Structural Blocks Retrieval in Macromolecules: Saliency and Precision Aspects

Virginio Cantoni[1] and Dimo T. Dimov[2]

[1] Department of Industrial and Information Engineering, Pavia University, Italy
virginio.cantoni@unipv.it
[2] Inst. of Inf. & Comm. Tech., Bulgarian Academy of Sciences, Sofia, Bulgaria
dtdim@iinf.bas.bg

Abstract. A structural *motif* is a compact 3D block of a few secondary structural elements (SSs) – each one with an average of approximately 5 and 10 residues for sheets and helices respectively – which appears in a variety of macromolecules. Several motifs pack together and form compact, semi-independent units called *domains*. The domain size varies from about 25 up to 500 amino acids, with an average of approximately 100 residues. This hierarchical makeup of molecules results from the generation of new sequences from preexisting ones, in fact motifs and domains are the common material used by nature to generate new functionalities. Structural biology is concerned with the study of the structure of biological macromolecules like proteins and nucleic acids, and it is expected to give more insights in the function of the protein than its amino acid sequence. In this paper we propose and analyze a possible performance of a new approach for the detection of structural blocks in large datasets such as the Protein Data Base (PDB).

Keywords: protein motif retrieval, protein structure comparison, protein secondary structure, protein data bases, secondary structure saliency, error analysis.

1 Introduction

In the last decade many approaches have been developed for retrieving a block (a motif, or a domain, …, up to an entire protein) within a protein, or within the entire PDB, by using 3D structural comparison [1-3].

As an example, starting from a traditional pattern recognition techniques – the Generalized Hough transform (GHough) [4] – a family of new approaches have been proposed on the basis of the 'primitives' complexity from which the voting process can rise. The smallest aggregate can be the single SS [5]; at another more effective level we adopted the occurrences of pairs of SSs [6]; in alternative we proposed terns of SSs occurrences [7]; up to the entire motif of m SSs, $m \geq 3$, for a complete exhaustive matching [8]. All these techniques are similar for what refers the basic process, and adopt the same Parameter Space (PS) – the protein volume – but differ consistently about the voting process and consequently in performances [9]. Nevertheless, in all

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 372–380, 2013.

these methods, in the PS, after the voting process, the points which have the expected number of votes are candidates as Reference Point (RP) locations of the searched motif. Note that, it is known the expected peak intensity and the composition: the number of occurrences of each SSs types in the motif model.

The computational complexity of the quoted GHough approach is limited, e.g. for the pairs (and terns) co-occurrences, being N the number of SSs of the macromolecule under analysis, the computational complexity is $O(N^2m^2)$: the number of protein SS pairs is $O(N^2)$ and the number of model motif SS pair is $O(m^2)$ and each protein pair is to be compared to each motif pair to eventually give a vote in the correspondent displacement of the RP location.

Nevertheless, for searching all the presences of a given motif in a large PDS, this approach is considered slow and thus impractical – even if, in the average, a three-SSs motif is detected in a given protein in about 5 μs! So, in this paper we proposed a new approach, derived by the quoted ones, but exploiting other than the 3D SSs distribution, also salient biochemical information on the SS composition. In this way we implement a new planning strategy in order to reduce consistently the computing time, but without losing the precision performance.

2 Our New Strategy

The SSs are usually represented as oriented linear segments, e.g. the axis for the α-helix and the best fit segment for a β-strand. The determination of the SSs of a protein is usually given by programs designed to standardize the SS assignment, such as the DSSP [10], or the STRIDE [11], which are both considered sufficiently precise (even if on the average 4.8% of the target residues were differently assigned). These programs support a rich information, such as: i) the total number of residues, ii) the number of chains, iii) the total number of hydrogen bonds, iv) the sequential residue number, v) the amino acid sequence, vi) a SS summary, vii) the type of helix (of 3-4-5 turns) or of β-bridge or the sheet label, moreover, viii) geometrical bend, ix) chirality, x) solvent accessibility, etc.

In our approach, the basic idea is initially to refer just to one SS, selected in the model motif for its saliency on the basis of the above detailed information. As for the GHough methods, it is also convenient to limit the displacement of analysis for a trivial reason related to precision, and thus to select the main SS as close as possible to the barycenter of the model motif. We will call S_0 this reference SS of the motif model hereinafter.

