Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Conformation families of protein fragments in multidimensional torsion-angle space

Frantisek Pavelcik[a,b]* and Pamela Pavelcikova[c]

[a]Department of Chemical Drugs, Faculty of Pharmacy, University of Veterinary and Pharmaceutical Sciences in Brno, 612 42 Brno, Czech Republic, [b]Department of Inorganic Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, 842 15 Bratislava, Slovak Republic, and [c]Institute of Molecular Biology, Slovak Academy of Sciences, 845 51 Bratislava, Slovak Republic

Correspondence e-mail: pavelcikf@vfu.cz, pavelcik@fns.uniba.sk

Received 11 April 2007
Accepted 18 September 2007

Protein conformation families for automatic model building were determined for dipeptidic, tripeptidic, tetrapeptidic and pentapeptidic fragments. Mapping in n-dimensional conformational space (n = 2, 4 and 6), a conformation-generator method, a deletion-sorting process and a verification procedure were used to calculate the conformational preferences. Torsion angles were harvested from PDB structures with resolutions better than 1.5 Å. Statistical weights were calculated for the conformation families.

1. Introduction

Protein conformations are mainly classified on the basis of hydrogen-bonding patterns and the two-dimensional Ramachandran plot (Ramachandran & Sasisekharan, 1963). Limited information is available for higher dimensional spaces (Pavelcik & Vanco, 2006; Sims *et al.*, 2005).

The properties of a conformation space for polypeptides have emerged in a step-by-step fashion during the history of protein structure, starting with secondary structures (Pauling & Corey, 1951; Pauling *et al.*, 1951) and continuing with the Ramachandran map (Ramachandran & Sasisekharan, 1963), the classification of turns (loops; Venkatachalam, 1968), hydrogen-bonding patterns (Kabsch & Sander, 1983) and many other contributions. A detailed analysis of two-dimensional space has recently been published (Hovmöller *et al.*, 2002). Results for tripeptides in four-dimensional torsion-angle space can be found in Pavelcik & Vanco (2006). Sims *et al.* (2005) have analyzed higher dimensional $(\varphi, \psi)_n$ maps. Useful information can be found on the PDBsum web pages (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/). A comprehensive classification of small motifs is available at http://www.ebi.ac.uk/msd-srv/msdmotif. The study of the similarity of protein fragments by cluster analysis is a closely related topic. In this approach, different descriptors, such as the positions of C^{α} atoms (see, for example, Micheletti *et al.*, 2000; Kolodny *et al.*, 2002), are used rather than torsion angles.

The allowed and disallowed regions of the Ramachadran map are similar to the rotameric nature of saturated organic molecules, protein side chains (Dunbrack & Karplus, 1994) and nucleic acids (Murray *et al.*, 2003). Generalization of these concepts leads to the notion of conformation families. The α_R, β_1, β_2 and α_L regions of the Ramachandran map can be regarded as examples of conformation families in protein structures. The concept of a conformation family is useful for the classification of macromolecular structures, structure verification and model building (both theoretical and experimental). It is believed that protein conformation space is

highly correlated (Micheletti *et al.*, 2000) and that protein structure can be reconstructed using a limited number of low-dimensional structural fragments (Jones & Thirup, 1986).

A simple method (Pavelcik & Vanco, 2006) to search for and to identify conformational families has recently been developed and this work is extended in scope within this article. A conformation family, within the concept of Pavelcik & Vanco (2006), is defined as a region of conformational space that is highly populated by experimentally observed conformations. A smoothed function of conformation density in this region has a local or global maximum. Conformation space is an infinite periodic torsion-angle space. A square $360 \times 360^\circ$ (or 1×1 in cycles) in two dimensions is a unit cell that resembles a crystallographic unit cell. If a molecular fragment is symmetrical, then even a plane (space) group and asymmetric unit can be assigned (however, with the exception of polyglycine, this is not the case for proteins). The principal idea of the method is that it is practically impossible to view multidimensional surfaces but it is relatively easy to identify maxima (or other critical points) on such surfaces. The distance between two conformations in multidimensional torsion-angle space is a second criterion for definition of the conformation family. The distance between two conformation families should not be shorter than a certain minimum (Pavelcik & Vanco, 2006).

