Analysis of Dihedral Angles Distribution: The Doublets Distribution Determines Polypeptides Conformations

RON UNGER,[1] DAVID HAREL,[1] SCOT WHERLAND,[2] and JOEL L. SUSSMAN[3]*

[1]Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel, [2]Department of Chemistry, Washington State University, Pullman, Washington 99164-4630, USA, and [3]Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel

SYNOPSIS

It is possible to construct fragments of protein structures by using the known values for the fixed bond lengths, bond angles, and torsion angles, and "dialing" in the dihedral angles ϕ and ψ. By choosing these angles in different ways, it is possible to create different populations of fragments and to investigate their properties. We analyzed the following populations: *Real fragments* taken randomly from known structures. *Reconstructed fragments,* which are constructed, using the "fixed geometry" assumption, from a set of consecutive pairs of dihedral angles drawn from known structures. *Random fragments* that are constructed from a random set of dihedral angles from known structures, and *doublet-preserving fragments,* which are constructed from a set of dihedral angles drawn at random from known structures in a way such that the distribution of two consecutive pairs of dihedral angles in this population is similar to that distribution in the known structures. We examine the fixed geometry assumption and demonstrate that even reconstructed fragments contain many atomic collisions. We show that random fragments have only slightly more interatomic collisions than the reconstructed fragments. Nevertheless, the population of random fragments is structurally different from the population of reconstructed fragments. On the other hand, we show that the doublet-preserving fragments exhibit properties that are similar to the real population. Thus the doublet preserving random population can be used to simulate the structure of short polypeptides.

INTRODUCTION

In 1963, Ramachandran et al.[1] showed that the main-chain dihedral angles ϕ and ψ about the C_α atoms in proteins are restricted to specific domains of allowed values. Over the past 25 years, this observation has been confirmed in proteins and peptides whose x-ray structures have been determined. In addition, the x-ray structures of small peptides supported the "fixed geometry" assumption, namely, that bond lengths and bond angles along a polypeptide chain are kept fixed with very small tolerance. Thus, it became evident that, to a certain accuracy, the ϕ, ψ angles are sufficient to describe a protein's main-chain structure. We are interested in the extent to which the observed distribution of the dihedral angles actually *determine* the structure of proteins. In other words, does the local tendency of the dihedral angles to take on only specific values dictate the structure of longer fragments of proteins, or are there some higher order effects that determine the structure?

METHODS

Our structural data was taken from the Brookhaven Protein Data Bank, as released in January 1987. We used only those structures for which x-ray data had been collected to 3.0 Å or higher resolution, and which had been refined against the observed x-ray data to an R value of less than 30%. In order not to

CCC 0006-3525/90/5-60499-10 $04.00
Biopolymers, Vol. 30, 499–508 (1990)

* To whom correspondence should be addressed.

Table I Refined Brookhaven Data Base[a]

1APR	1BP2	1CC5	1CCR	1CPV	1CTF	1ECA	1FB4h
1FBJl	1FC2d	1FDX	1HIP	1HMQa	1INSa	1INSb	1LZ1
1PP2r	1PPD	1SBT	1SN3	2ACT	2ALP	2APP	2AZAb
2CAB	2CCYa	2CTS	2CYP	2ESTe	2FD1	2INSa	2LHB
2LZM	2PABa	2PKAa	2PKAb	2RHE	2SGA	2SNS	2SODo
351C	3C2C	3PGM	3PTP	3RP2a	3RXN	3SGBe	3TLN
4ADH	4APE	4ATCa	4ATCb	4CYTr	4DFRa	4FXN	4HHBb
4HHBc	4HHBd	4SBVa	5CPA	5RSA	5RXN	7CATa	

[a] The first four characters indicate the Brookhaven Protein Data Bank file name (release of January 1987), and the chain indicator, when required, is the fifth lower case character.

include trivially homologous proteins we retain only polypeptides that do not share identical dodecamer sequences. This left us with a set of 63 peptide chains (10,803 residues) (see list in Table I), which we called "the refined Brookhaven data base." Kabsch and Sander's program DSSP[2] was used to extract the data base of ϕ and ψ values from the coordinate files. The plot of these values is shown in Figure 1. A Pascal program was written to generate the main-chain atomic coordinates from a set of dihadrel angles, assuming fixed geometry (see Table II). For the distance between two structures s and t of length n (as represented by the C_α coordinate vectors r^s and r^t), we use the following definition, called the rms deviation distance: We first align the structures to the greatest possible extent using the BMF (best molecular fit) algorithm of Kabsch.[3,4] We then calculate the difference in the positions of the corresponding C_α atoms by

$$\mathrm{RMS} = \left[\frac{\sum_{i=1}^{n}(r_i^s - r_i^t)^2}{n-2}\right]^{1/2}$$

