**Abstract**

**Motivation:** The secondary structure is a key element of architectural organization in proteins. Accurate assignment of the secondary structure elements (SSE) (helix, strand, coil) is an essential step for the analysis and modelling of protein structure. Various methods have been proposed to assign secondary structure. Comparative studies of their results have shown some of their drawbacks, pointing out the difficulties in the task of SSE assignment.

**Results:** We have designed a new automatic method, named P-SEA, to assign efficiently secondary structure from the sole Cα position. Some advantages of the new algorithm are discussed.

**Availability:** The program P-SEA is available by anonymous ftp: ftp.lmcp.jussieu.fr directory: pub/

**Contact:** E-mail: Labesse@lmcp.jussieu.fr

---

**Introduction**

The existence of regular segments in the structure of protein was predicted from theoretical analysis (Pauling and Corey, 1951; Pauling et al., 1951). The experimental determination of three-dimensional (3D) structures of proteins has confirmed the presence of these regular secondary structures. The comparative analysis of protein folds shows that secondary structure elements (SSE) are more conserved than the precise atomic structure (Mizuguchi and Go, 1995).

The precise delineation of SSE is an essential step in the study of protein structures. It becomes an important tool for molecular modelling, experimental determination of protein structure (by NMR and crystallography) and structure prediction.

Different methods have been designed to perform the assignment of SSE, among which three are the most widely used: DSSP (Kabsch and Sander, 1983), DEFINE (Richards and Kundrot, 1988) and P-CURVE (Sklenar et al., 1989).

They are based on distinct geometric criteria and algorithms. However, Colloc'h et al. (1993) have shown some discrepancies between the assignments, pointing out difficulties in the assignment. This prompted them to propose an assignment computed from a ternary consensus method (TCM) of the previous three methods (Colloc'h et al., 1993).

Various practical disadvantages make this approach time consuming and rather tedious. Furthermore, the need for a precise and universal assignment of SSE from even partial data (beginning of the structure determination, knowledge of the Cα trace alone), to help both surveying and modelling of proteins, leads us to re-analyse the assignment methods.

We have developed a new method, P-SEA (Protein Secondary Element Assignment), based solely on the Cα positions, making it more general. The parameters used were adjusted to perform the task as efficiently as the methods based on backbone analysis. The improvements due to our rapid new algorithm are discussed here.

**Algorithm**

The assignment of P-SEA is performed on any PDB formatted file. It uses only the Cα coordinates of the protein trace to assign the residue to one of the three states: helix (including mainly α-helices, but also 3.10 and π helices), strand (parallel and anti-parallel β strands) or coil (loops and turns).

First, the distances d2i, d3i and d4i between the (i - 1)th residue and the (i + 1)th, the (i + 2)th and the (i + 3)th, respectively, are computed from the cartesian coordinates of the Cα carbons, as well as the angle αi defined by the Cα carbon triplet (i - 1, i, i + 1) and quadruplet (i - 1, i, i + 1, i + 2), respectively. The assignment of SSE is determined from the satisfaction of either the distance (d2i, d3i, d4i) or angle (αi) criteria (Table I).

The use of only one type of geometrical criterion leads to inaccurate assignment (over- or under-assignment as shown in Figure 1).

The helices are first assigned to any five or more residue-long segment in which each position (Cα) possesses geometrical characteristics satisfying either the distance (d3i, d4i) or the angle (αi) helical criteria (shown in Table I). The former segment is lengthened by one residue at each end when d3i or αi satisfies helical criteria. Then, the β strands are defined similarly, according to specific criteria ((d2i, d3i, d4i) or (αi, αi)). Lenthening by one residue at each end is accepted if the d3i value corresponds to an extended structure. Small β strands (three residues long) are retained only when they are included in a β sheet (Figure 1). This
deduced from the number of contacts (distance between 4.2 and 5.2 Å) with spatially but not sequentially close neighbours. The observation of at least five contacts between the three Cα atoms (in the three-residue strand) and the surrounding Cα atoms leads to the assignment of a short β strand. The loop regions are defined by default. The first and last positions of a polypeptide chain are mainly assigned to the coil state.

