Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Conformational attractors on the Ramachandran map

Dirk Walther[a] and Fred E. Cohen[b,c]*

[a]Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, California 94143-0450, USA, [b]Department of Medicine, University of California at San Francisco, San Francisco, California 94143-0450, USA, and [c]Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0450, USA

Correspondence e-mail:
cohen@cmpharm.ucsf.edu

Received 24 August 1998
Accepted 7 October 1998

Frequency distributions of protein backbone dihedral angles φ and ψ have been analyzed systematically for their apparent correlation with various crystallographic parameters, including the resolution at which the protein structures had been determined, the R factor and the free R factor, and the results have been displayed in novel differential Ramachandran maps. With improved sensitivity compared with conventionally derived heuristic Ramachandran maps, such differential maps automatically reveal conformational 'attractors' to which φ/ψ distributions converge as the crystallographic resolution improves, as well as conformations tied specifically to low-resolution structures. In particular, backbone angular combinations associated with residues in α-helical conformation show a pronounced consolidation with substantially narrowed φ/ψ distributions at higher (better) resolution. Convergence to distinct conformational attractors was also observed for all other secondary-structural types and random-coil conformations. Similar resolution-dependent φ/ψ evolutions were obtained for different crystallographic refinement packages, documenting the absence of any significant artificial biases in the refinement programs investigated here. A comparison of differential Ramachandran maps derived for the R factor and the free R factor as independent parameters proved the better suitability of the free R factor for structure-quality assessment. The resolution-based differential Ramachandran map is available as a reference for comparison with actual protein structural data under *WebMol*, a Java-based structure viewing and analysis program (http://www.cmpharm.ucsf.edu/cgi-bin/webmol.pl).

1. Introduction

In protein structure determination and validation, the Ramachandran plot – the two-dimensional scatter plot of the protein main-chain torsion angles φ and ψ – provides a simple yet sensitive tool to evaluate the quality of the structure refinement, as the data can be compared to reference distributions of sterically allowed and energetically favorable regions (Ramachandran & Sasisekharan, 1968; Laskowski *et al.*, 1993; Kleywegt & Jones, 1996; Kuszewski *et al.*, 1996; Hooft *et al.*, 1997). Since various regions in the Ramachandran map are associated with canonical backbone conformations such as α-helices and β-strands, the Ramachandran plot also imparts a coarse picture of structural features found in the protein under investigation (Adzhubei *et al.*, 1987; Karplus, 1996). Evidently, the reference distribution, henceforth referred to as the Ramachandran map, bears critically on the conclusion drawn from such analyses. With the steadily increasing availability of high-resolution protein structural data, the heuristic derivation of Ramachandran maps from known protein structures instead of theoretical investigations has become standard, as in the popular program *PROCHECK* (Laskowski *et al.*, 1993). Statistics of φ/ψ distributions are gathered from a set of high-resolution crystal structures to estimate their underlying probability density-distribution function, which is then used to evaluate backbone conformations in proteins of interest (Hooft *et al.*, 1997). Naturally, the crystallographic resolution cut-off is a major determinant for the data-set generation and the reliability of the statistics obtained. Morris *et al.* (1992) observed that φ/ψ distributions become increasingly tightened as the resolution improves. Their analysis was based upon a coarse classification of the Ramachandran map into 'core', 'allowed', 'generous' and 'outside' conformational regions. With improving resolution, an increasing fraction of residues was found to fall into the defined core region, reported similarly by Kleywegt & Jones (1996), Hooft *et al.* (1997) and the EU 3-D Validation Network (1998).

In this study, we resolve this correlation with the crystallographic resolution to the level of detail of the Ramachandran map itself and for all secondary structural and amino-acid types separately. Instead of analyzing φ/ψ frequencies, their changes with respect to resolution and other crystallographic parameters are displayed in novel differential Ramachandran maps, thus essentially capturing 'evolutionary' features. Without the need to pre-define favorable backbone conformational states and in a quantitative manner, these maps automatically carve out the fine structure in the φ/ψ frequency distributions and reveal 'focal points' or 'conformational attractors' to which backbone conformations converge as resolution improves, as well as progressively depleted regions in the Ramachandran map. The differential Ramachandran maps may therefore serve as a valuable supplement to conventional maps, as they may help to identify suspicious backbone conformations for which a re-examination of the electron-density map may be indicated, may help in detecting strained backbone conformations and may guide protein model-building efforts. Differential maps were also compared between different refinement programs and for other relevant crystallographic parameters, such as the *R* factor and free *R* factor (Brünger, 1992*a*). Differences in their correlation with observed backbone angular frequencies, as reported here, may imply different merits of both parameters as structure-quality criteria.

2. Materials and methods

2.1. Data

A non-redundant subset of 880 protein crystal structures or protein chains exhibiting less than 30% pairwise sequence identity after optimally aligning their sequences and which were determined at 3.0 Å or better (Dunbrack, 1997) were selected from those stored in the PDB (Bernstein *et al.*, 1977). It was required that the selected proteins fulfill the basic quality criteria imposed by the program *DSSP* discarding, for instance, structures with no complete residue (Kabsch & Sander, 1983). This set contained a total of 209376 non-terminal amino acids involved in *trans*-peptide bond conformation to the corresponding preceding and succeeding amino acid along the polypeptide chain. Assignments of secondary-structural states were generated by the program *DSSP*. Crystallographic parameters, the nominal resolution, *R* factor, free *R* factor, year of determination and refinement program used were conveniently retrieved from the database *PDBFINDER* (Hooft *et al.*, 1998).

2.2. Ramachandran map

The main-chain torsion angles φ, defined by the backbone atoms $C_{i-1}-N_i-C_{\alpha,i}-C_i$, and ψ, defined by $N_i-C_{\alpha,i}-C_i-N_{i+1}$, were measured for each non-terminal and *trans*/*trans*-bonded amino-acid residue *i*. The φ/ψ diagram was subdivided into 72 × 72 bins or cells with each cell collecting data (counts) for itself and the eight cells surrounding it, thereby smoothing the data by applying a sliding interval of 15° shifted at a 5° register. For example, a 5 × 5° cell centered at $\varphi = 22.5°$ actually comprises data within the interval $15.0° \leq \varphi < 30°$ and the cell centered at $\varphi = 27.5°$ actually comprises data with the interval $20.0° \leq \varphi < 35°$, and similarly for the ψ direction. Neighboring cells exceeding the boundaries of −180° or 180° were circularly closed.

2.3. Differential Ramachandran maps

For every protein in the data set used, the absolute counts in each of the 72 × 72 bins or cells were converted to relative frequencies; *i.e.* divided by the total number of counts (φ/ψ measurements) for the given protein, thus representing the fraction $f_{\varphi/\psi}$ of amino-acid residues in that particular φ/ψ conformational state (bin), and stored together with characteristic crystallographic parameters. These included the nominal resolution (Res), the *R* factor and the free *R* factor ($R_{\rm free}$) and year of determination associated with that particular protein *p*.

The resulting 880-element vector of fractional frequencies $f_{\varphi/\psi}$ for a given φ/ψ bin was then correlated with specific crystallographic parameters Par(p), *e.g.* resolution, associated with each of the 880 protein chains by a linear regression with $f_{\varphi/\psi}(p) = m{\rm Par}(p) + n$, where p denotes the corresponding protein characterized by slope m and intercept n. This linear regression was effected for all 72 × 72 φ/ψ bins. The resulting matrix of linear correlation coefficients $r_{\varphi/\psi}$ associated with each linear regression was then plotted in a two-dimensional map and color coded according to the sign of $r_{\varphi/\psi}$, with its absolute value being encoded by color saturation. Blue-colored cells indicate negative values of $r_{\varphi/\psi}$, *i.e.* increasing relative frequency of a particular φ/ψ state with decreasing values of the chosen parameter (*e.g.* improving resolution). Red-colored cells indicate positive correlations, *i.e.* conformations less frequently observed at smaller values of the parameter. Only those values of $r_{\varphi/\psi}$ were plotted with 25 or more non-zero values of $f_{\varphi/\psi}(p)$; *i.e.* plotted φ/ψ states had to be observed in at least 25 proteins; all others were colored gray. The correlation coefficient rather than the slope of the linear regression was chosen for analysis, as it not only measures the overall trend but also reflects the significance of the correlation.

