 ± 0.02	0.28 ± 0.01
1acx	108	87	0.48 ± 0.03	0.44 ± 0.03
1ag6	99	78	0.38 ± 0.05	0.29 ± 0.03
1ame	66	54	0.36 ± 0.04	0.29 ± 0.03
1bfg	126	108	0.46 ± 0.03	0.32 ± 0.02
1wba	171	139	0.40 ± 0.03	0.29 ± 0.02
2ioo	187	157	0.37 ± 0.02	0.27 ± 0.01
4gcr	174	151	0.48 ± 0.03	0.32 ± 0.01

Table 4. Results for Differences between the Experimental and Closest Rotamer in Our Library for α/β Proteins in the Test Set

PDB	N	N'	$\overline{r}_3 \pm \Delta\overline{r}_3$ (Å)	$\overline{z}_3 \pm \Delta\overline{z}_3$
121p	166	150	0.36 ± 0.02	0.26 ± 0.01
1a3s	158	129	0.46 ± 0.03	0.33 ± 0.02
1agi	125	105	0.43 ± 0.04	0.30 ± 0.02
1ayd	101	85	0.47 ± 0.04	0.34 ± 0.02
1b1i	122	103	0.48 ± 0.04	0.34 ± 0.02
1bk7	190	162	0.36 ± 0.02	0.27 ± 0.01
1jhy	346	286	0.37 ± 0.02	0.29 ± 0.01
1pgb	56	50	0.30 ± 0.03	0.22 ± 0.02
2fox	138	119	0.37 ± 0.03	0.27 ± 0.02
7rat	124	103	0.37 ± 0.02	0.29 ± 0.02

Table 5. Results for Differences between the Experimental and Closest Rotamer in Our Library for Proteins without Defined Secondary Structure in the Test Set

PDB	N	N'	$\overline{r}_3 \pm \Delta\overline{r}_3$ (Å)	$\overline{z}_3 \pm \Delta\overline{z}_3$
1aap	56	46	0.40 ± 0.04	0.28 ± 0.03
1ab1	46	36	0.45 ± 0.05	0.39 ± 0.04
1bik	110	89	0.39 ± 0.03	0.29 ± 0.02
1eyt	83	69	0.34 ± 0.03	0.26 ± 0.02
1z3s	216	184	0.44 ± 0.02	0.33 ± 0.02
2hip	71	58	0.41 ± 0.03	0.31 ± 0.02

residues (N') included in the calculation of the mean value given in eq 3, and results of these averages obtained according to Test 3 ($\overline{r}_3$) are gathered for each protein structure in the corresponding table. The statistical errors of the average values are also reported. We have also included an additional reduced mean value, $\overline{z}_3$, which is calculated as follows

$$\overline{z}_3 = \frac{1}{N'}\sum_{i=2}^{N-1}\frac{r(i)}{r_{\text{cut}}^{m}} \qquad (4)$$

where r_{cut}^{m} depends on the amino acid type m and is calculated as $r_{\text{cut}}^{m} = d_m/2$, using as distance values for d_m the Park and Levitt values for the distances between α-carbons and side chain centroids in the different residue types m (see the last column in Table 1). It represents a measurement of the angular divergence between the reconstructed side chain centroid and the experimental value, a property that has taken an important role in the definition of our library, as described in the Methods section. According to this equation, values of z_3 for an individual residue above 1 indicate that the evaluated centroid for its side chain has been reconstructed outside the tolerance limit of our library, fixed in 30° as we have already detailed. The mean value ($\overline{z}_3$) calculated over the whole protein structure gives us an additional idea about the efficiency of the side chain centroid reconstruction and, specially, of the "complete" character of our rotamer library.

In Tables 2, 3, 4, and 5, similar results are observed for the average values among the different proteins inside each table and also among the different tables. This is a clear indication that our rotamer library performs equally well for different types of structural protein families. Both the averages on the deviation distances $\overline{r}_3$ and angular divergences $\overline{z}_3$ are rather low and with small statistical deviations, indicating that the reconstructed side chain centroids sit very close to the experimental ones. Neither the number of evaluated residues in each protein structure nor the type of protein according to its secondary structure seem to have any significant influence on the results.

