Multiple Sequence Alignment Based on Structural Properties

Bugra Ozer, Gizem Gezici, Cem Meydan, Ugur Sezerman
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey
bozer@sabanciuniv.edu,gizemgezici@su.sabanciuniv.edu,cemmeydan@su.sabanciuniv.edu,ugur@sabanciuniv.edu

Abstract— **A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences. Main idea behind multiple sequence alignment is to see the similarities between input sequences, to be able to make phylogenetic analysis and other evolutionary conclusions. We propose a multiple sequence alignment method based on contact maps derived from structural data and network properties. We show that such methods may be useful in creating multiple alignments that can identify domains and similar structures where sequence identity is low.**

Keywords- Multiple Sequence Alignment; MSA; graph properties; cluster.

I. INTRODUCTION

Finding the similarities between proteins has been one of the important issues in biological sciences. Different approaches have been tried to obtain valuable information based on the similarities of proteins. To reach such information, multiple sequence alignment methods have been widely used.

One of the first studies about multiple sequence alignment is based on dynamic programming method of pairwise alignment. First of all, hierarchical clustering of the protein sequences is obtained with the help of the pairwise alignment score matrix. Then, by using the hierarchal clustering and the pairwise scores from the matrix, close groups are aligned until one group which contains all sequences is reached [1]. The first studies were improved by adding further features to the original structure of the approach. CLUSTAL, an improved and widely used algorithm for multiple sequence alignment, uses hierarchal clustering and neighbor-joining method similarly with the original structure of multiple sequence alignment method. The progressive method for MSA (Multiple Sequence Alignment) is CLUSTAL which is the combination of the techniques used by the original method. CLUSTAL has three steps mainly. First step is to use a pairwise alignment method and to obtain a distance matrix. Then, a similarity tree (guide tree) by using a distance matrix is constructed and neighbor-joining method is applied as the second step. Lastly, alignments are combined by taking into account of the order of the guide tree. These are the main steps of CLUSTAL which is used for multiple sequence alignment. Yet, some new features were added to the progressive multiple alignment method, beside the CLUSTAL to make multiple alignments as accurate as possible. These improvements are sequence weighting and modifying gap penalties, resulting in CLUSTALW. The purpose is to reduce the weight of the most related sequences and to increase the weight of the most divergent ones by adding a feature as sequence weighting. Based on this idea, to be able to give such special weights to sequences, two main methods are examined. First step is to use the guide tree with the order. As the second step, value-sharing among the common branches should be done. The second special feature which is a more complicated one is to modify the gap penalties. To do this, two important gap penalties are taken into consideration which are GOP (initial gap penalty) and GEP (Gap extension penalty). Modifying gap penalties are carried out in four main titles. First one is residue specific penalties, differently speaking, aminoacid specific penalties. As an example, glycine and valine at a position will be given different gap penalties. Second one is hydrophilic gap penalties; in this area the aim is to increase the chances of gaps especially for some specific regions such as loop or random coil regions. In these regions, gaps are common. As the third one, gap separation distance approach is applied. In this approach, the goal is to decrease the chances of gaps which are too close to each other. Lastly, end gap separation parameter is taken into account. This is useful to modify gap penalties when there are end gaps which are not biologically meaningful. After the detailed steps of modifying gap penalty feature, its purpose could be summarized as follows, the goal of altering gap penalties is to raise gap penalties in specific regions (typically secondary structure elements) so that gaps are preferentially opened in the less well conserved regions (typically surface loops) [2]. These added features beside the CLUSTALW provide us a better multiple sequence element as an improvement. In addition to the technique of finding similarities between proteins with CLUSTALW (pairwise alignment + distance matrix + neighbor-joining method), as mentioned above multiple structural alignment of proteins has been tried, as well.

Moreover, one of the recent studies suggests an algorithm that could perform multiple structure comparison and motif detection, at the same time. Also, it is claimed that this algorithm considers all structures simultaneously, instead of initiating from pairwise comparisons [3].

