Pattern Recognition Letters 34 (2013) 559–563

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Protein motifs retrieval by SS terns occurrences

V. Cantoni [a,*], A. Ferone [b], O. Ozbudak [c], A. Petrosino [b]

[a] *University of Pavia, Department of Electrical and Computer Engineering, Via A. Ferrata, 1, 27100 Pavia, Italy*
[b] *University of Naples Parthenope, Department of Applied Science, Centro Direzionale Isola C4, 80133 Napoli, Italy*
[c] *Istanbul Technical University, Department of Electronics and Communication Engineering, 34469 Istanbul, Turkey*

ARTICLE INFO

Article history:
Received 20 July 2012
Available online 10 December 2012

Communicated by S. Sarkar

Keywords:
Protein motif retrieval
Secondary Structures
Similarity search
Generalized Hough Transform
Pattern recognition
Structural biology

ABSTRACT

This paper describes a new approach to the analysis of protein 3D structure based on the Secondary Structure (SS) representation. The focus is here on structural motif retrieval. The strategy is derived from the Generalized Hough Transform (GHT), but considering as structural primitive element, the triplet of SSs. The triplet identity is evaluated on the triangle having the vertices on the SS midpoints, and is represented by the three midpoints distances. The motif is characterized by the complete set of triplets, so the Reference Table (RT) has a tuple for each triplet. Tuples contain, beside the discriminant component (the three edge lengths), the mapping rule, i.e. the Reference Point (RP) location referred to the triplet. In the macromolecule to be analyzed, each possible triplet is searched in the RT and every match gives a contribution to a candidate location of the RP. Presence and location of the searched motif are certified by the collection of a number of contribution equal (obviously in absence of noise and ambiguities) to the RT cardinality (i.e. the number of motif triplets). The approach is tested on twenty proteins selected randomly from the PDB, but having a different number of SSs ranging from 14 to 46. The retrieval of all possible structural blocks composed by three, four and five SSs (very compact and completely distributed) have been conducted. The results show valuable performances for precision and computation time.

1. Introduction

Many evolutionarily and functionally meaningful links between proteins come to light through the analysis of their spatial 3D structures. Protein structure and morphology are significant to understand and predict their functionality (Shuoyong et al., 2007). Protein structure comparison is an important issue that helps biologists to understand various aspects of protein function and evolution. For this reason protein comparison and retrieval are basic issues that helps biologists to comprehend various aspects of the phylogenetic evaluation and of the tasks performed i.e. proteins role in the machinery of life.

The protein 3D structure is vitally important in many biological applications, such as rational drug design. The retrieval of a protein 3D structure can be achieved by different experimental and bioinformatics methods. To this aim, X-ray crystallography is a powerful tool although time-consuming, expensive, and not feasible for all proteins (e.g. so far very few membrane protein structures have been determined). Nuclear magnetic resonance (NMR) is another tool that can be employed to determine the 3D structures of membrane proteins, even though time-consuming and costly. In order to acquire the structural information in a timely manner, it is possible to adopt various bioinformatics tools (see, e.g. (Li et al., 2011; Ma et al., 2012; Wang and Chou, 2011; Chou et al., 1997; Wang and Chou, 2012) and a review Chou, 2005). The present study is devoted to develop a novel method to search a database of protein structures for 3D patterns of secondary structural elements.

Structural comparison and protein structure retrieval problems have been studied in the structural biology community. In most cases just representing the set of the protein by a set of SS elements. Can and Wang (2003) present a new method for conducting protein structure similarity searches and applies differential geometry knowledge on their 3D structure for extracting “signatures” such as curvature, torsion and SS type. Camoglu et al. (2003), to find similarities in protein database, build an indexing structure based on SS elements triplets by using R-tree. Chionh et al. (2003) propose the SCALE algorithm to compare protein 3D structures through matrices that utilizes angles and distances between SS elements. Krissinel and Henrick (2004) describe the Secondary Structure Matching (SSM) algorithm for comparison in 3D, including an original procedure for matching graphs built on the protein’s SS elements, that is followed by an iterative 3D alignment of protein backbone C_α atoms. Chi et al. (2004) design a fast system for protein structural block retrieval by using image based distance matrices and multidimensional indices. The 1D string

[*] Corresponding author. Tel.: +39 0382 985358; fax: +39 0382 985373.