To determine how the rotamer library depends on the experimental data (the fact that the library has been obtained from a training set of X-ray structures alone should be recalled here), we have selected 10 more proteins solved by NMR spectroscopy whose PDB files contain only one minimized average structure (Table 6). In this case, both distance averages and statistical deviations are clearly higher than those obtained for X-ray proteins commented above, although the values are still very reasonable at this average level. To check whether the structural averaging procedure may be blamed for the slightly larger deviations, in this table we have also included the results from our rebuilding procedure when we use the C^{α} coordinates for a single NMR model, taking the first one in the multimodel PDB file, when this is also deposited in the PDB as indicated in the table. As we can see, the differences between the results obtained using both sets of coordinates are inside the error bars of the corresponding averages, and therefore, they are not statistically meaningful.

As a final result, because the average values can be sometimes a too crude way of showing the actual performance of our library, we want to finish this section in a more detailed way, showing the results calculated on individual residues for one representative protein taken from the different subsets of our testing set. This way we present in each graph of Figure 7 the individual values of z_3 for each residue along the sequence of four different proteins, one of them (in the bottom graph) solved through two different experimental techniques. In the

Table 6. Results for Differences between the Experimental and Closest Rotamer in Our Library for NMR Proteins in the Test Set.

Minimized average structures[a]					First model[b]		
PDB	N	N'	$\bar{r}_3 \pm \Delta\bar{r}_3$ (Å)	$\bar{z}_3 \pm \Delta\bar{z}_3$	PDB	$\bar{r}_3 \pm \Delta\bar{r}_3$ (Å)	$\bar{z}_3 \pm \Delta\bar{z}_3$
1a23	189	166	0.67 ± 0.03	0.50 ± 0.02	1a24	0.62 ± 0.03	0.47 ± 0.02
1ak6	174	153	0.76 ± 0.04	0.58 ± 0.03	1ak7	0.78 ± 0.04	0.60 ± 0.03
1bbl	37	30	0.52 ± 0.05	0.41 ± 0.05	1bal[c]	0.73 ± 0.06	0.58 ± 0.05
1dem	60	50	0.73 ± 0.04	0.52 ± 0.03	1den	0.77 ± 0.05	0.55 ± 0.04
1d8v	263	243	0.58 ± 0.03	0.45 ± 0.02	not available in the PDB		
1ef5	88	82	0.73 ± 0.05	0.57 ± 0.04	not available in the PDB		
1f7w	144	121	0.52 ± 0.03	0.40 ± 0.02	not available in the PDB		
1hnr	47	39	0.63 ± 0.07	0.48 ± 0.06	1hns	0.64 ± 0.09	0.46 ± 0.07
1km7	100	88	0.72 ± 0.05	0.53 ± 0.03	1klv	0.80 ± 0.05	0.58 ± 0.03
2gb1	56	50	0.61 ± 0.05	0.49 ± 0.04	1gb1	0.63 ± 0.06	0.51 ± 0.04

[a] Minimized average structures deposited in the PDB. [b] First model from the multi-model NMR structure deposited in the PDB. [c] This structure has $N = 51$, $N' = 44$, which explains the larger deviations with respect to the minimized average structure 1bbl, with less residues due to the removal of the disordered parts of the protein for the structural averaging.

