BMB
reports

SABA (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures

Sang Youn Park[1], Min-Jae Yoo[1], Jaemin Shin[2] & Kwang-Hwi Cho[1,]*

[1]School of Systems Biomedical Science, Soongsil University, Seoul 156-743, [2]SBScience, Inc., Sungnam 463-943, Korea

Most widely used secondary structure assignment methods such as DSSP identify structural elements based on N-H and C=O hydrogen bonding patterns from X-ray or NMR-determined coordinates. Secondary structure assignment algorithms using limited Cα information have been under development as well, but their accuracy is only ~80% compared to DSSP. We have hereby developed SABA (Secondary Structure Assignment Program Based on only Alpha Carbons) with ~90% accuracy. SABA defines a novel geometrical parameter, termed a pseudo center, which is the midpoint of two continuous Cαs. SABA is capable of identifying α-helices, 3_{10}-helices, and β-strands with high accuracy by using cut-off criteria on distances and dihedral angles between two or more pseudo centers. In addition to assigning secondary structures to Cα-only structures, algorithms using limited Cα information with high accuracy have the potential to enhance the speed of calculations for high capacity structure comparison. [BMB reports 2011; 44(2): 118-122]

INTRODUCTION

Secondary structures in proteins refer to the highly regular, local sub-structures of helices and strands as suggested in 1951 by Pauling and Corey (1, 2). With the advent of ~65,000 X-ray, NMR and cryo-electron microscopy (EM) determined tertiary protein structures in the Protein Data Bank (PDB) (3), assigning secondary structure elements is still a prerequisite to the modern structural bioinformatic analysis of multiple protein structures or structure-based sequence alignments. Biologists can manually identify individual secondary structure elements by performing a visual check on tertiary structure coordinates. However, nowadays, automated programs assigning secondary structure elements allow one to bypass this tedious step. For example, DSSP (4), the oldest and the most widely used assignment method, identifies structural elements based on main-chain amide bond nitrogen (N-H) and carbonyl (C=O) hydrogen bonding patterns from the coordinates. Many other computational algorithms have since been developed, including STRIDE (5), DEFINE (6), P-SEA (7), KAKSI (8), P-CURVE (9), XTLSSTR (10), ECSTR (11), SEGNO (12), and VoTAP (13), all of which rely on geometrical features of the helices and the strands for identification. By comparing these programs, scientists have noticed that their results often disagree regarding the lengths of the assigned secondary structures. However, many acknowledge that the combination of algorithms produces reliable results that approach those of manual inspection (14, 15).

Due to two main reasons, there have been efforts to develop secondary structure assignment algorithms using limited Cα coordinate information. First, not all PDB coordinates include the entire atomic information necessary to reconstitute the geometrical features of helices and strands. For example, in the entire PDB as of April 2010, 618 chains are only in Cα and 1912 chains contain more than five continuing amino acids represented only as Cα. These coordinates commonly result from high flexibility in certain regions of the protein structure. Further, more and more Cα-only structures are continuously being reported from models generated from cryo-EM envelopes. The second reason is that as faster methods for generating secondary structure elements from protein structures are necessary for high capacity structure comparison studies, algorithms accurately identifying secondary structure elements using only Cα have the potential to enhance the speed of calculation.

Various methods such as DEFINE (6), P-SEA (7), and VoTAP (13) contain algorithms that have already been developed to assign secondary structures with only Cαs. DEFINE identifies secondary structures by comparing up to six parameters of inter-atomic difference distance matrices in structural fragments to an idealized reference distance typical for a particular secondary structure type. P-SEA uses Cα-Cα distances and dihedral angles between Cαs as the geometric criteria for defining secondary structure elements. Most recently, VoTAP utilizes a

*Corresponding author. Tel: 82-2-820-0454; Fax: 82-2-812-5762; E-mail: chokh@ssu.ac.kr
DOI 10.5483/BMBRep.2011.44.2.118

Received 1 November 2010, Accepted 17 December 2010

Keywords: α-helix, β-strand, Pseudo center, Secondary structure identification, 3_{10}-helix

geometrical tool based on three-dimensional Voronoï tessellation, which subdivides space over Cαs, to yield contact matrices used for secondary structure analysis. Despite the progress achieved, the resulting accuracies using only Cαs for secondary structure assignment approach only ~80% of that of DSSP [DEFINE: 73.0% (6); P-SEA: 80.2% (7); VoTAP: 83.2% (13)].