E-mail addresses: virginio.cantoni@unipv.it (V. Cantoni), alessio.ferone@uniparthenope.it (A. Ferone), ozbudak@itu.edu.tr (O. Ozbudak), alfredo.petrosino@uniparthenope.it (A. Petrosino).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.12.003

representation of local protein structure retains a degree of structural information. This type of representation can be a powerful tool for comparison and classification. Friedberg et al. (2006) described the use of a particular structure fragment library, denoted as KL-strings, for the 1D representation of protein structure and developed an infrastructure for comparing structures with 1D representation. Shuoyong et al. (2007) developed a program, ProSMoS (Protein Structure Motif Search) to find fold-level structural similarities and to search for the presence of structural motifs. This package searches a library of protein structures for user defined 3D patterns of SS elements. Also a web server to make a pattern-based search, using interaction matrix representation of protein structures (Shuoyong et al. (2009)), has been developed. Albrecht et al. (2008) propose a different approach and apply data reduction techniques directly to the protein structure and convert 3D data into 2D so accelerating the structural comparisons. Zotenko et al. (2007) propose an approach to speed up protein comparison by mapping a protein structure to a high-dimensional vector and approximating structural similarity by suitable distances between the corresponding vectors. Zhang et al. (2009) by a transition probability matrix and some structural characteristic vectors of proteins developed FDOD (Function of Degree of Disagreement) a score scheme to measure the protein similarity. Nguyen and Madhusudhan (2011) propose a new algorithm, CLICK, to capture such similarities. This method optimally superimposes a pair of protein structures independently of their topology and can generally be applied to compare any pair of molecular structures represented in Cartesian coordinates as exemplified by the RNA structure superimposition benchmark. Cantoni and Mattia (2012) and Cantoni et al. (2012) made a study for retrieving structural motifs by using GHT and range tree. This approach is completely new, because the analysis is based on the 3D spatial distribution of the SS.

In this paper, a new approach for structural block retrieval based on protein SS comparison is proposed. Here, triangles joining the middle points of the SS triplets are considered as “structural elements” and all the block triangles are compared with all the macromolecule triangles. The focus of the paper is on the retrieval of an existing structural block completely and precisely known. The block can be defined without constraints such as adjacency, distance limits, homogeneity, etc. The only constraints is that the SS components exist in the protein macromolecule.

The rest of the paper is organized as following. Section II introduces the GHT and the triangle approaches. Section III represents the experiments and their results. In the final session IV a brief discussion and the future works are described.

2. Methodology

In this paper a novel approach, GHT-based, for motif retrieval is proposed. The GHT is used for comparison and search of structural similarity between a given structural block (a motif or a domain or the entire protein) and the proteins of a database like the PDB. Note that, if the searched structure is just a component of a protein (like a structural motif or a domain) the same algorithm supports the detection and the statistical distribution of these components.The primitive patterns to which is applied the cumulative voting procedure are triplets of SSs, that is the structural elements are the triangles having the vertices in the middle point of the SS triplets.

2.1. The triangular structural elements

In this algorithm we use SS triplets for motif retrieval in protein macromolecule. In three-dimension, 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. 1).

The coordinates of the RP are determined with respect to this local reference system. A structural block, that in the sequel we name motif, is defined by a few SSs, and for each motif a RP is fixed in the center of gravity of the midpoints of these SSs. Being n the number of motif SSs, the number t of triplets/triangles is given by:

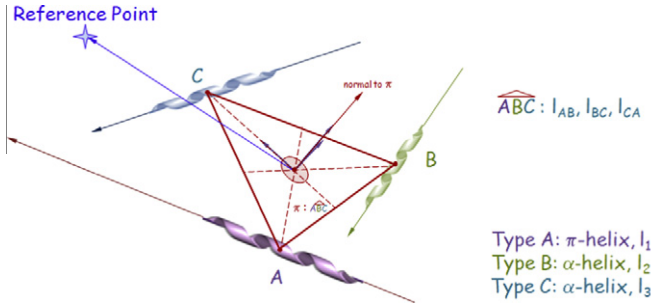

Fig. 1. Local reference system representation for the A, B, C triplet. The comparison parameters are the length of triangle edges (i.e. the mid-points distances). Other discriminant parameters can be considered such as: type of SS (π-helices, α-helices, β-strands, etc.), SS lengths (i.e. number of amino acids), types of amino, etc.