Also in this new approach, as for the mentioned GHough approaches, it is necessary to set up a Local Reference System (LRS) for the model representation, and it is convenient that the LRS origin is to be as close as possible to the barycenter. A suitable point for that is the midpoint of S_0. Moreover, it is convenient to put the y-axis of the LRS on the axis of S_0, meanwhile the x-axis is located on the plane defined by the y-axis and the midpoint of a second SS, in this connection the z-axis is obviously orthogonal (see figure 3). For the same reason of precision robustness, it is convenient that also this second SS would be well characterized and situated as far as possible

from S_0 , and as more as possible tilted with respect to S_0. Note that, the higher the distance the higher also the searching space for this second SS. We will call S_1 this second SS of reference for the motif.

For this SS pair (S_0, S_1), which characterizes the motif and fix the LRS, two parameters are discriminant: ρ_1, the Euclidean distance between the midpoints of the two SSs; φ_1, the angle between the axis of S_0 and the midpoint of S_1, as shown in figure 1.

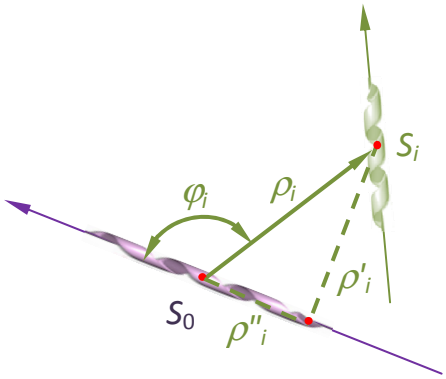

Fig. 1. The (ρ, φ) parameters definition of the main motif SSs-pair (S_0, S_i), i=1

The general target is to detect all possible instances of the motif in the given protein, or in a set of proteins, or in the whole PDB.

A preliminary search is devoted to look for all possible SSs, which have the same peculiarities (SS type, number of residues, amino acid sequence, etc.) of S_0, where we locate the origin of the LRS. Being n, $n \leq N$, the number of protein SSs, which match S_0, it would be necessary to analyze all the n-neighborhood NB(n) to validate the possible existence of the motif.

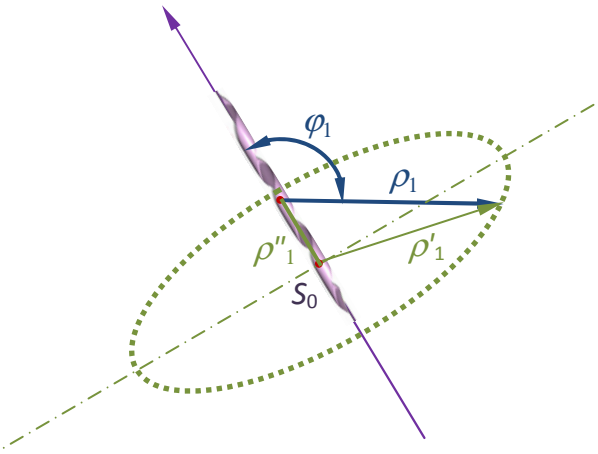

Fig. 2. The locus of candidate S_1 positions is a circle in PS, which is orthogonal to reference S_0 and centered on S_0 axis

In a second phase we have to introduce the LRS of the candidate motif in each NB(n). For this purpose we need to identify the possible locations of S_1. Obviously, these instances of S_1 must be compatible with the two parameters (ρ_1,φ_1), shown in figure 1, referred to S_0, other than compatible to the peculiarities of S_1 (its SS type, number of residues, amino acid sequence, etc.). From the geometrical point of view the locus of the candidate positions of the midpoints of S_1 is a circle belonging to a plane, normal to the axis of S_0, and having the center on this axis at a distance ρ.cos(φ) from the midpoint of S_0. Figure 2 details this locus. For each point of the dashed circle the possible existence of S_1 must be validated.

Let us call n_1 the number of compatible pairs (S_0, S_1) that are extracted in the above way, obviously $n_1 \leq n$. It is now necessary to examine the existence of all the other SSs of the motif in the proper locations of neighborhoods NB(n_1).

For this purpose all the SSs of the motif model are described in a Reference Table (RT) by their displacements (x, y, z) referred to LRS. The cardinality of the RT equals (m-2). In each of these m-2 positions, a SS of the motif must be located. Beside this geometrical validation, obviously, also the biochemical validation is better to be done (as usual in terms of SS type, number of residues, amino acid sequence, etc.). Figure 3 illustrates this process for the simple case of a motif model composed of just three SSs.

Saliency and precision of these matches is analyzed in the sequel.