Here, we have developed a novel secondary structure assignment method using criteria developed around a newly defined pseudo center. The pseudo center is an imaginary geometrical point, which is the midpoint of two consecutive Cαs. By using distances and dihedral angles between two or more pseudo centers as the cut-off criteria, and by including criteria for Cα-Cα distances, we were able to identify α-helices, 3_{10}-helices, β-strands, and random coils using only the Cα coordinates. When our algorithm was tested on a collection of previously defined coordinates (13), we achieved overall ~90% accuracy compared to that of DSSP, which is a ~10% improvement compared to other methods using only Cα information for secondary structure assignment.

RESULTS AND DISCUSSION

Pseudo center and development of SABA

If a pseudo center is defined as the midpoint of two consecutive Cαs, then the positions of N-H and C=O, which participate in backbone hydrogen bonding, are closer to the pseudo center than to the Cα (Fig. 1a). As a result, when the α-helices in the entire PDB were analyzed, the standard deviation of the distances between pseudo centers i' and i'+3 was smaller (0.27) than that between i and i+4 Cαs (0.32) (Fig. 1b). This suggests that the use of pseudo centers should be beneficial compared to the use of Cα based geometrical criteria as in P-SEA (7). Cut-off criteria for secondary structure elements solely based on this pseudo center distance resulted in ~90% accuracy compared to DSSP in assigning the helical structure elements; however, the criteria using dihedral angles from four consecutive pseudo centers as well as the Cα-Cα distances were further included for better assignment of the β-strands.

Various criteria of the distances between two pseudo centers, dihedral angles from four continuous pseudo centers, and Cα-Cα distances were all taken into consideration when setting the secondary structure element cut-off criteria, which gave the best matched result compared to identification by DSSP. To easily compare our results in reference to the most recent VoTAP (13) results, we used the same Statset and Checkset used by Mornon *et al*. for the parameterization and analysis of our pseudo center criteria results. An optimization program to set the converging boundary was programmed in Python using selected 230 PDB coordinates from the 282 coordinates previously defined as Statset (13) in order to give the best cut-off criteria for each secondary structure element (Table 1). Fifty-two coordinates of the original Statset were excluded due to an existing sequence gap or a diffraction resolution lower than 2 Å.

By using the various cut-off criteria for each secondary structure elements based on pseudo centers and Cα-Cα distances (see Methods section and Table 1), we have developed a program called SABA in Python to assign the secondary structure elements (SABA can be run from the website http://ebio.ssu.ac.kr/saba).

Example of SABA identification with nuclear transport factor 2

As an example test case, the secondary structure elements of nuclear transport factor 2 (PDB ID: 1OUN) from Statset (13)