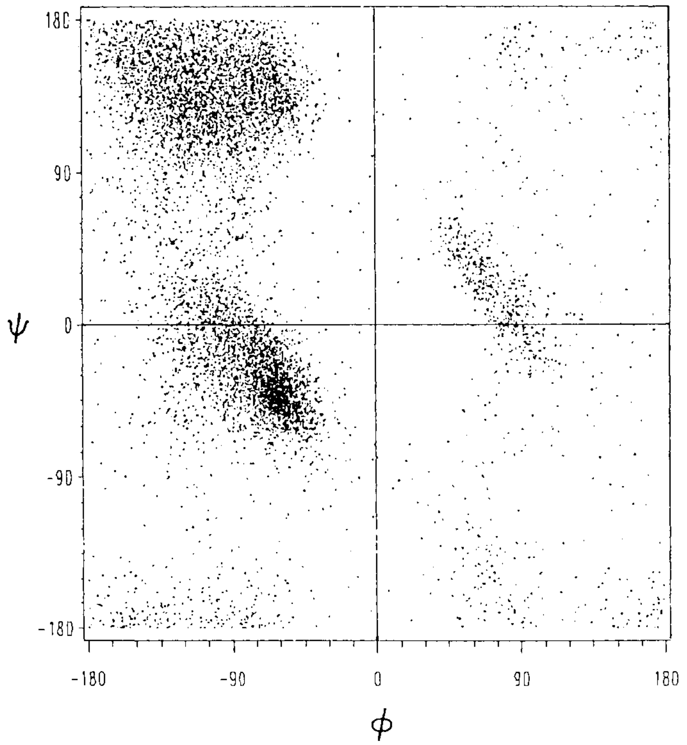

Figure 1. A plot showing the values of the dihedral values ϕ and ψ from the refined Brookhaven data base. All 10679 pairs of angles from a set of 63 proteins are shown.

Table II Fixed Geometry Values[a]

	Mean Value	SD	Fixed Value
Distance N-CA	1.47 Å	0.03 Å	1.47 Å
Distance CA-C	1.53 Å	0.04 Å	1.53 Å
Distance C-O	1.25 Å	0.03 Å	1.24 Å
Distance C-N	1.32 Å	0.04 Å	1.32 Å
Angle N-CA-C	110.82°	4.97°	110°
Angle CA-C-O	120.00°	4.29°	121°
Angle CA-C-N	116.10°	4.31°	114°
Angle O-C-N	123.53°	4.33°	125°
Angle C-N-CA	122.07°	5.02°	123°
Dihedral planner angle	179.63°	6.01°	180°

[a] The distances and angles of the *trans* peptide bond. The first and the second column show the mean values and standard deviations that were calculated over the refined Brookhaven data base. The last column shows the fixed values that were given by Pauling et al.[5] Note the good matching between these fixed values and the mean values observed in the refined Brookhaven data base. The standard deviations are relatively large, indicating that some proteins deviate significantly from the fixed geometry values.

RESULTS

Generating Valid Structures

In order to investigate the fixed geometry assumption, we first did a survey of the bond lengths and bond angles along the polypeptide chain as they are found in the 63 selected proteins (see Table II). Next, in order to evaluate the structural consequences of this assumption, the following experiment was done: Assuming fixed geometry values,[5] we chose a sequence of $N - 1$ original consecutive pairs of dihedral angles to *reconstruct* a fragment containing N residues, each consisting of the main-chain atoms N, O, C, and C_α. We then checked how many such fragments are *structurally valid,* and how many are not, due to collisions between atoms with distances less than the allowed van der Waals distances[6] (see Table III). A structure was invalidated if at least one pair of its atoms came too close. A sample of 3000 hexamers were tested, and 772 (25%) reconstructed structures were invalidated. The very high proportion of invalid reconstructed structures throws doubts on the fixed geometry assumption. A problem of a similar nature was described by Zwick[7] where creating the structure of myoglobin using fixed geometry and the original dihedral angles failed to reconstruct the original structure.

A careful analysis showed that for many of these structures, a few slightly nonplaner peptide bonds often accumulate to relieve collisions between atoms, especially for N and O atoms in helical and sharp turns structures.

In a similar way we created a random collection of protein main-chain fragments for which the distribution of dihedral angles was the same as their distribution in our refined Brookhaven data base. This collection was created by independently drawing at random pairs of ϕ, ψ values from the data base. For a set of 3000 such random structures, 660 (22%) were invalided. Thus there were about the same number of valid structures found among reconstructed fragments as in random structures.