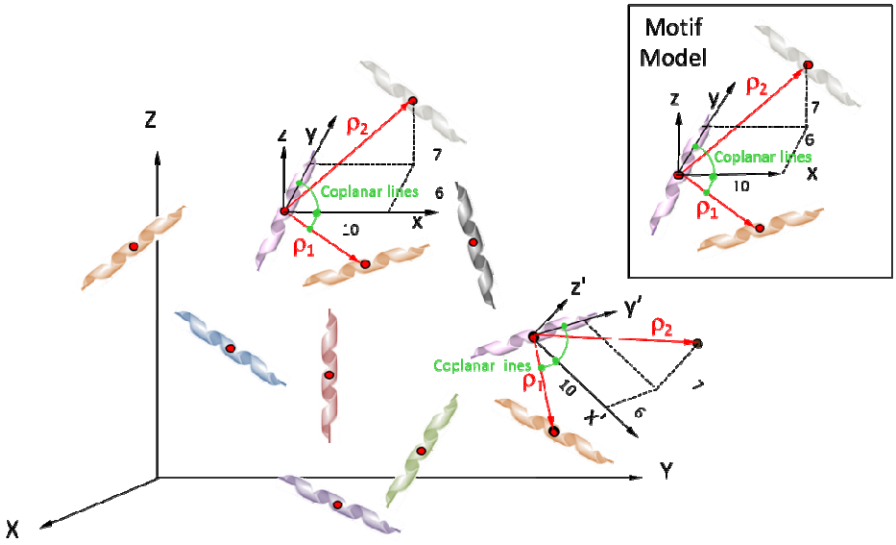

Fig. 3. Top-right: a 3 SSs motif model description. The local reference system is defined as follows: the y-axis coincides with the axis of S_0, the (x,y)-plane is defined by this axis and the midpoint of S_1. The motif existence confirmation process for two cases, in which the main pair (S_0, S_1) of helices is detected, is illustrated: top-left the complete motif is present; bottom-right the 3rd SS (S_2) of the motif is missing.

3 The Algorithm

A protein structure of *N* segments and a motif of $m, 0 < m \le N$ are considered. Each segment is determined in a 3D space by its start and end points, i.e. the directions are known. Additionally, the midpoints of the all segments are also considered known.

For next formula simplicity we will introduce $k, k = m-1,\ 0 < k < N$. Thus, the algorithm on the proposed strategy can be briefed as follows:

1 Choose a reference SS S_0. Consequently choose S_1 to build the LRS of the motif. Referring to LRS build the model motif description on the RT, containing the information of all the remaining SS: ρ and φ parameters, the displacement (x, y, z), and the biological information.

2 Detect all the reference segment candidates in the protein structure under analysis. The search rule is – each candidate should have similar biological characteristics like the motif reference segment. Let the number of the reference candidates is $n, n \le N$.

3 For each protein reference candidate, do
- look for a reference SS candidate S_1. Let the number of the these reference candidates is $n_1, n_1 \le N$.
- look for a *k*-long series protein SSs compatible to ρ and φ and other biological and/or geometrical parameters like in the motif.

4 End and result: a number *l* of appearances (instances) of the motif are found in the given protein, $0 \le l \le n_1$, where *l*=0 is the worst case of nothing found.

4 The Algorithm Errors Evaluation

We can model the errors of an appearance of the given motif into the given protein structure following the strategy described above, namely:

- The given motif consists of *k*+1 number of segments, $S_i, i = 0,1,...k$, where S_i plays for both, the segment itself and for its midpoint.
- As the initial (referencing) segment S_0 is already chosen, the midpoint of all the rest of segments $S_i, i = 1,...k$ determines the respective couple (ρ_i, φ_i), $\rho_i = D(S_0, S_i)$, $D(.,.)$ is the Euclidian distance operator, and φ_i is the angle between the segment S_0 and the line-cut $\overline{S_0, S_i}$.
- A few reference instances $\tilde{S}_0$ can be found in the given protein, such that $\tilde{S}_0 \sim S_0$, and to dispatch the task we will consider as $\tilde{S}_0$ only one of them. Thus, each segment $\tilde{S}_i, i = 1,...k$ of the motif appearance in the protein will determine the respective couple $(\tilde{\rho}_i, \tilde{\varphi}_i)$.
- For the errors between the motif and its appearance we can write down as follows: $\Delta\varphi_i = |\tilde{\varphi}_i - \varphi_i|$ and $\Delta\rho_i = |\tilde{\rho}_i - \rho_i|$, $i = 1,...k$, as well as:

$\Delta_i = \max\{\rho_i \sin(\Delta\varphi_i), \Delta\rho_i\} \approx \max\{\rho_i \Delta\varphi_i, \Delta\rho_i\}$, where Δ_i, $i = 1,...k$ is the respective distance error modeled by a sphere of radius Δ_i and center $\tilde{S}_i$, $i = 1,...k$, see also figure 4.

