21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Structural Analysis of Protein Secondary Structure by GHT

Virginio Cantoni
University of Pavia
virginio.cantoni@unipv.it

Alessio Ferone
University of Naples Parthenope
alessio.ferone@uniparthenope.it

Ozlem Ozbudak
Istanbul Technical University
ozbudak@itu.edu.tr

Alfredo Petrosino
University of Naples Parthenope
alfredo.petrosino@uniparthenope.it

Abstract

Structural biology is a branch of life science concerned with the study of the structure of biological macromolecules like proteins. The structure of a protein gives much more insight in its functions than that of its amino acid sequence. Protein structure comparison is important for understanding the evolutionary relationships among proteins, predicting protein functions, and predicting protein structures from the chemical composition. In this paper we propose a new approach for structural block retrieval based on the Generalized Hough Transform (GHT). A first technique uses as primitives the single Secondary Structure (SS), an alternative adopts co-occurrence of SSs couple, the third approach uses SSs triplets, and finally the primitive can be an entire block. In this paper we describe some experiments for the retrieval of elementary structural blocks consisting of four- and five-SSs.

1. A hierarchical structure

Proteins are formed by two basic regular 3D structural patterns called *secondary structure*: helices and sheets. A structural *motif* is a compact 3D protein structure referring to a small specific combination, which appears in a variety of molecules and is often called super SSs. While the spatial sequence of elements is the same in all instances of a motif, they may be encoded in any order: in this sense sequence is sometimes misleading and the structure analysis may give much more insight. Several motifs are packed together to form compact, local, semi-independent units called *domains* i.e. with more interactions inside than with the rest of the protein. Therefore, a domain forms a compact 3D structure, independently stable, and can be determined by two characteristics: its compactness and its extent of isolation. Moreover, many proteins consist of several domains to form multi-domain and multifunctional molecules. Many domains could have once existed as independent proteins. Multi-domain proteins are likely to have emerged from a selective pressure during evolution to create new functions. This hierarchical makeup of macromolecules is quite explicit the F. Jacob's aphorism: *Nature is a tinkerer and not an inventor*; that is new sequences are adapted from pre-existing ones rather than invented, in fact motifs and domains are the common material used by nature to generate new sequences.

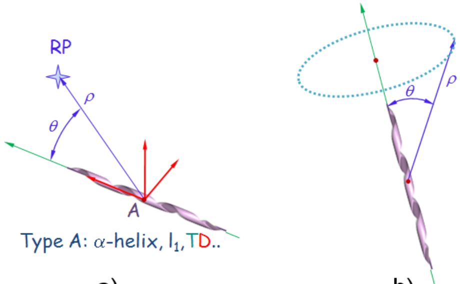
2. A new investigation playing field

In recent years many investigations have been made to analyze the various structural levels of proteins [1, 2, 7], for details see [3]. Starting from traditional pattern recognition techniques new approaches for retrieving structural blocks (motif, or domain, or an entire protein) within a protein or even within the Protein Data Base, are proposed in this paper. This proposal adopts the GHT and has been developed and experimented in various forms on the basis of the primitives' complexity from which the voting process can rise. The smallest aggregate can be the single SS; up to a direct use co-occurrences of two SSs for an exhaustive matching of the entire motif of $n \geq 3$ SSs. In

978-4-9906441-0-9 ©2012 ICPR

all methods the barycenter of the motif is assigned as Reference Point (RP) and in order to find the RP in a biomolecule a GHT voting process is applied. These techniques are similar for what refers the basic process and adopt the same Parameter Space (PS) but differ about the voting process. Moreover, in all methods, after the voting process, the points which have the expected number of votes are candidate as locations of the RPs of the searched motif (Note that it is known the expected peak intensity: the number of occurrences in the motif). To improve the robustness of these approaches the PS is scanned by a cubic mask (e.g. a unitary template) to integrate the votes in a neighborhood before searching the peaks.