While in the previous paper (Pavelcik & Vanco, 2006) we concentrated on method development, in this paper a comprehensive analysis of conformation families was carried out using high-resolution structures from the Protein Data Bank (Bernstein *et al.*, 1977; Berman *et al.*, 2000). Our calculations identified conformation families in protein structures and their approximate statistical weights.

The primary rationale for these calculations is to derive accurate conformation families for automatic interpretation of electron density in X-ray crystallography (Pavelcik *et al.*, 2002; Pavelcik, 2003, 2004, 2006*a*), particularly at lower resolutions. The top conformation families are used directly as search fragments in the phased rotation-conformation and translation function (PRCTF; Pavelcik, 2006*b*). The statistical weights may be useful in more sophisticated model-building methods. Less populated families can be used to extend regions of secondary structure.

2. Methods

The method described by Pavelcik & Vanco (2006) is based on multidimensional conformation maps (*n*-dimensional maps), generators of conformation families and a verification and averaging procedure. Only minor changes have been introduced into the method and into the computer program since the work described in Pavelcik & Vanco (2006).

2.1. Mapping

A grid is constructed in multidimensional space. A spherical search probe is positioned at each grid point and the number of experimental points that are within the sphere are considered as the conformation density assigned to that grid point. Peak-picking is used to locate maxima on the conformation density surface. No principal changes were made compared with the description in Pavelcik & Vanco (2006). Faster code was developed. A larger amount of experimental data has allowed us to reduce the radius of the search probe and to thereby obtain better resolution of the conformation families.

2.2. Verification and averaging

The radius of averaging is variable and is defined as

$$R = R_{\rm D}(0.5N)^{1/2},$$

where N is a dimension of the conformation search and $R_{\rm D}$ is the empirically found radius for the two-dimensional search. This increases the *n*-dimensional volume of the space covered, without the danger of obtaining overlapping families. The formula is empirical, but is related to a length of a body diagonal in a multidimensional cube ($d = aN^{1/2}$), and may help to scale distances in spaces of different dimensionality.

2.3. Sorting and deleting

This procedure was developed as an independent step in the algorithm. Two conformation families separated by a short distance are merged into one family by deleting the less populated family. The deletion distance is a function of the search dimension and is defined as 35° for two and three dimensions, 37° for four dimensions and 40° for six and eight dimensions. A deletion limit cannot be increased more greatly in proportion to the number of degrees of freedom because two conformation families might differ only in one torsion angle. The deletion limit was introduced as a tool for removing false maxima in high-density regions which can be caused by fluctuations in the conformation density arising from noise. The importance of the deletion limit is reduced if the number of experimental points is high and can be reduced or eliminated completely for smooth surfaces.

2.4. Generation of conformation families

The generator method, which was only mentioned in Pavelcik & Vanco (2006), is further tested and used in this work. Using this tool, multidimensional polypeptide conformation families can be generated, for example, as combinations of two-dimensional conformation families. Subspaces are analyzed by direct mapping and conformation families are determined. These low-dimensional conformation families are called generators. In principle, generators can come from different sources, such as theoretical calculation, stereochemical rules or the literature [for example, one generator could be the pair of torsion angles (-64, -41°) and another generator could be the quartet (-119, 138, -115, 135°) to generate the conformation family in six dimensions (-119, 138, -115, 135, -64, -41°) or (-64, -41, -119, 138, -115, 135°); we can use two triplets or three doublets for the same purpose]. From five two-dimensional generators, we can generate 25 conformation families in four dimensions or 125 families in six dimensions. The generated conformation family

Table 1
Conformation map for dipeptidic (two dimensions, AlphaD) fragment.

The number of experimental conformations is 466 713. Grid NG = 16. Search probe $R_1 = 30^\circ$. Verification radius $R_2 = 50.0^\circ$. Noise level 5. Deletion limit 35°. N is the number of experimental conformations in each family. P is the percentage probability. φ and ψ are averaged main-chain torsion angles for each family.