- Additionally, considering that each midpoint $\tilde{S}_i$, $i = 1,...k$, is chosen independently from the other ones, we can evaluate the maximal error Δ of choosing an appearance of the motif as follows:

$$\Delta = \max_{i=1,...k}\{\Delta_i\} = \max_{i=1,...k}\{\max\{\rho_i \Delta\varphi_i, \Delta\rho_i\}\} = \max\left\{\max_{i=1,...k}\{\rho_i \Delta\varphi_i\}, \max_{i=1,...k}\{\Delta\rho_i\}\right\} \leq$$
$$\leq \max\left\{\max_{i=1,...k}\{\rho_i\}\Delta\varphi, \Delta\rho\right\} = \max\{P.\Delta\varphi, \Delta\rho\}; \quad (1)$$

where $P = \max_{i=1,...k}\{\rho_i\}$ can be calculated preliminary, while $\Delta\varphi$ is the admissible angle error, and $\Delta\rho$ is the admissible distance error, both of them to be given outside, see also figure 5.

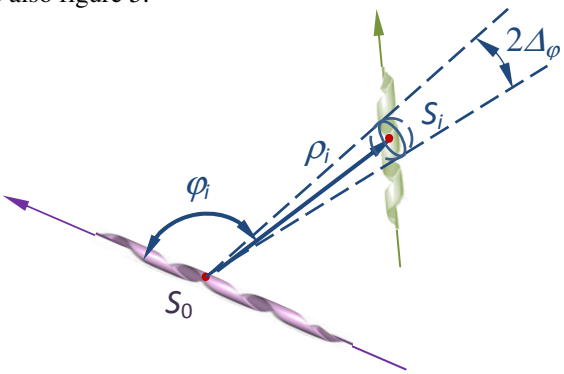

Fig. 4. Errors model of choosing the midpoint of a segment

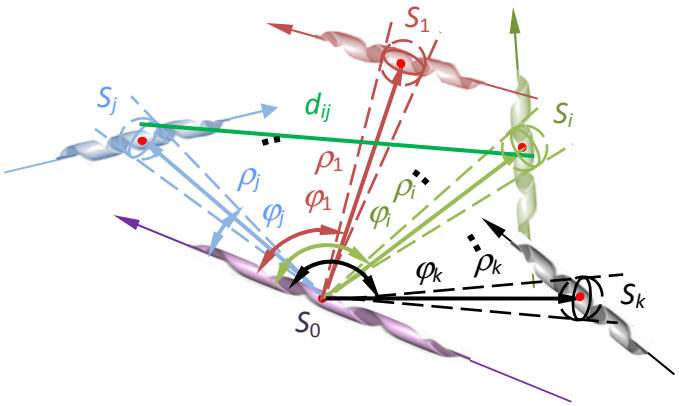

Fig. 5. Errors model of choosing the all segment midpoints of a motif

- We can rewrite (1) as follows:

$$\Delta = \max\{P.\Delta\varphi, \Delta\rho\} \quad \Leftrightarrow \quad \Delta = \begin{cases} \Delta\rho, & \Delta\varphi \le \Delta\varphi_0 \\ P.\Delta\varphi, & \Delta\varphi > \Delta\varphi_0 \end{cases}, \tag{2}$$

where $\Delta\varphi_0 = \dfrac{P}{\Delta\rho} = \max_{i=1,\ldots k}\{\rho_i\} \Big/ \Delta\rho$.

Besides, $\Delta\varphi$ should be given in radians, while $\Delta\rho$ should be given relatively to the size of the pixels (voxels) in the searching algorithm (see sections 2 and 3).

- Additionally, we can use the error evaluation (1) to check the precision of the given appearance of the motif, considering the triangle distance inequality among each triplet of points $(S_0, S_i, S_j), 1 \le i < j \le k$ (see also figure 5):

$$\left| D(\tilde{S}_i, \tilde{S}_j) - D(S_i, S_j) \right| \le \Delta \tag{3}$$

If (3) is fulfilled for every couple $(\tilde{S}_i, \tilde{S}_j), 1 \le i < j \le k$, i.e. $k(k-1)/2$ number of check-ups, then the appearance has all chances to be correct.

- To accomplish a full check of the given appearance also the directions and the lengths of each $\tilde{S}_i$, $i = 1,\ldots k$ can be compared with the respective S_i of the motif. And the latter can be performed using a similar model of errors like the above described, see also figure 6, where $d_i/2$ and γ_i are to play instead of ρ_i and φ_i, $i = 1,\ldots k$.