Table 1
Algorithm for the retrieval of all possible r motifs contained in a set of M proteins.

```
Input: Protein DSSP 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 $N_i : r = C(N_i, m)$
  for $j$ = 1 to $r$ do
    Find the motif barycenter RP
    Calculate the number of motif triangles: $c = C(m, 3)$
    Calculate the number of protein triangles: $p = C(N_i, 3)$
    for $k$ = 1 to $c$ do
      Compute the edge lengths of motif triangle: $d1_k, d2_k, d3_k$ //RT
   constituents
    for $l$ = 1 to $p$ do
      Compute the lengths of protein triangle: $d1_l, d2_l, d3_l$
      for $k$ = 1 to $c$ do
        if match($d1_k, d2_k, d3_k$ and $d1_l, d2_l, d3_l$) then $A_{RPl} = A_{RPl} + 1$
    Compute the peaks in HS
    Assign the position with the expected votes as candidate RP
```

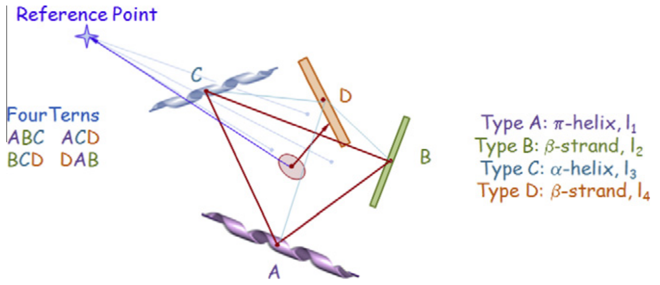

Fig. 2. A heterogeneous motif composed of two helices A and C, and two strands B and D. In this case $t = 4$, the corresponding triangles are shown on the left of the figure. In detail, it is represented the center of gravity of triangle ABC and it is shown the correspondence displacement, i.e. the RP position. If the motif is completely contained in the macromolecule the corresponding RP location receive one contribution for each of the four triangles, as shown (the other three contributions are just sketched).