No.	N	P (%)	φ (°)	ψ (°)	Type
1	209063	44.79	−67	−35	A
2	108908	23.34	−122	136	B
3	79397	17.01	−72	142	C
4	20919	4.48	72	21	G
5	3530	0.76	94	178	X
6	2646	0.57	63	−136	E

is evaluated by the number of experimental conformations that are within a multidimensional sphere constructed around a point belonging to the generated conformation family. The radius of the sphere is approximately the average of the probe radius used in mapping and the verification radius. Low-population conformation families are deleted. If two generated families are too close (separated by a short multidimensional distance) the less populated family is deleted. Surviving conformations are optimized.

2.5. Maximization (optimization) procedure

The generator usually represents a local maximum in a torsion-angle subspace. However, the combined conformation may not be the local maximum in n-dimensional torsion-angle space. For example, it may be a multidimensional ridge on a slope close to the local maximum. A simple procedure was designed to move a generated conformation towards the maximum. This procedure can eventually lead to merging of some close conformations. The procedure is based on repeated averaging within a sphere defined by a maximization (optimization) radius. The averaged conformation defines a new position for a conformation. In a sphere with an asymmetric distribution about the centre, the averaging moves the conformation to a more populated region (towards the maximum). The radius of averaging should be sufficiently small to prevent undesirable merging and to prevent escape from a local maximum (particularly to a nearby dominant conformation type). The averaging radius was selected to be equal to the mapping radius. The process is iterative. Cycles of averaging are mixed with cycles of deletion. Final families are verified again and averaged within the verification radius in the same way as used for mapping.

3. Results and discussion

Calculations were carried out for dipeptidic, tripeptidic, tetrapeptidic and pentapeptidic fragments. These are related to model-building blocks AlphaD, AlphaT, AlphaQ and AlphaP, respectively (Pavelcik, 2006*b*). The chemical connectivity of these fragments can be formulated as AlphaD, C^α–CO–Ala–N–C^α; AlphaT, C^α–CO–Ala–Ala–N–C^α; AlphaQ, C^α–CO–Ala–Ala–Ala–N–C^α; AlphaP, C^α–CO–Ala–Ala–Ala–Ala–N–C^α.

Table 2
Conformation map for tripeptidic (four dimensions, AlphaT) fragment.

The number of experimental conformations is 458 741. Grid NG = 16. Search probe $R_1 = 30^\circ$. Verification radius $R_2 = 70.7^\circ$. Noise level 5. Deletion limit 37.0°. N is the number of experimental conformations in each family. P is the percentage probability. φ and ψ are averaged main-chain torsion angles for each family. The cumulative probability for the families listed is <95%.

No.	N	P (%)	φ_1 (°)	ψ_1 (°)	φ_2 (°)	ψ_2 (°)	Type
1	164850	35.94	−64	−37	−68	−34	AA
2	81392	17.74	−119	138	−115	135	BB
3	42265	9.21	−82	142	−76	142	CC
4	25841	5.63	−75	140	−73	−28	CA
5	17586	3.83	−95	−5	−78	139	DC
6	14521	3.17	−92	−21	−143	153	DB
7	13309	2.90	79	13	−90	144	GC
8	12821	2.79	−77	−21	−119	53	AZ
9	12404	2.70	−132	150	−73	−24	BA
10	11684	2.55	−88	93	−139	156	ZB
11	10219	2.23	−93	−3	74	22	DG
12	5328	1.16	−60	136	80	4	CG
13	4136	0.90	72	20	−83	−22	GD
14	3206	0.70	−123	138	57	40	BG
15	2519	0.55	−93	−9	88	−173	DX
16	2195	0.48	56	36	75	11	GG
17	2052	0.45	102	−173	−76	141	XC
18	1633	0.36	59	−131	−85	−2	ED
19	1257	0.27	−139	156	99	−174	BX
20	1200	0.26	−127	127	61	−130	BE
21	1201	0.26	95	173	−70	−27	XA
22	1150	0.25	89	−166	−135	150	EB
23	1100	0.24	−156	179	152	119	??
24	1071	0.23	−77	146	102	−169	CX
25	366	0.08	87	−170	−99	79	XZ
26	201	0.04	106	10	61	42	DG
27	127	0.03	73	−155	91	−167	EE
28	114	0.02	88	−176	61	36	XG
29	91	0.02	92	163	73	−162	XE

