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 Park1, Min-Jae Yoo1, Jaemin Shin2 & Kwang-Hwi Cho1,*

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

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...
geometrical tool based on three-dimensional Voronoi tessellation, which subdivides space over Ca's, to yield contact matrices used for secondary structure analysis. Despite the progress achieved, the resulting accuracies using only Ca's 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 Ca's. By using distances and dihedral angles between two or more pseudo centers as the cut-off criteria, and by including criteria for Ca-Ca distances, we were able to identify α-helices, 3_{10}-helices, β-strands, and random coils using only the Ca 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 Ca 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 Ca's, then the positions of N-H and C=O, which participate in backbone hydrogen bonding, are closer to the pseudo center than to the Ca (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 Ca's (0.32) (Fig. 1b). This suggests that the use of pseudo centers should be beneficial compared to the use of Ca 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 Ca-Ca 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 Ca-Ca 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 Ca-Ca 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)