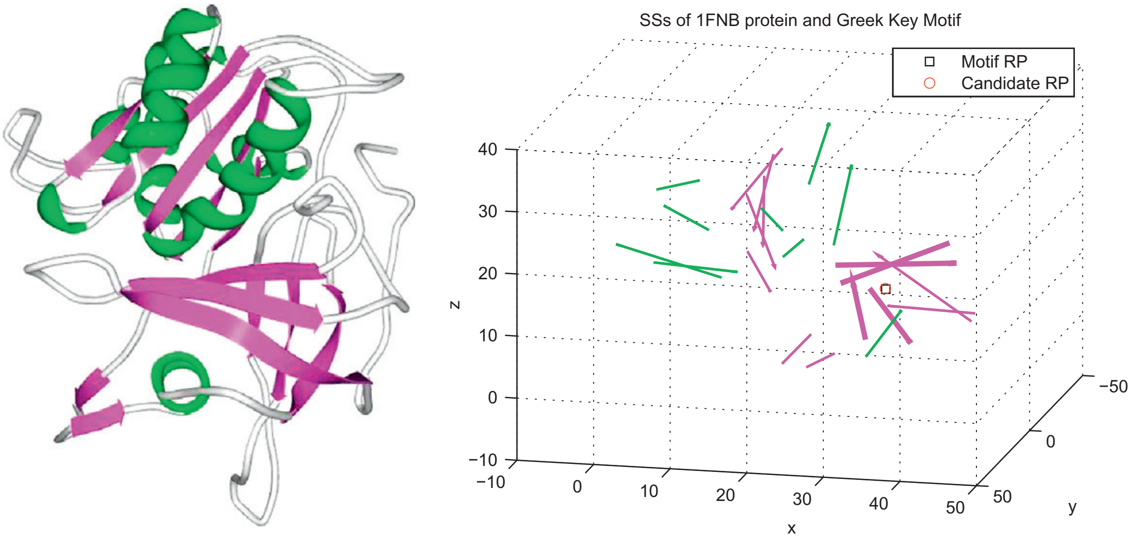

Fig. 3. SSs of the 1FNB protein. Green lines are α-helices and pink lines are β-strands. Bold lines form the four-SS motif. RP and Max. vote coordinates are coincident. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

$$t = C(n,3) = \frac{n!}{(n-3)!3!} \quad (1)$$

For each triangle in the motif the three length values are used to characterize and identify the triplets. So the tuples of the Hough RT is composed by the three edge length parameters (for matching discrimination) and by the three coordinates of the RP expressed in the corresponding local triangle reference system (mapping rule for locating the candidate position of the RP). The cardinality of the RT is the number of motif triangles t.

Being N the number of SSs in the protein macromolecule, the number T of triplets is given by: $T = C(N,3)$. For each of the T triplets, the edge lengths are computed. Then motif triangles t and macromolecule triangles T are compared. For every match a vote is given according to the displacement defined by the candidate RP position of the matching triplet. Table 1 shows a sketch of this algorithm for searching all possible motifs in a set of M proteins. Fig. 2 shows an example of four heterogeneous SSs motif. Generally, as established above, the RP is fixed on the motif barycenter, instead in Fig. 2 it has been located outside for evidencing graphically the voting process. Fig. 3 shows just an example of search of the Greek Key motif composed of four β-strands on the protein 1FNB containing 22 SSs.

3. Experiments and performances

The aim of this experiment is to test precision and computation time of the proposed method.

In order to assess the statistical performances the following three cross-validation methods are often used: independent dataset test, subsampling (or K-fold cross validation) test, and jackknife test (Chou and Zhang, 1995). In particular, the jackknife test is considered less arbitrary in that it always produces a unique result for a given dataset. The rationale is: (i) for the independent dataset test, the selection of the test samples could be quite arbitrary unless the number of independent samples is sufficiently large, thus leading to potentially different results (Chou and Zhang, 1995); (ii) for the subsampling test, the cross-validation is usually employed, but the number of possible selections in dividing a dataset can grow quite fast even for a very simple dataset, as proved in (Chou, 2011). Therefore, only a small fraction of the possible selections can be taken into account. Also in this case, different selections can lead to different results even for the same dataset; (iii) in the jackknife test, all the samples in the benchmark dataset are singled out one-by-one and tested by training through the remaining samples. This test can exclude the ”memory” effect and also the arbitrariness problem because the result is always unique for a given benchmark dataset. As a consequence, the jackknife test has been increasingly used to assess the performance of various predictors (see, e.g. (Chen et al., 2012; Esmaeili et al., 2010; Chou, 2001; Guo et al., 2011; Hayat and Khan, 2012; Mei, 2012; Zou et al., 2011; Chou et al., 2012; Wu et al., 2012)).

In this connection a set of proteins has been randomly selected among the PDB 82160 structures[1] having a number of SSs ranging from 14 to 46 (a number of residue from 174 to 496). All possible structural blocks having three, four and five SSs each, have been retrieved; note that each motif (searched structural block) could be homogeneous or heterogeneous (i.e. constituted by an arbitrary number of helices h or splines s the only rule is that $h+s$ equals the number of motif SSs). Table 2 reports the number of experiments (i.e. column three: number of motifs) and the cumulative and average time performances (the given computation times are related to a desktop computer with a processor Intel Core 2 Duo 6600, 2.4 GHz, 2 GB RAM).

