I. INTRODUCTION

Structural genomics projects have been started in several countries. It is estimated that the solution of approximately 10,000 carefully chosen protein structures would be sufficient to represent each major family of folds [19]. Then it should be possible to model the structures of almost all the proteins. However the models provided today by homology modelling do not have an accuracy approaching those obtained by crystallography or NMR. It is expected that while the required structures are solved, modelling techniques will improve so that the data can be used effectively.

In the approach used in the laboratory for homology modelling, the first step is to find homologous proteins of known structure which may be done by aligning the sequence of amino acids of the protein with sequences of all the proteins from the Protein Data Bank. Then it is possible to predict the structurally conserved parts of the structure of the new protein by copying the structures corresponding to similar sequences of residues. The links between these structures can be determined by ab initio calculations or by copying loops from databases [5,6,18]. Nevertheless the models produced by this method often contain errors, which may be due to a wrong alignment, incorrect modelling or movement of secondary structural elements in the target relative to the parents. Additionally the quality is dependant on the number of parents that can be used.

In order to improve the quality of the models it is necessary to understand how the backbone of a structure changes with a change in the sequence, especially in secondary structures. It is thus important to be able to accurately predict these secondary structures in known proteins to study the influence of a change in the sequence on the structures. Before this may be studied, secondary structure must be assigned from the three dimensional coordinates.

Secondary structures are characterized by a certain geometry which is the consequence of a network of hydrogen bonds between the >C=O group of residue n and the >N-H group of another residue m. For example in an alpha helix m = n+4 and in a 310 helix m = n+3. The presence of hydrogen bonds has often been exploited to develop algorithms assigning secondary structure elements based on the calculation of hydrogen bond energies [8,10]. Some other programs use geometric recognition of secondary structures [11,15,20,21]. The geometric features employed are numerous and quite different. The program xtlsstruc [11] for example uses the angles between three consecutive C_α, the dihedral angle between two consecutive carbonyl groups and distances between atoms to determine helices and strands. The program P-curve [20] is based on an algorithm defining an axis along the protein and determines the structures using parameters relative to this axis. However the programs which are the most widely used are Stride and especially DSSP based both on the calculation of hydrogen bonds.

The program DSSP [10] calculates energies of hydrogen bonds using a classical electrostatic function. The residues are assigned in a secondary structure category depending on the type of hydrogen bonding they take part in. The obligation to be involved in two hydrogen bonds for a residue in the middle of a helix is very restrictive and may cause breaks in helices. To reduce this problem the algorithm gathers two helices which have an overlap according to the definition of helices by the algorithm. The result of this is the definition of a single, highly distorted helix when, in fact, a human would define two distinct helices. What is more, there is no restriction on the backbone torsion angles Φ and Ψ and many of residues assigned in a secondary structure have Φ, Ψ angles incompatible with this conformation (see figure 1).

The program Stride [8] also calculates energies of hydrogen bonds but uses a different energy function taking the backbone torsion in account with a term dependent on the probability for a residue to have such (Φ,Ψ) in a given secondary structure (for alpha helix or beta strand). This results in the elimination of many of the false positives due to a wrong (Φ,Ψ) for alpha helices and beta strands, but not for the 3_{10} residues for which no restriction is applied to the backbone angles. Although Stride can be considered as an improving on DSSP, it also defines many residues as being in secondary structure when, in fact, they are not.

The assignments provided by all these programs are more or less equivalent and globally give an account of the real secondary structures. But in the detail there are substantial differences between all the assignments. Thus to improve the accuracy of the recognition of secondary structure we have several possibilities. The first is to make a consensus of the assignments provided by each program [3]. The second solution is to develop a new algorithm of secondary structure assignment.

We have decided to develop a new algorithm based on geometric features to assign secondary structures. We think that this geometric approach can produce improvements on the previous methods. We have based our algorithm on the assumption that the backbone of each type of secondary structure may be fitted inside a sum of small cylinders along an axis, which was to be determined. Additionally Φ and Ψ are used. This is incorporated in a new program SEGNO.

II. METHODS

Brief introduction

The program Segno is based on an algorithm using geometric parameters to define strands and helices. The first task was thus to define which parameters were the most able to distinguish the secondary structures. Strands and helices have several geometric characteristics [16] but some of them are poor at distinguishing different types of structure. Additionally it is desirable to use as few parameters as possible.

The parameters have been chosen by visual examination of protein structures and the optimisation of the cut-offs has been made over a dozen proteins. Firstly secondary structures were defined according to geometric parameters linked with an approximate axis (this approach is derived from the one used by Richardson and Richardson [17] to define the ends of helices). The axis of the structures is approximated by linking the mean positions of the C_α along the protein (the axis position related to residue n is defined by $A_n = \frac{1}{4}\sum_{j=n-1}^{n+2} C_{\alpha j}$). Two parameters allow the definition of secondary structure elements: the radius, distance between the C_α of the residue and the axis, and the angle τ formed by the radius and the axis of the structure (see figure 2). Additionally we calculate the dihedral angles between different amide planes, according the structure we want to define. For helices, we can distinguish 3_{10} helices from alpha helices by comparing the dihedral angle ω_3 between the amide plane n and the amide plane n+3 with the dihedral angle ω_4 between the amide plane n and the amide plane n+4. For beta strands the dihedral angle used is the one between the carbonyl group n and the carbonyl group n+1 (this angle is called ω_1). We have also used other parameters such as Φ and Ψ (see figure 1), the backbone torsion angles, in order to eliminate some false positives (residues incorrectly assigned as helical or strand). The table 1 recapitulates all the parameters used by Segno to define secondary structure elements.

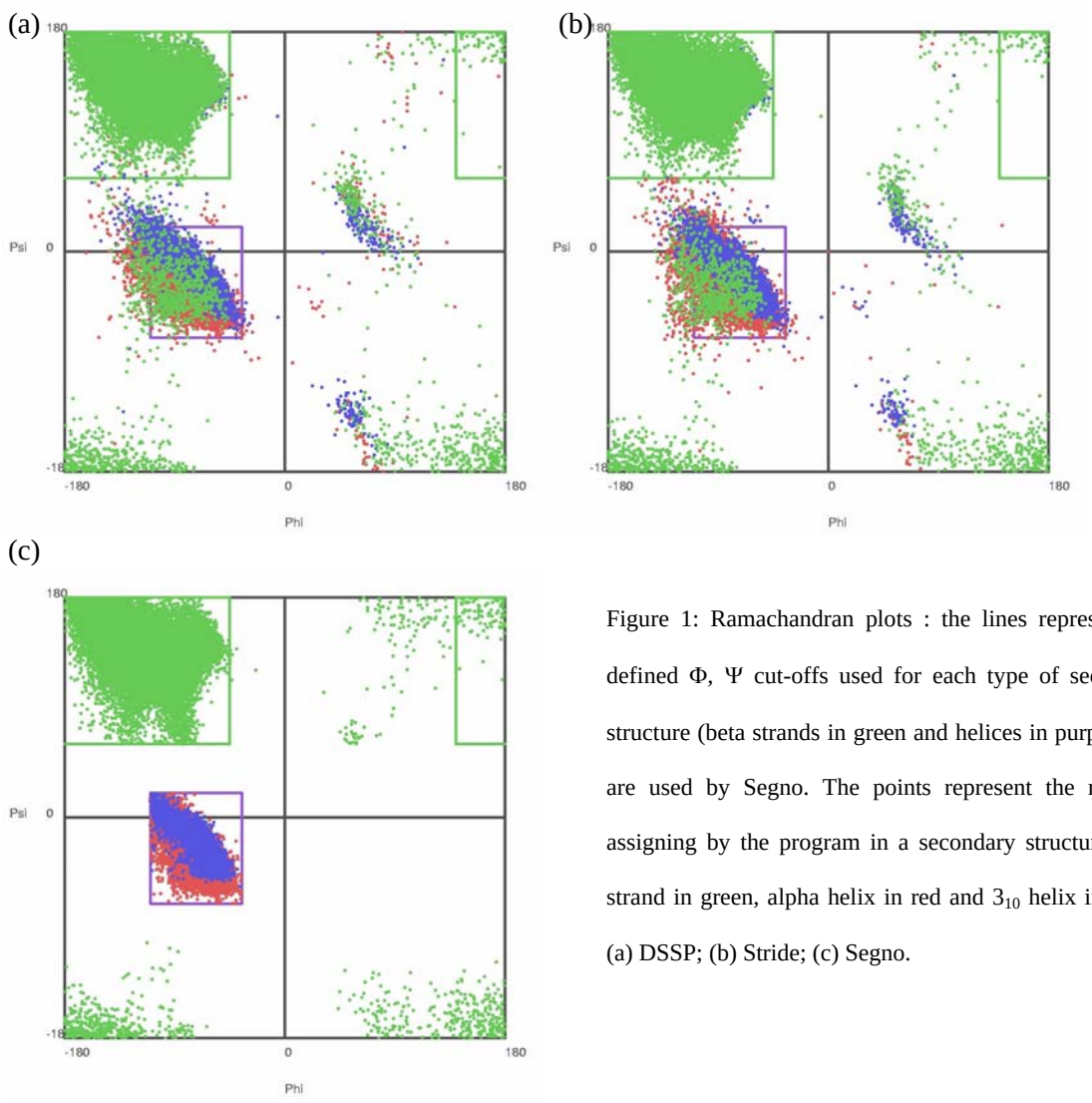


Figure 1: Ramachandran plots : the lines represent the defined Φ, Ψ cut-offs used for each type of secondary structure (beta strands in green and helices in purple) that are used by Segno. The points represent the residues assigning by the program in a secondary structure (beta strand in green, alpha helix in red and 3_{10} helix in blue). (a) DSSP; (b) Stride; (c) Segno.