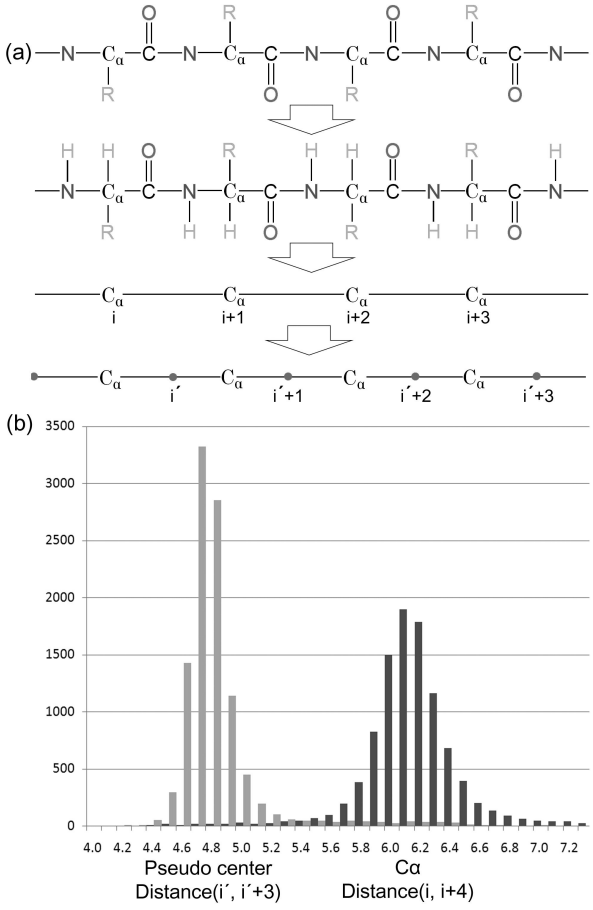

Fig. 1. Pseudo center definition for a polypeptide chain and distribution of pseudo center distances and Cα-Cα distances between hydrogen bonding atoms in the α-helices of the entire PDB. (a) When a pseudo center is defined as the midpoint of two consecutive Cαs, then N-H and C=O, which participate in backbone hydrogen bonding, are closer to the pseudo center than to Cα. (b) When all α-helices in the entire PDB were analyzed, the standard deviation of the distances between i' and i'+3 pseudo centers was smaller (0.27) than that between i and i+4 Cαs (0.32). Hence, pseudo center can be a good candidate for the parameterization of secondary structure elements.

Table 1. List of distance cut-off criteria defined for assigning each secondary structure element

Secondary structure	Conditions* and cut-offs
α-helix†	4.21 Å < pseudo center (i', i'+3) < 5.23 Å
	43.5° < pseudo center (i', i'+1, i'+2, i'+3) < 78.3°
3_{10}-helix†	pseudo center (i', i'+2) < 4.82 Å
	pseudo center (i'+1, i'+3) < 5.24 Å
	5.14 Å < pseudo center (i', i'+3) < 9.12 Å
	42.1° < pseudo center (i', i'+1, i'+2, i'+3) < 119.5°
Parallel β-sheet‡	2.58 Å < pseudo center (i', j') < 5.18 Å
	4.34 Å < pseudo center (i'-1, j'-1) < 5.03 Å
Anti-parallel β-sheet§	4.36 Å < pseudo center (i', j') < 5.19 Å
	4.16 Å < pseudo center (i'+1, j'-1) < 5.27 Å
	1.42 Å < Cα (i+1, j) < 5.99 Å
	No β-sheet relation for i-1 and j+2 residues, when pseudo center (i'-2, j'+2) > 5.64 Å or 3.00 Å < Cα (i-2, j+3) < 6.70 Å
	No β-sheet relation i+2 and j-1 residues, when pseudo center (i'+2, j'-2) > 6.26 Å or 1.42 Å < Cα (i+3, j-2) < 5.99 Å

*Pseudo center (i', j'): distance between i' and j' pseudo centers; pseudo center (i', j', k', l'): dihedral angle of j', j', k' and l' pseudo centers; Cα (i, j): distance between i and j Cαs, †Prolines are excluded in helical structure identification, ‡i and j residues should be more than four residues apart, §Some exclusions are applied at the end residues for the identification of anti-parallel β-sheets.

