Comparison of GHT-Based Approaches to Structural Motif Retrieval

Alessio Ferone[1] and Ozlem Ozbudak[2]

[1] University of Naples Parthenope, Department of Applied Science,
Centro Direzionale Napoli - Isola C4, 80143, Napoli, Italy
alessio.ferone@uniparthenope.it

[2] Istanbul Technical University, Department of Electronics and Communication Engineering, 34469, Istanbul, Turkey
ozbudak@itu.edu.tr

Abstract. The structure of a protein gives important information about its function and can be used for understanding the evolutionary relationships among proteins, predicting protein functions, and predicting protein folding. A *structural motif* is a compact 3D protein block referring to a small specific combination of secondary structural elements which appears in a variety of molecules. In this paper we present a comparison between few approaches for motif retrieval based on the Generalized Hough Transform (GHT). Performance comparisons, in terms of precision and computation time, are presented considering the retrieval of motifs composed by three to five SSs for more than 15 million searches. The approaches object of this study can be easily applied to the retrieval of greater blocks, up to protein domains, or even entire proteins.

Keywords: Hough transform, Protein motif retrieval, Protein structure comparison.

1 Introduction

Proteins are central molecules in biological phenomena because they form the functional and structural cell components of every organisms and their function is determined, to a large extend, by their spatial structures. Starting from the linear sequence of amino acid given in Protein Data Bank (PDB) [1], two basic regular 3D structures can be envisaged [11], called SSs: *helices and sheets*. Small specific combinations of SSs, which appear in a variety of molecules, are called *motifs*, and can be considered as super-SSs [12].

Several motifs are packed together to form compact, local, semi-independent units, i.e. with more interactions within it than with the rest of the protein, called *domains*. As consequence, a structural domain forms a compact 3D structure, independently stable, and can be determined by two characteristics: its compactness and its extent of isolation.

From the quantitative view-point, a structural motif is a 3D structural block appearing in a variety of molecules and usually consists of just a few SSs, each

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 356–362, 2013.

one with an average of approximately 5 and 10 residues for sheets and helices respectively. The size of individual structural domains varies from about 25 up to 500 amino acids, but the majority (90%) has less than 200 residues with an average of approximately 100 residues. A protein in the average has 15 SSs or equivalently about 300 residues [10].

2 GHT-Based Approaches to Protein Structural Analysis

In recent years, many investigations have been made to analyze proteins at various structural levels [2,9,14], for more details see [3]. In particular, we developed various approaches for retrieving a structural block (a motif, or a domain, or ..., or an entire protein) within a protein or within the entire (PDB), by a 3D structure comparisons based on traditional pattern recognition techniques [7].

A central strategy is to exploit the Generalized Hough Transform (GHT) to implement blocks (of various sizes) retrieving through an exhaustive matching of structural elements. The searched block (let us call it model and m the number of its SSs) is in general decomposed in primitives consisting of a suitable subset of SSs. The subset can contain one, two, three, ..., up to m the entire block to be searched. The barycenter of the block model is usually assigned as Reference Point (RP) and the problem is the detection and the location of the RP in the macromolecule under scrutiny. The basic process is then a GHT voting process on the Parameter Space (PS) which is the 3D protein space.

In this work we compare the performance of four subsets consisting of the following primitive aggregates [5]: the single SS, the SS couples of the model, the SS triplets of the model and the entire model.

These subsets of primitive aggregates of the model are compared with all equivalent instances in the macromolecule or protein. For every correspondence, a vote is given to the candidate barycenter location, which is figured out with a special mapping rule determined from the RP position referred to the matched primitive aggregate of the model.

After the voting process, the points in PS which have the expected number of votes are candidate as location(s) of the RP(s) of the searched motif. Note that it is known the expected peak intensity: the number of occurrences of the primitive aggregates in the motif. In Tab. 1 a program sketch is given for searching all possible motifs in a set of M proteins.

2.1 Single Secondary Structure (SSS)

This method [13] adopts as primitive for the voting process the single SS. The SS being an helix or a sheet is represented by a straight segment on the regression line from all the C_α atoms of the segment. The extremes are determined by the projection of the terminal C_α atoms. The selective component of the Reference Table (RT) consists of two parameters, ρ and θ; ρ is the segment length between RP and SS midpoint A, and θ is the angle between SS axis and the segment $\overrightarrow{A-RP}$. The mapping rule which determines the candidate RP locations, for