Recognition of helical residues

There are several sorts of helices in proteins, for example α helices, 3_{10} helices and π helices. However they have some similar characteristics. It is therefore sensible to assign them first as helical, and then to distinguish between the different types. Segno currently assigns only alpha and 3_{10} residues but will assign the much rarer π residues in the near future. The parameters used to define helical residues are the radius, τ, Φ, and Ψ (see figure 2). The cut-offs are then adjusted in order to include all the different kind of residues in a first step: (1) the radius (noted r) must be between 1,7 and 3,0 Å, (2) τ must be between 75 and 120 degrees, (3) Φ must be between –110 and –35 degrees, (4) Ψ must be between –70 and 20 degrees. All of these cut-offs were determined empirically.

Though these cut-offs are not overly strict we have additional problems at the C-termini of the helices. These problems have two different origins. The first problem is a computational problem: at the end of a helix the axis defined by the mean position of C_α carbons is not as close to the real axis as it is in the middle of the helix because it contains information from non-helical residues. Thus the angle made by the radius and the axis for the three last helical residues may not be in the range of the cut-offs.

The algorithm therefore calculates the complementary angle of τ (noted τ_{-1}), which must define a set of complementary cut-offs. The second problem is that the C-termini ends of helices are more variable than the other helical residues. The reason for this is that the four last helical residues often participate at only one hydrogen bond, whereas the other helical residues participate in two, including the first residues that are very often engaged in a hydrogen bond with side chains. We therefore used less constrained cut-offs for the last three residues of the helix in order to assign them correctly ($50 \leq \tau_{-1} \leq 112$ which corresponds to $68 \leq \tau \leq 130$).

Once the program has defined the helical residues it has to distinguish between α and 3_{10} helices. This distinction is made differently according to the length of the helix because certain parameters cannot be calculated for short helices.

If the length of the helix is only of three residues, then it can only be a 3_{10} helix or a coil, as α helices require n to n+4 hydrogen bonds. Therefore the program checks whether there can be a hydrogen bond between the first and the last residue by calculating the distance between the oxygen of the first residue (n) and the nitrogen of the last residue (n+3): the maximum length for a hydrogen bond in this case is defined as 3.5 Å. If this is the case all the residues of the helix are assigned as 3_{10} residues if their Φ and Ψ are compatible with it.

For helices with four residues, the program checks if a hydrogen bond is more likely to appear between the first residue and the third or between the first and the fourth. This is made by testing if the distance between the oxygen of the first residue (n) and the nitrogen of the third residue (n+3) - this distance is indicated as hbond3 on the figure 2 - is shorter than the distance between the oxygen of the first residue (n) and the nitrogen of the fourth residue (n+4) - this distance is labelled hbond4 on the figure 2 - minus 0.3 Å. If this is the case all the residues are assigned as 3_{10} residues if their Φ and Ψ are compatible with a 3_{10} helix, otherwise the residues are assigned as α helical if their Φ and Ψ are compatible with an alpha residue.

For longer helices the program applies the same criteria as for helices of 4 residues length, but there is another parameter which is used to distinguish the two types of helix: the comparison between ω_3 and ω_4. In the case of a 3_{10} residue ω_3 is closer to 180 degrees (see figure 3). Thus to be a 3_{10} residue a helical residue must have $\omega_3 > \omega_4$. As for the determination of helical residues, these tests are made forwards and backwards along the peptide chain (by looking at the residues n-3 and n-4 for the residue n) so that every residue can be tested, included the last residues in the helices.

Finally when all the 3_{10} residues have been assigned, the rest of the residues except the last residue of each helix are assigned as α residues if their Φ and Ψ are compatible. Indeed, as discussed above, the last residue of a helix is more likely to have unusual Φ or Ψ; and so less stringent cut-offs are required.

The helices are then reassigned so that a 3_{10} residue cannot be alone between two α residues and vice versa. The program then eliminates 3_{10} helices with less than three residues and α helices with less than four residues.

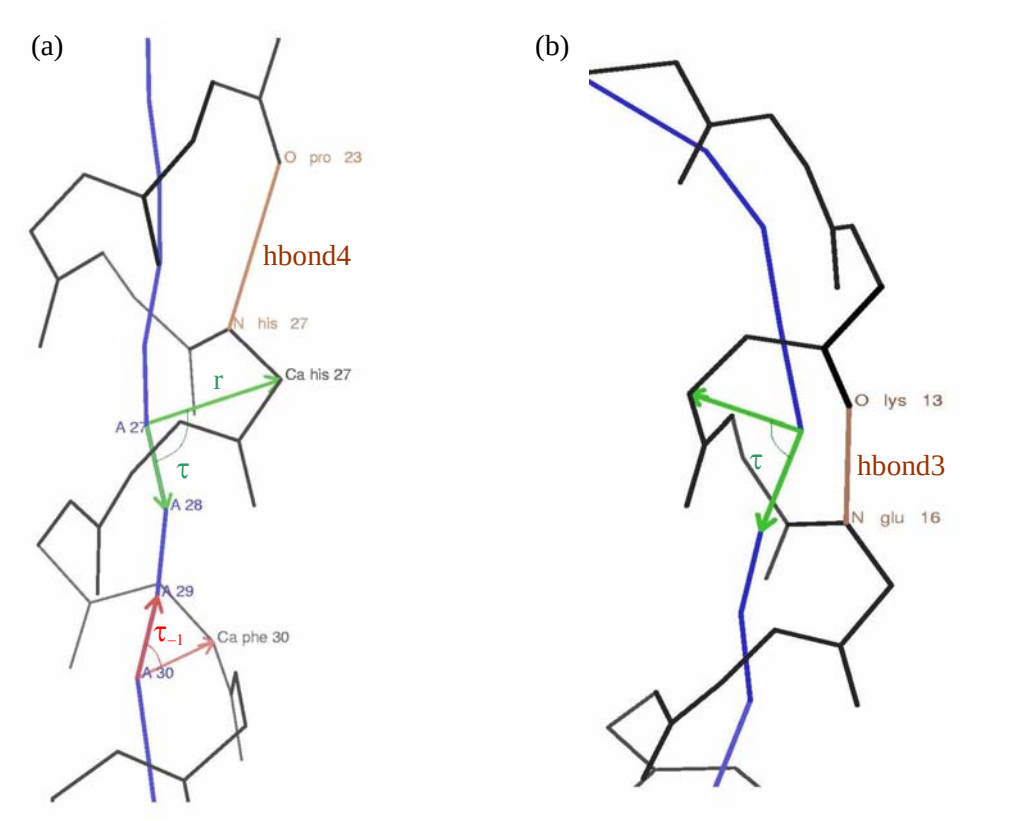


Figure 2: (a) α helix taken from the file 1gdj (res 23 - res 31); (b) 3_{10} helix taken from the file 1bgf (res 11 - res 18). The axis are shown in blue. r is the green vector linking A_{27} and $C_{\alpha 27}$. τ is the angle between the vectors in green and τ_{-1} is the angle between the vectors in red. The distances hbond3 and hbond4 are shown in brown.

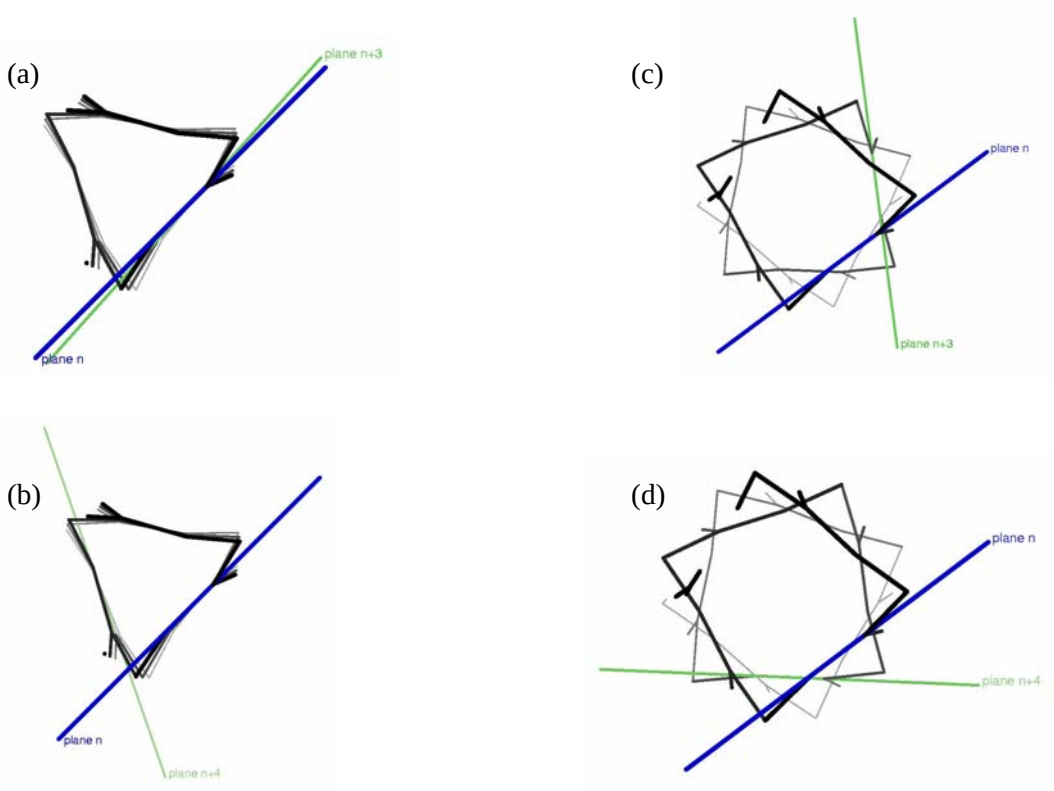


Figure 3: Parameters ω_3 and ω_4: (a) and (b) are pictures of a regular 3_{10} helix looked along its axis. (c) and (d) are pictures of a regular α helix looked along its axis, both produced by the program Sybyl.