As our aim was to investigate populations of valid fragments, we felt we should not invalidate so many reconstructed structures. Eliminating the invalidated structures, mostly helics and turns, would significantly bias our population. The energy minimization technique is one sophisticated way to weaken the fixed geometry constraints so as to eliminate most of the collisions. We decided to chose a simpler approach that maintain most of the structures by ignoring accidental collisions and forbid only major

Table III Minimal van der Waals Distances[a]

Interaction	Minimal Distance (Å)
O–O	2.6
O–N	2.6
O–C	2.7
N–N	2.6
N–C	2.8
C–C	2.9

[a] The minimal allowed distances between nonbonded atoms in a protein main chain.[6]

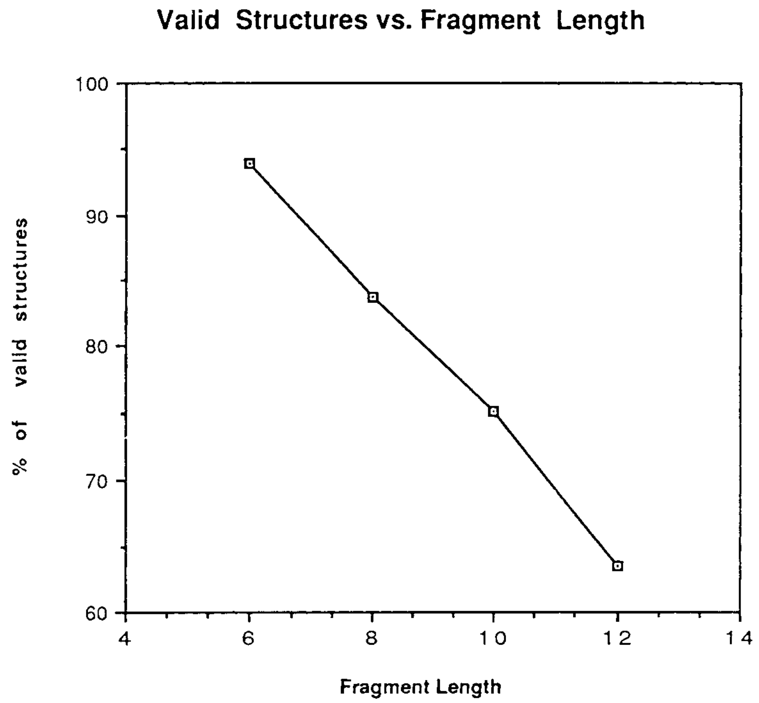

Figure 2. The percentage of valid random structures (i.e., structures in which no pair of the N, C, and C_α atoms are closer than the allowed van der Waals distances) as a function of the structure size. The random structures have the same distribution of dihedral angles as in the real data base. As the selection of consecutive ϕ, ψ angles is *not* dependent on the previous values, it is clear that longer structures have more chance of containing collisions.

collisions. Thus we decided to ignore collisions in which the carbonyl oxygen atoms were involved. As these oxygens are branching out from the main chain, collisions involving them may be just a side effect of the fixed geometry assumption. Only structures that have main-chain collisions, i.e., between N, C, and C_α atoms, are considered invalid. Repeating the same experiment with this new definition of valid structure, we found the following results: For a sample set of 3000 reconstructed hexamers, 113 structures (3.7%) were invalid, while for a set of 3000 random hexamers, 206 (6.7%) were invalid.

Clearly, longer fragments have more chance of containing collisions. The proportion of valid structures (according to the new definition) as a function of the lengths of fragments are shown in Figure 2.

One sees that a surprisingly high proportion of random fragments are valid, ranging from 93% for fragments of length 6 and decreasing to 63% for fragments of length 12. Thus, in a way, it is possible to create a set of structural fragments in this random manner. By inspecting the distribution of the number of residues apart between colliding atoms, a clear peak was found between atoms that are separated by six or seven residues (see Figure 3). This finding is in agreement with the observation[8] that most naturally occurring cyclic peptides are six or seven residues long. Another finding was that the invalidity of a structure was usually not due to a single forbidden interaction but, rather, to many such interactions.

We repeated the experiment, adding the C_β atoms to the constructed chain. This had only a marginal effect on the proportion of valid structures (about 3% less). We assume that adding full side-chain atoms will not force many additional collisions as the side chains have additional degrees of freedom that enable them to adopt to the available volume of space.