The distance and angle criteria (Table I) were first derived from visually assigned SSE. Then the parameters were slightly refined to assign the same SSE as the crystallographer or the TCM (Colloc'h et al., 1993). Similar values for the angles, named 2i and ai here, have been independently proposed for the detection of helical or extended conformations from the Cα trace (Oldfield and Hubbard, 1994), confirming the validity of our approach.

The comparison of the new method with ones currently in use (DSSP, DEFINE, P-CURVE and TCM) is discussed in the following section.

Results

P-SEA has been designed to be easy to use, rapid and efficient. It allows the identification of the SSE from the Cα coordinates alone.

The parameters used are shown in Table I. Starting from the mean value observed in the perfect helical and extended stretch of refined crystal structures, we adjust our criteria to allow a variation of ~10% around the mean value for all distance and angle parameters but the ai value. The dihedral angle defined by four consecutive Cα carbons is the more variable of the geometrical parameters used, but is also a highly discriminating one. Our algorithm performs a rapid assignment of SSE, enabling the definition of the topology of any 3D protein structure.

In order to check the validity of our method, we compared our results with the assignments performed by the three popular assignment methods (DEFINE, P-CURVE and DSSP) and the consensus method (TCM) derived from the former three independent algorithms. The precise comparison of the six algorithms is given (see Table II) for each type of SSE (helices, strands and coil) and for the global analysis of a non-redundant protein database including most of the currently known folds. This database contains 226 polypeptide chains from the PDB (Bernstein et al., 1977), corresponding to 43 489 residues.

As already observed for DEFINE, P-CURVE and DSSP assignments (Colloc'h et al., 1993), the major differences between the distinct methods, including P-SEA and STRIDE (Frishman and Argos, 1995), involve the precise definition of the SSE ends; the few SSE assigned only by one method while missed by the others are mostly short (three or four amino acids long) strands. The helices are the better assigned SSE according to the agreement (>90%; see Table II) of all the methods tested here. This could be explained by the precise and distinct geometry of the helical conformation as seen on a

![Fig. 1. Example of P-SEA assignment. The SSE of the dihydropteridine reductase (PDB ID HR) (Varughese et al., 1992) were assigned using the program P-SEA. Some SSE are missed when only single geometrical criteria (distances or angles) are used (strand in italics). Loosening of one of the criteria leads to over-assignment, as in DEFINE. The use of a consensus assignment using simultaneously distance and angle criteria allows a better identification of helices, strands and coils. Small strands are assigned when present in sheets (underlined strand).](http://bioinformatics.oxfordjournals.org/)

<table>
<thead>
<tr>
<th>Assignment parameters</th>
<th>Secondary structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle $\alpha$ (°)</td>
<td>89 ± 12</td>
</tr>
<tr>
<td>Angle $\alpha$ (°)</td>
<td>50 ± 20</td>
</tr>
<tr>
<td>Distance d2 (Å)</td>
<td>5.5 ± 0.5</td>
</tr>
<tr>
<td>Distance d3 (Å)</td>
<td>5.3 ± 0.5</td>
</tr>
<tr>
<td>Distance d4 (Å)</td>
<td>6.4 ± 0.6</td>
</tr>
<tr>
<td>Helix</td>
<td>124 ± 14</td>
</tr>
<tr>
<td>Strand</td>
<td>−170 ± 45</td>
</tr>
<tr>
<td>Helix</td>
<td>6.7 ± 0.6</td>
</tr>
<tr>
<td>Strand</td>
<td>9.9 ± 0.9</td>
</tr>
<tr>
<td>Helix</td>
<td>12.4 ± 1.1</td>
</tr>
</tbody>
</table>
Table II. Comparison of SSE assignment methods. Comparison of the agreement of the various assignments of the SSE, residue by residue, of 226 protein chains from the PDB database. The P-SEA assignment is compared to DSSP, P-CURVE, DEFINE, STRIDE and TCM for each of three states (helix, strand and coil) and the complete set.