Recognition of beta strands

The recognition of beta strands is based on four parameters: the angle τ, the dihedral angle ω_1, Φ, and Ψ. As for helices the algorithm makes two series of tests forwards and backwards along the peptide chain but in this case the two tests are a slightly different.

To determine if the residue n is in a β strand with the residue n+1, τ must be greater than 110 degrees, ω_1 must be between 123 and 210 degrees, and Φ(n+1) and Ψ(n) must be inside the region of a beta strand in the Ramachandran plot (see figure 1). To determine if the residue n is in a strand with the residue n-1, τ_{-1} must be less than 80 degrees, ω_{-1} (the dihedral angle between the carbonyl group n and the carbonyl group n-1) must be between 123 and 210 degrees, and Φ(n) and Ψ(n) must be inside the region of a beta strand in the Ramachandran plot. This difference in the tests, determining which Φ is tested comes from the fact that Φ(n) involves the residue n-1, whereas Φ(n+1) involves the residue n+1 (see figure 4). When this first test has been made, a second test is made at both ends of the strands with less constrained parameters. Finally the strands with less than three residues are eliminated.

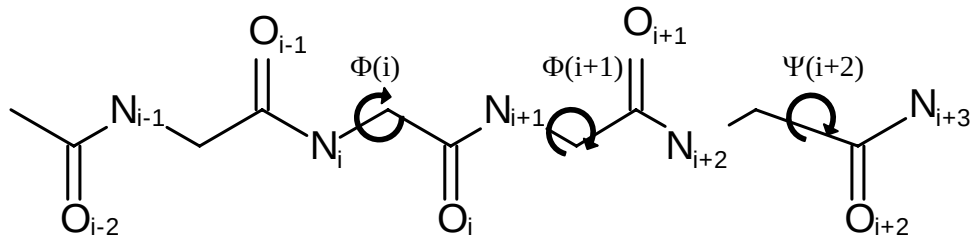


Figure 4: Angle Φ and Ψ: Φ(i) is the dihedral angle around the bond N_i-Cα_i involving the atoms C_{i-1}, N_i, Cα_i and C_i; Φ(i+1) is the dihedral angle around the bond N_{i+1}-Cα_{i+1} involving the atoms C_{i-1}, N_i, Cα_i and C_i. Ψ(i+2) is the dihedral angle around the bond $C_{\alpha i+2}$-C_{i+2} involving the atoms N_{i+2}, $C_{\alpha i+2}$, C_{i+2} and N_{i+3}.

Once all the strands have been found the program determines if they are in a sheet. Two strands are linked in a sheet if they fulfil the following conditions: first there must be at least two hydrogen bonds between the two strands (distance between an oxygen of one strand and a nitrogen of the other strand less than 4 Å); then there must be at least three consecutive distances between two C_α from each strand ($dC_{\alpha 1}$, $dC_{\alpha 2}$, and $dC_{\alpha 3}$) less than 6 Å. Lastly the dihedral angles between these three consecutive $C_\alpha C_\alpha$ vectors (γ_1 and γ_2) must be greater than 135 degrees. The parameters used to define strands and sheets are shown on figure 5.

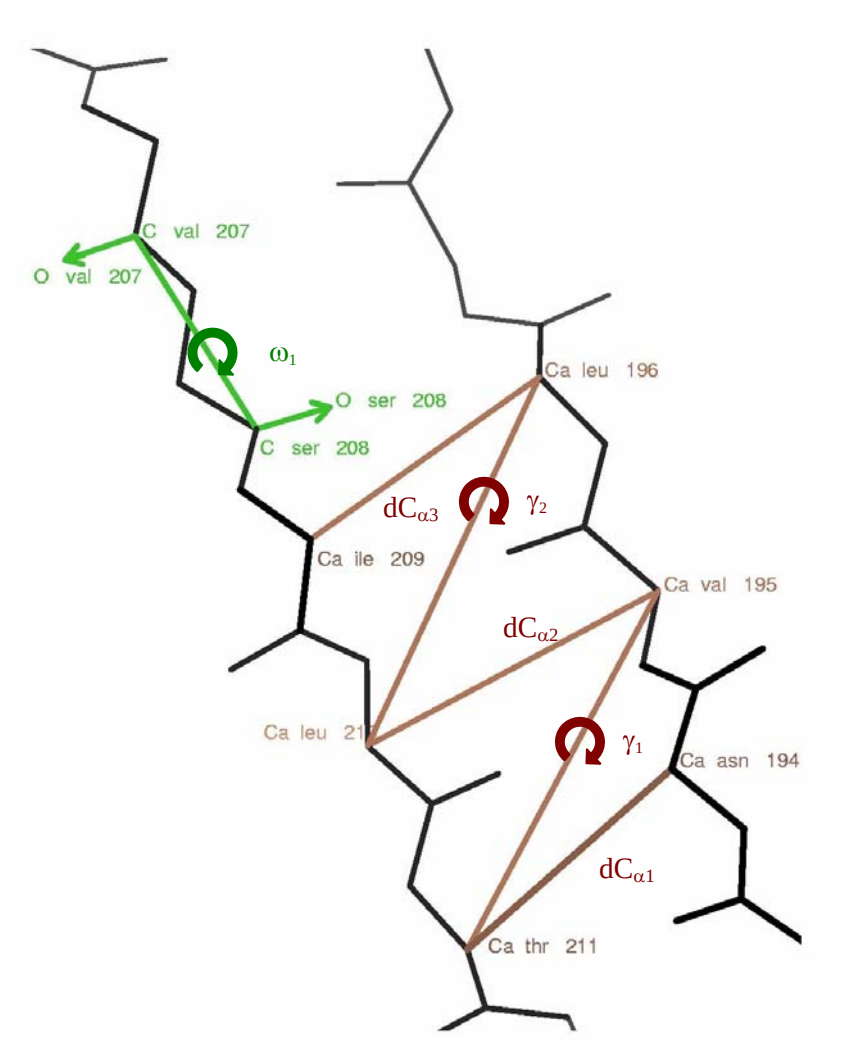


Figure 5: β strands in an antiparallel β sheet taken from the file 1hmp (res 193 - res 198 and res 208 - res 211).

Parameter	Definition	Figure
r	Distance between the axis and the C_α	2
τ	Angle between the radius and the axis	2
τ_{-1}	Complementary of τ	2
Hbond3	Distance between O_n and N_{n+3}	2
Hbond4	Distance between O_n and N_{n+4}	2
ω_3	Dihedral angle between the amide planes n and n+3	3
ω_4	Dihedral angle between the amide planes n and n+4	3
ω_1	Dihedral angle between the carbonyl groups n and n+1	5
$DC_{\alpha i}$	Distance between two C_α from two different strands	5
γ_I	Dihedral angle between two $C_\alpha C_\alpha$ vectors	5
Φ	Dihedral angle around the bond N_n-Cα_1	4
Ψ	Dihedral angle around the bond $C_{\alpha i}$-C_i	4

Table 1: Parameters used by Segno to define the secondary structures

Dataset

We have used a database of 500 structures of better than 1.8 Å resolution which has been developed for a study on the backbone torsion angles (Lovell, S.C., Word, J.M., Richardson, J.S. & Richardson, D.C., unpublished). This database is based on a previous database of 240 structures of 1.7 Å resolution or better [12], augmented by structures of resolution better than 1.8 Å taken from the 30% homology cut-off PDB Select list [9] from February 2000 and new structures of 1.5 Å resolution or better released from February to May, 2000. A selection has then been made, the structures with best combination of clashscore (number of van der Waals overlaps ≥ 0.4 Å per 1000 atoms [22] and resolution if they were related but not identical. The first chain of the protein was chosen if multiple identical chains were present, unless the header indicated that another was better-ordered. The database has been updated with new high-resolution structures (better than 1.5 Å) that has been added if not already in the database, or has replaced structures if they were solved to higher resolution and had better clashscore. Files with multiple, non-homologous chains have been split only if each formed a separate compact unit.

Some other filters have been applied to he database thus obtained. Specifically, structures were rejected if they had a clashscore ≥ 22 for those atoms with B<40, if they had a large number of distorted main chain bond angles (defined as ≥10 main chain bond angles per 1000 atoms ≥5 standard deviations from standard geometry [7], if they had unusual amino acids with main chain substitutions, or if they were subjected to free-atom refinement. Wild type was preferred to mutant if they were otherwise equivalent. Large numbers of B-values ≤1, which is an indication of the use of U^2 rather than B, or unrefined B-values have been looked for; for this data set, however, none were found. This resulted in a data set of 148 files from the previous database, 329 from the PDB Select list and 23 new recently solved files, giving a total of 500.