978-1-4244-5969-8/10/$26.00 ©2009 IEEE

Beside these methods, a new approach has been examined recently. Using CLUSTALW with a different approach about the distances idea is applied. Instead of using actual distances, utilizing the network properties of the structures of the proteins is the main idea in this method. This method is ultimately based on network properties and dynamic programming with affine gap penalty. This approach stems from the idea that similar protein structures give similar network contacts captured by graph (network) properties. These network properties are connectivity, second connectivity, clustering coefficient, path length, betweenness, closeness centrality, graph centrality and lastly stress centrality. Each network property has a function and owing to these functions, values are calculated instead of actual distances. Then these network properties can be used for global alignment of protein structures and to obtain functionally preserved contact networks. This idea has been used for pairwise alignment of proteins. The main root is here that using structures of proteins by converting them to contact graphs and after obtaining information from the graph using this information to align two structurally similar proteins [4].

Our approach is the same with the above method that network properties are used instead of actual distance values with CLUSTALW method. Yet, we added some other network properties to obtain detailed profile of proteins. Hence, we believe that finding similarity between proteins and aligning them could be more accurate. Furthermore, we do not make pairwise alignment; instead we make multiple alignments of proteins by using the contact maps of proteins.

II. METHODS

A. Contact Map Generation

Using contact maps is a highly preferred method to represent the proteins structures in the literature [5]. In these contact maps used as graphs, nodes correspond to the residues of the proteins and the contacts mean the links of the graph. Proteins are treated as networks of interacting aminoacids. We use the term residue network to differentiate them from protein networks which are used to describe systems of protein interactions [6]. A contact map is a graph created from the residues that are closer to each other than a predefined threshold. The graph is constructed by the coordinates of C_β atoms in residues with 7 Å cutoffs. The contact map A is created as

$$A_{ij} = \begin{cases} H(r_c - r_{ij}), & i \neq j \\ 0, & i = j \end{cases}$$

where r_{ij} is the distance between the i^{th} and j^{th} nodes, r_c is the maximum distance between atoms to be considered in contact (i.e. cut-off distance), and H(x) is the function given by H(x) = 1 for x > 0, and H(x) = 0 for x ≤ 0.

B. Graph Properties

After the construction of contact maps, the values of network properties of the nodes are calculated in the graphs which correspond to the residues of the proteins.

Totally, we have twenty two properties in this study, which are connectivity, second connectivity, clustering coefficient, characteristic weak path, weighted characteristic weak path length, weak closeness centrality, weak betwenness centrality, weak distinct betweenness, centrality, weak special betweenness centrality, weak betwenness graph centrality, weak closeness graph centrality, weak stress centrality, characteristic strong path, weighted characteristic strong path length, strong closeness centrality, weak betwenness centrality, weak distinct betweenness, centrality, strong special betweenness centrality, strong betwenness graph centrality, strong closeness graph centrality, strong stress centrality.

The connectivity calculates the neighbors of a residue in the contact map, whereas second connectivity calculates the number of neighbors of the neighbors of a residue. Another main property is clustering coefficient. It shows how well the neighbors of a node are connected to each other [1].

$$\mathrm{C_n} = \frac{2\mathrm{E_n}}{\mathrm{k(k-1)}}$$

Closeness centrality measures how long the information takes to spread from a given node to other reachable nodes and is given in equation [1].

$$\mathrm{C_C(i)} = \frac{1}{\sum_{t=V} \sigma(i,t)}$$

Another main centrality property is betweenness centrality, which is the quantitative measure of a node or an edge that describes the degree in between other nodes [1].

$$C_B(i) = \sum_{s \neq i \neq t \epsilon V}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

The stress centrality measures the total number of shortest paths that passes over node i [1].

$$\mathrm{C_s(i)} = \sum_{s \neq i \neq t \epsilon V}^{N} \sigma_{st}\,(i)$$

Moreover, we have more network properties as a new improvement in comparison to the first study about using contact maps as graphs to represent the structures of proteins. First of all, graph centrality property is made as more detailed. We have three graph centralities instead of one graph centrality network property. These are closeness graph centrality; betweenness graph centrality and degree graph centrality. In addition to these, we have two different betweenness centrality properties such as distinct betweenness centrality and special betweenness centrality. These two different betwenness centrality properties are added because it

is known that betweenness of a node in the graph is a highly important property. If a node is the most between node in the graph, that node is a significant node, so if a node in the graph of the representation of a protein has high betweenness centrality score then we can say that that residue in the protein is important in that graph. Also, we can compare that residue with another residue in another protein's representation graph.