Population of Random Fragments

The population of random fragments constructed according to the procedure described above was further analyzed. Our procedure guaranteed that there was a similar distribution of the dihedral angles in the random population and in the population of real

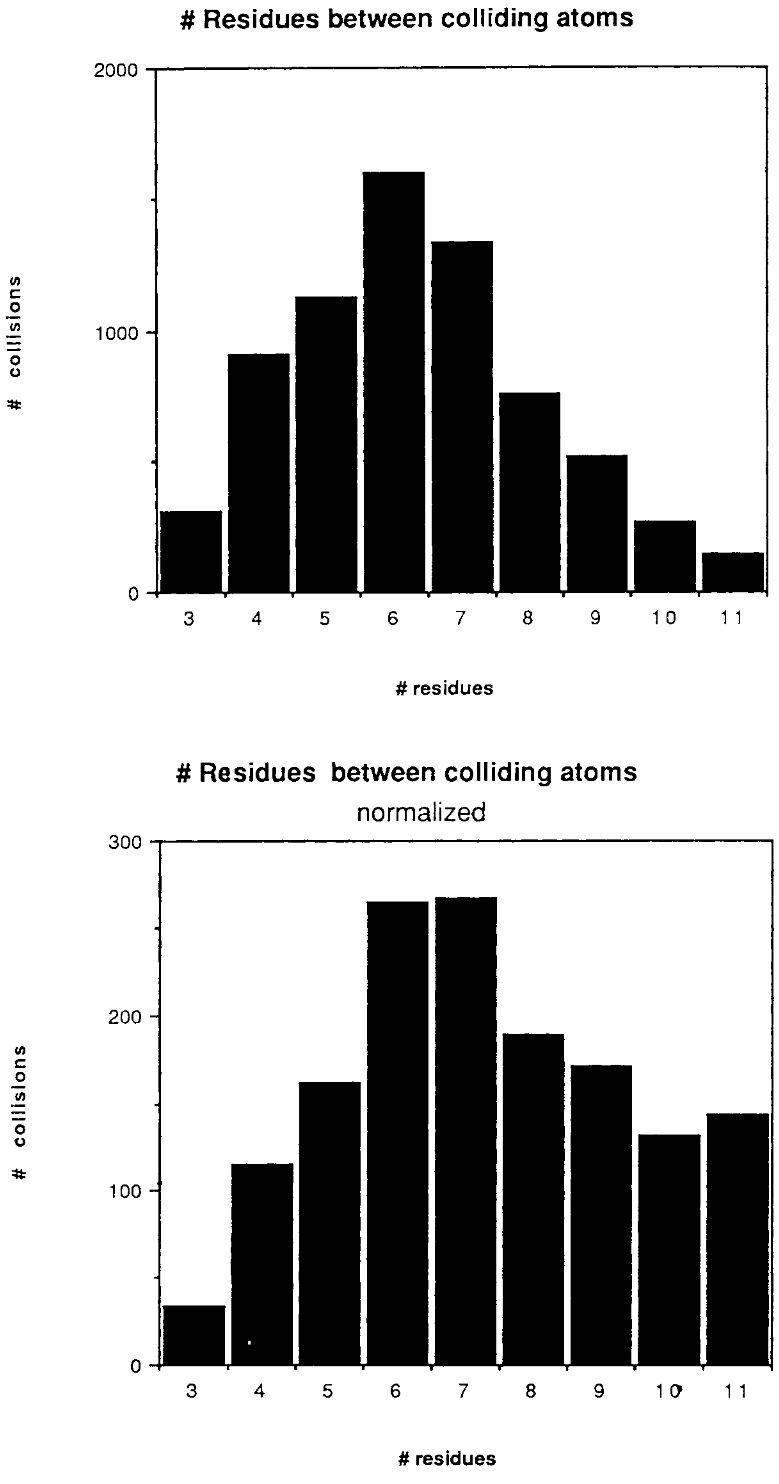

Figure 3. Histogram (a) shows, for the population of invalid structures of size 12, the distribution of the number of residues between atoms that collide. [For example, a collision between atom from residue 2 and atom from residue 8 is recorded in column 6 (8-2).] For each distance, there is a different number of possible interactions; for example, in a structure of length 12 there are 10 interactions of distance 2 (residues 1 and 3, 2 and 4, etc.) but only 1 interaction of distance 11 (residues 1 and 12). Thus, in histogram (b) we show the number of collisions normalized by the number of interactions. In both histograms there is a peak showing a tendency of atoms, that are 6–7 residues apart, to collide.

fragments found in proteins. We wanted to compare the two populations to see whether the similarity on the level of the dihedral angles implied a similarity of the structures in the two populations. It is not clear a priori how to measure the similarity between two populations of structures. If the structures were three-dimensional points distributed in a cube, then we would split the cube into many small cubes, comparing the number of points from each population that lie in each of the small cubes. Integrating these differences over all the small cubes would indicate the average similarity between the two populations. Since the structural space is multidimensional, we decided to generalize this idea.