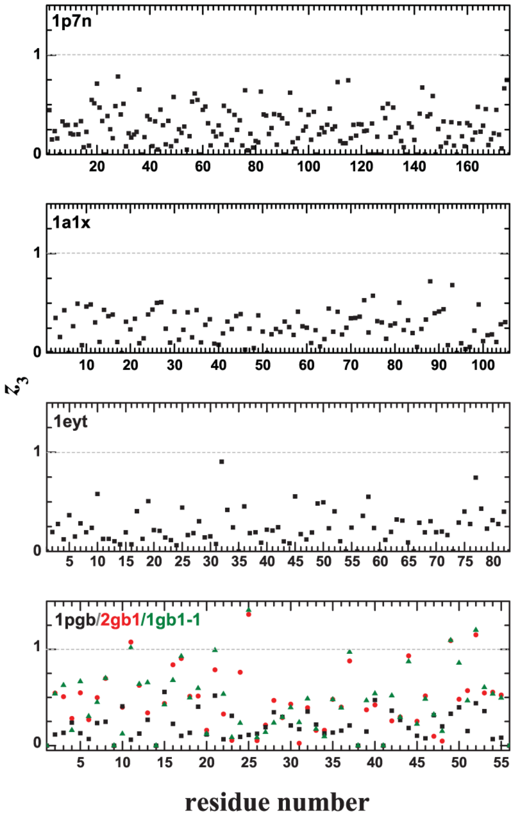


Figure 7. Comparison of the relative angular displacement profiles (z_3, as defined in eq 4) for every residue of several representative X-ray and NMR protein structures. The first three graphs are examples of X-ray structures: 1p7n, a mainly α protein; 1a1x, a mainly β protein; and 1eyt, a protein without defined secondary structure. In the last panel, we show the results for protein GB1, an α/β protein, under the PDB codes 1pgb (X-ray structure, black symbols), 2gb1 (minimized average NMR structure, red symbols), and 1gb1-1 (first model from those in the 1gb1 file, green symbols).

cases of 1p7n, 1a1x, and 1eyt (mainly α, β, and coiled proteins, respectively), there is no value of z_3 above 1. As a matter of fact, the values for most of the residues lie well below 0.5 Å, and as a consequence, the average values of $\bar{z}_3$ for these proteins, collected in Tables 2, 3, and 5, do not exceed 0.3 Å. This proves the correct capabilities of the rotamer library along the full protein sequence and explains the small statistical errors for the average values reported in the results tables. In the last graph of Figure 7, we have compared the results on three different structures deposited for the immunoglobulin-binding domain of the *streptococcal* protein G because its structure can be found under the PDB codes 1gb1,[36] 2gb1,[36] and 1pgb,[37] depending on the experimental methods by which they were resolved (NMR and X-ray crystallography, respectively). In this case, we can use for the NMR structure, as we did before, the minimized average structure (2gb1) or the first model from the total of 60 individual models under the PDB code 1gb1. As it happened with the results in Table 6, the deviations obtained for individual residues are very similar between the minimized average structure and the single model. From the data obtained either with the minimized NMR structure 2gb1 (colored in red) or with the first model from 1gb1 (in green), we find four residues with a value of z_3 above 1. As is evident, these residues (Thr11, Thr25, Thr49, and Phe52) are reconstructed much closer to the native side chain centroid in the X-ray structure (1pgb, black spots) than in 2gb1 or 1gb1. We have not found anything particularly relevant for these four residues. They are neither specially exposed in the protein surface nor show any other distinct feature. As a matter of fact, there are several other residues in this NMR structure that, being below the threshold $z_3 = 1$, lie close to it. These results are important because the intrinsic flexibility of a NMR determined structure, even blurred by the averaging process used to get a single set of coordinates or by selecting just one of the available models, can still be present and, therefore, appear as larger deviations from our rotamer library, which as already stated is derived from X-ray structures.

A detailed analysis per type of amino acid has allowed us to quantify the percentage of residues from the full set of testing structures that have not been located as close to the native side chain centroid as we have set in our criterion of maximal divergence, i.e., they show $z_3 > 1$. The results are shown in Table 7 as a function of the residue type. We have independently analyzed the X-ray and the minimized NMR protein structures looking for the residues that have not been properly placed. In the X-ray structures, as it could be guessed from all the previous results, the percentages are always well below the 5% threshold used to define our rotamers, and actually, they become almost negligible in a number of cases. The misplaced residues are, expectedly, more abundant in

Table 7. Percentage of Analyzed Residues That Were Not Properly Set in the Test Proteins According to the Divergence Criterium Explained in the Text[a]

amino acid	r_{cut} (Å)	X-ray (%)	NMR (%)
Ala	0.75	2.5	4.5
Arg	2.05	0.8	0.0
Asn	1.25	1.9	11.5
Asp	1.25	0.0	7.4
Cys[b]	1.00	–	–
Gln	1.55	0.0	2.1
Glu	1.55	0.8	0.0
His	1.55	2.1	10.0
Ile	1.15	0.5	0.0
Leu	1.30	0.9	14.9
Lys	1.75	0.4	2.1
Met	1.50	0.0	3.8
Phe	1.70	2.3	19.0
Pro	0.95	1.9	2.2
Ser	0.95	1.1	1.5
Thr	0.99	0.8	30.0
Trp	1.95	1.5	0.0
Tyr	1.90	1.8	4.2
Val	1.00	1.0	9.8

[a] r_{cut} is defined just below eq 4. [b] No Cys are present in our blind testing data set.