This analysis was also performed separately for only certain types of secondary-structural states characterized by specific main-chain–main-chain hydrogen-bonding networks as defined in *DSSP* (Kabsch & Sander, 1983) and for each of the 20 amino-acid types. In such analyses, only those amino-acid residues belonging to that particular secondary-structural type or amino-acid type were considered.

2.4. Cross-correlations between Ramachandran maps

Differential Ramachandran maps obtained for two different crystallographic parameters, *e.g.* resolution and R factor, and other φ/ψ-related property maps (*e.g.* B factor) were cross-correlated by determining the cross-correlation coefficient, $r_{\rm cross}$, where $-1 \leq r_{\rm cross} \leq 1$, between all equivalent 72×72 bin values in each individual differential Ramachandran map, on condition that only those bins were considered for which both maps contained more than 25 examples (proteins).

2.5. Temperature *B*-factor analysis

Temperature B factors (or Debye–Waller factors) for atomic positions were contained in 865 of the 880 PDB files examined. All B factors were rescaled with a new mean value of zero and a unit standard deviation with $B_{\rm new} = (B_{\rm pdb} - \langle B_{\rm pdb}\rangle)/\sigma_{B_{\rm pdb}}$, where $\langle\rangle$ denotes the mean and σ is the standard deviation.

2.6. Statistical significance, *P* value

For uncorrelated values x and y and a large number of measurements ($N >> 3$), the linear correlation coefficient r (Pearson's correlation coefficient) is distributed normally with a zero mean and a standard deviation σ of $1/N^{1/2}$. The probability of obtaining a correlation coefficient equal to or greater in magnitude than a certain measured coefficient r in a sample distribution can therefore be estimated from $P = {\rm erfc}[|r|(N/2)^{1/2}]$, where erfc is the complementary error function (Press *et al.*, 1988) and $0 \leq P \leq 1$. It has to be borne in mind, however, that as the distributions of parameters investigated here (*e.g.* resolution) cannot be considered random, the calculated P values only represent an estimate.

3. Results

The heuristic Ramachandran map derived from the protein set used in this study, with assignments for the distinct peaks and regions to secondary-structural types and designations for

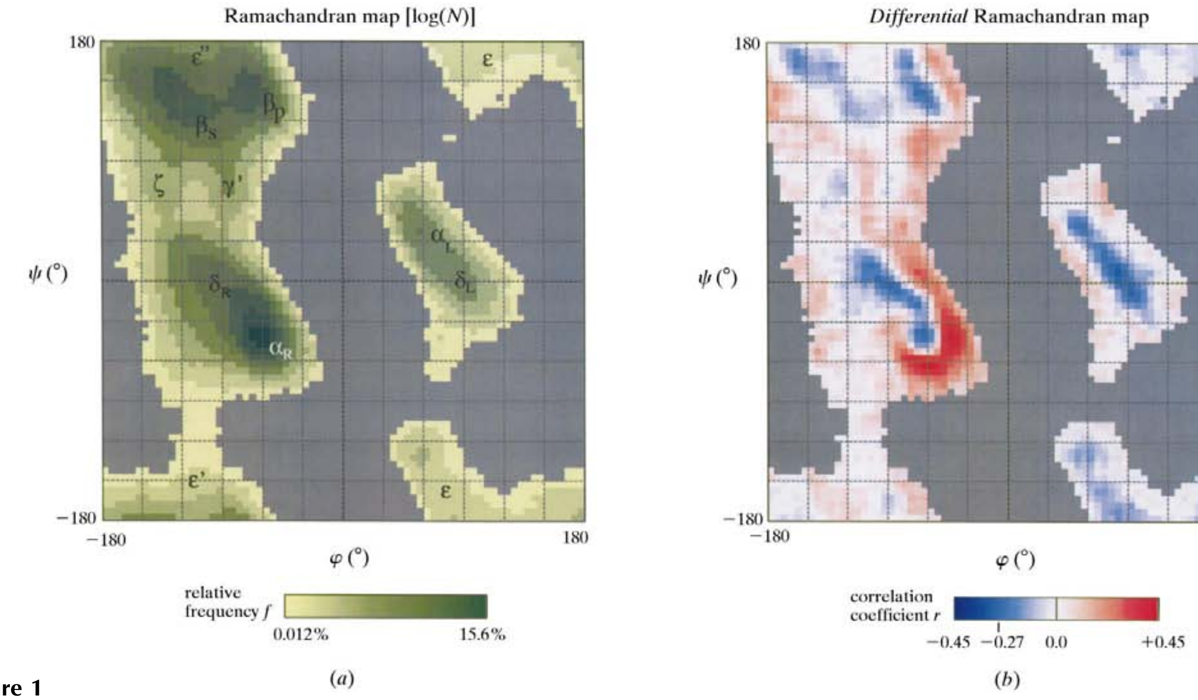


Figure 1
Heuristic and differential Ramachandran map. (*a*) The database-derived Ramachandran map obtained from all non-terminal *trans*/*trans*-bonded amino-acid residues in the protein data set used in this study. Logarithmic absolute frequencies N are encoded by the color spectrum displayed underneath the diagram, with the observed upper and lower limits of the relative frequencies associated with each color also indicated. φ/ψ regions (bins) with occurrences in only 25 or fewer proteins are colored gray. The designation of the distinct maxima and regions in the plot are adapted from Karplus (1996) with α_R, right-handed α-helix; α_L, mirror-image (left-handed) of α_R; β_S, β-strand residues involved in antiparallel and parallel β-sheet formation; β_P, extended polyproline helices (Adzhubei & Sternberg, 1993); δ_R, right-handed region commonly referred to as the bridge region – region encompasses α-helical residues, residues in 3_{10}-helical, turn, β-bulge as well as random-coil conformation; δ_L, mirror image of δ_R; γ', inverse γ turn (Milner-White, 1990); $\varepsilon - \varphi > 0°$ and $\psi \simeq \pm 180°$; ε' and ε'', mirror images of ε, with ε'' denoting the region overlapping with β_S and β_P; ζ, conformational region preferably occupied by residues preceding prolines (Karplus, 1996). (*b*) The differential Ramachandran map (see §2.3) derived from the same data set. The color spectrum refers to the correlation coefficients r obtained from linear regressions of fractional frequencies with crystallographic resolutions. 100% color saturation would correspond to r values of +0.45 and −0.45, respectively. The actual observed maximal r obtained for a φ/ψ bin was +0.45 and the greatest negative value −0.27.

other structural substates, is shown in Fig. 1(*a*). The color spectrum used to encode the backbone angular frequencies is based on a logarithmic scale, since the α-helical conformation ($\varphi \simeq -64^\circ$ and $\psi \simeq -40^\circ$) dominates the map when a linear scale is used, essentially masking all other conformations (peak value of 15.6% as compared with 2.2 and 2.1% for β_P and β_S, respectively, ranked next).

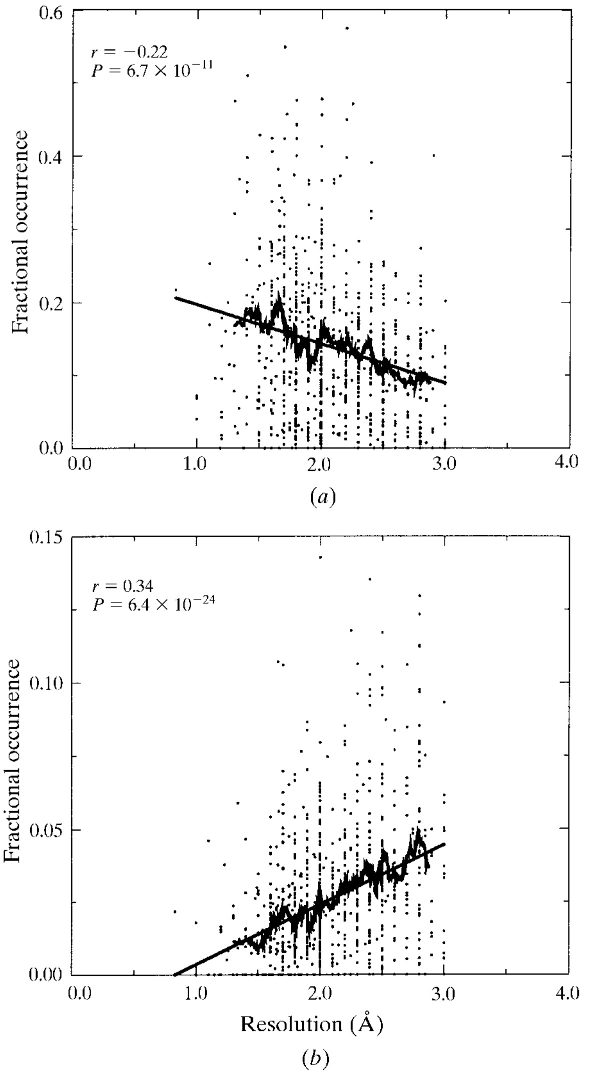
3.1. Resolution-based differential Ramachandran maps