Using the spirit of the cube approach, we applied two practical tests. In the first, we counted, for each structure in one population, the number of similar structures it has in the two populations, and compared them. For example, if, for a given structure s, there are $n1$ structures in population $p1$ with rms distance from it that is less than some threshold value, and in the $p2$ population there are $n2$ such structures, then we say that s has a difference of $|n1 - n2|$ in the two populations. The average of these differences over all structures in the population was used as the distance between the two populations. Obviously, if the structures in the two populations are distributed similarly, then their difference will be small. For distributions with a high difference, many of the structures will have a significantly different number of neighbors in the two populations.

We created two sets of 1000 random hexamers (fragments of six residues) and compared them to two sets of 1000 real hexamers taken from the refined Brookhaven data base. The "radii" of the neighborhood around each structure was set to 1 Å rms. The results (see Table IV) show that the two populations are significantly different. While the distances between two sets from the same population were small (in the range of 3–5), the distances between sets from different populations were much higher (20–25). That the distances between sets from the same population were small is a result of their common distribution; this is true even for the two random populations, whose small distance (3.9) demonstrates the robustness of this test. On the other hand, the large distance between sets from different populations testifies to their different distributions.

Our second test is illustrated in Figure 4. In earlier work[9] we found that most hexamers occurring in protein structures can be represented by a relatively small master set of typical hexamers, which we called *building blocks*. Each one of our test sets was *mapped onto* this master set of building blocks in the following way: Each hexamer from the test set is matched to its most similar (in rms distance) building block. The distribution pattern of the hexamers onto the master set (i.e., the number of hexamers each building block was matched with) turns out to be very different for random and real hexamers. Some very rare real hexamers occur very frequently in the set of random hexamers, while the most frequent naturally occurring hexamers are much less dominant in the random populations. The specific selection of the master set of hexamers is not crucial in our case, and similar results were obtained for different ones.

Table IV Distances Between Hexamer Populations[a]

	RN1	RN2	CN1	CN2	PN1	PN2
RN1	■	3.9	21.0	23.9	25.5	26.3
RN2		■	20.4	22.8	24.9	25.7
CN1			■	4.5	6.9	7.7
CN2				■	5.3	6.1
PN1					■	4.1
PN2						■

[a] Two sets of 1000 hexamers were randomly chosen from each of the three types of populations: RN1 and RN2 from a population of random hexamers that maintains the same overall ϕ, ψ distribution as in the data base; CN1 and CN2 from a population of real hexamers; and PN1 and PN2 from a population that maintains the distribution of doublets (two consecutive pairs) of dihedral angles. The distance between two sets is the averaged difference in the number of neighbors, of each fragment, from the two populations in a radius of 1 Å (see text). The distances between sets from the same population is much smaller than the distances between sets from different sets. The distances between the doublet-preserving sets and the real sets are much smaller than the distances between the former and the sets that only maintain the single dihedral distribution.

Population of Random Doublets of Dihedral Angles

We showed that the population of random hexamers was very different from the population of real hexamers, although the distribution of their dihedral angles was the same. We next constructed a set of random hexamers whose distribution of *doublets* of dihedral angles was similar to that distribution in the real structures. We defined a set of dihedral angles to be *doublet-preserving* set if the chance of getting any specific consecutive pairs of (ϕ_1, ψ_1), (ϕ_2, ψ_2) is similar to the chance of getting it in a real set