III. RESULTS

Segno assigns each residue of a protein into one of the four following categories: α helix, 3_{10} helix, β strand, or coil. These residues form secondary structure which can be divided into five categories: α helices, 3_{10} helices, mixed helices (containing α and 3_{10} residues), isolated β strands, and β strands belonging to a beta sheet. The α helices contain at least four residues, the 3_{10} helices and the strands at least three residues.

Once the program was written we benchmarked it. This task is not straightforward because the recognition of secondary structure is not clearly defined and somewhat subjective. Specifically, the exact ends of secondary structural elements can be difficult to define. We have therefore used a number of indirect tests in order to evaluate the performances of our program Segno relatively to the most widely used programs of assignment of secondary structure: Stride and DSSP.

General comparison between the three programs

The first way of looking to the differences between Segno, Stride and DSSP is to compare the assignments provided by each program to the assignment made by the authors of crystallographic structures.

	Segno	Stride	DSSP
Helical residues	93.1%	94.1%	94.4%
Beta strand residues	89.3%	94.7%	94.8%
Global agreement	82.0%	89.3%	89.7%

Table 1: Results of the comparison between authors' assignments and assignments provided by Segno, DSSP and Stride.

For helical residues (α and 3_{10} are not distinguished) the agreement between each program and the authors' assignments is equivalent and good. For strand residues the differences are larger. One first explanation is that Segno's assignment includes isolated single strands. This first result shows that the assignment provided by each program is globally correct. However in spite of the fact that the results of each program are quite similar concerning the comparison with the authors' assignments, it appears that the agreement between each program is less good. Indeed the general agreement between Segno and DSSP is 82.35 % while the agreement between Segno and Stride is 84.10 %. This is the sign that each program assigns the secondary structures quite differently.

It should be noted that, although originally secondary structure was assigned by the authors of a structure, more recently DSSP has often been used to make automated assignments, and therefore a good agreement between DSSP and the PDB assignment is expected. For this reason we have used less direct methods to determine the accuracy of the programs.

C-capping of α helices

The aim of this first test is to find the position of the C-caps of α helices and to correlate it with the assignments given by the three programs. At the C-termini of helices specific sequence and structural motifs often occur [13]. Thus a good correlation between the position of these motifs and the end of the helices is a good indicator of a correct assignment of helices by the various programs.

The C-cap has been defined as the last residue at the C-terminus of the helix whose C_α belongs to the cylinder of the helix by Richardson and Richardson [17]. The motifs at the C-terminus end of the helix have been described for the first time in 1988 [14,17] and consist of hydrogen bonding between the >C=O groups of the last residues of the helix and the >N-H groups of the turn following the helix, which stabilise the helix. Aurora and Rose [1] have published a more comprehensive study of the different C-capping motifs in 1997. Our purpose here is not to identify all the known motifs of C-capping but to produce a test allowing us to judge the results of the three programs. We have therefore chosen to look for the feature which is the most common in the different motifs of C-capping, that is the first residue after the C-cap has a positive Φ.

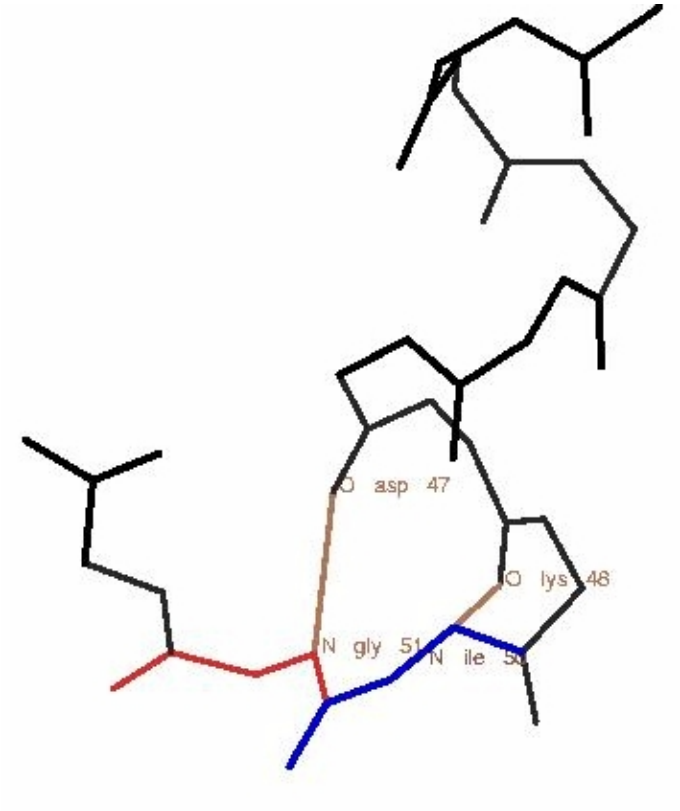


Figure 6: Example of a C-capping motif taken from the file 119l (res 42 - res 52): the C-cap residue is shown in blue and the residue with a positive Φ is shown in red. The two hydrogen bonds of this motif appear in brown.

To find the C-cap we thus have looked for the first residue with a positive Φ at the end of each helix. The C-cap is then the immediately preceding residue and should be the last residue assigned as alpha helical by the programs. The results are summarized in table 3.

Position of the Φ > 0 residue	-3 H	-2 H	-1 H	0 C	1 C	2 C	3 C
Segno	0 (0%)	0 (0%)	0 (0%)	2458 (81%)	277 (9,1%)	182 (6%)	119 (3,9%)
Stride	3 (0,1%)	0 (0%)	1 (0%)	2335 (84,8%)	183 (6,6%)	124 (4,5%)	107 (3,9%)
DSSP	13 (0.4%)	14 (0.5%)	30 (1%)	1911 (63,9%)	709 (23.7%)	193 (6,5%)	119 (4%)

Table 3: Determination of the C-caps of alpha helices: position of the first residue with Φ > 0. The letter under the position number corresponds to the assignment of the residue relatively to the helix concerned. Thus the position -1 is the last helical residue and must correspond to the C-cap, while the position 0 must correspond to the first residue with a positive Φ.

The distribution of the position of the first residue with a positive Φ shows in the three cases a peak for the position 0. However that peak is sharper for Segno and Stride, suggesting a more reliable definition of C-terminal ends of helices.

We can see that we have a much greater number of helices where the C-cap residue is counted as the last residue of the helix for Segno and Stride, while with DSSP the position of the C-cap residue is more widely spread over the different positions at the end of the helix.

Moreover we can see that with DSSP a proportion of residues with a positive Φ are found in helices which is incompatible with the backbone dihedral angles of a residue in an α helix. In contrast there are no residues with positive Φ at the C-terminus end of helices with Segno or Stride α-helical assignments.

N-Capping of α helices

In this second test we are now looking for the N-caps of α helices. The N-cap of an α helix has been defined in 1988 by Richardson and Richardson [17] as the last non-helical residue at the N-terminus of the helix. There are several N-capping motifs reviewed by Presta and Rose [14]. The most common motif consists of a hydrogen bond between the oxygen of the side chain of the N-cap residue (n) and the >N-H group of the residue n+3 (the third residue of the helix). An example of a N-capping motif is given in figure 7. The residues which can adopt the correct geometry to form the hydrogen bond are serine, threonine, asparagine or aspartate. We therefore looked at the beginning of each alpha helix for the presence of one of these three residues involved in a hydrogen bond with the >N-H group of the residue located three residues later. This residue, if it makes the correct hydrogen bond, is the N-cap residue of the helix.

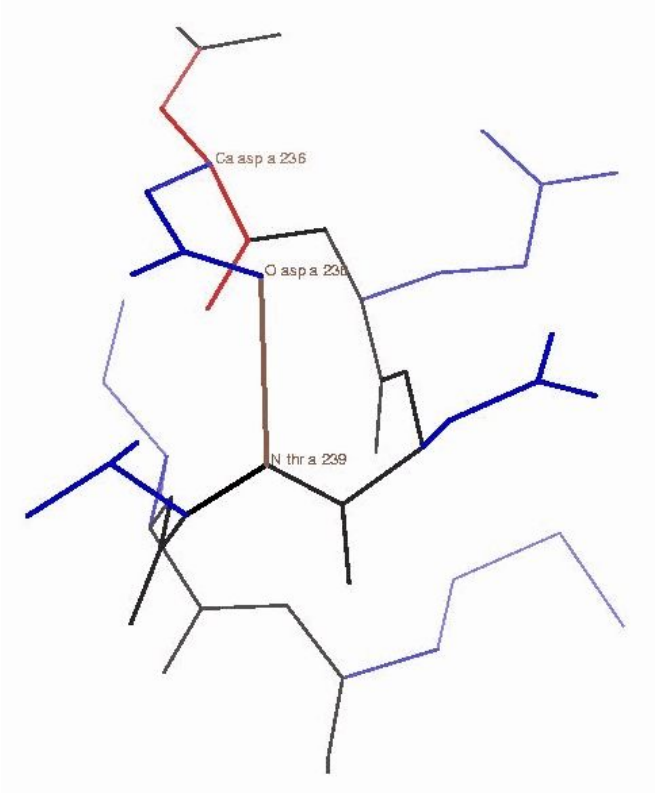


Figure 7: Example of a N-capping motif taken from the file 1bqc (res 236 - res 242): the hydrogen bond between the side chain oxygen of the N-cap residue and the >N-H group of the third helical residue is shown in brown.