This conventional Ramachandran map imparts an integral view, irrespective of the resolution at which protein structures contained in the data set have been determined. The fidelity of such maps can be improved by including only high-resolution structures. Still, an arbitrary resolution cut-off has to be defined, possibly limiting the available amount of data significantly. More critically, no insight about where specifically interpretations of low-resolution electron-density maps may be consistently misled during the structure model-building and refinement process can be gained from any single such map with a fixed resolution cut-off. Alternatively, we introduce the differential Ramachandran map shown in Fig. 1(*b*): instead of plotting absolute φ/ψ frequencies, their changes or sensitivity with respect to crystallographic parameters (*e.g.* resolution) are measured by means of correlation coefficients and mapped onto the φ/ψ-parameter space (see §2.3). The sign (encoded by color) and magnitude (encoded by color saturation) of the correlation coefficients obtained immediately delineate the extent to which particular φ/ψ frequencies appear to depend upon the particular parameter, as well as their overall tendencies. Such differential maps can be imagined as a static visualization of a time series of several conventional Ramachandran maps, each obtained from proteins falling within consecutive intervals of the particular parameter of interest. What is captured in the differential map is the significance and direction of change in φ/ψ regions in such a 'flicker movie'.

It would be expected that some regions of the Ramachandran map become progressively less populated while others become increasingly more populated as resolution – taken as the independent parameter – improves. In the differential map, such behavior is reflected by different signs of the obtained correlation coefficients, encoded by a red or blue color scale. Parameter-indifferent behavior would result in small correlation coefficients, with white color indicating zero correlation.

The resolution-based differential Ramachandran map shown in Fig. 1(*b*) indeed reveals a substantial 'evolution' in φ/ψ distributions. It demonstrates that the differences in the frequencies observed in the integral Ramachandran map (Fig. 1*a*) are not only caused by natural deviations from the mean but also by systematic shifts correlated with the resolution at which the proteins have been determined crystallographically. In particular, the distinct red region south-east of the right-handed α-helical region stands out noticeably, indicating substantially lowered frequencies as resolution improves, with obtained correlation coefficients for individual bins as high as +0.45 and corresponding P values of 1.2×10^{-40}. The blue regions highlighting φ/ψ combinations with increased relative frequencies at higher resolutions largely coincide with the maxima observed in the 'conventional' Ramachandran map (Fig. 1*a*). Thus, the frequency shifts reflect convergence rather than migration of the maxima itself. By contrast, some φ/ψ regions do not display any obvious correlation with resolution; *e.g.* regions designated by ε'' and parts of ε with nearly white colored cells.

Capturing differential aspects effectively sharpens the view on conventional Ramachandran maps and the fine structure of the φ/ψ frequency distribution becomes readily visible. Distinct φ/ψ angular combinations are identifiable to which observations of backbone dihedral angles appear to converge



(*a*)

(*b*)

Figure 2
Fractional occurrences *versus* resolution. Fractional occurrences of two selected φ/ψ conformational regions plotted *versus* the resolution at which the corresponding proteins harboring those conformations were determined crystallographically; *i.e.* each dot represents one protein. Both regions were positioned close to the canonical α-helical φ/ψ values with (*a*) $-65 \le \varphi < -50^\circ$ and $-45 \le \psi < -30^\circ$ and (*b*) $-65 \le \varphi < -50^\circ$ and $-65 \le \psi < -50^\circ$. Linear-regression lines with the corresponding correlation coefficients r as well as 50-point running averages (fluctuating solid line) for the abscissa-ordered raw data (dots) are shown.

as resolution improves (blue regions). For instance, the ζ and γ regions in Fig. 1(a) appear as separated islands or conformational attractors and not as extensions of the β-strand conformational region, demonstrating the improved sensitivity to structural detail of our analysis. Also, the β_S region separates into the two attractors known for parallel and antiparallel sheets (Chothia, 1973).

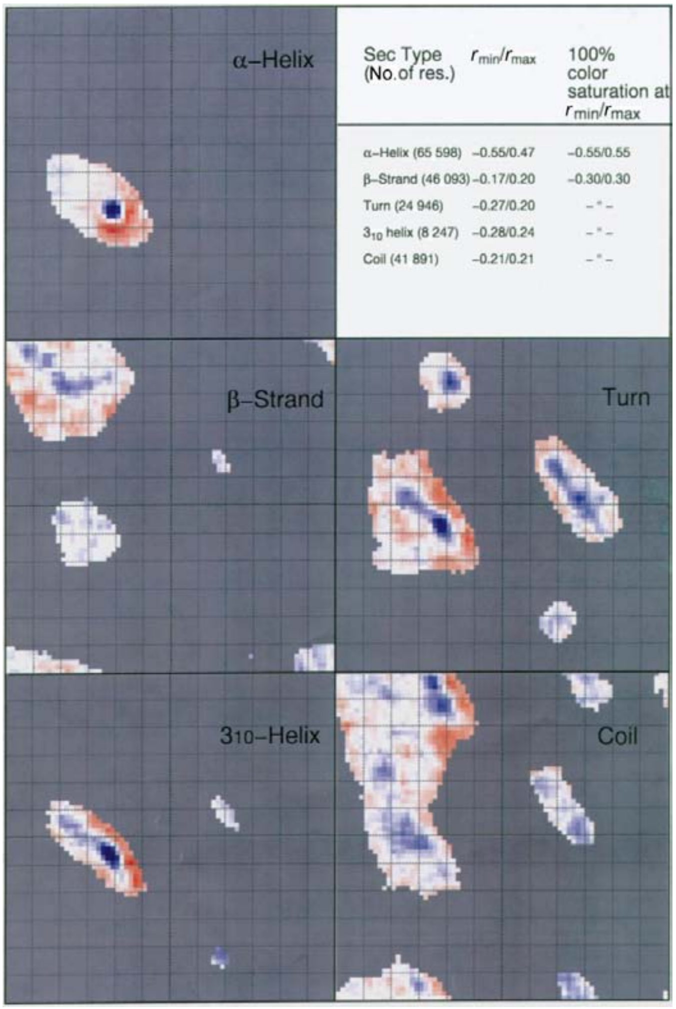
We use the term 'conformational attractor' for the description of such backbone dihedral angle conformations that can be assumed to correspond to local energy minima, rendering them preferred conformational states. Backbone dihedral angles in real protein and at a given temperature will always show fluctuations around such lowest energy states, and we wish the term 'attractor' be understood as a 'basin of attraction' rather than a singular point.

Fig. 2 illustrates how the fractional frequencies of backbone dihedral angles for two selected φ/ψ bins were actually observed to depend upon resolution. The overall trends are captured by a linear-regression analysis and the corresponding correlation coefficients. Although there is substantial scatter in the data, the corresponding P values are extraordinarily small. Therefore, it seems highly unlikely that these correlations originated from random fluctuations. It might have been expected that the fractional frequencies reach a plateau value below a certain level of resolution, as was observed for the fraction of core residues ['core' with respect to favorable regions in the Ramachandran map, as defined by Morris *et al.* (1992)]. Our results suggest that this may not be generally true for all local regions in the Ramachandran map (Fig. 2). The running averages of the two distributions display a continued upward or downward slope, suggesting that even in protein sets determined at relatively high resolution (<2.0 Å), the limiting or ultimate fractional frequencies appear as not yet determined.