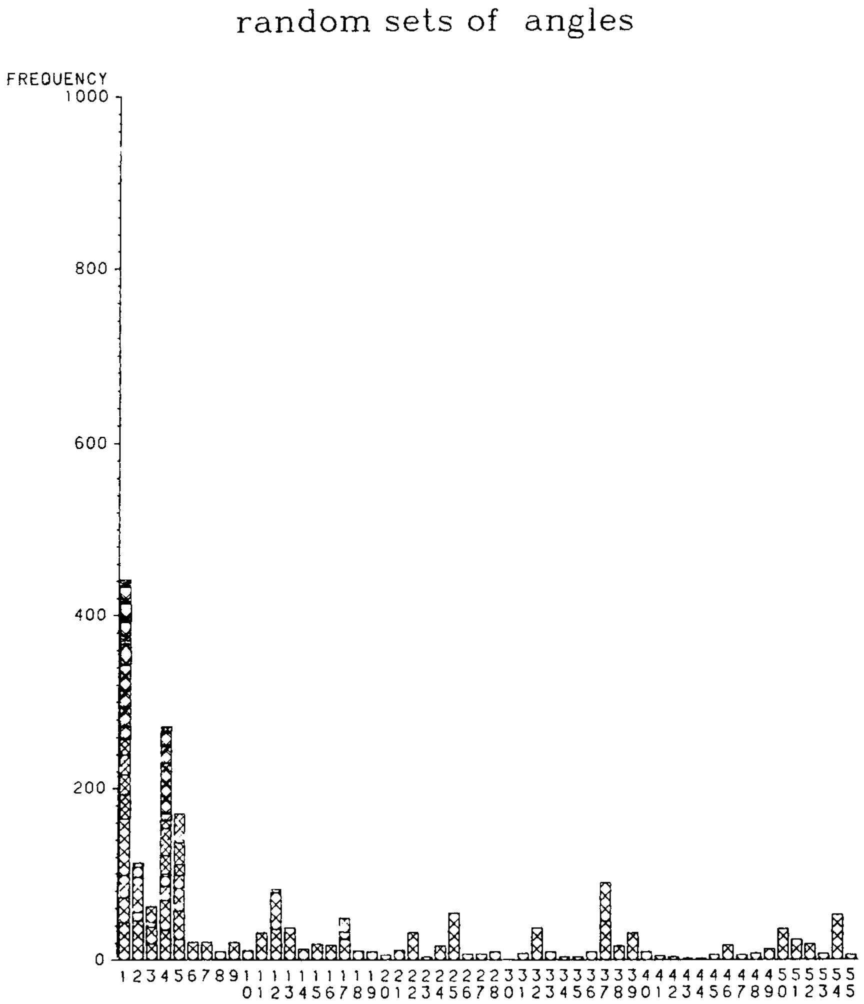

Figure 4. Distribution of populations into building blocks. The mapping of sets of 2000 hexamers to a master set of typical building blocks.[9] Fifty-five shapes are listed along the horizontal axis; the subdivision into specific building blocks within each shape is represented by different texture of bin in the histogram columns. Each hexamer is matched with its most similar building blocks. The distribution of the hexamers into building blocks is shown for three types of sets: (a) set of random hexamers with the same dihedral angle distribution as the real data base; (b) set of real hexamers; and (c) set of doublet-preserving hexamers. It can be seen that the pattern for the single random hexamers (a) is very different than the pattern for the real hexamers (b). The pattern for the set that maintain the doublet distribution (c) is almost indistinguishable from the set of real hexamers.

of dihedral angles. Technically, the doublet-preserving sets of dihedral angles were chosen in the following way: We randomly selected consecutive pairs of ϕ, ψ values and concatenated them, by overlapping the second pair in the first selection with the first pair in the next selection, but only if their values are similar enough. We considered a distance of 15° [that is, $\sqrt{(\phi_1 - \phi_2)^2 + (\psi_1 - \psi_2)^2} < 15$] to be close. We then constructed two sets of 1000 doublet-preserving hexamers and matched them with the other sets of hexamers, using the two tests discussed in the previous section. In the first test, in which the average difference in the number of neighbors was calculated, we found that the doublet-preserving population had a distances ranging from 6 to 8 from real hexamer population and a distances ranging from 22 to 26 from the random sets maintaining the single dihedral distribution. In the second test, the distribution of the mapping into the set of typical hexamers was similar to the distri-