In all the about 7.5 million cases, the matching of candidates terns with the RT tuples has been verified with an edge length tolerance $\varepsilon = 1\%$. In all cases, the collected RP locations had exactly the expected number of votes/contributions (one, four and ten respectively for three, four and five SSs per motif). Moreover, no spurious peaks have been detected, and no displacement from the true RP position could be measured: the motif location (just the one where the model was defined) perfectly coincides with the true RP location. Details on the number of tests and the average search time per motif and for proteins are given in Figs. 4 and 5, respectively.

The average searching time is confined to about a millisecond for small proteins (ranging from 0.75 to 1.5 for about 10 SSs) to a range among 6 ms to 8 ms for the greatest proteins (about 50 SSs, obviously the lower limit corresponds to motif composed of three SSs).

[1] http://www.rcsb.org/pdb/.

Table 2
Proteins and a few important parameters.

Number of proteins	Number of SSs per motif	Number of motifs	Total searching time (sec)	Average searching time per motif (msec)
20	3	105,971	768.508	7.3 [0.9–11.7]
20	4	918,470	10303.806	11.2 [1.2–16.9]
20	5	6,455,009	111809.428	17.3 [1.4–24.4]

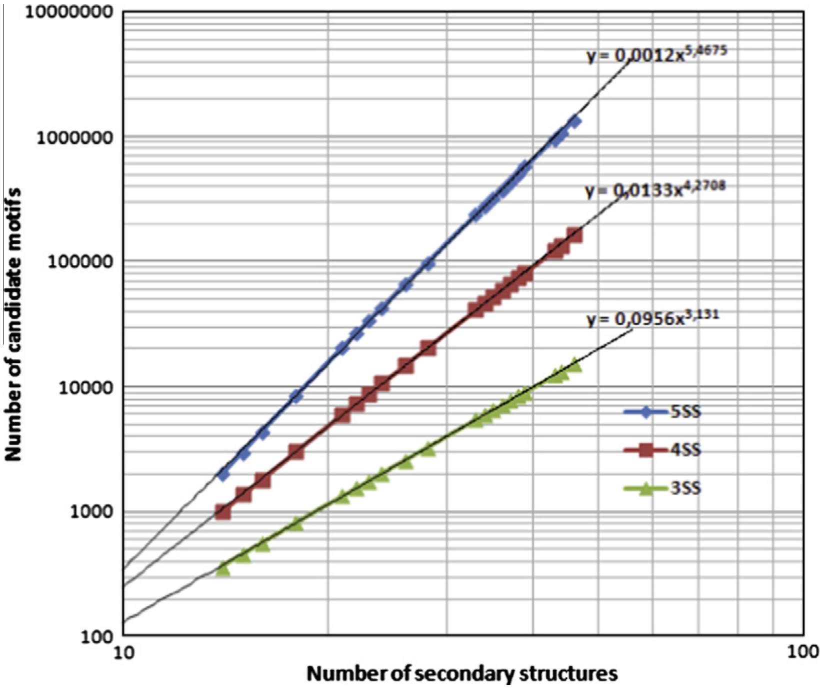

Fig. 4. Number of candidate motifs tested for each protein of the benchmark. All possible combination of three, four and five SSs have been tested. The protein set covers a range of 10 to 50 secondary structures.