The methodology of differential Ramachandran maps can also be applied to specific residue subsets to reveal trends hidden in the overall picture. Fig. 3 juxtaposes the resolution-based differential Ramachandran maps obtained for common secondary-structural types and residues in random-coil conformations. Again, distinct blue regions identifying conformational attractors are contrasted by red regions indicating depleted regions, as well as white regions for conformations with frequencies uncorrelated with resolution. The differential Ramachandran map for α-helices shows all three such qualities in well separated φ/ψ areas. The increased confinement of α-helical backbone conformations with $\varphi \simeq -64°$ and $\psi \simeq -40°$ is obvious from the graph, as is the pronounced trend out of regions south-east of the α-attractor. By contrast, regions north-west of the focal point are evenly populated throughout all resolutions. A more detailed analysis of these findings is presented below.

β-strand conformations appear to move out of areas beneath the main diagonal $\psi = -\varphi$ into regions above this demarcation line. As this line marks the distinction between left- (beneath the diagonal) and right- (above the diagonal) handed β-twists, the preference for right-handed conformations (Chothia, 1973) is demonstrated. In better resolved structures, left-handed conformations are progressively less common. The observations of β-residues falling into the α-helical region originate from residues in β-bulge conformation. No obvious conformational attractor is evident for this structural motif.

Residues in turn as well as in 3_{10}-helical conformations converge to similar conformational attractors. Since both are defined by main-chain–main-chain hydrogen bonds between residues at relative sequence positions i and $i + 3$, this similarity is not surprising. Although generally perceived as less regular in their φ/ψ values (Barlow & Thornton, 1988), the differential Ramachandran map identifies a weak attractor for 3_{10}-helices near $\varphi \simeq -70°$ and $\psi \simeq -20°$. Observations with



Sec Type (No. of res.)	r_{min}/r_{max}	100% color saturation at r_{min}/r_{max}
α–Helix (65 598)	–0.55/0.47	–0.55/0.55
β–Strand (46 093)	–0.17/0.20	–0.30/0.30
Turn (24 946)	–0.27/0.20	– " –
3_{10} helix (8 247)	–0.28/0.24	– " –
Coil (41 891)	–0.21/0.21	– " –

Figure 3
Differential Ramachandran maps for common secondary-structural types and random-coil conformations. The table summarizes the obtained statistics and the minimal and maximal observed correlation coefficient for each graph, as well as the color-spectrum boundaries [color saturation for red (positive) values of r and blue (negative) values of r]. 'Coil' residues refer to those residues in proteins with no assignment to any secondary or any other structural type (*e.g.* 'bend') in *DSSP* (Kabsch & Sander, 1983), *i.e.* the corresponding column in the *DSSP* file record was a blank character. As in Fig. 1, the minimal required number of occurrences for a given φ/ψ bin was 25 proteins. Horizontal axes measure φ, vertical axes ψ, each within a (−180, 180°) interval.

Table 1
Cross-correlations between differential Ramachandran maps.

		Residues in α-helical conformation			
	Parameter	Resolution	$R_{\rm free}$	R factor	Year
All residues/ all secondary structural elements	Resolution	1	0.79	0.82	−0.06
	$R_{\rm free}$	0.69	1	0.75	−0.13
	R factor	0.79	0.67	1	−0.29
	Year	−0.03	−0.07	−0.26†	1

† Note that for resolution, R and free R factor 'smaller' means 'better', whereas for year larger values may be expected to correspond to 'better' data.

Table 2
Range of correlation coefficients obtained for differential Ramachandran maps derived for various crystallographic parameters.

All residues/all secondary structural states

Parameter	$r_{\rm min}$	$r_{\rm max}$	Standard deviation†	Number of proteins
Resolution	−0.28	0.43	0.102	880
$R_{\rm free}$	−0.29	0.51	0.121	279‡
R factor	−0.21	0.26	0.078	880
Year	−0.11	0.09	0.028	880

Residues in α-helical conformation

Parameter	$r_{\rm min}$	$r_{\rm max}$	Standard deviation†	Number of proteins
Resolution	−0.55	0.47	0.157	812
$R_{\rm free}$	−0.43	0.49	0.163	268
R factor	−0.33	0.32	0.108	812
Year	−0.10	0.09	0.032	812

† Only φ/ψ bins with 25 or more examples (proteins) were considered. ‡ The free R factor (Brünger, 1992*a*) is not listed or measured for all proteins.

$\varphi > 0$ can be attributed predominantly to glycines positioned at the N-terminus of 3_{10}-helical segments. These conformations allow the particular main-chain hydrogen bond to form but their φ/ψ angles clearly differ from the canonical 3_{10}-attractor. We suggest that the assignment for those residues should be revised or interpreted with caution.

For main-chain conformations associated with residues in random-coil state; *i.e.* residues not involved in regular intra-protein hydrogen-bond networks, the polyproline(II)-helical state (Fig. 1*a*) was identified as the strongest attractor supplemented by other rather diffuse regions corresponding to the ζ-region and a region with $\psi \simeq 0°$. As the polyproline(II) helix is not stabilized by main-chain–main-chain hydrogen bonds, the origin of favorable energies in this state has been investigated by several research groups without reaching a consensus (Eisenhaber *et al.*, 1992; Adzhubei & Sternberg, 1993; Maccallum *et al.*, 1995). Its identification as a pronounced conformational attractor emphasizes its relevance as a distinct secondary-structural state.

The ζ-state (Fig. 1*a*) was found to be adopted almost exclusively by residues preceding proline (MacArthur & Thornton, 1991; Hurley *et al.*, 1992; Karplus, 1996). In our data set, approximately 70% of all ζ occurrences were associated with this sequence motif.

Differential Ramachandran maps were also generated separately for each of the 20 amino acids. However, the statistical significance of correlations was weaker and largely coincided with the tendencies observed for the secondary-structural type for which the particular amino acid is known to have a preponderance. For example, maps for alanine and glutamic acid, common participants in α-helices, were observed to converge to the α-helix attractor. Corresponding differential maps for all 20 amino-acid types can be accessed on the World Wide Web (see §5).

Systematic changes in the φ/ψ distributions correlated with resolution would of course lose their meaning if produced by a bias towards over-representation of a particular folding class, *e.g.* all-α, as resolution improves. However, the fractional content of any secondary structural type defined in *DSSP* was observed to be uncorrelated with resolution, with corresponding correlation coefficients of $r < 0.03$ for each secondary structural type, including random coil, interpreted as a separate structural category. Thus, the frequency shifts observed in Fig. 1(*b*) did not originate from an unbalanced data set.

3.2. Other crystallographic parameters

Differential Ramachandran maps can also be obtained for other crystallographic quantities as independent parameters. Differential Ramachandran maps derived for the R factor and the free R factor (Brünger, 1992*a*) revealed near-identical patterns of depletion and enrichment of φ/ψ populations as obtained for the resolution-based maps (Fig. 1*b*). No significant correlations were found between conformational preferences and the year of structure determination. These observations are summarized in Table 1 by means of cross-correlation coefficients (see §2.4) between the different types of general differential maps (all residues/all secondary structural states) and maps of α-helical residues only, with high cross-correlation coefficients indicating coherent behavior.

Although qualitatively similar, the maps for the R factor and $R_{\rm free}$ differ in their magnitudes of correlation with respect to φ/ψ relative frequencies. In both the general map and the map derived for α-helical residues only, the determined maximal and minimal correlation coefficients were significantly smaller for the R factor than for $R_{\rm free}$. While minimal and maximal values of any distribution may correspond to extreme outliers within otherwise similar bulk distributions of data, the smaller standard deviations of correlation coefficients in R-factor-based maps also implies a truly smaller range. As the R factor itself is minimized in the structure-refinement process, this weaker correlation is comprehensible, yet has implications for its value as a structure-quality indicator (see §4.4). The correlation coefficients obtained for year-based maps were not instructive (Table 2) and were generally not correlated with any other crystallographic parameter (Table 1).

4. Discussion

Given current crystallographic technologies on one hand and the nature of proteins and solvent-containing protein crystals on the other, it is often not possible to obtain diffraction data at a precision that would allow an unambiguous interpretation of experimentally phased electron-density maps. An understanding of the apparent evolution of Ramachandran maps with resolution as investigated here may therefore help to eliminate errors in the interpretation of low-resolution structures.