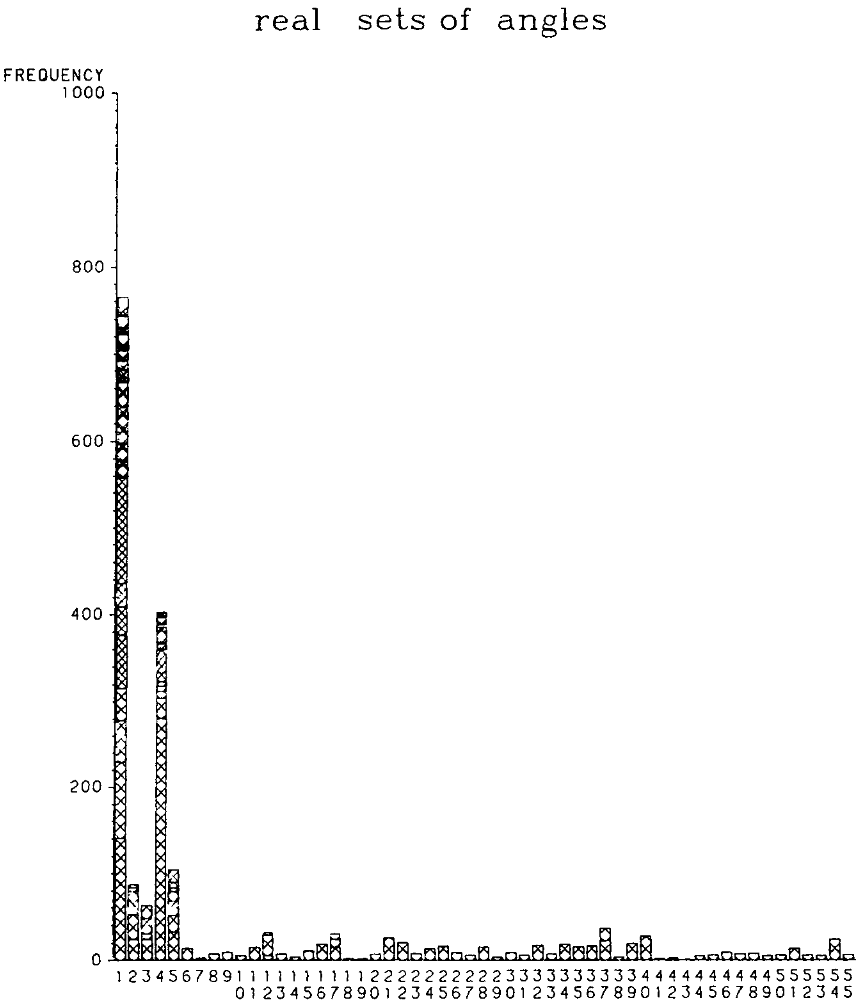

Figure 4. *(Continued from the previous page.)*

bution of real hexamers and very different from the distribution of the hexamers maintaining the single dihedral values. These results, summarized in Table IV and Figure 4, indicate that the doublet-preserving populations are much more related to the sets of real hexamers than to the sets that only maintain the single dihedral angles distribution. The similarity between real and doublet-preserving structures is kept, to a lesser extent, even for longer fragments. We applied the "counting neighbors" test for dodecamers (fragments of 12 amino acids) using a radius of 2 Å (see Table V). The small distances between sets from the same population show that the integrity of each population is maintained. The random population is different from the real population (11.2–14.2) and from the doublet-preserving population (17.2–21.6), while the distance between the real and the doublet-preserving population is smaller (8.1–10.5).

DISCUSSION

We found that using the fixed geometry assumption and original observed set of dihedral angles, many of the reconstructed structures contains collisions. This indicate that using this assumption is not appropriate for high-quality "reproduction" of even short protein fragments. On the population level, we bypass this problem by ignoring oxygen collisions. We have shown that it is possible, by using random sampling of existing dihedral angles, to construct a collection of random sets of short structures that are structurally valid and contain no other collisions. This finding strengthens Ramachandran's claim that the main source of colliding atoms are adjacent residues in the polypeptide chain, and that the values of consecutive pairs of dihedral angles are not significantly restrained due to short van der Waals distances between atoms.