The fragments reflect the fact that because of its planarity the peptide group is a better basic building block for protein modelling than an amino-acid residue. One has to realise that, for example, a tripeptidic AlphaT contains two pairs of (φ, ψ) and represents an extended dipeptide, while the total number of atoms is more compatible with a tripeptide. The conformational flexibility of AlphaT can be expressed as $(\varphi, \psi)_2$ (Sims *et al.*, 2005).

Protein structures were filtered from the Protein Data Bank based on 90% sequence similarity (mainly to remove equivalent structures and mutants) using the PDB server advanced search (http://www.rcsb.org/pdb/search/advSearch.do), a minimal chain length of eight residues (to remove synthetic polypeptides) and resolutions of 0.5–1.5 Å (February 2007, 1471 files). PDB files containing DNA or RNA were removed (mainly for computational reasons). The codes of the PDB structures can be found in Table 1 of the supplementary material[1]. No further analysis was performed with respect to R factors or the numbers of duplicate chains in each structure. The bias introduced by multiple copies of the same protein in

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: SX5071). Services for accessing this material are given at the back of the journal.

Table 3
Conformation map for tetrapeptidic (six dimenions, AlphaQ) fragment.

The number of experimental conformations is 451 218. Grid NG = 16. Search probe $R_1 = 30^\circ$. Verification radius $R_2 = 86.6^\circ$. Noise level 5. Deletion limit 40°. N is the number of experimental conformations in each family. P is the percentage probability. φ and ψ are averaged main-chain torsion angles for each family.

No.	N	P (%)	φ_1 (°)	ψ_1 (°)	φ_2 (°)	ψ_2 (°)	φ_3 (°)	ψ_3 (°)	Type
1	133602	29.61	−64	−39	−65	−39	−68	−36	AAA
2	45415	10.06	−118	139	−116	136	−114	135	BBB
3	18182	4.03	−85	143	−81	142	−79	141	CCC
4	15009	3.33	−86	142	−61	−32	−71	−25	CAA
5	12304	2.73	−96	140	−81	143	−70	−29	CCA
6	10154	2.25	−68	−21	−93	−5	−83	140	ADC
7	8157	1.81	−67	−29	−79	−19	−121	47	CDZ
8	8048	1.78	−68	−30	−92	−2	75	22	ADG
9	7785	1.73	−94	−4	−85	143	−82	140	DCC
10	6982	1.55	−94	−22	−143	151	−121	138	DBB
11	6654	1.47	−84	134	−91	−26	−148	151	CDB
12	6566	1.46	−92	104	−140	155	−76	142	ZBC
13	6436	1.43	−93	−3	77	21	−91	143	DGC
14	6024	1.34	78	13	−93	145	−88	137	GCC
15	5630	1.25	−94	−5	−78	137	−75	−25	DCA
16	4148	0.92	−80	141	−90	95	−138	157	CZB
17	3814	0.85	−80	−21	−121	59	−66	−27	DZA
18	3477	0.77	−94	39	−97	−46	−64	−36	ZDA
19	3003	0.67	−140	153	−135	147	−157	−168	BB?
20	2915	0.65	−140	153	−142	−177	−67	143	B?C
21	2911	0.65	−92	−8	−111	−50	−109	−9	D?D
22	2856	0.63	−89	−20	−143	159	−75	−24	DBA
23	2755	0.61	−88	144	−60	137	81	3	CCG
24	2683	0.59	−71	140	−160	170	−127	177	C??
25	2683	0.59	−78	134	−97	−4	−80	137	CDC
26	2653	0.59	−58	134	84	−2	−87	147	CGC
27	2620	0.58	−165	174	−164	169	−143	152	??B
28	2540	0.56	−119	169	−84	−1	−92	−7	BDD

the asymmetric unit (NCS) is probably minimal with this amount of data. Disordered or incomplete residues were removed from the calculation of torsion angles. The number of torsion angles harvested from the PDB was almost half a million.