Table 1. Algorithm for the retrieval of all possible r motifs contained in a set of M proteins. v is equal to 2 and 3 for couples and triplets respectively. p and p' are Md, Ad and φ for couple and direct matching, meanwhile are l_1, l_2, l_3 for triplets. r and s are respectively m and p for couples and terns and q and m for direct matching.

```
Input  : Protein .nss files; $N_i$: number of protein SSs; $m$: number of motif SSs
Output : Locations of candidate motifs in the accumulator $A_{RP}$, representing the
         parameter space.
for $i$=1 to $M$ do
    Calculate all m combinations of $N_i$: $P_q = C(N_i, m)$
    for $j$=1 to $P_q$ do
        Find the motif barycenter RP
        Calculate the number of motif primitives: $P_r = C(m, v)$
        Calculate the number of protein primitives: $P_s = C(N_i, v)$
        for $k$=1 to $P_r$ do
            Compute the three parameters:$p_1, p_2, p_3$ //RT constituents
        for $l$=1 to $P_s$ do
            Compute the three parameters: $p_l^{'}, p_2^{'}, p_3^{'}$
            for $k$=1 to $P_r$ do
                if $(p_1, p_2, p_3)$ matches with $(p_l^{'}, p_2^{'}, p_3^{'})$ then $A_{RP_l} = A_{RP_l} + 1$
        Compute the peaks in HS
        Assign the position with the expected votes as candidate RP
```

a given SS, is a circle on a plane perpendicular to the axis of the SS, with radius $r = \rho \sin \theta$, having the center along the SS axis and with a displacement $d = \rho \cos \theta$ from midpoint A. Each SS of the protein under scrutiny contributes on a circular locus on the PS. The candidate RP locations are detected as the points of intersections of these circles and, in ideal conditions, the number of intersection is just S_1.

2.2 Secondary Structure Couple Co-occurrences (SSCC)

An SS couple setup a local reference system, having the origin in the middle point of the first SS, the y-axis on its SS axis, and the x-axis on the plane defined by the y-axis and the mid-point of the second SS, then the z-axis is orthonormal to the previous two. In this reference system, the motif RP coordinates are determined, and for each couple of SSs of the protein under scrutiny that matches a motif couple, the candidate RP location is uniquely fixed [3].

The number of motif couples and protein couples is given by 2-combinations of m and N respectively: $C(m, 2)$, and $C(N, 2)$.

For every couple in the motif, a tuple is introduced in the RT where the selective component that characterizes the couple co-occurrence is composed by three parameters [10]: Md, the Euclidean distance between the middle points of the two SSs; Ad, the shortest distance between the two SSs axis; φ, the angle between the two SSs translated to present a common extreme. For each motif couple the mapping rule is reduced to a single location.

2.3 Secondary Structure Triplet Co-occurrences (SSTC)

In 3D, middle points of three SSs can be joined and an imaginary triangle is composed. So, through the SS triplets a local reference system is setup [6], e.g. having the origin in the triangle barycenter, the y-axis passing through the farthest vertex, the x-axis laying on the triangle plane and orthonormal to y-axis, and the z-axis following the triangle plane normal. With this reference system the motif RP coordinates are determined, and also in this case for each triplet of SSs of the protein under scrutiny that matches a motif triplet, the candidate RP location is uniquely fixed [4].

For every triplet in the motif, a tuple is introduced in the RT where the selective component that characterizes the triplet is composed by three parameters represented by the lengths of the triangle edges. For each motif triplet the mapping rule is reduced to a single location. The numbers of motif triplets and protein triplets are given by 3-combinations of m and N respectively: $C(m, 3)$, and $C(N, 3)$.

2.4 Entire Motif

This approach [8] consists on an exhaustive Motif Direct Matching (MDM) among the motif and all possible blocks (B) of the biomolecule under scrutiny having the same number of motif SSs. Let N and P_q be the number of SSs in the macromolecule and the cardinality of B respectively, i.e. P_q is the m-combinations in N, computed as $P_q = C(N, m)$. Each element of B must be compared with the motif.

So, for each couple of SSs in both biomolecule and motif, the terns Md, Ad and φ are computed. As in SSCC the RT tuples are composed for the discriminant component of the quoted set of motif terns, combined to the relative RP location as mapping rule. 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 block barycenter.

3 Experimental Results

The aim of these experiments is the evaluation of precision and computation time of the proposed approaches. A set of proteins has been randomly selected among the PDB 91939 structures having a number N_i of SSs ranging from 14 to 46 (a number of residue from 174 to 496). All possible structural blocks with m equal to three, four and five, have been retrieved for the SSCC and SSTC approaches. For the MDM approach, due to high computation time, the experimentation has been limited to just one thousand randomly cases selected. Due to the evident poor performance regarding both computation time and precision the SSS has been experimented just in a few cases. Table 2 reports the number of experiments for the SSCC and SSTC cases $\sum_{i=1}^{M} C(N_i, m)$ (column three: $C(N_i, m)$) and the cumulative and average time performances.