<table>
<thead>
<tr>
<th>SSE</th>
<th>DSSP (%)</th>
<th>P-CURVE (%)</th>
<th>DEFINE (%)</th>
<th>STRIDE (%)</th>
<th>TCM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helix</td>
<td>93.8</td>
<td>92.4</td>
<td>83.6</td>
<td>93.2</td>
<td>94.2</td>
</tr>
<tr>
<td>Strand</td>
<td>78.4</td>
<td>74.4</td>
<td>66.8</td>
<td>77.5</td>
<td>79.4</td>
</tr>
<tr>
<td>Coil</td>
<td>79.3</td>
<td>80.6</td>
<td>84.4</td>
<td>81.2</td>
<td>85.1</td>
</tr>
<tr>
<td>Total</td>
<td>83.4</td>
<td>82.4</td>
<td>79.0</td>
<td>84.1</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>12856</td>
<td></td>
<td></td>
<td>10582</td>
<td>20051</td>
</tr>
</tbody>
</table>

Ramachandran plot. The agreement for strand assignment is lower (between 67 and 80%; see Table II). The presence of extended but distorted stretches in the protein structures may impair improvement of the β strand assignment. Similar results were previously described for a smaller dataset (Colloc’h et al., 1993).

As the main discrepancies are observed at the terminal ends of the SSE, the differences between the agreement of the assignment for helices and strands may be correlated to the mean length of these elements. The average length (>12 amino acids) of the helices may explain the higher correlation observed between the different assignments. In contrast, the mean β strands are seven amino acids long [only five amino acids long in (β/α) structures]. It is even more critical as the shortest accepted SSE are usually four amino acids long. In order to assign the short β strands (three amino acids) that may actually belong to a β sheet, we add the survey of the neighbouring Ca carbons.

The shortening or lengthening of SSE from one assignment to another is mainly due to threshold artifacts. We believe that a consensus method such as TCM or P-SEA decreases this risk. While tuning the parameters of P-SEA, we carefully examined the assignment of the terminal ends of SSE. This was performed by surveying the correlation between the assignment and the residue conformations as monitored by the trace geometry or the backbone dihedral angles (ϕ, χ).

Table III. Comparison of SSE assignment methods. Agreement of the various assignments (DSSP, P-CURVE, DEFINE, STRIDE, TCM and P-SEA) performed on the 226 protein chains from the PDB database and measured on a residue-by-residue basis.

<table>
<thead>
<tr>
<th>Versus</th>
<th>DSSP (%)</th>
<th>P-CURVE (%)</th>
<th>DEFINE (%)</th>
<th>STRIDE (%)</th>
<th>TCM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSP</td>
<td>100</td>
<td>79.2</td>
<td>74.6</td>
<td>95.1</td>
<td>89.6</td>
</tr>
<tr>
<td>P-CURVE</td>
<td>74.8</td>
<td>94.2</td>
<td>78.9</td>
<td>89.6</td>
<td>84.9</td>
</tr>
<tr>
<td>DEFINE</td>
<td>74.5</td>
<td>74.5</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>STRIDE</td>
<td>78.9</td>
<td>89.6</td>
<td>74.5</td>
<td>84.9</td>
<td>88.1</td>
</tr>
<tr>
<td>P-SEA</td>
<td>83.4</td>
<td>82.4</td>
<td>79.0</td>
<td>84.1</td>
<td>86.5</td>
</tr>
</tbody>
</table>

While also based on the Ca coordinate analysis, the program DEFINE used only distance criteria to assign SSE. The comparison of SSE assignment by DEFINE with those of DSSP and P-CURVE showed the deficiencies of DEFINE parameterization. The most important deviation noticed is the assignment of too many SSE (mainly β strands; Colloc’h et al., 1993). Its agreement with TCM is lower than that of P-SEA with TCM (see Table III) despite the fact that DEFINE contributes to TCM computation. Comparison of DEFINE with DSSP or P-CURVE shows only 75% agreement. P-SEA compares favourably with either DSSP or P-CURVE (83% agreement with both). Thus, P-SEA, while also using only the Ca trace coordinates, seems to improve the recognition of SSE. The two other commonly used programs, DSSP and P-CURVE, do not outperform P-SEA. Their agreement with the visual survey of 3D structures or TCM assignment is as good as that of P-SEA. The comparison with TCM shows 89.6% agreement with both DSSP and P-CURVE, respectively, versus 86.5% with P-SEA. TCM is the consensus assignment from these methods, but not P-SEA. Furthermore, they require the complete backbone to be known. The STRIDE assignment has also been compared, but its results appeared rather identical (at 95%) to those of DSSP (agreement with TCM and P-SEA: 88.2 and 84.1%, respectively).