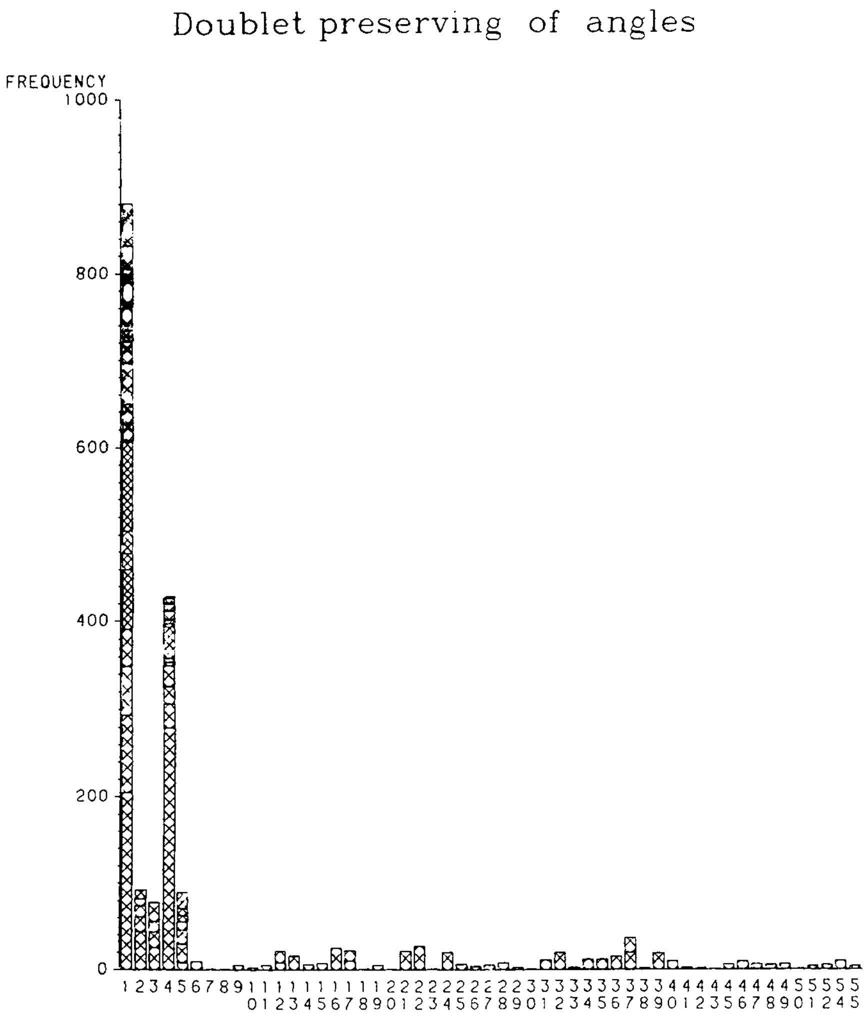

Figure 4. (*Continued from the previous page.*)

Table V Distances Between Dodecamer Populations[a]

	RN1	RN2	CN1	CN2	PN1	PN2
RN1	■	1.1	11.2	12.1	21.3	17.2
RN2		■	14.2	12.6	21.6	17.2
CN1			■	2.2	10.4	8.9
CN2				■	10.5	8.1
PN1					■	6.2
PN2						■

[a] Two sets of 1000 dodecamers were randomly chosen from each of the three types of populations: RN1 and RN2 from random population that maintains the overall real ϕ, ψ distribution in the data base; CN1 and CN2 from a population of real hexamers; and PN1 and PN2 from a population that maintains the distribution of doublet dihedral angles. For dodecamers, a radius of 2 Å was used. (See text and Table IV.) Again the distances between sets from the same population is smaller than the distances between sets from different sets. The distances between the doublet-preserving sets and the real sets are smaller than the distances between the former and the sets that only maintain the single dihedral distribution.

Nevertheless, the population of structures that were produced in this way is significantly different from the population of real structures, indicating that the distribution of single dihedral angles does not dictate structure. This finding implies that the values of consecutive pairs of ϕ, ψ are dependent. This dependency is apparently not used to avoid collisions, but more likely, reflects the tendency of protein fragments to be organized in structural motifs.

When we restricted ourselves to random structures that maintain the doublet dihedral angles distribution we found that, on the population level, these structures are almost indistinguishable from sets of real structures, and are very different from the population of single random fragments. We conclude that the distribution of doublets of dihedral values reflects most of the other parameters that determine the structure of short fragments of protein. Thus, the population of structures that were built in this way is very similar to the real population of structures.

It is clear that this approach of analyzing of structure can only be applied to relatively short structures, which explains the difference in the results between hexamers and dodecamers. When the fragment becomes too long it is exposed to many other global constraints, like solvent-exposed areas, sheet complementarity, disulfate bridges, etc., which are beyond the scope of this analysis. We feel a range of 6–12 residues seems to be short enough, on the population level, not to be affected by these global constraints. This leads us to propose that producing doublet-preserving structures is a simple way of generating random, short, protein-like structures. This method can supply a useful simulation tool in analyzing polypeptide structure where random protein-like fragments are needed.

REFERENCES

1. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963) *J. Mol. Biol.* **7,** 95–99.
2. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
3. Kabsch, W. (1976) *Acta Cryst. B* **32,** 922–923.
4. Kabsch, W. (1978) *Acta Cryst. A* **34,** 828–829.
5. Pauling, L., Corey, R. B. & Branson, H. R. (1951) *Proc. Natl. Acad. Sci.* **37,** 205–211.
6. Ramachandran, G. N. & Sasisekharan, V. (1968) *Adv. Protein Chem.* **23,** 283–437.
7. Zwick, M. (1968) Ph.D. thesis, MIT.
8. Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985) *Adv. Protein Chem.* **37,** 1–109.
9. Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989) *Proteins* **5,** 355–373.

Received November 21, 1989
Accepted January 10, 1990