The results of the research of the N-cap are summarized in table 4.

Position of the N-cap	-2 C	-1 C	0 H	1 H	2 H
Segno	13 (1.2%)	1049 (98.8%)	0 (0%)	0 (0%)	0 (0%)
Stride	9 (0,9%)	1036 (99,1%)	0 (0%)	0 (0%)	0 (0%)
DSSP	22 (2.1%)	1023 (97.7%)	0 (0%)	1 (0.1%)	1 (0.1%)

Table 4: Research of the N-cap residue. The letter under the position number corresponds to the assignment of the residue relatively to the concerned helix. Thus the position 1 corresponds to the first helical residue and the N-cap residue must correspond to the position -1.

The results for the 3 programs show a very sharp peak for the N-cap in the position -1 (which means the last non helical residue). This suggests that the three programs provide a good assignment at the N-termini of helices.

Bending of helices and strands

Idealized helices and strands are straight, but in the reality secondary structures which occur in proteins have a number of distorsions, including bends. These bends may be due to many factors (steric interaction between side chains, interaction with solvent molecules [2]). However bends are very rarely large in size. In contrast, misassignment of secondary structures can result in apparently large bends, for example in helices where a helix-turn-helix combination is assigned as a single helix. Thus a large number of extreme bends should be viewed with suspicion. Visual inspection is required to distinguish genuinely distorted structures from misassignments.

a) Helices

In this test we calculate bends of α helices at each residue. The bend at residue n is defined as the angle between the axis of an ideal alpha helix of three residues fitted on the residue n-1 and the axis of an ideal helix of three residues fitted on the residue n+1 (see figure 8).

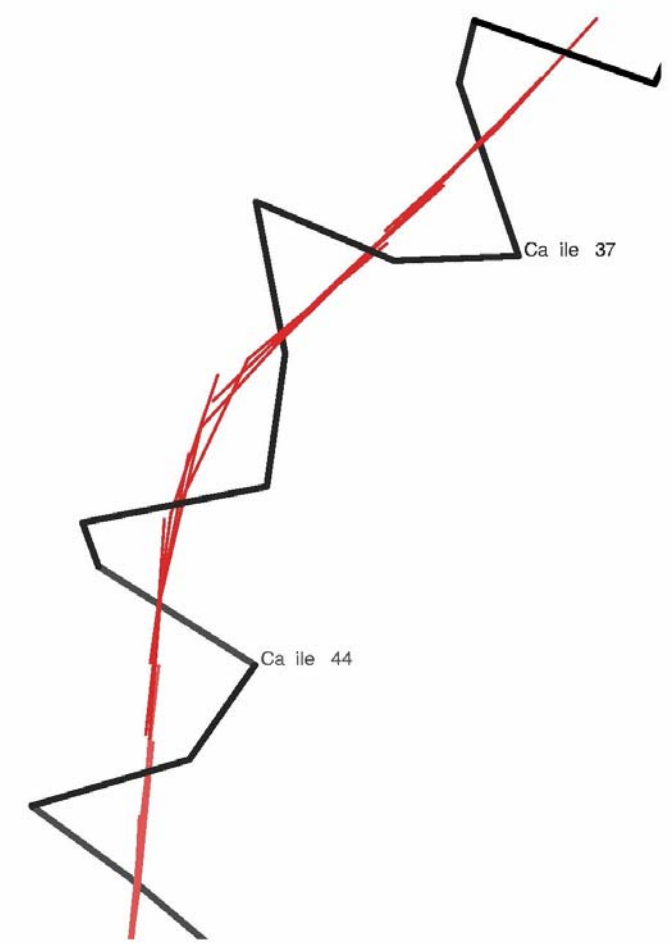


Figure 8: Calculation of the bend of helices: the black line represents the C_α carbons of the chain (taken from the file 1php (res 34 - res 47)). The red segments represent the axis of each short helix fitted on the residues of the helix.

The results of the calculation are shown on figure 9 and 10.

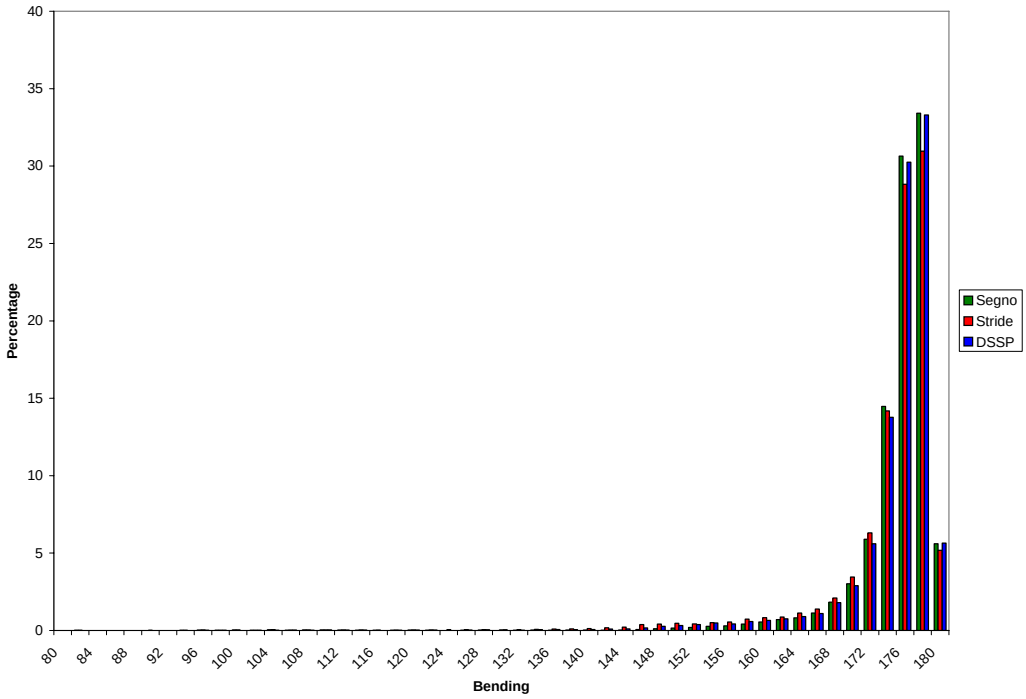


Figure 9: Distribution of the bends in helices.

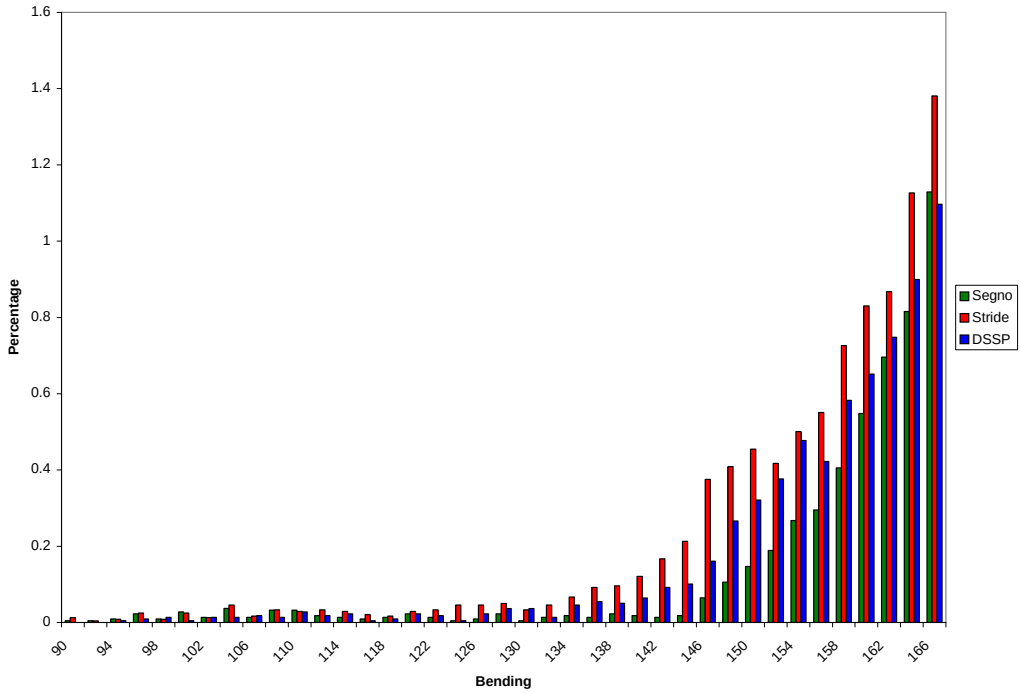
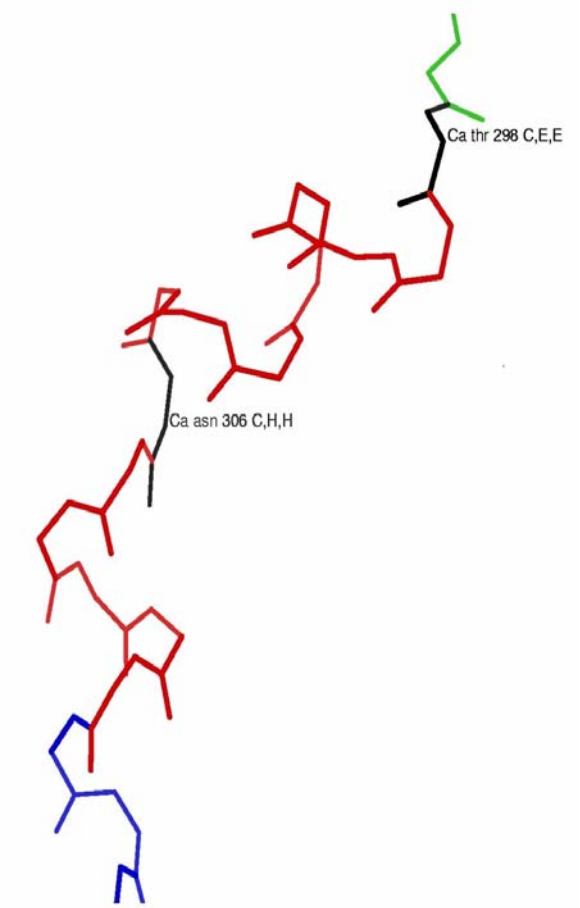