NMR structures, exceeding the 5% permitted, in principle, by our methodology. As a conclusion of the evidence observed for these structures and for the $\overline{z}_3$ values collected in the tables presented above, it can be said that the success of our rotamer library reconstructing side chain centroids from the X-ray backbones is clearly higher than from the NMR structures, having found no correlation with the size of the protein or with the predominant type of secondary structure. Anyway, in spite of some larger deviations and having taken into account the abundance of the different amino acids in proteins, the side chain centroid reconstruction procedure is also pretty good for NMR structures, with average values of $\overline{r}_3$ and $\overline{z}_3$ for the NMR proteins studied that lie, on average, just around 0.35 Å above the X-ray results.

■ SUMMARY AND CONCLUSIONS

In this work, we have developed a new backbone-dependent rotamer library for coarse-grained models that can be potentially used in the molecular modeling and simulation of the protein-folding problem. With this new library, we have tried to cover the gap between atomistic models and C^α coarse-grained models, describing in detail the building of a simple representation of a side chain as one single bead centered at the centroid of its real atoms. To keep the model as simple and computationally efficient as possible, the library only uses the coordinates of the α-carbons in the polypeptide chain because they are usually kept in most of the coarse-grained models to provide a geometric reference system easy to compute in every step of a simulation procedure. We have focused on the geometric characteristics of the side chain centroids, and therefore, we have not tried to define a volume or at least a radius under the assumption of a spherical shape for every rotamer. As a matter of fact, these properties would be related to an interaction potential and more specifically to its repulsive part, which would then control the packing of the side chains in the model. Although this is a very useful piece of information, it cannot be readily extracted from the pure geometric analysis we have carried out in this work. Moreover, different implementations of an interaction potential would then lead to different "sizes" of the coarse-grained side chains. Therefore, our library can accommodate different shapes or sizes of the represented side chains built on the centroids provided here.

We have tried to include in our library a number of rotamers for every residue that is large enough to rebuild as accurately as possible the position occupied by the side chain centroid from the C^α coordinates taken from PDB files. The library takes into account both the chemical identity of the considered residue and the local structure of the backbone around the residue in which it is located. This goes beyond other simple and efficient approaches previously described.[19,20] Actually, we have checked that the role of the backbone geometry is very important to determine the number of rotamers for a given residue, as well as the relative populations among them. In our simple model, the backbone conformation is taken into account through the distance between α-carbons $i-1$ and $i+1$ around residue i, whose side chain is being located. We have divided the possible values of this distance in three different intervals, according to the distribution function for this property that we have found in the structural database of folded proteins.

Even at the level of one single bead to represent the side chain, the rotamer library reproduces the internal flexibility of the different side chains. This is first shown by the presence of two different preferred distances from the α-carbon to the side chain centroid in residues with large and flexible side chains. The angular dependence of the side chain position with respect to the backbone, which defines the rotamers themselves, includes 95% of all the possibilities that we have found in our statistical analysis over a large training set of well-resolved protein structures. The definition of the rotamers has been difficult for several residues because the angular distributions are widely spread around a large orientational area, and therefore, any clustering procedure could severely limit the quality of the library results. In order to avoid this fact, we have used an iterative procedure, increasing the number of rotamers in the library so that the real side chain centroids whose clustering defines a given rotamer have a narrow distribution around their average. The resulting library is in this way larger than other previous versions. The number of rotamers defined for every residue type is collected in Table 8. As a comparison, we have also included in the last column of this table the number of rotamers from the library of Rainey and Goh.[20] These authors use the terminal atoms for the definition of relevant side chain positions, so the numbers in the last column of Table 8 are the sum over the different atoms for a given residue (for example, the 10 rotamers in Asn come from the sum of five cases for terminal atom $O^{\delta 1}$ and another five for atom $N^{\delta 2}$). In other cases where coarse-grained rotamer libraries have been defined with a certain backbone dependence[31] (based on the local secondary structure), the number of rotamers is limited to a maximum of three for a given residue type and backbone geometry. Thus, our number of rotamers for a given residue type is always larger than in libraries previously published. This is specially true for several residues with large internal flexibility in their side chains, which in our α-carbon reference system show highly spread orientations, as shown in Figure 4. Given the available large memory of present computers, our library can be perfectly loaded in the computer memory along the execution of a molecular simulation program employing a coarse-grained representation. Therefore, we have