All conformation maps were calculated with a grid NG = 16 (this is equivalent to 360/16 = 22.5°). The noise level of the conformation density map was estimated to be five experimental conformations per mapping sphere belonging to the grid point. In the text, all angles and all distances (in angular space) between conformations are given in degrees. P is the probability of appearance of the conformation and is reported as a percentage.

The following letters are used to classify two-dimensional conformations and typical values (φ, ψ) are given in parentheses. A is α-helix (−64, −41), B is β1 (−121, 128), C is β2 (−66, 137), D is δ (−90, 0), G is γ (50, 25), Z is a bridge region (−100, 70), X is (90, 180), E is (60, −140) and '?' is an unspecified conformation (mainly extended glycine conformations).

Two-dimensional and four-dimensional maps, which were also evaluated in the previous study (Pavelcik & Vanco, 2006), were recalculated with the additional experimental data (a 20-fold increase). Full results are given in Tables 1 and 2.

No new conformation family appeared in the new dipeptidic map compared with the original analysis (Pavelcik & Vanco, 2006). There are only small changes in the populations of individual conformations. There are six conformational families in two dimensions, with 9% of the total observations being in unspecified conformations using spherical conformation regions (given by the verification radius) in the current method. In principle, a more detailed topological analysis specifying boundaries of the conformation families can be performed with the near-half-million torsion angles available.

From a stereochemical point of view, a maximum in conformation density may contain several close but overlapping conformation types. These types depend upon the conformations of neighbouring residues and on their structural role (*e.g.* helix or loop). To distinguish conformation families from these subtypes, we will call the latter 'conformational primitives'.

More tripeptidic conformations (φ_1, ψ_1, φ_2, ψ_2) = (φ, ψ)$_2$ were generated than in the previous study (Pavelcik & Vanco, 2006), but all of the 17 conformation families calculated in Pavelcik & Vanco (2006) are among the top families. The most important change is that the β region is split into two independent conformation families: BB and CC. Two new conformations, (−77, −21, −119, 53), P = 2.8, AZ type, and (−88, 93, −139, 156), P = 2.6, ZB type, are related to a bridge region (Hovmöller *et al.*, 2002). Z (inverted classical γ-turn) is not an independent conformation family in the two-dimensional map. Among the top 20 ($P > 0.26$) conformations, these are the only additional families that emerge from the new analysis.

Four new conformation families ($0.23 < P < 0.26$) emerged from the noise region. There is a sharp drop in conformation density beyond conformation 24. The absolute number of conformations could be estimated as 24–25, with 5% of the remaining conformations regarded as unspecified and in a noise region. There may still be some minor conformation families in the noise region. Several conformation families belonging to β-turns clearly showed (−90, 0°) as an independent conformation primitive, which is overlapped in two dimensions by a dominant α_R region. The overlap of the α_R region with the β-turn type I is reflected in a slightly reduced ψ angle for conformation 1 compared with 'pure' α-helix (−64, −41°) (Hovmöller *et al.*, 2002). All major β-turns (I, II, VIII, I′, II′; Hutchinson & Thornton, 1994) can be found among the conformation families.

Two-dimensional and four-dimensional calculations were extended by direct tetrapeptidic mapping. The calculation of a six-dimensional map took approximately 5 d on an Intel Core2 Duo processor (8 h per section, 16 sections) with 0.4×10^6 experimental conformations and 16×10^6 grid points. The top conformation families (representing approximately 75% of the conformation space) are given in Table 3. The full tetrapeptidic table is deposited as supplementary material.