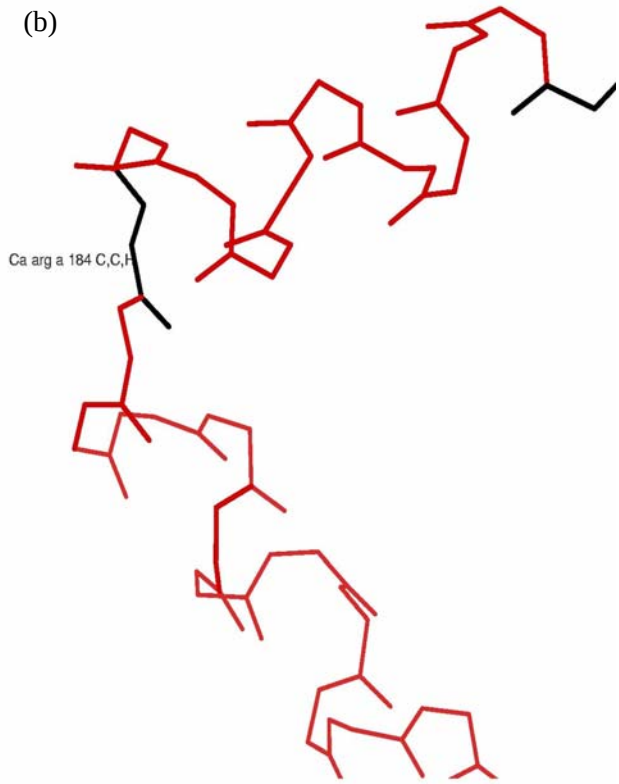
Figure 10: Zoom on the previous graph for the region [90 ; 168]

It can be seen from these graphs that Segno assigns helices with less extreme bends than Stride and DSSP. This is particularly remarkable in the region of bends between 125 and 165 degrees, which correspond to very bent helices. Some examples of the difference in assignment for bent helices between the three programs are shown in figure 11.

(a)



(b)



(c)

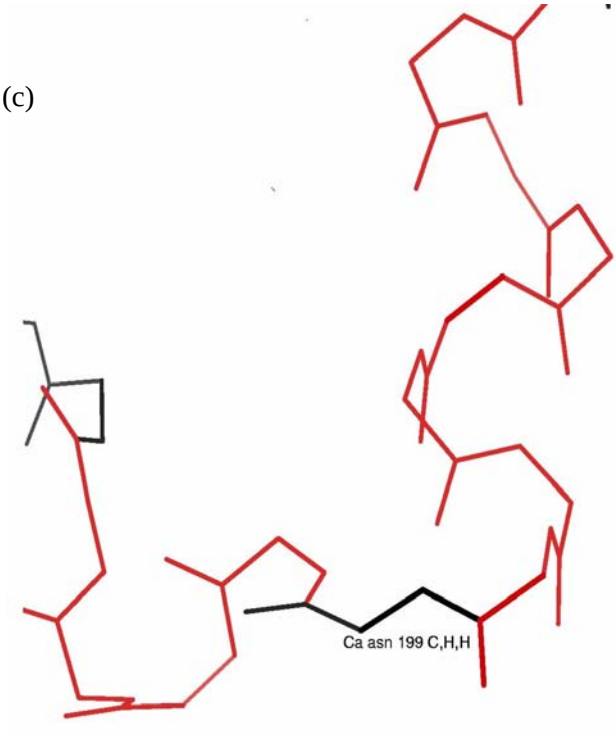


Figure 11: (a) Picture taken from the file 1hpm (res 297 - res 313); (b) Picture taken from the file 1uro (res 174 - res 196); (c) Picture taken from the file 2baa (res 191 - res 204). The residues in red are assigned as α helical by Segno, those in green are assigned as 3_{10} helical, those in green are assigned as strand, and those in black are assigned as coil. The residues labelled are those where the assignment is different for the three programs: the first letter correspond to Segno's assignment, the second to DSSP's assignment and the third to Stride's.

It is interesting to note that all the highly bent helices are not eliminated by Segno. By examining the structures we have seen that such helices are in fact due to 3_{10} residues at the end or the beginning of helices which have been counted as α residue by Segno (see figure 12). Such an assignment is not strictly correct, but the distinction between 3_{10} and α residues is very arguable at the end of helices. The most important is thus to define the helical residues, all the more since lots of studies do not use the distinction between 3_{10} and α residues.

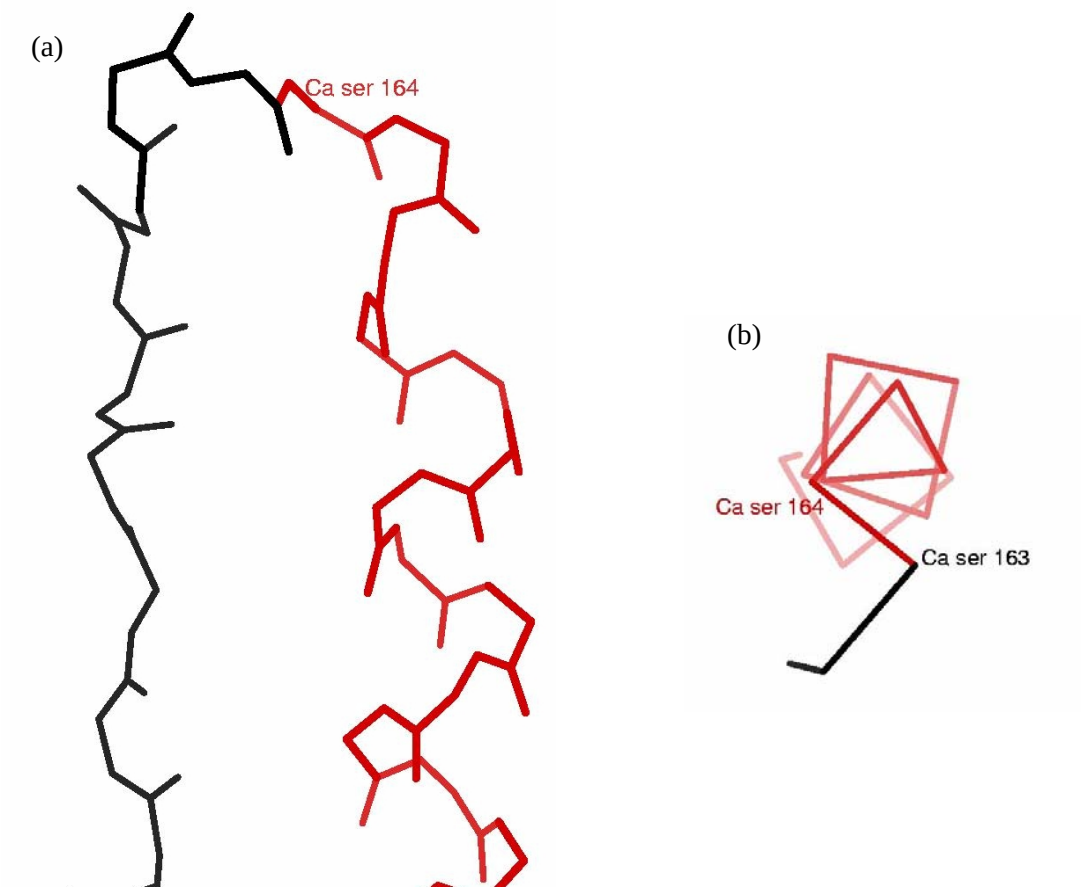


Figure 12: Pictures taken form the file 1aop (res 155 - 177 for (a) and res 162 - res 175 for (b)). The residues assigned by Segno as α helical are shown in red. The picture (a) shows clearly that the helix defined by Segno is not bent although the calculation suggests the contrary. This is due to the fact that the first helical residues have a strong 3_{10} character as it can be seen on the picture (b).

We have also noted that a small number of the bent helices that are assigned as multiple helices by Segno contain a π residue in their middle. This residue is then counted as non-helical by Segno, which leads to break the helix (see figure 13). It would be more appropriate to assign this as a single helix, but it is important to note that Segno is very consistent in its assignment, so that a π helix is never counted as α or 3_{10} residue. It should therefore be easy to modify Segno to recognize π helices.