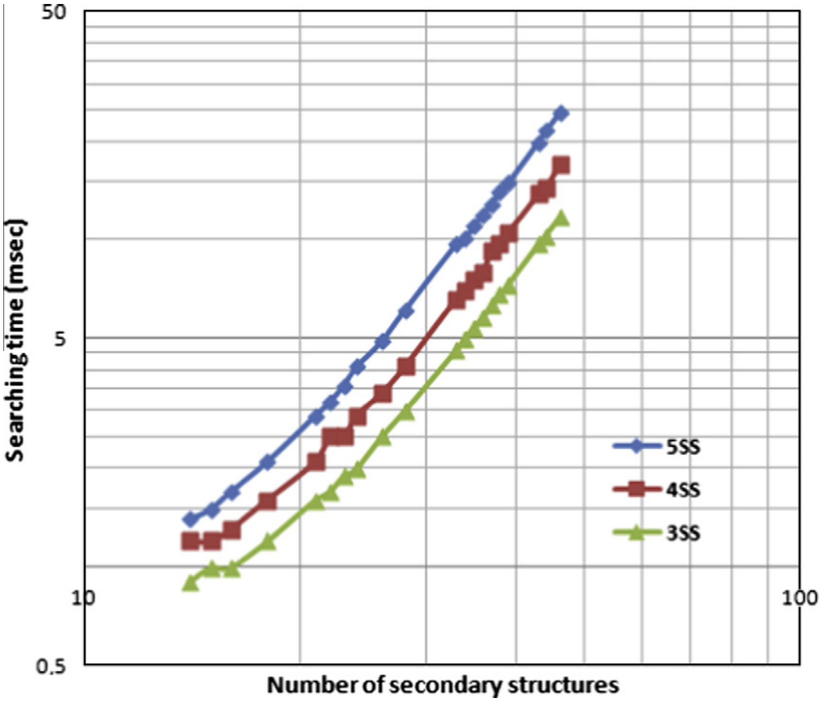

Fig. 5. The resulting searching time for each motif of three, four and five SSs as versus the protein size i.e number of protein secondary structure.

4. Conclusions

Comparing protein structures and retrieving motif remain an active area of development in structural biology. The new approach refers to the structural analysis of the 3D distribution of SSs. In this paper the problem of combining SS triplets for searching general motifs (details are given for the cases of three, four and five SSs), in protein structure datasets is considered. The comparison is conducted, by considering triangles as primitives (or, as basic structural elements) using motif and macromolecule triplets. We form the imaginary triangles joining the middle points of three SSs and use edge lengths as discriminant parameters. Then, comparing the motif triangles to macromolecule triangles and, for the resulting matchings, by voting a candidate RP location – through a particular GHT implementation – the motif existence can be established.

The results show that the candidate RP is located with very high precision and the motif is retrieved from the macromolecule with the expected number of contributions. In the experimented cases, due to the high precision in the edge model lengths, the integration on a neighborhood of the vote was completely unuseful (in fact the motif is defined directly on the macromolecule data).

It can be concluded that the proposed approach based on the GHT is very effective for protein motif matching and retrieval. This new approach to compare motif and protein represented by SSs is simple to implement, robust, computationally efficient, and very fast with respect to the other implementations, even with GHT approach (Cantoni and Mattia, 2012).

In this paper we discussed a new tool, in which the structural components were represented by oriented 3D segments (with possible detailed attributes such as a type attribute of the segment, i.e., in proteins, helix or spline or even a more detailed description). Here we identify directly the motif on the macromolecule under test and we show how effective and efficient is the technique. In a more application oriented paper, we will discuss the modeling statistics and the adoption of a suitable metric for standard protein motifs (that is for a homologous structural blocks common to different proteins), so the focus will be a common component for a set of proteins that performs a similar function, not restricted to a block instance of a given protein.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009), we provide the SSTerns Occurrences package to experiment our proposal in: <http://vision.unipv.it/bioinformatics/tools.php>.

References

Albrecht, B., Grant, G.H., Sisu, C., Richards, W.G., 2008. Classification of proteins based on similarity of two-dimensional protein maps. Bioph. Chem. 138 (1–2), 11–22.

Camoglu, O., Kahveci, T., Singh, A., 2003. PSI: Indexing protein structures for fast similarity search. Bioinformatics 19 (suppl. 1), 81–83.

Can, T., Wang, Y.F., 2003. CTSS: A robust and efficient method for protein structure alignment based on local geometrical and biological features. In: Proc. IEEE Computer Society Conf. on Bioinformatics, pp. 169–179.

Cantoni, V., Mattia, E., 2012. Protein structure analysis through Hough transform and range tree. New tools and methods for pattern recognition in complex biological systems. Nuovo Cimento C 35 (5, suppl. 1).

Cantoni, V., Ferone, A., Petrosino, A., 2012. Protein motif retrieval through secondary structure spatial co-occurrences. New tools and methods for pattern recognition in complex biological systems. Nuovo Cimento C 35 (5, Suppl. 1).