Results for the tetrapeptidic fragments in Table 3 show the high conformational flexibility of protein structures and the importance of the Ramachandran plot. Again, many conformations can be classified in terms of basic two-dimensional 'Ramachandran' conformations. Tripeptidic mapping demonstrated that we also have to consider D (−90, 0°) and Z (−104,

70°) conformations as independent conformation primitives for classification. Tetrapeptidic mapping supports this idea (for example, conformations 6, 8, 9; 7, 12, 16). The tripeptidic mapping also showed that there are no BC or CB peaks representing mixed $\beta1$ and $\beta2$ structure. Again, only two peaks representing all-β structure (BBB, $P = 10.1\%$; CCC, $P = 4.1\%$) and not eight (BBB, BBC, BCB, CBB, CCB, CBC, BCC, CCC) are found by tetrapeptidic analysis. Tetrapeptidic mapping reveals several highly extended conformations that may help to classify or subdivide the large β region (for example, conformations 19, 20, 24 or 27) and conformation families 19 and 20 suggest (−150, −170°) as another conformation primitive.

In a previous analysis employing multidimensional scaling, Sims *et al.* (2005) found only eight representative conformations of $(\varphi, \psi)_3$, which could be an artifact of their method and the limited (only 6000) number of structure fragments used. A completely different picture has emerged from this analysis. There are three highly populated conformations (AAA, P = 29.6%; BBB + CCC, P = 14.1%) that cover 44% of conformation space. The top four, well separated, families AAA, BBB, CCC and CAA ($P = 3.3\%$) represent almost half of the conformation space. The other conformation families (starting with CCA, $P = 2.7\%$, and ADC, $P = 2.3\%$) represent a continuous sequence of families with slowly decreasing probabilities of appearance and no sharp changes. These families represent the other half of conformation space.

If we consider six conformation families in two dimensions, we theoretically have 36 conformational families in four dimensions and 216 combinations in six dimensions. Comparable numbers are found by n-dimensional mapping. We found 24–25 tripeptidic $(\varphi, \psi)_2$ conformation families. It is difficult to determine the number of conformation families in higher dimensions because of the long tail of noise conformations and the arbitrary cutoff. The number of conformation families appears to increase with the number of available data. One can define the relative number of conformations for fixed occupancy of the conformation space. For 90% occupancy we obtain five $(\varphi, \psi)_1$, 13 $(\varphi, \psi)_2$ and 83 $(\varphi, \psi)_3$ conformation families. For 95% we obtain six, 29 and 133 conformations, respectively. Similar data have also been obtained by cluster methods [for example, 28 and 202 by Micheletti *et al.* (2000) for $(\varphi, \psi)_2$ and $(\varphi, \psi)_3$, respectively]. The view of Kim's school (Sims *et al.*, 2005) appears to be over-optimistic with respect to Levinthal's paradox (Levinthal, 1968). From our analysis, the number of conformations can be estimated as 4^N–5^N, where N is the number of residues in the protein. Cumulative probability graphs for each fragment type are given in Figs. 1, 2 and 3.

Pentapeptidic families (eight-dimensional space, six C^α atoms, AlphaP fragment) were obtained using the generator method. The eight-dimensional map contains 4×10^9 grid points (for NG = 16) and this number is not comparable with the number of experimental data (0.2×10^6). Calculations for direct mapping in conformation space would take several years on a single processor.

The generator method was tested in four and six dimensions. Dipeptidic conformation families (two dimensions; Table 1) were used to generate tripeptidic and tetrapeptidic conformation families. These results were compared with those of direct four-dimensional and six-dimensional mapping. Results for AlphaT can be found in the supplementary material. From these analyses several conclusions could be reached.

(i) The top generated families are consistent with those of the mapping.

(ii) The number of generated families is smaller than those of the mapping.

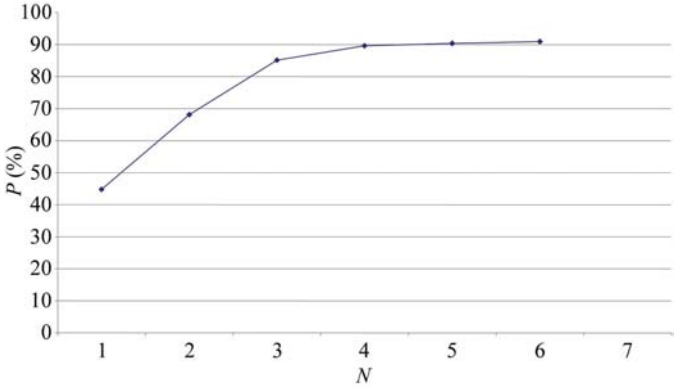

Figure 1
Cumulative probability *versus* family number for dipeptidic fragment.