Table 8. Number of Rotamers Defined in Our Library for Every Residue in the Different Backbone Intervals, $I^{(x)a}$

amino acid	$I^{(a)}$	$I^{(b)}$	$I^{(c)}$	results from ref 20
Ala	1	1	1	1
Arg	20	19	17	12
Asn	5	3	3	10
Asp	5	4	3	8
Cys	4	4	3	6
Gln	12	9	8	9
Glu	12	16	11	8
His	3	4	4	10
Ile	5	12	11	4
Leu	2	3	3	6
Lys	15	16	11	7
Met	9	12	11	4
Phe	3	3	3	3
Pro	2	2	1	4
Ser	3	4	3	3
Thr	2	3	3	6
Trp	7	8	6	10
Tyr	5	6	4	3
Val	2	3	4	6

[a]The number of residues in another coarse-grained rotamer library[20] is given in the last column as a comparison.

opted for a larger library that can better accommodate the different possibilities that could be found in the conformations appearing along the sampling procedure.

To verify that we have indeed built a complete enough rotamer library, we have checked that for any residue in the training database, we can rebuild all the side chain centroids with a divergence toward the experimental values that is below 0.5 Å, as an average over the residues in a given protein. It is also interesting to notice that worse results are obtained if the rotamer with the highest probability is used in this comparison (Figure 6). This is not surprising because the rebuilding procedure does not take into account the position of the residues in the core or the surface of the protein, two situations that can be rather different from the point of view of the available conformational freedom, as it has been already pointed out.[20,28] This is one of the reasons why the choice of the most probable rotamer is not necessarily the best one when trying to assess certain aspects about the quality of a rotamer library. The relatively large number of rotamers in our library is enough to properly locate the different experimental side chains in the native structures close enough to one of the possibilities included in the library, independent of the situation of the corresponding residues.

As an additional blind test, we have checked the library using a set of testing proteins different from those included in the training database. Again, the results presented here correspond to the locations of the best rotamers in the library for the reasons stated above. If the most probable rotamer were preferred for any reason, the results in Figure 6 show that the use of the most probable rotamer would yield deviations that are on average less than 1 Å larger than those corresponding to the best rotamer; this is a result that still can be considered as rather impressive, depending on the desired level of detail and the complexity of the rotamer library.

According to our limit of tolerance, for any side chain that is reconstructed with a maximum divergence of 30° from the PDB side chain centroid, we have proved that our library presents a high capability to rebuild in a very reasonable way the amino acid side chain centroids from the C^{α} coordinates alone, no matter the dominant secondary structure or the chain length for the protein considered. The results are very good for X-ray solved structures and slightly worse for those proteins whose structure has been solved by NMR spectroscopy. In these latter cases, however, the centroid positioning is still of a high quality for many of the residues, with average deviations that are more than reasonable at the level of coarse graining we have designed for our library.

Thus, we have proved that the new backbone-dependent rotamer library proposed in this work is perfectly able to reconstruct in a very efficient manner the side chain centroids in a coarse-grained representation of the different amino acids in a polypeptide chain and can therefore be a good tool to improve the coarse-grained models for protein folding.

The library is set as a couple of text files and is available from the authors upon request.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jsbach@quim.ucm.es.

Present Address

†M. Larriva: Dept. Farmacología y Toxiocología, Universidad de Navarra, E-31008 Pamplona, Spain.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (Grant No. FIS2009-13364-C02-02), and Comunidad Autónoma de Madrid (Grant No. S2009/PPQ-1551). M.L. acknowledges a former Scholarship from Spanish Ministerio de Educación (FPU program).

■ REFERENCES

(1) Creighton, T. E. *Proteins: Their Structure and Molecular Properties*; 2nd ed.; Freeman: New York, 1993.

(2) Kolinski, A.; Skolnick, J. Reduced models of proteins and their applications. *Polymer* **2004**, *45*, 511−214.