Chen, C., Shen, Z.B., Zou, X.Y., 2012. Dual-layer wavelet svm for predicting protein structural class via the general form of Chou's pseudo amino acid composition. Protein Pept. Lett. 19, 422–429.

Chi, P.H., Scott, G., Shyu, C.R., 2004. A fast protein structure retrieval system using image based distance matrices and multidimensional index. Internat. J. Software Eng. Knowl. Eng. 15 (3), 527–545 (Special Issue on Software and Knowledge Engineering Support in Bioinformatics).

Chionh, C.H., Haung, Z., Tan, K.L., Yao, Z., 2003. Augmenting SSEs with Structural Properties for Rapid Protein Structure Comparison. In: Proc. Third IEEE Symposium on Bioinformatics and Bioengineering, pp. 341–348.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genetics 43, 246–255 (Erratum: Chou, K.C. 2001. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, Genetics, vol. 44, 60).

Chou, K.C., 2005. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J. Proteome Res. 4, 1681–1686.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247 (50th Anniversary Year Review).

Chou, K.C., Shen, H.B., 2009. Review: Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92. http://dx.doi.org/10.4236/ns.2009.12011 <http://www.scirp.org/journal/NS/>.

Chou, K.C., Zhang, C.T., 1995. Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Jones, D., Heinrikson, R.L., 1997. Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett. 419, 49–54.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosystems 8, 629–641.

Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.

Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z., Godzik, A., 2006. Using an alignment of fragment strings for comparing protein structures. Bioinformatics 23 (2), 219–224.

Guo, J., Rao, N., Liu, G., Yang, Y., Wang, G., 2011. Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. J. Comput. Chem. 32, 1612–1617.

Hayat, M., Khan, A., 2012. Discriminating outer membrane proteins with Fuzzy K-nearest neighbor algorithms based on the general form of Chou's pseaac. Protein Pept. Lett. 19, 411–421.

Krissinel, E., Henrick, K., 2004. Secondary-structure matching, (SSM), a new tool for fast protein structure alignment in three dimensions. Acta. Cryst. D60, 2256–2268.

Li, X.B., Wang, S.Q., Xu, W.R., et al., 2011. Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method. PLoS One 6, e28111.

Ma, Y., Wang, S.Q., Xu, W.R., et al., 2012. Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach. PLoS One 7, e38546.

Mei, S., 2012. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. J. Theor. Biol. 293, 121–130.

Nguyen, M.N., Madhusudhan, M.S., 2011. Biological insights from topology independent comparison of protein 3D structures. Nucleic Acids Res. 39 (14), 1–16.

Shuoyong, S., Zhong, Y., Majumdar, I., Krishna, S.S., Grishin, N.V., 2007. Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. Bioinformatics 23 (11), 1331–1338.

Shuoyong, S., Chitturi, B., Grishin, N.V., 2009. ProSMoS server: A pattern-based search using interaction matrix representation of protein structures. Nucleic Acids Res. 37 (suppl. 2), 526–531.

Wang, J.F., Chou, K.C., 2011. Insights from modeling the 3D structure of New Delhi metallo-beta-lactamase and its binding interactions with antibiotic drugs. PLoS ONE 6, e18414.

Wang, J.F., Chou, K.C., 2012. Insights into the mutation-induced HHH syndrome from modeling human mitochondrial ornithine transporter-1. PLoS One 7, e31048.

Wu, Z.C., Xiao, X., et al., 2012. iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein Pept. Lett. 19, 4–14.

Zhang, S., Yang, L., Wang, T., 2009. Use of information discrepancy measure to compare protein secondary structures. J. Mol. Struct.: THEOCHEM 909 (1), 102–106.

Zotenko, E., Dogan, R.I., Wilbur, W.J., O'Leary, D.P., Przytycka, T.M., 2007. Structural footprinting in protein structure comparison: The impact of structural fragments. BMC Struct. Biol. 7 (1), 53.

Zou, D., He, Z., He, J., Xia, Y., 2011. Supersecondary structure prediction using Chou's pseudo amino acid composition. J. Comput. Chem. 32, 271–278.