Single SS method. This method uses as primitive for the voting process the Single SS (SSS). Two parameters, ρ and θ, are calculated to built the Reference Table (RT). As shown in Fig. 1a, ρ is the segment length between RP and SS midpoint and θ is the angle between SS axis and the quoted segment. The mapping rule which determines the locations of candidate RP, compatible with a given SS, in this case is a circle defined by the parameter ρ and θ (see Fig. 1b). So each SS of the protein analyzed increments on the PS a circle. The candidates RP locations are the points of max intersections of these circles.



a) b)

Figure 1. a) RT parameters ρ and θ; b) mapping rule: circle of candidate locations

SS co-occurrences. In this approach SSs co-occurrences (SSC) set up a local reference system, e.g. having the origin in the middle point of the first SS, the *y*-axis on the SS and the *x*-axis on the plane defined by the *y*-axis and the mid-point of the other SS, and *z*-axis orthonormal (see Fig. 2). In this system the motif RP coordinates are determined. For every couple in the motif, three parameters are calculated: *Md*, the Euclidean distance between middle points of two SSs; *Ad*, the shortest distance between two SSs axis; φ, the angle between two SSs translated to present common extreme. These parameters are stored in the RT. For each motif couple the mapping rule is reduced to a single location.

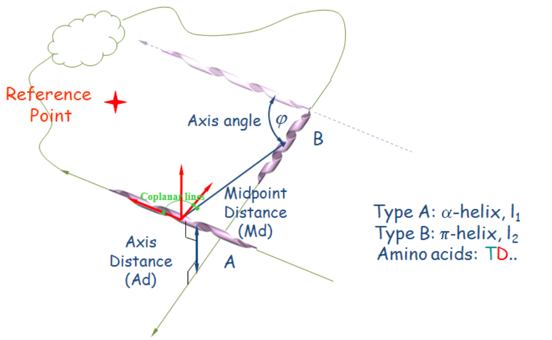


Figure 2. RT parameter terns for SSC: Md, Ad, φ

SSs triplets. In this method the primitives are SS Triplets (SST). In 3D, middle points of three SSs are joined and an imaginary triangle is composed. So, through the SS triplets a local reference system is set up, e.g. having the origin in the triangle barycenter, the *y*-axis passing through the farthest vertex, the *x*-axis on the triangle plane, and the *z*-axis following the triangle plane normal (see Fig. 3). The RT parameters are the lengths of the triangle edges. As in the previous case, the mapping rule is reduced to a single location.

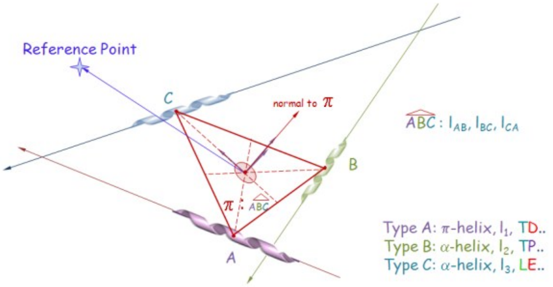

Reference Point
C
normal to π
ABC : l_AB, l_BC, l_CA
B
π : ABC
A
Type A: π-helix, l₁, TD..
Type B: α-helix, l₂, TP..
Type C: α-helix, l₃, LE..

Figure 3. RT triplets parameters: l_{AB}, l_{BC}, l_{CA}

Motif direct matching. This approach consists on an exhaustive Motif Direct Matching (MDM) among the motif and all possible blocks having the same number of motif SSs in the biomolecule. For each couple of SSs in both structures the tern *Md*, *Ad* and φ is calculated. The RT is composed of the set of motif terns and, for each tern, the relative RP location. For every correspondence between an SS motif couple and a couple of the candidate block a vote is given to the location of the candidate barycenter. If these locations collect the expected number of votes which corresponds to the number of motif couples, the motif is established.

3. Experimental Comparisons

Two sets of experiments are illustrated. In the first a motif composed of five SSs selected randomly inside the protein 7FAB (see Fig. 4) is searched in the same biomolecule. Figures 5 and 6 represent, after the voting process, the PSs respectively for the SSS implementation and for the SSC-SST-MDM approaches. The results in terms of location precision and computation time are given in Tab. 1.