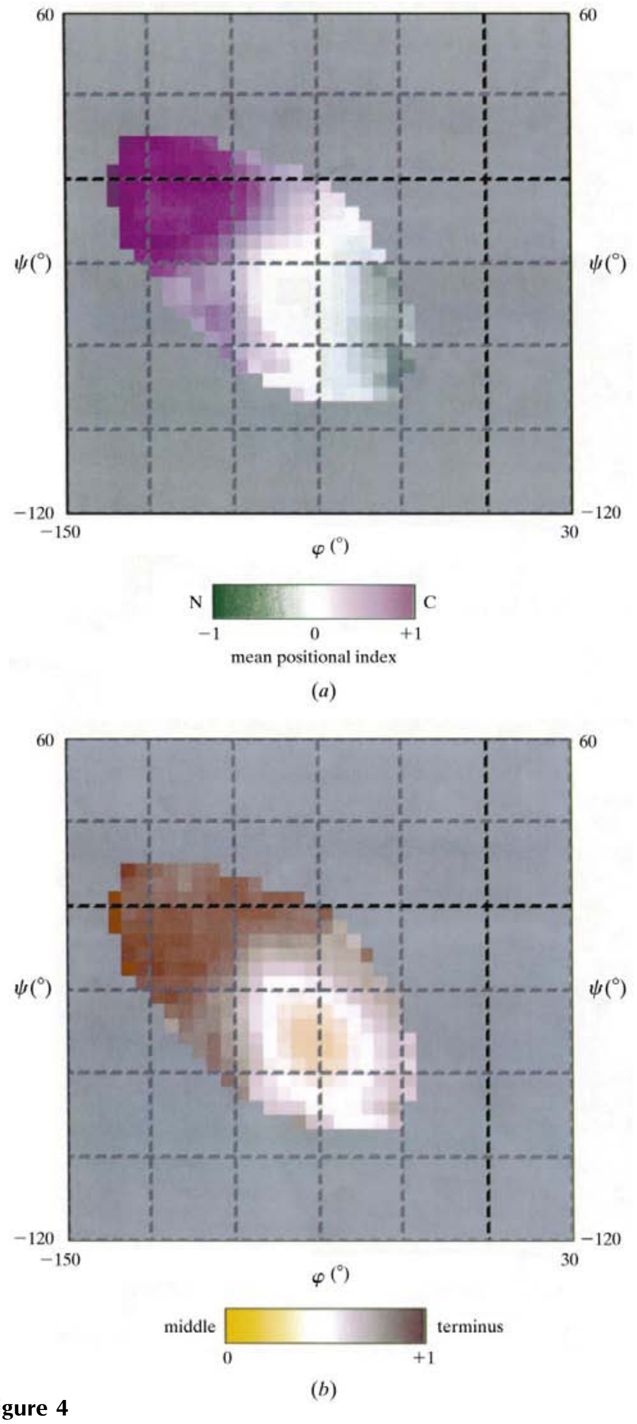


Figure 4
Conformational preferences for α-helical residues at terminal or middle positions along the helix path. Maps show mean values of positional indices for α-helical residues as a function of φ/ψ encoded by the color spectrum displayed underneath the maps. In (*a*) the four N-terminal helical residues were assigned a positional index of '−1', the four C-terminal helical residues '+1' and all others '0'; in (*b*) the four N- and four C-terminal residues were assigned the index '+1' and all others (central positions) '0'. Only helices with nine or more residues were considered in both analyses. Data represent statistics for 15 × 15° bins shifted at a 5° register with at least 25 measurements contained in each cell.

The shifts in conformational preferences that we have identified may originate from a variety of sources, including refinement protocols and structural flexibility. As the differential Ramachandran map for α-helices (Fig. 3) shows the strongest dependence upon resolution, we first discuss trends in α-helices in greater detail.

4.1. Focusing on the α-helix

One plausible origin for non-canonical α-helical φ/ψ angles (red regions in Fig. 3, α-helix) could be the introduction of bends into straight helices. However, φ/ψ values for residues in curved helices were not observed to specifically fall into the detected depleted φ/ψ regions. Next, we examined the relative position of such 'outlier' residues along the helix path. Figs. 4(*a*) and 4(*b*) indicate that the N-terminal, middle and C-terminal residues occupy distinct φ/ψ regions. By compar-

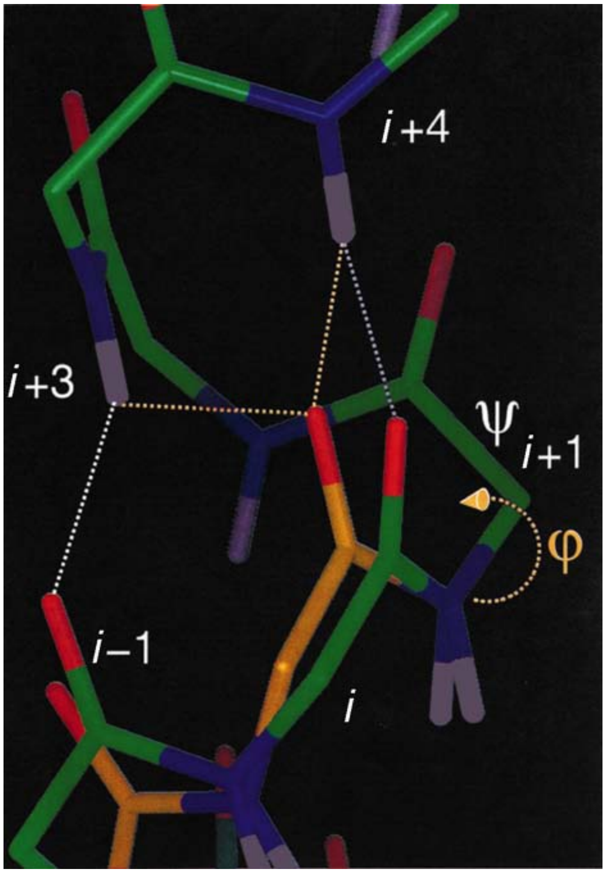


Figure 5
Illustration of the structural consequences of smaller absolute φ-dihedral angles in α-helices. A perfectly regular α-helix with $\varphi = -65^\circ$ and $\psi = -40^\circ$ is shown by the green backbone C atoms. Upon adjusting the indicated φ towards smaller absolute values ($\varphi = -42^\circ$), a kink is introduced bringing the carbonyl O atom of residue *i* closer to the amide H atom of residue *i* + 3, with the new structure represented by the orange carbon-backbone trace. O atoms are colored red, N atoms blue and H atoms white. This figure was produced with *InsightII* (Biosym).

