IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11333-3

Colloquia: PR PS BB 2011

Protein motif retrieval through secondary structure spatial co-occurences

V. Cantoni([1]), A. Ferone([2]) and A. Petrosino([2])

([1]) *University of Pavia - Pavia, Italy*
([2]) *University of Naples "Parthenope" - Naples, Italy*

ricevuto il 30 Settembre 2011; approvato l' 1 Dicembre 2011

Summary. — The Generalized Hough Transform (GHT) allows to recognize general patterns once defined a model to be recognized, a reference point (RP) rigid with the model, and a mapping rule. This rule establishes the contributions in the parameters space; this space, generally speaking, is given by the parameters of a rigid motion leading to overlap a model item with an equal item detected on the unknown pattern. In this paper we discuss a particular implementation of the GHT applied to structural blocks retrieval into a protein data base. The spatial distribution of rigid arrangement of protein secondary structures (SSs) constitutes the items supporting the contributions. Starting from the co-occurrence of two not necessarily homogeneous SSs (two helices, one helix and one strand, or a β-sheet composed of two or more β-strands parallel or antiparallel) the approach can be generalized easily up to an entire motif composed of a few SSs. The main characteristic of this approach is that even for a simple couple of SSs, the mapping rule is reduced to a single location for the RP for each analogous couple found in the unknown pattern. This reduces very much the contributions (and then the signal-to-noise ratio) on the parameter space and simplifies the implementation and data structure, obviously with the drawback of a more elaborated pre-analysis.

PACS 87.18.Xr – Proteomics.
PACS 87.85.mk – Proteomics.
PACS 87.15.B- – Structure of biomolecules.
PACS 87.15.bd – Secondary structure.

1. – Preliminary statements

The importance of the study of structural building blocks, their comparison and their classification are instrumental to the study on evolution and on functional annotation of proteins, and has brought many methods for their identification and classification in proteins of known structure. These procedures, often automatic or semi-automatic, for reliable assignment are essential for the generation of the databases (especially as the number of protein structures is every time increasing) and the reliability and precision of the taxonomy is a very critical subject. This also because there is no standard definition of what a structural motif, a domain, a family, a fold, a sub-unit, a class, etc. really is, so that assignments have varied enormously, with each researcher (other than for each DB) using its own set of criteria.

It is quite explicit the aphorism: "Nature is a tinkerer and not an inventor" [1] by F. Jacob, 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.

In proteins, a structural motif is a three-dimensional structural element which appears in a variety of molecules and usually consists of just a few elements. Several motifs packed together to form compact, local, semi-independent units are called domains. The size of individual structural domains varies from between about 25 up to 500 amino acids, but the majority, 90%, has less than 200 residues with an average of approximately 100 residues. The term family as it is used in taxonomy should not be confused with protein family which is a group of evolutionarily related proteins, that is: proteins in a protein family descend from a common ancestor and typically have similar three-dimensional structures, functions, and significant sequence. Note that it is also often used the term super-*, where * can stand for motif, or domain, or family, or fold, or class.

There are several methods for defining protein secondary structure, but the Dictionary of Protein Secondary Structure (DSSP) [2] method is the most commonly used. The DSSP defines eight types of secondary structures, nevertheless, the majority of secondary prediction methods simplify further to the three dominant states: Helix, Sheet and Coil. Namely, the helices include 3_{10}-helix, α-helix and π-helix; sheets or strands include extended strand (in parallel and/or antiparallel β-sheet conformation); finally, coils include hydrogen bonded turn, bend, and amino acid residues which are not in any of the previous types. The structural analysis for protein recognition and comparison is conducted mainly on the basis of the two most frequent components [3]: the α-helices and the β-strands.

There are several DBs for structural classification of proteins; among them the most commonly used are Structural Classification Of Proteins (SCOP) and Class Architecture Topology and Homologous super families (CATH). They differ in domain and class definition and also because the former is more based on human expertise meanwhile the latter is a semi-automatic classifier. Another well-known DB is Families of Structurally Similar Proteins (FSSP), which is purely automatic [4].

2. – The Hough approach

The approach that we propose follows the modality of the Generalized Hough transform [5, 6] (G-Hough) that has been never applied in this context (something loosely similar is applied in eHiTS [7]) which can integrate methods for the comparison that exploits, at various levels of abstraction, various proteins representations: namely, at the atomic level, at the level of secondary structures such as structural motifs (both for proteins and RNAs) and at tertiary structures (such as domains) and even at entire protein level.