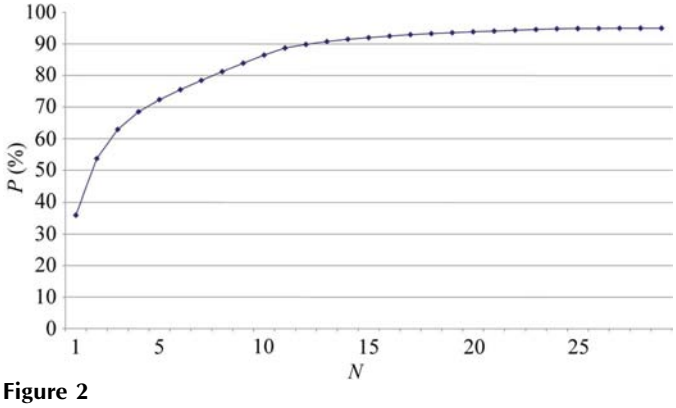

Figure 2
Cumulative probability *versus* family number for tripeptidic fragment.

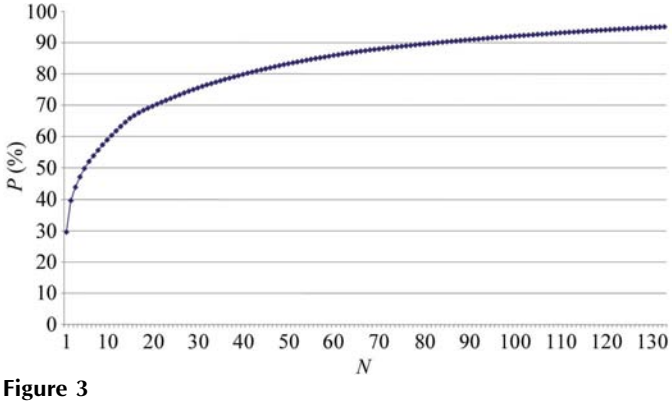

Figure 3
Cumulative probability *versus* family number for tetrapeptidic fragment.

(iii) Generators are the only method to study high-dimensional conformation spaces (pentapeptides, hexapeptides).

(iv) Some generated families are regions of high conformation density rather than local maxima.

(v) It is difficult to classify large flat density regions (like the β region in proteins). The primitive optimization procedure is not able to move the conformation to the nearest maximum.

(vi) It is difficult to detect conformation families close to a dominant conformation family.

(vii) In higher dimensions the signal-to-noise ratio is small and only top conformation families can be reliably detected (the noise region is a region where the number and position of conformation families depend on the number and selection of structures used for the analysis).

The pentapeptidic conformation families were generated by two methods. The first was by using tripeptidic/tripeptidic combinations. The second method is based on combination of tetrapeptidic and dipeptidic families. Because of the polarity of the polypeptide chain both combinations were considered: dipeptidic/tetrapeptidic and tetrapeptidic/dipeptidic. There were approximately 600 combinations above the noise level, which in this case was reduced to 3. By applying ten optimization cycles, the number of conformation families converged to 210–220. The conformation families generated by all methods were merged and the merging resulted in 269 conformation families ($P > 0.01\%$). An extensive table for AlphaP is deposited as supplementary material. Only 194 conformation families are needed to cover 95% of the experimental conformations. The number of conformation families covering 90% of the conformation space is 131. Whether this relative drop in the number of conformation families is a property of the pentapeptidic group (all main structure features are already sufficiently reflected by tetrapeptidic AlphaQ) or is an artifact of the generator method remains unanswered. The overall features of pentapeptidic families are the same as for tetrapeptidic families. For obvious reasons, no principally new basic conformation types are detected. Nevertheless, the conformation probabilities may be important for statistically grounded model building.