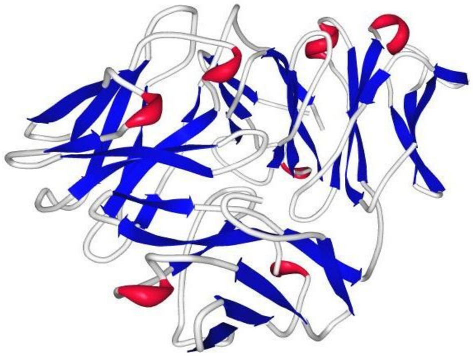


Figure 4. SSs of the 7FAB protein. Red lines: α-helices and blue lines: β-strands.

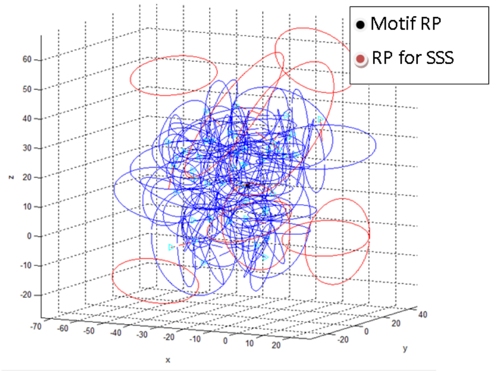


Figure 5. PS for SSS method. Five SSs motif, protein 7FAB.

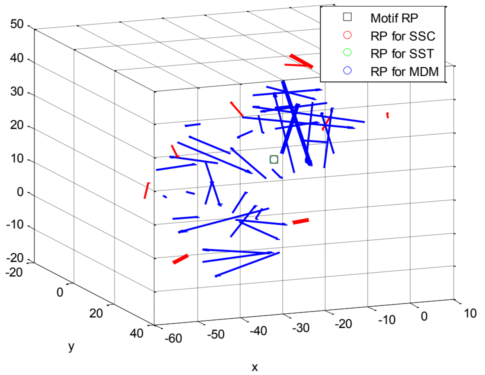


Figure 6. PS for SSC-SST-MDM. Five SSs motif, protein 7FAB. Bold lines represents motif SSs.

Table 1. Performances searching five SSs in 7FAB.

	Motif RP	Candidate RP	ER (%)	Time (sec)
SSS	x:-17.59 y: 9.51 z:15.21	x:-17.56 y: 9.46 z:15.17	0.28	108
SSC	x:-17.48 y: 9.17 z:15.48	x:-17.40 y: 9.14 z:15.48	0.34	43
SST	x:-17.48 y: 9.17 z:15.48	x:-17.48 y: 9.17 z:15.48	0.00	49
MDM	x:-17.48 y: 9.17 z:15.48	x:-17.45 y: 9.16 z:15.50	0.15	112

In the second set a well known motif, the Greek Key, composed of four β-splines, is searched in the 1FNB (see Fig. 7) protein, that contain one instance of the Greek Key, selected as motif. The results according to the taxonomy of the previous experiments are shown in Figs. 8 and 9, and in Tab. 2. From these results it is at a glance evident that SST technique over-performs the other methods. For the computing time point of view the best performance is of SSC (slightly better than SST) and the two worst solutions are the SSS and the MDM. This is certainly given, in the first case by the cumbersome mapping rule which complicates both the voting process and the peaks detection on the PS. For the MDM instead, being and exhaustive matching, the number of comparisons grows with polynomial complexity M^m where M is the number of candidate instances in the macromolecule and m is the number of SSs in the motif). For the precision view point the best performance, by far, is given by the SST, second the MDM, and the worst cases the SSS and SSC methods. Note that for the SST it is not always necessary to scan the PS by the cubic mask, to integrate the votes in a neighborhood, before searching the matching peaks. This happens because the PS has sometimes a sufficiently good signal to noise ratio. But at a first analysis it looks critical for this method the precision on the length evaluation (in [4] an in depth discussion about this problem is given); in fact, the triplet matching is based on a trivial discriminant function computed on side length differences and the mapping rule is slight. In all other approaches an essential role is played by the SS orientation which is a computation demanding parameter but add a significant contribution to the pure distances among SS.