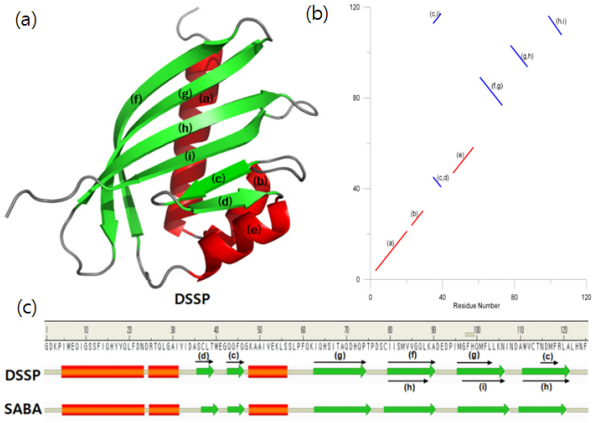

Fig. 2. Pseudo center distance plot and comparison between SABA and DSSP results in an example case. (a) Secondary structure elements of nuclear transport factor 2 (PDB ID:1OUN) were identified using DSSP and SABA. Colors and alphabets identify α-helices (red) and β-strands (green), with the pairing strands drawn above and below the strands. (b) Pseudo center distances that satisfy distance criteria for an α-helix (red) and a β-strand (blue) are plotted for the secondary structure identification. Symmetrical interactions were omitted for clarity. Lines were connected when more than two residues satisfied the pseudo center distance criteria for an α-helix and a β-strand. In the case of the β-strands, interacting β-strands resulting in the formation of β-sheets are noted together [eg. (c, d) and (f, g)]. (c) Comparison of SABA result to the DSSP result indicates high correlation between the two methods in assigning α-helices (orange) and β-strands (green).

were identified using SABA and subsequently compared with the DSSP results (Fig. 2). 1OUN was only chosen since it contains all of the secondary structure elements (α-helix, parallel β-sheet, anti-parallel β-sheet), except for a 3_{10}-helix. In SABA, all of the pseudo centers for the protein were defined, and the intra-distances between them were calculated from the coordinate information. Amino acid pairs, which satisfied the pseudo center distance cut-offs for the secondary structure elements (Table 1), are plotted (Fig. 2b). In the plot, α-helices (a, b, e) are positioned at the center with a positive slope due to the systemic i and i+4 residue interaction. Parallel β-sheets (c, i) are positioned at the corner of the plot with a positive slope, and anti-parallel β-sheets are positioned at the center with a negative slope due to the respective nature of the parallel or anti-parallel alternating β-strands. In this test case featuring nuclear transport factor 2, SABA shows 94.9% accuracy compared to DSSP in assigning the secondary structure elements (Fig. 2c).

SABA identification on Checkset

Using SABA, secondary elements were identified using a separate set of 183 PDB coordinates previously defined as Checkset (13) in order to minimize skew bias resulting from the analysis of the same coordinates that were used to set the geometric cut-offs. Ten coordinates from the original 193 in Checkset were excluded from our test set due to existing sequence gaps. The accuracies of the results were compared to DSSP. The comparison of the overall accuracy of the structural element identification between SABA and DSSP of the Checkset is shown in Table 2, and the individual accuracy results of Statset and Checkset is shown in the Supplementary Table. The overall accuracy of SABA in comparison to DSSP was 90.6%, which was also checked reciprocally to account for any bias from over-identification (Table 2), as false-positive over-identification of α-helices or β-sheets can inadvertently increase the accuracy.

Table 2. Comparison of overall accuracy of secondary structure element identification results between SABA and DSSP on Checkset

	SABA (%) (Compared against DSSP)	DSSP (%) (Compared against SABA)
Helix	94.4	93.5
Sheet	90.0	88.3
Coil	89.1	89.1
Total	90.6	90.6

In conclusion, we have developed a novel method for identifying secondary structure elements using distances and dihedral angles from pseudo centers and Cα-Cα distances. SABA operated with ~90% accuracy compared to DSSP using only the Cα information, which is a 7-10% improvement compared

to other similar methods such as VoTAP, P-SEA, and DEFINE. Given the accuracy of SABA based only on Cα coordinates, this method will assist the rapid identification of secondary structure elements in high-throughput studies whose aims are to compare vast numbers of proteins.