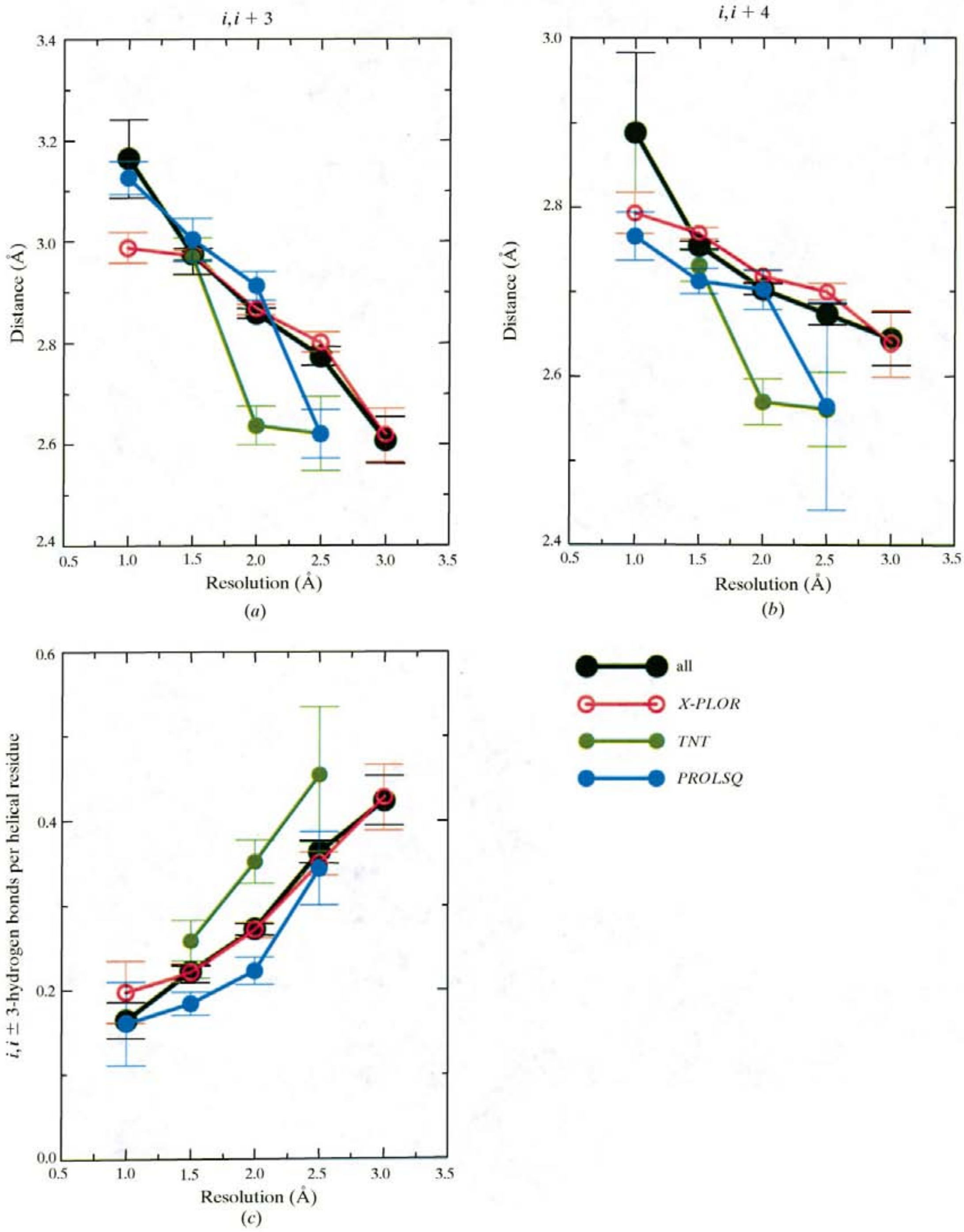


Figure 6

Mean minimal hydrogen-bond donor–acceptor distance between residues *i* and residues *i* + 3 (*a*) and *i* + 4 (*b*) in α-helices as a function of resolution. For each protein in the database, the shortest distance between any carboxyl O atoms (acceptor) of residue *i* and the main-chain amide N atom of residues *i* + 3 and *i* + 4 were determined with both residues required to be in α-helical conformation. To ensure that the intervening residues are also in α-helical conformation and belong to the same α-helix as the examined residue pair, it was required that the carbonyl bond (C=O) vector of each intervening residue opened an angle to its preceding C=O bond vector no greater than 70° (tco data column in *DSSP* datafiles). Raw data (minimal donor–acceptor distance in a given protein) were binned into 0.5 Å wide resolution intervals. Error bars correspond to the observed standard error of the mean in each bin. (*c*) Number of *i*, *i* + 3 main-chain–main-chain hydrogen bonds per helical residue as a function of resolution. Definitions of hydrogen bonds were adopted from McDonald & Thornton (1994). At helix termini, detected hydrogen bonds were counted if at least three of the four-residue 'ring' closed by a *i*, *i* + 3 hydrogen bond were assigned α-helical by *DSSP*. For the N-terminal residue of a given α-helix, the carbonyl O atom of the preceding non-helical residue was considered as belonging to this N-terminal helical residue because this carbonyl O atom is greatly influenced by this residue's φ-dihedral angle (see Fig. 5). Likewise, the NH group of the first non-helical residue following the C-terminus of an α-helix was considered as belonging the C-terminal helical residue as it is influenced directly by the last helical residue's ψ-dihedral angle. Hydrogen bonds per residue were counted for both the potential donor (N—H) and the potential acceptor (C=O) atoms. Raw data (mean number of *i*, *i* + 3 hydrogen bonds per helical residue per protein determined at a given resolution) were binned into 0.5 Å wide resolution intervals. Error bars correspond to the observed standard error of the mean in each bin.

ison with Fig. 3 (α-helix), it can be concluded that residues at the N-terminus of α-helices are more likely to fall into the red (increasingly depleted) sub-regions. By contrast, central residues are most regular and adopt conformations consistent with the α-attractor. C-terminal positions prefer different main-chain angles (north-west of α-attractor) and fall into an area with no obvious correlation with resolution.

What are the structural consequences of non-canonical φ values ($\varphi > \varphi_{\alpha\text{-helix attractor}}$) that have been observed to occur most frequently at helical N-termini? Fig. 5 illustrates that such φ angles introduce a kink in the helix disrupting the regular $i, i+4$ hydrogen-bond pattern and thereby terminating the helix when not compensated by adjustments in preceding backbone angles. Greater φ angles also bring the carbonyl O atom of the preceding residue (residue i in Fig. 5) closer to the H atom donated by the main-chain N atom of residue $i + 3$, establishing an additional hydrogen bond. At helical N-termini this polar H atom would normally not be saturated by another helical main-chain carbonyl O atom but by specific capping interactions (Aurora & Rose, 1998) or water. The described effect raises the question as to whether the over-representation of greater (smaller absolute values) φ angles in helices is introduced by an overestimation of electrostatic attraction during refinement of structures determined at low resolution. The utilization of atomic force fields (including non-covalent terms) is a matter of some controversy amongst crystallographers, as the danger of biases encoded by the parameters of the force field appears realistic. We compared the statistics of intrahelical main-chain–main-chain hydrogen bonds for three refinement packages: *X-PLOR* (Brünger *et al.*, 1987; Brünger, 1992*b*), *TNT* (Tronrud *et al.*, 1987) and *PROLSQ* (Hendrickson & Konnert, 1980). *X-PLOR* allows the refinement of protein models against an emperical energy function including electrostatic terms, whereas the latter two are pure least-squares refinement packages. Observed minimal hydrogen-bond donor–acceptor distances for residues in relative sequence positions i and $i + 3$, and i and $i + 4$ are shown in Figs. 6(*a*) and 6(*b*), respectively. It would naturally be expected that even with constant mean values corresponding minimal distances increase with better resolution and the standard deviation of donor–acceptor distances can be expected to drop simultaneously. However, the extent to which donor–acceptor distances, in particular for i, $i + 3$ hydrogen bonds, increase with improved resolution is remarkable. Accordingly, the frequency of $i, i + 3$ main-chain–main-chain hydrogen bonds per helical residue also decreases sharply with improved resolution. All three individual packages seem to converge to similar values at better resolution, despite the apparent divergence below 1.5 Å, which possibly results from data scarcity in this resolution range. Hydrogen-bond donor–acceptor distances were clearly not observed to be distinctly shorter in models refined with *X-PLOR*, refuting the suspicion of biasing influences arising from over-weighted attractive electrostatic potentials during refinement. The same conclusion can be drawn from a direct comparison of differential Ramachandran maps derived from structures refined with either of the three different packages examined here. Regardless of the refinement program used, residues in α-helical conformation in protein structures converged to the same α-helix attractor ($\varphi \simeq -64°$ and $\psi \simeq -40°$), and the cross-correlation coefficients between the individual differential Ramachandran maps were high (Table 3). This consistent behavior of the refinement programs documents the absence of any significant bias encoded by the parameters used to optimize protein backbone geometries.

Table 3
Cross-correlations between resolution-based differential Ramachandran maps obtained from proteins refined with different refinement programs and for residues in α-helical conformation only.

	X-PLOR†	*TNT*‡	*PROLSQ*§
Number of proteins	487	66	105
X-PLOR	1	0.75	0.74
TNT		1	0.69
PROSLQ			1

† Brünger (1992*b*). ‡ Tronrud *et al.* (1987). § Hendrickson & Konnert (1980).