The aim of using the G-Hough is for the comparison and the search for structural similarity between a given protein and the proteins of a database (*e.g.* the Protein Data Base — PDB). Note that, if the searched structure is just a component of a protein (like a structural motif or a domain) the same method supports the detection and the

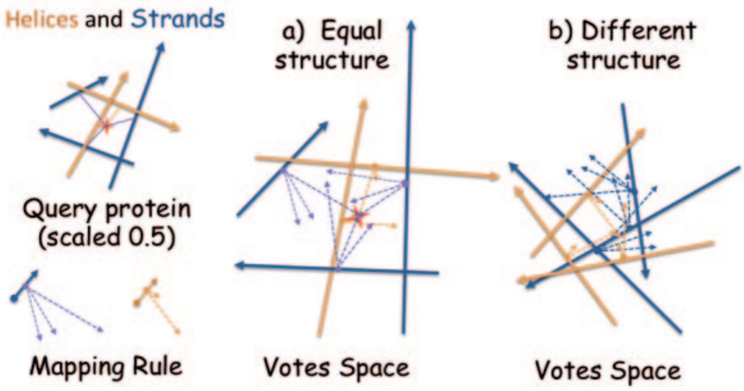

Fig. 1. – The principle, in 2D, of applying the Hough transform to protein block recognition. Top, left: model protein; bottom, left: mapping rules for strands and helices respectively, representing for each instance the compatible positions of the RP; center a): voting space with the model configuration, the peak collects 3 strands contributions plus 2 contributions from helices; right b): voting space with a completely different structure, votes are isolated.

statistical distribution of these components, possibly answering questions of permanence among different species and phylogenetic questions related to biological homology.

If we want to extract proteins similar to a given one (or also of a protein component, that we will call the model) from a bank of proteins, let we reference, for example, at the level of SSs, the approach is the following: every element (*e.g.* α-helix or β-strand) of the protein under examination (extracted from the DB) is superposed through a rigid motion (that is by roto-translation of the model) with each of the elements which possibly corresponds on the model. Then, for each possible correspondence a vote is given (that is a contribution with a convenient weight) to a particular candidate position of the model (defined by the parameters of the rigid motion that allows the overlapping in a suitable parameter space or Votes Space). In this way, every detail on the examined protein votes, with a weighted contribution, for a possible presence of the searched model. Having the accumulation of all the contributions of all the secondary components of an unknown molecule, if a particular attendance of the model obtains a sufficient number of contributions (it is obviously known the number of contributions that the model would obtain on itself), the similarity is detected (see fig. 1 in which a red star represent the position of a models Reference Point RP).

For a detailed description and 3D implementation of this approach see [8, 9] in this book. Summarizing, in a 3D space the mapping rule for each SS has a rotational symmetry: in fact it results in incomplete definition of the position of the point to vote, making voting of an entire circumference indispensable. This circumference is the rim of a cone centered in the object proteins secondary structure. If the secondary structures of the object proteins have a similar spatial arrangement as the ones of the model protein as many circumferences as the number of secondary structures in the model protein will all intersect, in the voting space, in one highly voted point, which can be used as a score for the comparison. This will not happen whether the two proteins have different structures, resulting in a low score.

Nevertheless, the richer and more reliable it is the given evidence (*e.g.* the type of helix — 3_{10}-helix, α-helix and π-helix —, the number of amino acid residues or length in atoms of the SS, the type of the amino acid residue, etc.) the more precise can be the