Fig. 2. Evolution of the SSE assignment with structure resolution. Comparison of the assignment of the SSE by P-SEA and the TCM method in the N-terminal part (first 50 amino acids) of the arabinose binding protein crystal structures at 2.5 Å (PDB1ABP) and 1.8 Å (PDB5ABP) resolution (Quiocho et al., 1989). P-SEA already sees strand at low resolution, while other methods fail (underlined strand). 'a', 'b' and 'c' refer to helix, strand and coil assignment, respectively.
The agreement of P-SEA with the various other methods shows that we were able to identify all the useful characteristics independently developed in these methods (e.g. helix regularity in P-CURVE). We may observe that the helical segments are better detected than extended ones. The average length of the helices or strands assigned by P-SEA is comparable to that of TCM or DSSP (data not shown). The main differences between the various assignments lie in the precise ends of the SSE. Only a few SSE are missed by P-SEA compared to TCM and these are mainly short distorted ones. Visual survey showed that most SSE assigned by P-SEA, but not by TCM, are true ones despite being distorted or short ones.

Implementation

P-SEA has been written in ANSI C language and has been implemented under various UNIX operating systems (HP, SGI, SUN, IBM). The user may ask for the assignment of a single (or a list) of PDB-formatted file(s). The assignment of a non-redundant protein database comprising 553 PDB entries is also available at the address: http://www.lmcp.jussieu.fr/~labesse/SEA.html.

Discussion

While the SSE assignment is an essential step for the analysis of protein fold, a simple, rapid and efficient method remains to be designed. We have developed a new program for the detection of SSE from the Cα trace positions. This new algorithm has been compared favourably to other methods, including TCM or DEFINE. The assignment made by DEFINE is also based on the analysis of the Cα atom coordinates, but it agrees less with visual assignment or TCM. The agreement of our new method with the methods based on the full backbone geometry analysis is higher than that of DEFINE. The use of new geometrical criteria explains the observed improvement.

Using only Cα, P-SEA may be used to analyse partial data such as in an initial experimental model. The choice of other criteria, e.g. hydrogen bonding, would decrease the recognition of strand at early steps of structure determination.

Using the refinement of the crystal structure arabinose binding protein (ABP) from *Escherichia coli* as an example, we show that P-SEA identifies SSE before DSSP and P-CURVE, but not DEFINE. The assignment of the 1.8 Å resolution structure (PDB5ABP; Quiocho et al., 1989) closely resembles that of P-SEA made at a lower resolution (2.5 Å; PDB1ABP). The other methods miss the central β strand they identify at higher resolution (Figure 2). The better agreement of the SSE assignments for protein structures solved at a higher resolution was noted previously (Colloc'h et al., 1993). At low resolution (2.5–3.5 Å), accurate assignment will facilitate the determination of new structures.

We have previously used P-SEA to analyse the SSE of new 3D structures to enlighten new structural similarities and fold conservation among distinct enzymes. In the case of the Reductase/Epimerase/Dehydrogenase (RED) family (Labesse et al., 1994), analysis of the SSE was essential for the observation of the structural similarities the RED share. It also led to the identification of a new evolutionary pathway among the dinucleotide enzyme superfamily. As only the Cα trace coordinates from the UDP-glucose epimerase (PDB1UDP, one of the RED) were available, P-SEA was a convenient tool for the comparative studies of the SSE.

The use of the minimal set of atoms to define SSE makes P-SEA a more convenient and more universal tool than previously developed ones. It is currently used to derive protein structure databases for guided homology modelling by fragment matching. We are working on the precise and rapid delineation of β turn, β bulge and Ω loop.

Acknowledgements

We wish to thank D.Chacé, J.Chomilier and C.Etchebest for helpful discussion, and J.D.Barton for critical reading of the manuscript. This work was supported by the CNRS and the Universités Paris 6 and Paris 7. G.L. is a fellow of the Ministère de l'Enseignement Supérieur, de la Recherche et de l’Insertion Professionnelle.

References


Received on September 19, 1996; revised on December 30, 1996; accepted on January 3, 1997