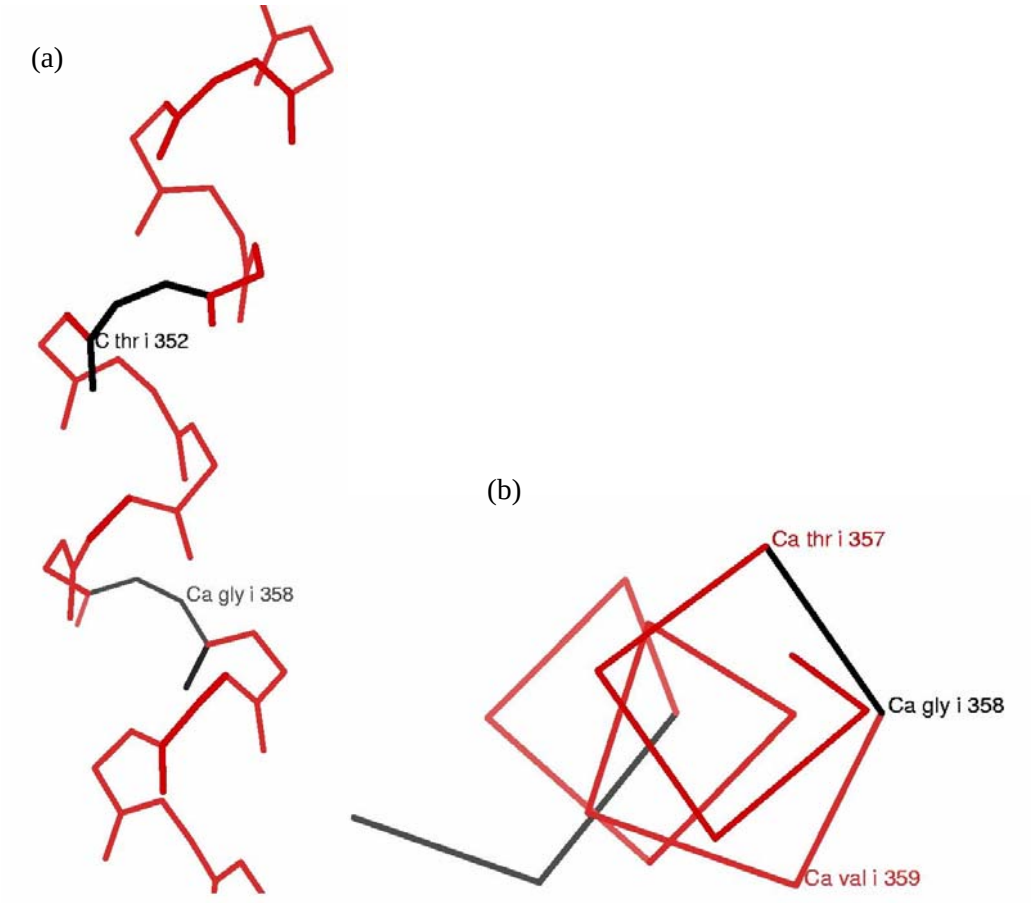


Figure 13: Pictures taken from the file 1yveI (res 346 - res 362 for (a) and res 354 - res 362 for (b)). The residues assigned by Segno as α helical are shown in red. Intuitively the picture (a) shows a single helix, whereas Segno assigns three. The picture (b) shows that this helix contains in fact π residues which are not assigned by Segno. So this helix cannot be correctly assigned by Segno up to now.

This test shows clearly that Segno is more able to distinguish between a single helix with a bend in its middle and two different helices than Stride or DSSP.

b) β strands

To determine the bends of the strands the same technique as for the helices has been applied. For each residue an ideal strand has been fitted over the real structure and the angle made by the axis of these strands gives us the bend of the strand for this residue. The results are shown in the graph 14.

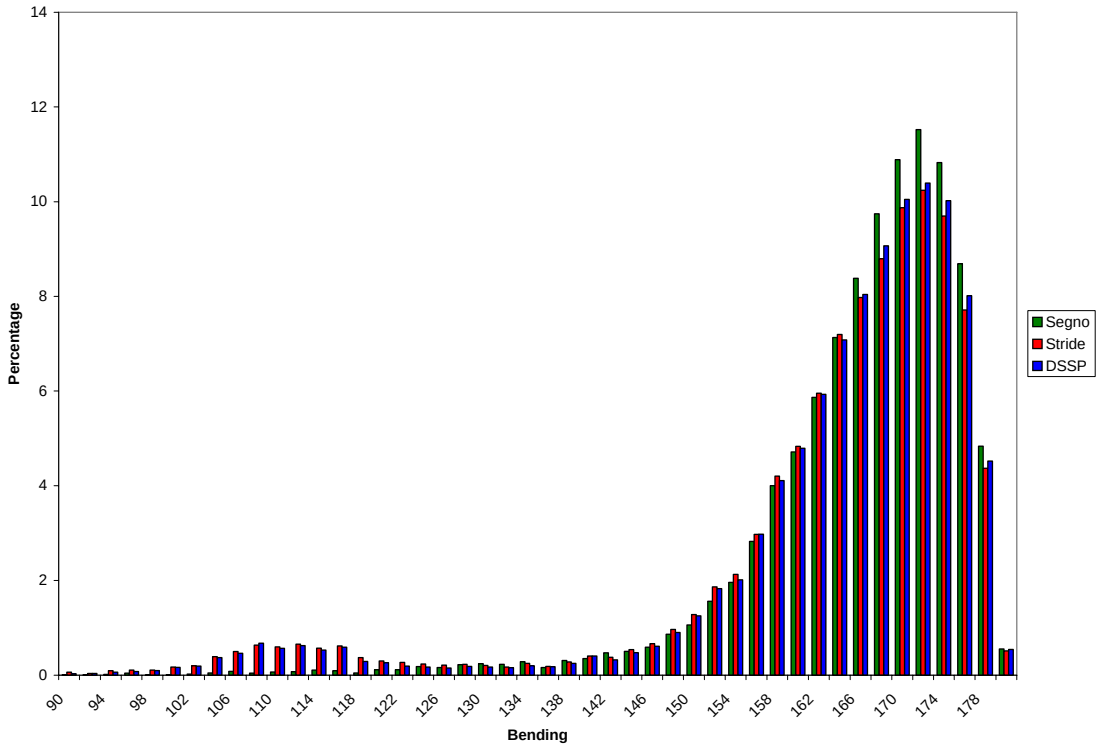


Figure 14: Distribution of the bends in beta strands.

We can see that the three distributions are not similar. Indeed we clearly see a secondary peak centered around 112 degrees. This peak corresponds to highly bent strands and does not appear in the distribution of Segno.

This again is due to misassignments by DSSP and Stride. Some examples of differences in assignment for bent strands are shown on figure 15.

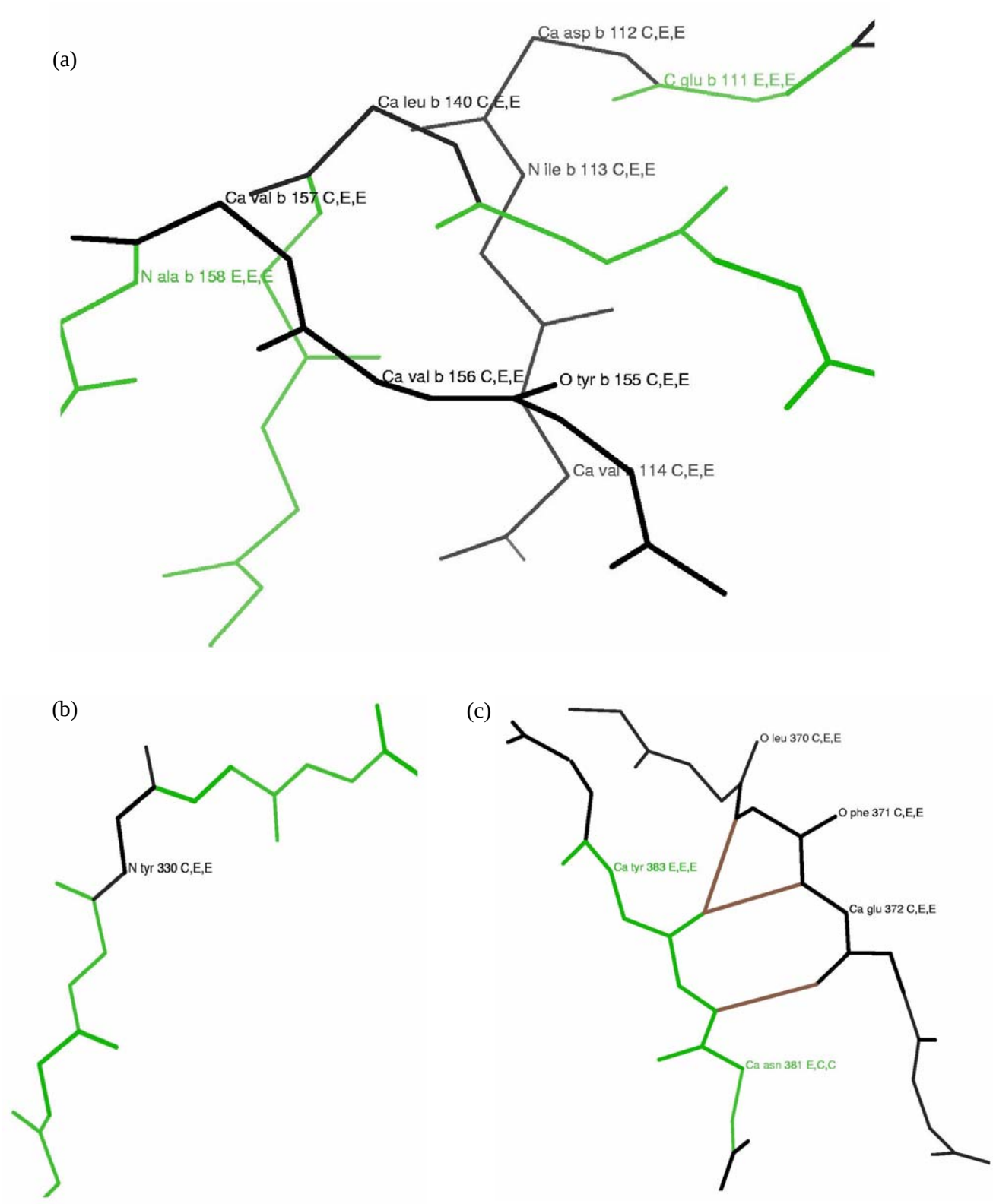


Figure 15: (a) Picture taken form the file 1tc1B (res 109 - res 114, res 137 - res 143, res 154 - res 158); (b) Picture taken from the file 2myr (res 327 - res 333); (c) Picture taken from the file 2myr (res 369 - res 374, res 380 - res 384). The residues in green are those assigned as strand residues by Segno and those in black those assigned as coil. The letters in the labels correspond to: Segno's assignment, DSSP's assignment and Stride's assignment in this order.