MATERIALS AND METHODS

α-helix

Pseudo center distance and dihedral angle cut-offs between two spatially proximal residues were used to determine the α-helices. In a typical α-helix, hydrogen bonding between residues i(C=O) and i+4 (N-H) results in i' and i'+3 pseudo centers in a close range. As an example, hydrogen bonding between C=O (residue 4) and N-H (residue 8) produced pseudo centers 4' and 7' proximal (Supplementary Fig. a). Distance cut-off between these two pseudo centers (i', i'+3; 4', 7') along with the dihedral angle of the i', i'+1, i'+2, and i'+3 pseudo centers (4', 5', 6', and 7'), which predicts the approximate directions of N-H and C=O, were used in the following form for identification of α-helices. [Pseudo center distance (i', i'+3) between 4.21-5.23 Å; pseudo center dihedral angle (i', i'+1, i'+2, i'+3) between 43.5-78.3°] α-helices were not defined for proline residues since they lack the main-chain N-H necessary for hydrogen bonding. Other criteria such as Cα-Cα distance did not show any improvement in the structural assignment results.

3_{10}-helix

Similar to α-helix, pseudo center distance and dihedral angle cut-offs between two spatially proximal residues were used to determine the 3_{10}-helices. In a 3_{10}-helix, hydrogen bonding between residues i(C=O) and i+3 (N-H) results in i' and i'+2 pseudo center in a close range. As an example, hydrogen bonding between C=O (residue 4) and N-H (residue 7) produced pseudo centers 4' and 6' proximal (Supplementary Fig. b). As was the case for α-helix, both pseudo center distance (i', i'+2; 4', 6') and dihedral angle (i', i'+1, i'+3, i'+4; 4', 5', 6', 7') were evaluated for defining 3_{10}-helices. Inclusion of additional distance criteria between other nearby pseudo centers such as (i'+1, i'+3) and (i', i'+3) resulted in an improved identification [Pseudo center distance (i', i'+2) $<$ 4.82 Å; pseudo center dihedral angle (i', i'+1, i'+2, i'+3) between 42.1-119.5°; pseudo center distance (i'+1, i'+3) $<$ 5.24 Å; pseudo center distance (i', i'+3) between 5.14-9.12 Å] 3_{10}-helices were not defined for proline residues for the same reason as for the α-helix. Other criteria using Cα-Cα distance did not show any improvement in the structural assignment results.

β-strands

For β-strands, two β-strands in close proximity with each other, forming either a parallel or an anti-parallel β-sheet, were a prerequisite to identification. Further, due to the limited information of using only Cα coordinates, it was important that the pairs of hydrogen bonding residues of the β-sheets are correctly assigned. Hence, it was necessary to pair the two closest residues in two or more β-strands and then subsequently pair the rest prior to using any pseudo center criteria.

Parallel β-sheet

To identify the parallel β-sheet relationship between two β-strands, pseudo center distances of (i', j') and (i'-1, j'-1) were used to define the parallel β-sheet relationships between the i-1/i and j-1/j residues. For instance, two pseudo center distances of (14', 64') and (13', 63') had to be within the cut-off range in order to be assigned a parallel β-sheet (Supplementary Fig. c). [Pseudo center distance (i', j') between 2.58-5.18 Å; pseudo center distance (i'-1, j'-1) between 4.34-5.03 Å; i and j residues must be more than four residues apart.] Other criteria such as pseudo center dihedral angles or Cα-Cα distances did not show any improvement in the structural assignment results.