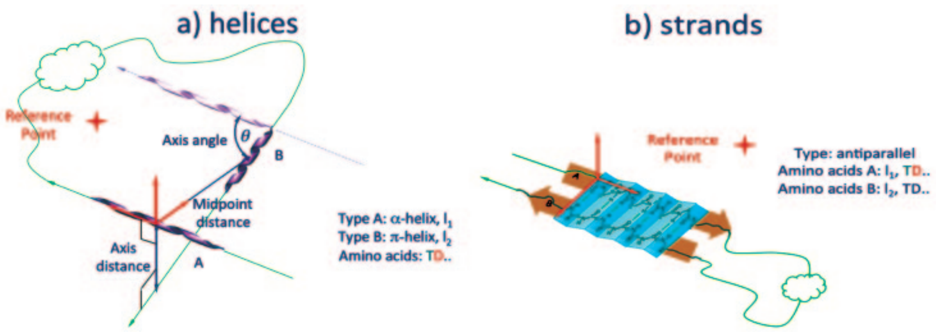

Fig. 2. – A couple of SSs set up a local reference system, *e.g.* having the origin in the middle point of the first SS, the y-axis on the SS, and the x-axis on the plane defined by the y-axis and the midpoint of the other SS. The couple parameters can be stored in the Reference Table (RT) in which the model couples are defined through: axes angle θ, Midpoint Distance MD, and Axes Distance AD. What is stored in the RT is in particular the displacement d of the Reference Point RP of the model (red cross) in the local reference system. In a) and b) the cases of homogeneous helices and strands are sketched. Obviously also heterogeneous couples can contribute.

contribution the simpler the analysis of the resulting votes becomes (in the parameters space in which every point corresponds to a possible presence of the model).

A different strategy is here proposed: instead of consider each SS isolated we can base our analysis on the co-occurrences of multiple SSs. Even with just two SSs the mapping rule is in general reduced to just one compatible location of the RP. In fact, two SSs are characterized by a displacement, as shown in fig. 2, defined by three parameters [10]: axes angle θ, Midpoint Distance MD, and Axes Distance AD. Multiple location mappings are possible if there are couples having equal parameter terns (or collinear SSs, but these contributions can be easily discarded).

3. – The G-Hough and SSs co-occurences

Around the axis of a SS a local reference system can rotate but fixing an external point (*e.g.* the middle point of a selected second SS) no degree of freedom remains and the RP position is unambiguously fixed (see fig. 2). The solution is then implemented in a few steps: for each SS of the unknown molecule the neighborhood is investigated for co-occurrences: that is, the neighborhood is analyzed to discover if there are SSs compatible with the parameter terns of the Reference Table of the couples of SSs of the model; for each co-occurrence a contribution is given for the possible existence of searched motif in the compatible location(s). In fig. 3 a sketch of this process is given.

Being n, the number of SSs, and k the number of occurrences requested (in this case 2), the number of possible co-occurrences (when order does not matter) is given by

$$C_{n,k} = \frac{n!}{(n-k)!k!} = \binom{n}{k}. \tag{1a}$$

With a motif of 4 SSs and considering just the couples, the number of co-occurrences is 6, unlike 12 permutations. As an example herewith is given the search a well-known

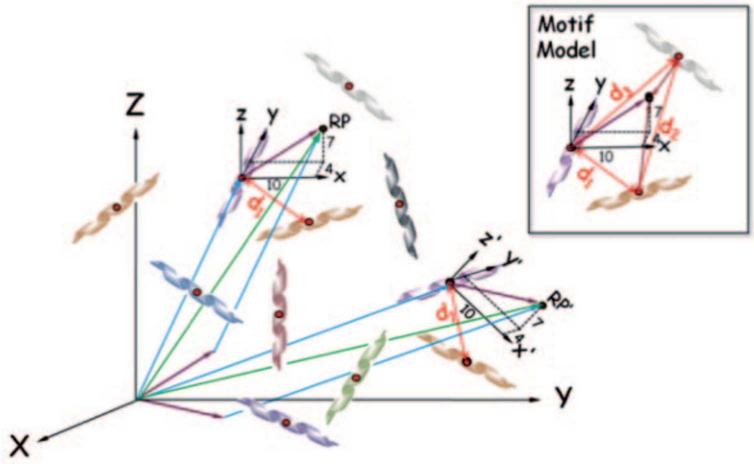

Fig. 3. – The voting process for a couple of helices. On the top right a sketch of the motif model containing just three couples having displacements δ_1, δ_2, δ_3 and midpoint distances d_1, d_2, d_3. On the left the unknown molecule, in which two couples of helices, having a midpoint distance d_1, have been detected. Each one supplies a contribution for the possible position of its RP with the displacement δ_1. If the complete motif model is present in its locations three contributions will be cumulated (it is just the case for the top RP, meanwhile RP on the left has just one vote).

motif, the Greek-key one represented in fig. 4, in a molecule of the protein 1FNB shown in fig. 5 containing an instance of this motif. When the couples discriminator is the complete tern of parameters θ, MD and AD the contributions are totally cumulated in the location corresponding to the RP of the valid instance. If an ambiguity is introduced, as for example only one of the tern parameters, is considers others spare contributions can appear.