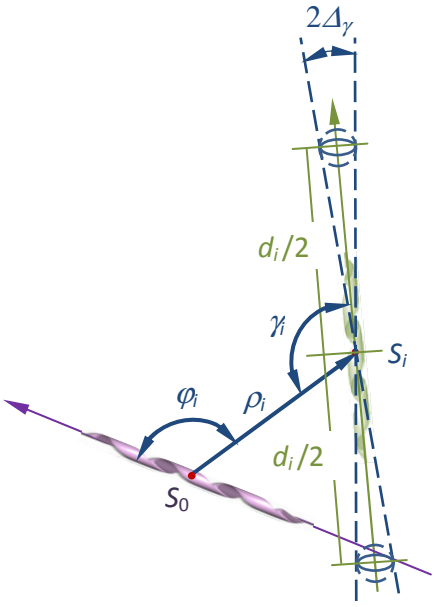

Fig. 6. Errors model of checking the direction and length of a segment

5 Conclusions

The spatial protein structure determines to a large extend the protein functionalities. It is thus very important to learn the structure-function relationship of proteins, and to compare their structures for retrieving of recurrent structural blocks such as motifs, domains, and units.

This paper aims to motif retrieval in single protein and in protein dataset, up to the whole PDB.

In order to retrieve motifs many approaches have been proposed in literature. In this paper we discussed a new tool that supports a planning strategy, targeting efficient analysis and robustness to error.

The efficiency is related to the capability to exploit the salient data, usually supported and made available, beside their geometrical features – such as the length of the structure in terms of number of amino acids, the biochemical properties, the sequence of the amino, etc., by known reliable programs like DSSP and STRIDE. Robustness is analyzed on the basis of a predefined maximum tolerable error, and on a rigorous analysis of the consequence of this constraint.

We are now planning an intensive quantitative analysis of the effectiveness of this new approach for practical problems such as alignment and structural block retrieval at different level of complexity: from basic motifs composed of a few SSs, up to entire units.

Moreover, in a more application oriented paper, we will analyze the motif modeling exactness or flexibility, with the opportunity of introducing a suitable metric for standard protein motifs (that is for structural blocks belonging to various protein families). Thus the focus will be on common components for a set of proteins that perform similar function, and will not relate to a metric block model extracted from a given protein as we experimented up to now.

Acknowledgements. This research is partly supported by the project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions).

References

1. Zotenko, E., Dogan, R.I., Wilbur, W.J., O'Leary, D.P., Przytycka, T.M.: Structural footprinting in protein structure comparison: The impact of structural fragments. BMC Structural Biology 7(1), 7–53 (2007)
2. Camoglu, O., Kahveci, T., Singh, A.: PSI: Indexing protein structures for fast similarity search. Bioinformatics 19(1), 81–83 (2007)
3. Chionh, C.H., Huang, Z., Tan, K.L., Yao, Z.: Augmenting SSEs with structural properties for rapid protein structure comparison. In: The Third IEEE Symposium on Bioinformatics and Bioengineering, pp. 341–348. IEEE Press (2003)
4. Illingworth, J., Kittler, J.: A Survey of the Hough Transform. Comp. Vision, Graphics, and Image Proc. J. 44, 87–116 (1988)

5. Cantoni, V., Mattia, E.: Protein Structure Analysis through Hough Transform and Range Tree. Nuovo Cimento della Società Italiana di Fisica. C. Geophysics and Space Physics 35, 39–45 (2012)
6. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Motif Retrieval by Exhaustive Matching and Couple Co-occurrences. In: Computational Intelligence Methods for Bioinformatics and Biostatistics, Houston, Texas, USA, June 12-14 (2012)
7. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Protein Motifs Retrieval by SS Terns Occurrences. Pattern Recognition Letters 34, 559–563 (2013)
8. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Searching Structural Blocks by SS Exhaustive Matching. In: Peterson, L.E., Masulli, F., Russo, G. (eds.) CIBB 2012. LNCS, vol. 7845, pp. 57–69. Springer, Heidelberg (2013)
9. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Structural Analysis of Protein Secondary Structure by GHT. In: The Int. Conf. on Pattern Recognition, Tsukuba, Japan, November 11-15, pp. 1767–1770. IEEE Computer Society (2012)
10. Kabsch, W., Sander, C.: Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. Biopolymers 22, 2577–2637 (1983)
11. Heinig, M., Frishman, D.: STRIDE: a Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. Nucl. Acids Res. 32, W500–W502 (2004)