(3) Clementi, C. Coarse-grained models of protein folding: Toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(4) Chan, H. S.; Zhang, Z.; Wallin, S.; Liu, Z. Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models. *Annu. Rev. Phys. Chem.* **2011**, *62*, 301−326.

(5) Friesner, R. A., Ed.; *Computational Methods for Protein Folding*; Wiley: New York, 2002.

(6) Brown, S.; Fawzi, N. J.; Head-Gordon, T. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10712−10717.

(7) Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.

(8) Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166−171.

(9) Larriva, M.; de Sancho, D.; Rey, A. Evaluation of a mean field potential with different interaction centers. *Phys. A* **2006**, *371*, 449−462.

(10) Enciso, M.; Rey, A. Simple model for the simulation of peptide folding and aggregation with different sequences. *J. Chem. Phys.* **2012**, *136*, 215103 1−9.

(11) Zhou, H.; Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2002**, *11*, 2714−2726.

(12) Kazmierkiewicz, R.; Liwo, A.; Scheraga, H. A. Addition of side chains to a known backbone with defined side-chain centroids. *Biophys. Chem.* **2003**, *100*, 261−280.

(13) Oliveira, L. C.; Schug, A.; Onuchic, J. N. Geometrical features of the protein folding mechanism are a robust property of the energy landscape: A detailed investigation of several reduced models. *J. Phys. Chem. B* **2008**, *112*, 6131−6136.

(14) Fogolari, F.; Esposito, G.; Viglino, P.; Cattarinussi, D. Modeling of polypeptide chains as C alpha chains, C alpha chains with C beta, and C alpha chains with ellipsoidal lateral chains. *Biophys. J.* **1996**, *70*, 1183−1197.

(15) Kolinski, A.; Galazka, W.; Skolnick, J. On the origin of the cooperativity of protein folding: Implications from model simulations. *Proteins* **1996**, *26*, 271−287.

(16) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rachovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and oarameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997**, *18*, 849−873.

(17) Sun, W.; He, J. From isotropic to anisotropic side chain representations: Comparison of three models for residue contact estimation. *PLoS One* **2011**, *6*, e19238.

(18) Gopal, S. M.; Mukherjee, S.; Cheng, Y.-M.; Feig, M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins* **2010**, *78*, 1266−1281.

(19) Park, B.; Levitt, M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* **1996**, *258*, 367−392.

(20) Rainey, J. K.; Goh, M. C. Statistically based reduced representation of amino acid side chains. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 817−830.

(21) Chandrasekaran, R.; Ramachandran, G. N. Studies on the conformation of amino acids, XI. Analysis of the observed side group conformations in proteins. *Int. J. Protein Res.* **1970**, *2*, 223−233.

(22) Bahar, I.; Jernigan, R. Coordination geometry of non-bonded residues in globular proteins. *Fold. Des.* **1996**, *1*, 357−370.

(23) Philippopoulos, M.; Lim, C. Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-ray structures of *Escherichia coli* ribonuclease HI. *Proteins* **1999**, *36*, 87−110.

(24) Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56*, 249−265.

(25) Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* **1997**, *267*, 1268−1282.

(26) Al-Lazikani, B.; Jung, J.; Xiang, Z. X.; Honig, B. Protein structure prediction. *Curr. Opin. Chem. Biol.* **2001**, *5*, 51−56.

(27) Pokala, N.; Handel, T. M. Protein design: Where we were, where we are, where we're going. *J. Struct. Biol.* **2001**, *134*, 269−281.

(28) Dunbrack, R. L., Jr. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431−440.

(29) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins* **2000**, *40*, 389−408.

(30) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694−698.

(31) Keskin, O.; Bahar, I. Packing of sidechains in low resolution models for proteins. *Fold. Des.* **1998**, *3*, 469−479.

(32) Carr, J. M.; Wales, D. J. Global optimization and folding pathways of selected α-helical proteins. *J. Chem. Phys.* **2005**, *123*, 234901 1−12.

(33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliand, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(34) Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets. *Protein Sci.* **1992**, *1*, 409−417. Updated databases available at the PISCES Web site, http://dunbrack.fccc.edu/PISCES.php

(35) Pain, R. H., Ed.; *Mechanisms of Protein Folding*, 2nd ed.; Oxford University Press: Oxford, 2000.

(36) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Chari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **1991**, *253*, 657−661.

(37) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **1994**, *33*, 4721−4729.