Anti-parallel β-sheet

Similar to the parallel β-sheet, pseudo center distances of (i', j') and (i'+1, j'-1) were used in conjunction with Cα-Cα distance criteria between the i+1 and j residues, resulting in optimal anti-parallel β-sheet relationships between the i/i+1 and j-1/j residues. For instance, two pseudo center distances of 16'/65' and 17'/64', as well as Cα distances between 17/65, had to be within the cut-off range in order to be assigned an anti-parallel β-sheet (Supplementary Fig. d). [Pseudo center distance (i', j') between 4.36-5.19 Å, pseudo center distance (i'+1, j'-1) between 4.16-5.27 Å, Cα distance (i+1, j) between 1.42-5.99 Å].

However, unlike the parallel β-sheet, further distance criteria had to be applied at the two N- and C-termini of the anti-parallel β-sheet, which showed the highest inconsistency between its SABA and DSSP results. This probably was due to the residues frequently diverging at the ends of the anti-parallel β-sheets. Diverging residues in β-strands often cause incorrect N-H and C=O orientations, which disrupt hydrogen bonding. However, whether or not the residues do so was difficult to predict by just using the Cα coordinates. To take this into account, the distance cut-off for the end residues was further considered after the anti-parallel relationship had been set based on the above criteria. If the pseudo center distance of one end (i'-2, j'+2) was larger than 5.64 Å or the Cα distance of (i-2, j+3) was between 3.00-6.70 Å, then residues i-1 and j+2 were not assigned as an anti-parallel β-sheet. Likewise, if the pseudo center distance of the other end (i'+2, j'-2) was larger than 6.26 Å or the Cα distance of (i+3, j-2) was between 1.42-5.99 Å, then residues i+2 and j-1 were not assigned as an anti-parallel β-sheet.

For example, residues 14 and 68 forming an anti-parallel β-sheet relationship depends on whether residues 13 and 69 form a hydrogen bonding pair (Supplementary Fig. d). Since the existence of hydrogen bonding pairs is especially difficult

to verify at the beginning and the end of a β-sheet just based on the Cα coordinates, we provided more strict criteria at these termini region. When the pseudo center distance of (13', 68') was larger than 5.64 Å or the Cα distance (13, 69) was between 3.00-6.70 Å, residues 14 and 68 were not assigned as an anti-parallel β-sheet. Likewise, when the pseudo center distance of (20', 61') was larger than 6.26 Å or the Cα distance (21, 61) was between 1.42-5.99 Å, residues 20 and 62 were not assigned as an anti-parallel β-sheet.

Secondary structure elements other than those mentioned above were classified as a coil.

Acknowledgements
This work was supported by a Korea Research Foundation Grant [KRF-2009-0076401].

REFERENCES

1. Pauling, L., Corey, R. B. and Branson, H. R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **37**, 205-211.
2. Pauling, L. and Corey, R. B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **37**, 251-256.
3. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.
4. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
5. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566-579.
6. Richards, F. M. and Kundrot, C. E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* **3**, 71-74.
7. Labesse, G., Colloc'h, N., Pothier, J. and Mornon, J. P. (1997) P-SEA: a new efficient assignment of secondary structure from Cα trace of proteins. *Comput. Appl. Biosci.* **13**, 291-295.
8. Martin, J., Letellier, G., Martin, A., Taly, J. F., Brevern, A. G. D. and Gibrat, G. F. (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol.* **5**, 17.
9. Sklenar, H., Etchebest, C. and Lavery R. (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **6**, 46-60.
10. King, S. M. and Johnson, W. C. (1999) Assigning secondary structure from protein coordinate data. *Proteins* **3**, 313-320.
11. Fodje, M. N. and Al-Karadaghi, S. (2002) Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Eng.* **15**, 353-358.
12. Cubellis, M. V., Cailliez, F. and Lovell, S. C. (2005) Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics*. **6**, S8.
13. Dupuis, F., Sadoc, J. F. and Mornon, J. P. (2004) Protein secondary structure assignment through voronoi tessellation. *Proteins* **55**, 519-528.
14. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. and Mornon, J. P. (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* **6**, 377-382.
15. Zhang, W., Dunker, A. K. and Zhou, Y. (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* **71**, 61-67.