This is the case of figs. 6, 7 and 8 in which only the midpoint distance constraint is applied. In this case other than the expected peak in the proper location of 12 contributions (unlike permutations) the existence of another couple at the same relative distance generates other 12 single contribution distributed in the voting space.

Fig. 4. – A well-known Greek-key motif: a series of four consecutive β-strands. On the right the representation in a picture generated by PyMOL on PDB file 1FNB with residues 56-116 displayed and everything else masked; in the middle a topology diagram; in the right a couple of decorational patterns used in ancient Greek vases at the origin of this motif name [11].

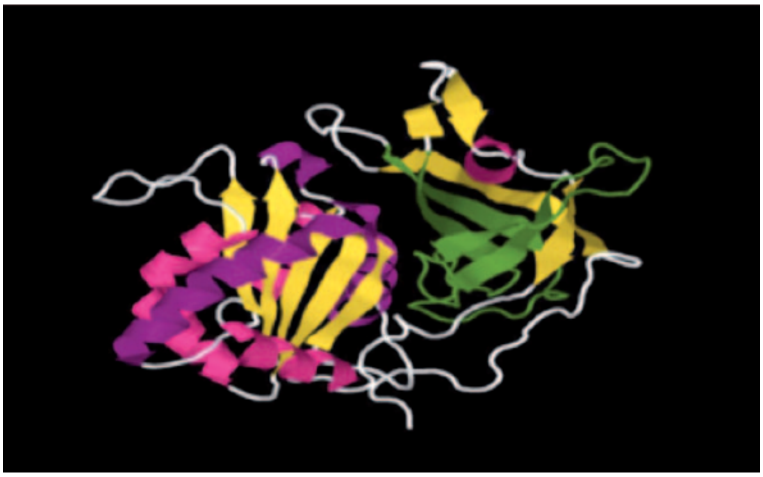

Fig. 5. – A picture generated by PyMOL on PDB file 1FNB rotated by $\pi/2$ for format reasons. In green the Greek-key motif (residues 56-116).

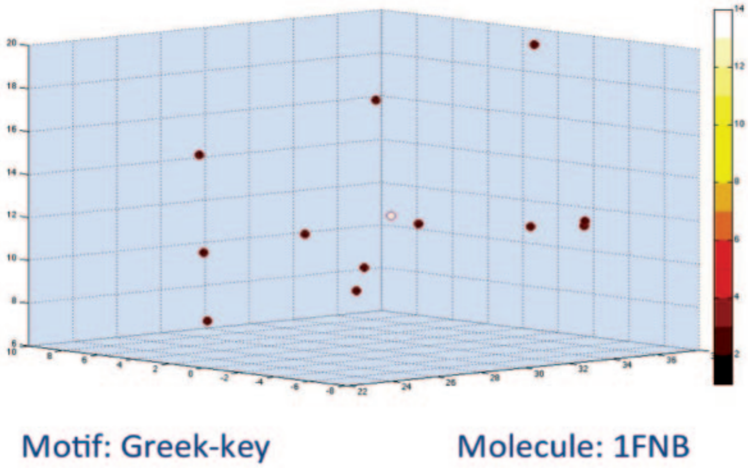

Fig. 6. – The votes space after searching the Greek key in the 1FNB protein. The couple constraint applied is limited to the MD parameter, and then there is an augmented possibility to introduce spare votes. In fact, in the RP of the motif position a total of 12 contributions (white circle) are gathered, while a distribution of 12 spare votes in red is present.

4. – Conclusion

It is worth pointing out that the G-Hough transform is indeed suited for parallel implementation (for example at the protein level, *e.g.* more blocks of a protein can "vote" at the same time, but also at the model level, *e.g.* helices and strands can vote at the same time), then the technique can be easily implemented on parallel machines, thus reducing the required computation time.

Obviously, this method can only supply an approximate solution, because it is based on the SS (that with the usual packages like DSSP or STRIDE on the average 4.8% of the