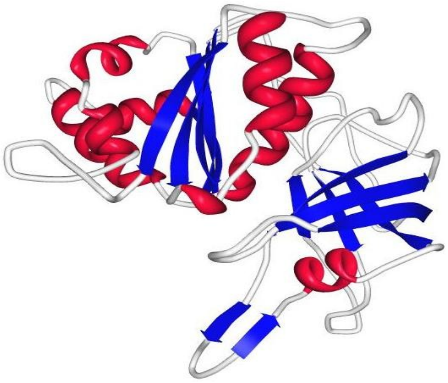


Figure 7. SSs of the 1FNB protein. Red lines: α-helices and blue lines: β-strands.

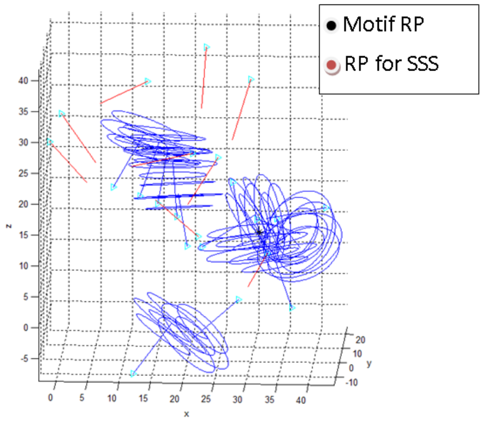


Figure 8. PS for SSS method. Greek Key motif, protein 1FNB.

We can conclude that the SST method is simple to implement and then computationally efficient, but for what refers robustness with respect to the other approaches we need to experiment on more complex structures, and with an extended statistical performance evaluation.

References

[1] E. Zotenko, R.I. Dogan, W.J. Wilbur, D.P. O'Leary and T.M. Przytycka. Structural fottprinting in protein structure comparison: The impact of structural fragments. *BMC Structural Biology*, vol.7, no.1, 7:53, 2007.

[2] O. Camoglu. T. Kahveci and A. Singh. PSI: Indexing protein structures for fast similarity search. *Bioinformatics*, vol.19, suppl.1, pp.81-83, 2003.

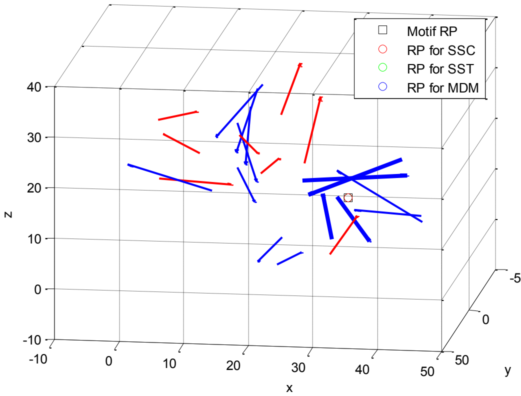


Figure 9. PS for SSC-SST-MDM. Greek Key motif, protein 1FNB. Bold lines represents motif SSs.

Table 2. Performances searching four SSs in 1FNB.

	Motif RP	Candidate RP	ER (%)	Time (sec)
SSS	x: 31.33 y: 1.14 z: 12.01	x: 31.41 y: 1.16 z: 11.94	0.32	35
SSC	x: 31.38 y: 1.08 z: 11.69	x: 31.33 y: 1.08 z: 11.79	0.33	4
SST	x: 31.38 y: 1.08 z: 11.69	x: 31.38 y: 1.08 z: 11.69	0.00	6
MDM	x: 31.38 y: 1.08 z: 11.69	x: 31.40 y: 1.12 z: 11.66	0.16	8

[3] V. Cantoni, A. Ferone, O. Ozbudak and A. Petrosino. Motif retrieval by exhaustive matching and couple co-occurrences, *CIBB'12*, accepted.

[4] V. Cantoni, A. Ferone, O. Ozbudak and A. Petrosino. Search of protein structural blocks through secondary structure triplets, *IPTA'12*, accepted.

[5] D. Eisenberg. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins, *Proc. of the National Academy of Sciences of the United States of America*, vol.100, no.20, pp.11207-11210, 2003.

[6] F. Jacob. Evolution and tinkering, *Science*, vol.96, no.1, pp.1161-1166, 1977.

[7] C.H. Chionh, Z. Huang, K.L. Tan and Z. Yao. Augmenting SSEs with structural properties for rapid protein structure comparison, *Proc. of the third IEEE Symposium on Bioinformatics and Bioengineering*, pp.341-348, 2003.