Table 2. Performances and protein parameters of the experimented set

Number of motif SSs: m	Number of motifs: P_q	Average search time and range per motif for SSCS (msec)	Average search time and range per motif for SSTS (msec)	Average search time and range per motif for MDM(msec)
3	105971	1.1 [0.6-1.5]	7.3 [0.9-11.7]	21.1 [2.5-42.7]
4	918470	1.4 [0.5-1.8]	11.2 [1.2-16.9]	310.1 [9.1-1039.6]
5	6455009	1.7 [0.5-2.2]	17.3 [1.4-24.4]	10647.5 [36.7-69353.3]

In all the nearly 15 million cases, the matching of candidates motifs with the RT tuples has been verified with a tolerance in the comparison parameters of $\epsilon = 1\%$. Figure 1 shows just an example of search of a motif composed by mixed helices and strands on the protein 7FAB containing 46 SSs. In all cases, the collected RP locations had exactly the expected number of votes/contributions (three, six and ten respectively for three, four and five SSs per motif). Moreover, no spurious peaks have been detected for the SSCC and SSTC cases; meanwhile for MDM case the detected spurious peaks follow the above mentioned rules. In details, the sets of second peaks have a ratio with the first peak of 1/3, 1/2, and 3/5 as expected.

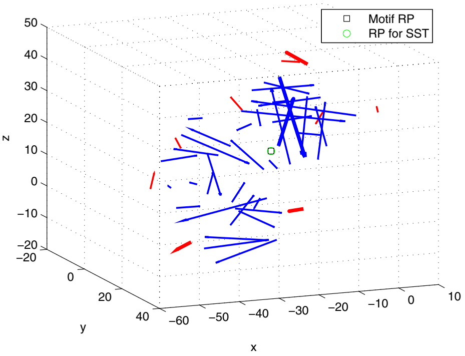

Fig. 1. Results obtained on searching a five SSs motif on the 7FAB protein. Red lines are α-helices and blue lines are β-strands. Bold lines form the five-SS motif (three α-helices and two β-sheets). RP and Max. vote coordinates are coincident.

No displacement from the true RP position could be measured: the motif location (just the one where the model was defined) perfectly coincided to the detected RP location.

From the computational time point of view the two worst solutions are the SSS and the MDM. This is certainly due, in the first case, to the cumbersome mapping rule which complicates both the voting process and the peaks detection on the PS. For the MDM instead, being an exhaustive matching, the number of comparisons grows with the polynomial complexity given above.

From the precision point of view we get good performances by the SSCC and SSTC, and also MDM, and the worst cases for the SSS that in the few experiment location precision, it was under 0.32%.

4 Conclusion

Important functionalities of proteins are determined by their 3D structure, so protein structures comparison and motif retrieving are areas of increasing interest in structural biology. This paper aims at comapring GHT-based approaches for retrieving a structural block on the basis of the 3D distribution of SSs.

All the analyzed approaches result effective for protein motif matching and retrieval. The approaches of SSCC and SSTC to compare motif and protein represented by SSs are simple to implement, robust, computationally efficient, and very fast with respect to the other implementations, even with GHT approach.

References

1. http://www.rcsb.org/pdb/
2. Camoglu, O., Kahveci, T., Singh, A.K.: Psi: indexing protein structures for fast similarity search. Bioinformatics 19(suppl. 1), i81–i83 (2003)
3. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Motif retrieval by exhaustive matching and couple co-occurrences. In: 9th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB (2012)
4. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Search of protein structural blocks through secondary structure triplets. In: 3rd International Conference on Image Processing Theory, Tools and Applications, IPTA (2012)
5. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Structural analysis of protein secondary structure by ght. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1767–1770 (2012)
6. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Protein motifs retrieval by ss terns occurrences. Pattern Recognition Letters 34(5), 559–563 (2013)
7. Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A.: Protein structural motifs search in protein data base. In: Proceedings of the 13th International Conference on Computer Systems and Technologies, CompSysTech 2012, pp. 275–281. ACM, New York (2012)
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. Chionh, C., Huang, Z., Tan, K., Yao, Z.: Augmenting sses with structural properties for rapid protein structure comparison. In: Proceedings of the Third IEEE Symposium on Bioinformatics and Bioengineering, pp. 341–348. IEEE (2003)
10. Dror, O., Benyamini, H., Nussinov, R., Wolfson, H.: Mass: multiple structural alignment by secondary structures. Bioinformatics 19(suppl. 1), i95–i104 (2003)
11. Eisenberg, D.: The discovery of alpha-helix and beta-sheet, the principal structural features of principal structural features of proteins. Proc. of the National Academy of Sciences of the United States of America 100, 11207–11210 (2003)

12. Singh, M.: Predicting Protein Secondary and Supersecondary Structure. Computer and Information Science Series. Chapman & Hall CRC (2005)
13. Cantoni, V., Mattia, E.: Protein structure analysis through hough transform and range tree. Biological Systems, Nuovo Cimento C 35(suppl. 1), 39–45 (2012)
14. Zotenko, E., Dogan, R., Wilbur, W., O'Leary, D., Przytycka, T.: Structural footprinting in protein structure comparison: The impact of structural fragments. BMC Structural Biology 7, 53 (2007)