IV. DISCUSSION AND CONCLUSION

Segno has been developed in order to solve the problems of assignment observed with other algorithms, especially Stride and DSSP that are the most widely used programs for secondary structure assignment. The most important issue is to define the ends of the secondary structures as precisely as possible.

In the majority of the cases the assignments provided by Segno, Stride and DSSP are similar (more than 80% agreement). However a further analysis of the results has revealed that this apparent agreement hides many differences particularly in the definition of the end of the structures.

For helices, three tests have been made. In two of them (concerning directly the C-capping and the N-capping) we have found that Segno and Stride were equivalent and better than DSSP. The third test concerning the bending of helices has shown a great superiority of Segno compared with the two other programs. Thus we can conclude that Segno is an improvement over the currently most widely used methods.

The comparison of these three programs of secondary structure recognition, two of which are based on hydrogen bonds (Stride and DSSP), suggests that the consideration of the geometry is an advantage when defining secondary structure. The use of hydrogen bonds is not sufficient to precisely define a helix and can lead to an incorrect assignment. For example one missing hydrogen bond in a helix would result in the splitting of it into two entities. The introduction of torsion angles Φ and Ψ (Stride) allows to eliminate residues engaged in a hydrogen bond but who are not helical. However the addition of these parameters is not sufficient to solve all the problems because non-helical residues can have a (Φ,Ψ) combination in the helical part of the Ramachandran plot. The contribution of these parameters is thus limited. Our approach does not deny the existence of the hydrogen bonds between the residues in a helix, but we consider that the geometry of a helix is deeply linked with the network of hydrogen bonds. Thus the helical geometry can only appear if the appropriate hydrogen bonds form.

For strands, we have been able to develop only one test to compare the three programs due to the lack of clear capping motifs in strands. Nevertheless this test has shown that Segno assigns strands more reliably than the other programs. What is more, the geometric approach to define strands allows the definition of single strands which are not involved in a beta sheet. Indeed these strands are not assigned by Stride and DSSP because there are not involved in hydrogen bonding. However, Segno provides a different assignment for single beta strands and beta strands in a sheet so that either both, or only one sort of strand may be used.

The advantages of Segno are numerous. Some of them have been described above such as the ability to assign a residue as helical despite the lack of a hydrogen bond, or the possibility to define single β strands. One other interesting particularity of the assignment provided by Segno is the possibility to define mixed helices. It is indeed possible to find helices in which we can distinguish parts of α helices and parts of 3_{10} helices. This provides new information in order to study the distortions in helices due for example to an insertion or a deletion of a residue in the amino acid sequence.

However this new algorithm to define secondary structures is not perfect and contains some disadvantages relative to the use of geometric parameters. The first of this disadvantages and the most important according to us is the number of cut-offs that need to be optimised. Though we have tried to reduce to the minimum the number of parameters, the number of cut-offs stays important. For example we need 12 different cut-offs to define the limits of the domains of the structures in the Ramachandran plot (4 for each type of structure). The second consequence of the use of geometric parameters is the fact that short helices are not defined with the same accuracy as longer ones. Indeed the definition of the approximate axis of the helices is less close to the real axis for short helices (as for the end of the helices) and as a consequence some short helices may be missed by Segno. What is more, even when these short helices are detected by Segno the distinction between α and 3_{10} helices is less accurate because we cannot use all the parameters (in particular the comparison between ω_3 and ω_4). However the problem of definition of short helices is common to the three programs. Indeed the assignment of short helices is somewhat subjective.

Segno is still in a phase of development and some improvements should be brought in the near future. The first of these improvements is to add the recognition of π helical residues. In spite of the fact that these residues are relatively rare in a protein, we have found that they contribute a large fraction of the cases where the Segno assignment is incorrect. We can, therefore, expect improvements. Moreover the addition of the π residues may imply some modifications in the cut-offs, particularly the maximal radius for the recognition of helical residues.

The results obtained by Segno show an improvement in the definition of secondary structures compared with the results of Stride or DSSP. These results are very encouraging and lead us to think that this new algorithm may be used in order to determine the effect of a change in the amino acid sequence on the backbone in secondary structure elements.

Moreover we think that its accuracy may be useful for researchers to find new features (such as a residue preferences at the ends of helices), which may be useful to predict secondary structures only from the amino acid sequence, in order to improve the algorithms already existing [4,23].

REFERENCES

1. Aurora, R. & Rose, G.D. (1998). Helix capping. *Protein Science* **7**, 21-38.

2. Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). Solvent -induced distorsions and the curvature of α-helices. *Nature* **vol.306**, 281-283.

3. Colloc'h,N., Etchbest, C., Thoreau, E., Henrissat, B. & Mornon, J.P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Engineering* **vol.6 no.4**, 377-382.

4. Cuff, J.A. & Barton, G.J. (1999). Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction. *PROTEINS: Structure, Function and Genetics* **34**, 508-519.

5. Deane, C.M. & Blundell, T.L. (2000). A Novel Exhaustive Search Algorithm for Predicting the Conformation of Polypeptide Segments in Proteins. *PROTEINS: Structure, Function and Genetics* **40**, 135-144.

6. Deane, C.M. & Blundell, T.L. (2001). CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Science* **10**, 599-612.

7. Engh, R.A. & Huber, R. (1991). Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta. Cryst.* **A 47**, 392-400.

8. Frishman, D. & Argos, P. (1995). Knowledge-Based Protein Secondary Structure Assignment. *PROTEINS: Structure, Function and Genetics* **23**, 566-579.

9. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science* **3**, 522-524.

10. Kabsch, W. & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **vol.22**, 2577-2637.

11. King, S.M. & Johnson, C.W. (1999). Assigning Secondary Structure From Protein Coordinate Data. *PROTEINS: Structure, Function and Genetics* **35**, 313-320.

12. Lovell, S.C., Word, J.M., Richardson, J.S. & Richardson, D.C. (2000). The Penultimate Rotamer Library. *PROTEINS: Structure, Function and Genetics* **40**, 389-408.

13. Penel, S., Morrison, R.G., Mortishire-Smith, R.J. & Doig, A.J. (1999). Periodicity in α-Helix Lengths and C-Capping Preferences. *J. Mol. Biol.* **293**, 1211-1219.

14. Presta, L.G. & Rose, G.D. (1988). Helix Signals in Proteins. *Science* **vol.240**, 1632-1641.

15. Richards, F.M. & Kundrot C.E. (1988). Identification of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *PROTEINS: Structure, Function and Genetics* **3**, 71-84.

16. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Advances in protein chemistry* **vol.34**, 167-338.

17. Richardson, J.S. & Richardson, D.C. (1988). Amino Acid Preferences for Specific Locations at the Ends of α Helices. *Science* **vol.240**, 1648-1652.

18. Rufino, S.D., Donate, L.E., Canard, L.H.J. & Blundell, T.L. (1997). Predicting the Conformational Class of Short and Medium Size Loops Connecting Regular Secondary Structures: Application to Comparative Modelling. *J. Mol. Biol.* **267**, 352-367.

19. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Structural Biology* **5**, 1029-1032.

20. Sklenar, H., Etchebest, C. & Lavery, R. (1989). Describing Protein Structure: A General Algorithm Yielding Complete Helicoidal Parameters and a Unique Overall Axis. *PROTEINS: Structure, Function and Genetics* **6**, 46-60.

21. Webber, C.L. Jr., Giuliani, A., Zbilut, J.P. & Colosimo, A. (2001). Elucidating Protein Secondary Structures Using Alpha-Carbon Recurrence Quantifications. *PROTEINS: Structure, Function and Genetics* **44**, 292-303.

22. Word, J.M., Lovell, S.C., LaBaen, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S. & Richardson, D.C. (1999). Visualizing and Quantifying Molecular Goodness-of-fit: Small-probe Contact Dots with Explicit Hydrogen Atoms. *J. Mol. Biol.* **285**, 1711-1733.

23. Zhu, Z.Y. & Blundell, T.L. (1996). The Use of Amino Acid Patterns of Classified Helices and Strands in Secondary Structure Prediction. *J. Mol. Biol.* **260**, 261-276.