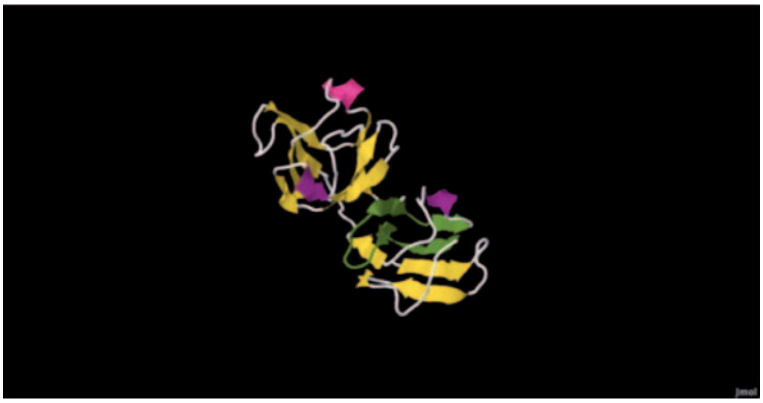

Fig. 7. – A picture generated by PyMOL on PDB file 4GCR rotated by $\pi/2$ for format reasons. In green the Greek-key motif (residues 34-62).

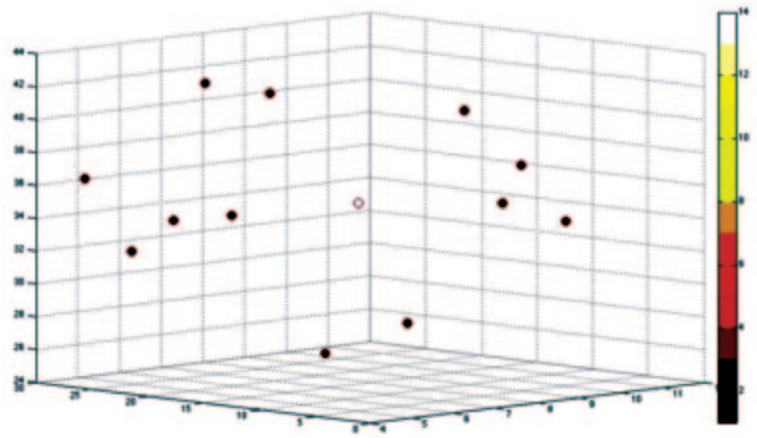

Fig. 8. – The votes space after searching the Greek key in the 4GCR protein. The couple constraint applied is limited to the MD parameter, and then there is an augmented possibility to introduce spare votes. In fact, in the RP of the motif position a total of 12 contributions (white circle) are gathered, while a distribution of 12 spare votes in red is present.

target residues were differently assigned, this number reaching 12% for certain targets), since the chemistry of the amino acid residue is not considered, because the proteins are considered rigid. This last hypothesis is often only partially verified [12]: the packing of the protein is usually much tighter in the interior than in the exterior producing a solid-like core and a more flexible surface. Nevertheless, the results of this approach will identify a limited subset for a sub-sequent phase of refining, restricted only to a few proteins, to which the analysis can be conducted at a very sophisticated level or even in an experimental way. Moreover, it is possible to widen the solution by taking care also of semi-rigid objects, so covering also the cases in which the secondary structures are acting like hinges of a door, allowing an opening and closing motion to occur [13].

These are obviously preliminary results, and extended experimentation is now required to properly validate this new approach, but, as it has been demonstrated, the results more than just testify the feasibility, look very promising.

REFERENCES

[1] Jacob F., *Science*, **196** (1977) 1161.
[2] Kabsch W. and Sander C., *Biopolymers*, **22** (1983) 2577.
[3] David E., *Proc. Nat. Acad. Sci. U.S.A.*, **100** (2003) 11207.
[4] Day R., Beck D. A. and Daggett V., *Protein Sci*, **12** (2003) 2150.
[5] Ballard D. H., *Pattern Recognition*, **13** (1981) 111.
[6] Sloan K. R. and Ballard D. H., *Experience with the generalized Hough transform, Proceedings of the 5th International Conference on Pattern Recognition & Image Processing* (1980).
[7] Zsoldos Z., Reid D., Simon A., Bashir S. S. and Johnson A. P., *J. Mol. Graph. Modell.*, **26** (2007) 198.
[8] Mattia E., *Protein structure comparison and motif retrieval by the generalized Hough transform*, IUSS Thesis, Pavia, Italy, 2010.
[9] Cantoni V. and Mattia E., *Protein structure analysis through Hough transform and range tree*, these proceedings.
[10] Dror O., Benyamini H., Nussinov R. and Wolfson H., *Bioinformatics*, **19** (2003) 95.
[11] Li M., *CS 882 Course Notes*, University of Waterloo.
[12] Hayward S., *Proteins*, **36** (1999) 425.
[13] Zhou Y., Vitkup D. and Karplus M., *J. Mol. Biol.*, **285** (1999) 1371.