Instead, the over-representation of certain helical backbone angles at low resolution may be related to anisotropies of intrahelical steric interactions. A qualitative van der Waals potential energy as a function of the backbone conformation of a central helical residue is mapped onto the φ/ψ diagram in

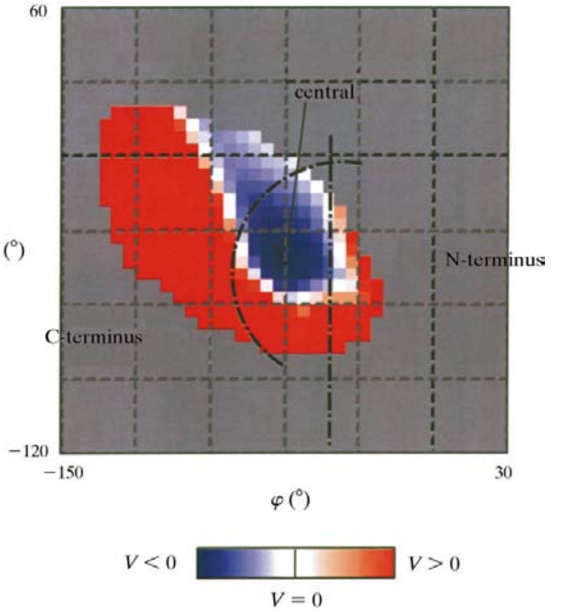


Figure 7
van der Waals energy of kinked α-helices. A perfectly regular 11-mer polyalanine α-helix was created with the standard settings of the builder module of *InsightII* (Biosym) with $\varphi = -65.0$, $\psi = -40.0$ and $\omega = 180.0°$. Backbone dihedral angles φ and ψ were then systematically scanned for residue six along the helix-peptide path and the van der Waals energy of the peptide calculated for each φ/ψ combination consistent with allowed α-helical conformations in Fig. 3 (α-helix panel). An arbitrary and unitless Lennard–Jones type potential V was defined to qualitatively measure the van der Waals energy as a function of backbone conformation of residue six with $V = (\sigma/\rho)^{12} - (\sigma/\rho)^{6}$, where ρ is the distance between any two non-covalently bonded heavy atoms and σ is the separation distance corresponding to the pairwise energy minimum, with σ arbitrarily set to 2.8 Å. The segmentation of the α-helical Ramachandran area into C/N-terminal and middle regions follows the results of Fig. 4(*a*).

Fig. 7. The determined α-helical attractor (Fig. 3) corresponds to low van der Waals energy, while surrounding regions are less favorable and may even result in steric conflicts (encoded red). Backbone conformations north of the canonical α-helix attractor, however, do not cause steric overlaps and are thus tolerable even as unique occurrences and in central positions in otherwise regular helices. Conversely, φ/ψ combinations causing steric conflicts strictly require compensations in other helical backbone dihedral angles or may occur only at helical termini (C-terminus in particular, φ/ψ values 'west' of α-helix attractor). When compensatory effects are the prerequisite for non-canonical φ/ψ values to occur, it is clear that such necessary compensations are more likely to be introduced as imprecisions in low-resolution structures. We conclude that closest allowed distances between atoms have a substantial impact on tolerable φ/ψ backbone angles, and that the differences between refinement packages observed in Fig. 6 may also be attributable to subtle differences in the treatment of such closest allowed interatomic distances.

4.2. Conformational flexibility

It is possible that high-resolution structures display convergence of their backbone conformations, as only structures with this property can diffract to such high resolution. In contrast, broader distributions for less well resolved structures might reflect true conformational flexibility while precluding collection of high-resolution data.

The standard measure of conformational flexibility on the atomic level is the B factor or Debye–Waller factor. Fig. 8 shows the mean standardized B factor (see §2.5) as a function of backbone conformation. Hydrogen-bonded structures (α-helices, β-sheets and other smaller regions) are characterized by lower B factors, *i.e.* greater rigidity, while, for instance, the polyproline-II region (Fig. 1*a*), although identified as a conformation attractor (Fig. 3), displays only medium B factors. A comparison with the general resolution-based differential Ramachandran map (Fig. 1*b*) reveals a coarsely similar variation with backbone conformation. The measured cross-correlation coefficient (§2) between both maps was determined as $r_{\rm cross} = 0.38$. Similar magnitudes of cross-correlation coefficients were obtained between differential Ramachandran maps and mean B-factor maps for sub-classifications of residues according to Fig. 3: $r_{\rm cross,\ helix\ \&\ strand} = 0.33$, $r_{\rm cross,\ 3_{10}\text{-helix}} = 0.35$ and $r_{\rm cross,\ no\ secondary\ structure} = 0.46$, respectively. These correlation coefficients support the view that backbone segments with higher B factors are more likely to be assigned conformations outside the identified conformational attractors. It is not clear, however, whether higher B factors indeed correspond to true flexibility or merely reflect positional errors. Thus, the relationship between conformational flexibility and conformational preferences has to remain inconclusive.

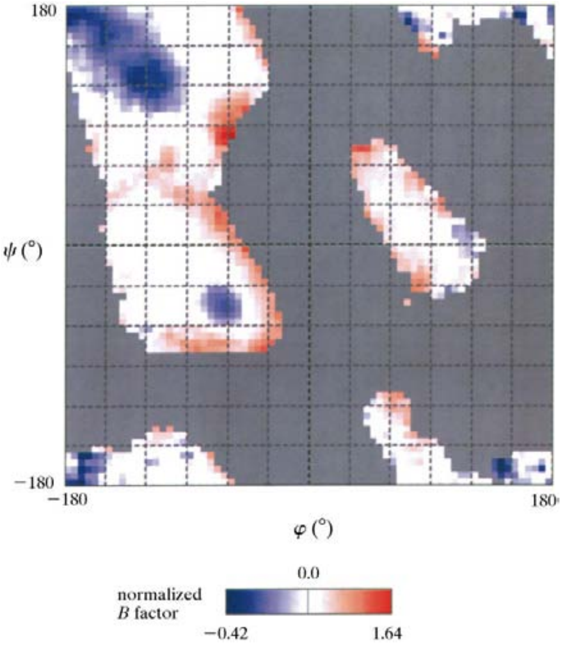


Figure 8
Mean standardized temperature B factor as a function of backbone conformation. Data represent statistics of B factors for the Cα atomic position of a each residue in the database with backbone conformation φ and ψ and compiled in 15 × 15° bins shifted at a 5° register with at least 25 measurements contained in each cell.

4.3. Peptide-bond planarity

Traditionally, the backbone dihedral angles φ and ψ have been assigned greatest importance in determining the conformation of polypeptide chains, whereas bond angles and distances can be assumed as largely constant. The third distinct dihedral angle along a polypeptide chain, ω, which measures the planarity of the peptide bond, is commonly assumed as 180° (for *trans*-peptide bonds) in agreement with the partial double-bond character of the peptide bond. However, significant variations in ω have been observed in high-resolution structures with a mean value slightly below 180° and a standard deviation of about 6° (MacArthur & Thornton, 1996). In low-resolution structures, peptide bonds which may in reality be non-planar may initially be modeled into the protein chain as planar. Adjustments to maintain an agreement with the electron-density map may then be propagated to φ or ψ, possibly explaining trends out of certain φ/ψ regions. In a survey on peptide-bond planarity, MacArthur & Thornton (1996) reported an unexpected dependence of ω on φ and ψ. In Fig. 9, we provide an analysis of that dependency for both the preceding and the succeeding ω angle relative to φ and ψ associated with a central residue. Clearly, the mutual dependence of main-chain dihedral angles cannot be ignored. Furthermore, there is a remarkable directionality in the $\varphi/\psi{-}\omega$ relationship, as the map for the preceding ω differs substantially from the map for the succeeding ω. Within α-helices, ω values smaller than 180° can be rationalized with regard to carbonyl bonds then pointing more outwards from the helix axes and allowing the formation of additional hydrogen bonds to water or other intra-protein hydrogen donors. An explanation for consistent deviations from peptide-bond planarity for other regions of the Ramachandran map is less obvious and a separate investigation appears worthwhile. In conclusion, strictly planar peptide bonds may often be in error and may contribute to the φ/ψ frequency shifts reported in this study.

4.4. *R* factor *versus* free *R* factor

We observed that the convergence of backbone dihedrals to distinct conformational attractors with improved resolution is not equally sensed by the *R* factor as compared with the free *R* factor as a structure-quality indicator (Table 2). In our integrated approach of analyzing general trends in the Protein Data Bank, this observation adds to the reports on failures to correctly recognize structural errors in individual protein structures (Kleywegt & Jones, 1995). The *R* factor can be diminished artificially by overfitting the data, for instance by placing solvent molecules or by excluding reflections, without actually improving the accuracy of the protein model (Kleywegt & Jones, 1997). The better suitability of the free *R* factor as a quality criterion has shown by several studies (Brünger, 1997) and is documented here from a different more generalized perspective.

Finally, it needs to be cautioned that non-canonical φ/ψ values may reflect real strained backbone conformations (Gunasekaran *et al.*, 1996; Karplus, 1996) which cannot be relaxed as other stabilizing interactions dominate, and may not be identified as errors automatically. The recognition of such real strained conformation does, however, require a thorough examination of each individual case.