4. Conclusions

Several conclusions can be drawn from the present analysis for automated model building. Information from higher dimensional spaces can be used for rational sampling of torsion-angle space in lower dimensions. Some less frequent dipeptidic conformations that cannot form an independent peak on a two-dimensional map in the proximity of a more dominant conformation family simply accumulate as independent families in higher dimensions because they are usually found in a building block with a specific structural function. From analysis of tripeptidic fragments it has become clear that the D and Z (bridge) conformations should now be considered to be conformational primitives. Tetrapeptidic families add further information on how to sample the flat β region [B, C, Z and $(-150, -170^{\circ})$]. The extended set of search dipeptidic fragments is A, B, C, G, X, E, D and Z.

The AlphaT fragment, defined by the tripeptidic conformation families, appears to be a suitable search fragment for automatic model building at lower resolutions. The number of conformation families is moderate (24–25; only 13 families are needed to cover 90% of the conformation space), while 16 conformers can easily be handled by PRCTF in the present version of the program *NUT* (Pavelcik, 2006*b*). Reconstruction of the protein structure with conformation families is further facilitated by the flexible-fragment concept and torsion-angle refinement (Pavelcik, 2003).

Building protein structure by PRCTF with the tetrapeptidic AlphaQ and pentapeptidic AlphaP fragments would be more complicated. Only top conformations could be used for the direct location of tetrapeptidic fragments in the electron density because PRCTF is a computationally intensive procedure. There is also a problem with varying fragment radii. On the other hand, tetrapeptidic and pentapeptidic tables are useful for loop building and connecting partial chains to larger units by chain extension directly into electron density.

An advantage of the mapping method over clustering is that the results are in torsion angles. Any molecular-modelling program can generate the atomic coordinates of the fragment from its torsion angles. Reconstruction of fragment coordinates from C^{α} atoms in the cluster method is more complicated and may not be unique. A disadvantage of the present method is that direct mapping is computationally forbidden for large fragments and the less reliable generator tool has to be used.

In summary, the relative importance of individual conformation families for biomacromolecular model building has been established. The calculated results will be implemented into new versions of the biomacromolecule model-building programs *NUT* and *HEL* (Pavelcik, 2006*a*,*b*). Tripeptidic conformation families are also suitable candidates for computer programs developed to check protein structures.

The research was supported by GACR grant 204/06/1007 (Czech Republic) and grant VEGA 1/2333/05 (Slovak Republic).

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, J. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

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

Dunbrack, R. L. Jr & Karplus, M. (1994). *Struct. Biol.* **1**, 334–340.

Hovmöller, S., Zhou, T. & Ohlson, T. (2002). *Acta Cryst.* D**58**, 768–776.

Hutchinson, E. G. & Thornton, J. M. (1994). *Protein Sci.* **3**, 2207–2216.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. (2002). *J. Mol. Biol.* **323**, 297–307.

Levinthal, C. (1968). *J. Chim. Phys.* **65**, 44–45.

Micheletti, C., Seno, F. & Maritan, A. (2000). *Proteins*, **40**, 662–664.

Murray, L. J. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.
Pauling, L. & Corey, R. B. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 729–740.
Pauling, L., Corey, R. B. & Branson, H. R. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 205–211.
Pavelcik, F. (2003). *Acta Cryst.* A**59**, 487–494.
Pavelcik, F. (2004). *Acta Cryst.* D**60**, 1535–1544.
Pavelcik, F. (2006*a*). *J. Appl. Cryst.* **39**, 287.
Pavelcik, F. (2006*b*). *J. Appl. Cryst.* **39**, 483–486.
Pavelcik, F. & Vanco, J. (2006). *J. Appl. Cryst.* **39**, 315–319.
Pavelcik, F., Zelinka, J. & Otwinowski, Z. (2002). *Acta Cryst.* D**58**, 275–283.
Ramachandran, G. N. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.
Sims, G. E., Choi, I.-G. & Kim, S.-H. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 618–621.
Venkatachalam, C. M. (1968). *Biopolymers*, **6**, 1425–1436.