5. Conclusions and availability

Differential Ramachandran maps have been introduced which correlate the relative frequency of main-chain torsion angles with crystallographic parameters, in particular with the resolution at which the proteins were determined. Despite the conceptual simplicity of this approach, the following insightful information was yielded from this analysis.

(i) As the resolution of protein structure determination improves, backbone conformations are increasingly drawn towards only a few conformational attractors revealed in the differential Ramachandran maps. As this limits the conformational space, computational methods for modeling protein structures should benefit.

(ii) Frequency distributions of φ/ψ angles were observed to still be subject to change, even at the present state-of-the-art resolution level. Therefore, the true frequency distribution remains to be determined.

(iii) Helical N-termini preferably adopt regular α-helical conformations in high-resolution structures.

(iv) The polyproline(II)-state represents a strong conformational attractor for residues not involved in regular main-chain–main-chain hydrogen-bond networks.

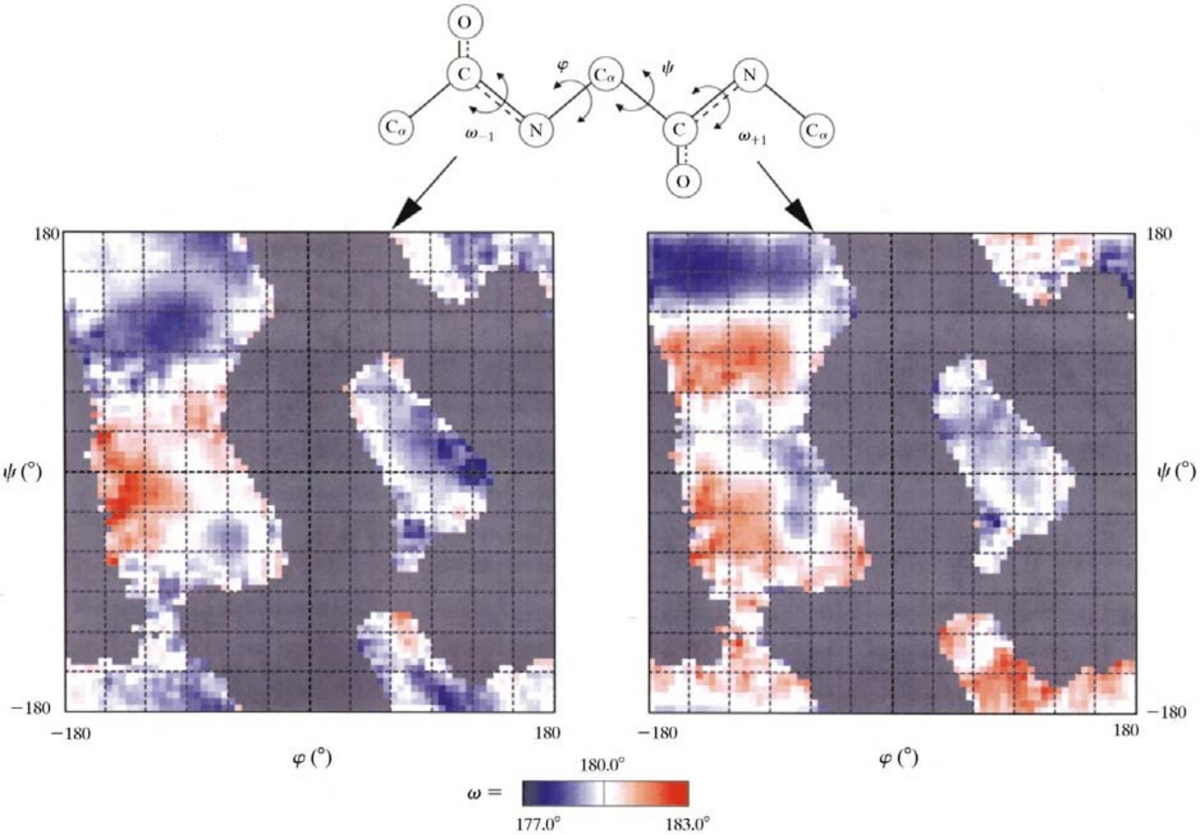


Figure 9
Mean main-chain dihedral angle ω as a function of φ and ψ of the succeeding (left panel) or preceeding (right panel) amino-acid residue. ω is defined as the dihedral angle created by the successive atomic positions of $C_{\alpha,i}$, C_i, N_{i+1} and $C_{\alpha,i+1}$, where *i* denotes the residue number. Data represent statistics for 15 × 15° bins shifted at a 5° register with at least 25 measurements contained in each cell.

(v) The free R factor provides a more sensitive measurement of structural quality than the conventional R factor.

The general resolution-based differential Ramachandran map (Fig. 1*b*) has been integrated into *WebMol*, an online Java-based PDB viewing and analysis program (Walther, 1997; http://www.cmpharm.ucsf.edu/cgi-bin/webmol.pl). As actual backbone conformations can be compared with reference to this map, it may facilitate the identification of suspicious backbone conformations and may therefore prove a helpful tool during the structure-determination process and in protein model building. Differential Ramachandran maps for all 20 amino-acid types are available at http://www.cmpharm.ucsf.edu/~walther/map.html.

The work was supported by a grant from the National Institutes of Health. DW was supported by a post-doctoral fellowship from the DFG (Deutsche Forschungsgemeinschaft). We thank Sebastian Doniach for helpful comments and suggestions and Clayton Springer and Jenifer Bartle for critically reading the manuscript. We also gratefully acknowledge consultations with Gerald Kleywegt and Dale Tronrud.

References

Adzhubei, A. A., Eisenmenger, F., Tumanyan, V. G., Zinke, M., Brodzinski, S. & Esipova, N. G. (1987). *J. Biomol. Struct. Dyn.* **5**, 689–704.

Adzhubei, A. A. & Sternberg, M. J. E. (1993). *J. Mol. Biol.* **229**, 472–493.

Aurora, R. & Rose, G. D. (1998). *Protein Sci.* **7**, 21–38.

Barlow, D. J. & Thornton, J. M. (1988). *J. Mol. Biol.* **201**, 601–619.

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

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR.* Yale University, Connectiut, USA.

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Chothia, C. (1973). *J. Mol. Biol.* **75**, 295–302.

Dunbrack, R. (1997). *Culling the PDB by Resolution and Sequence Identity*, http://www.fccc.edu/research/labs/dunbrack/culledpdb.html.

Eisenhaber, F., Adshubej, A. A., Eisenmenger, F. & Esipova, N. G. (1992). *Biofizika*, **37**, 62–67. In Russian.

EU 3-D Validation Network (1998). *J. Mol. Biol.* **276**, 417–436.

Gunasekaran, K., Ramakrishnan, C. & Balaram, P. (1996). *J. Mol. Biol.* **264**, 191–198.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.26. Bangalore: Indian Academy of Sciences.

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *CABIOS*, **13**, 425–430.

Hooft, R. W. W., Scharf, M., Sander, C. & Vriend, G. (1998). *The PDBFINDER Database*, http://www.sander.embl-heidelberg.de/pdbfinder.

Hurley, J. H., Mason, D. A. & Matthews, B. W. (1992). *Biopolymers*, **32**, 1443–1446.

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

Karplus, P. A. (1996). *Protein Sci.* **5**, 1406–1420.

Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.

Kleywegt, G. J. & Jones, T. A. (1996). *Structure*, **4**, 1395–1400.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1996). *Protein Sci.* **5**, 1067–1080.

Laskowski, R. A., MacArthur, M. W., Moss, M. W. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

MacArthur, M. W. & Thornton, J. M. (1991). *J. Mol. Biol.* **218**, 397–412.

MacArthur, M. W. & Thornton, J. M. (1996). *J. Mol. Biol.* **264**, 1180–1196.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995). *J. Mol. Biol.* **248**, 374–384.

McDonald, I. K. & Thornton, J. M. (1994). *J. Mol. Biol.* **238**, 777–793.

Milner-White, E. J. (1990). *J. Mol. Biol.* **216**, 385–397.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.

Press, H. W., Flannery, B. P., Teukolsky, S. A. & Vetterlin, W. T. (1988). *Numerical Recipes in C.* Cambridge University Press.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283–438.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B.W. (1987). *Acta Cryst.* A**43**, 489–501.

Walther, D. (1997). *Trends Biochem Sci.* **22**, 274–275.