Furthermore, two versions of each centrality measures were created, one of them is calculated by using weak shortest path algorithm, and the other by strong shortest path algorithm. In strong shortest path algorithm, there is a small difference in the edge relaxation phase of the algorithm. Weak path is the path that minimizes the total cost between two nodes, whereas strong path aims to minimize the maximum distance. As a result, we have two closeness centrality properties as strong closeness centrality and the weak closeness centrality, which are calculated from strong shortest path algorithm and strong shortest path algorithm respectively.

Besides the different implementation of the calculation of shortest paths, we add two more betweenness centrality properties as mentioned before. This is because betweenness is one of the most important properties for a node in the graph so it is important for a residue in the representation graph of a protein. One of these properties is distinct betweenness centrality. This favors the nodes such that occur on the different paths. In other words, different from the conventional betweenness centrality, this calculates the number of the distinct paths which include the node we are studying on. In other words, we are not interested in the total number of paths but which nodes are connected through this node. This favors the nodes such that occur on the different paths. In other words, different from the conventional betweenness centrality, this calculates the number of the distinct paths which include the node we are studying on.

Another betweenness property, named as special betweenness, favors another feature of a node. First of all, it determines two nodes, source and destination nodes. Then, it calculates the all number of shortest paths after that, it calculates the occurrences of the specified node on that paths. The ratio of occurrences of that node on these shortest paths to the all number of shortest paths gives us special betweenness centrality score of that specified node. In this property, the important thing is that how many different nodes that node connects. For instance, if a node k occurs in eight of the shortest paths of the all nine between node i and j; whereas node m occurs three shortest paths of these nine shortest paths between node i and j, then node k is more important than node m according to this special betweennees network property. Because, if this node k is eliminated from the graph then node i and node j will be connected with just one path; therefore this special betweenness property favors nodes like node k.

C. Multiple Alignment with Affine Gap Penalty

At multiple sequence alignment based on structural properties, every sequence is represented as a graph and a corresponding matrix with size [n x n] are created; where n represent the length of the sequences. Optimized gap penalties and similarity scores between amino acids are used to create the alignment matrix. In order to account for affine gap penalties (differentiate between first space of gap and gap extensions) three distinct arrays are used, which can be represented as following;

A[i, j] = maximum score of an alignment between S[1..i] and T[1..j] that ends in S[i] matched with T[j].

B[i, j] = maximum score of an alignment between S[1..i] and T[1..j] that ends in a space matched with T[j].

C[i, j] = maximum score of an alignment between S[1..i] and T[1..j] that ends in S[i] matched with a space.

The entries (i, j) of these arrays depend on previous entries according to the following formulas, valid for $1 \le i \le m$ and $1 \le j \le n$:

$$A[i, j] = p(i, j) + Max \begin{cases} A[i-1,\ j-1] \\ B[i-1,\ j-1] \\ (C[i-1, j\ -\ 1] \end{cases} \quad (1)$$

$$B[i,j] = Max \begin{cases} (h\ +\ g)\ +\ A[i, j-1] \\ g\ +\ B[i, j-1] \\ (h\ +\ g)\ +\ C[i, j\ -\ 1] \end{cases} \quad (2)$$

$$C[i,j] = Max \begin{cases} (h\ +\ g)\ +\ A[i, j-1] \\ g\ +\ C[i, j-1] \\ (h\ +\ g)\ +\ B[i, j\ -\ 1] \end{cases} \quad (3)$$

In order to understand the formulas (1, 2, 3); A[i,j] is represented as p(i,j), which here indicates the score of a matching between S[i] and T[j] ,plus the best score of an alignment between the prefixes S[1..i-1] and T[1.. j-1] [7]. Whenever the algorithm chooses and traverses along the array b, it is known that last column will contain a gap. At this point, it should be considered, whether this gap is the first cell of a gap or an extension is very important as they have different gap penalty scores. The mechanism is so that, if there is a gap prior to our current position, then it will be penalized with extension penalty instead of opening penalty. If the previous index was of matrix A or matrix C, then the new coming cell should include gap opening penalty. A similar argument also explains the formula for C[i, j].

The initialization of the arrays also requires important care. The entries that need initialization are those with indices of the form (i, 0) for $0 \le i \le m$ or (0, j) for $0 \le j \le n$. The Initializations has to be done according to properties of each arrays. For array A;

$$A[0,0] = 0,$$
$$A[i,0] = -\infty, for\ 1 \le i \le m\ and$$

$$A[0,j] = -\infty, for\ 1 \le j \le n.$$

For array B;

$$B[i,0] = -\infty\ , for\ 0 \le i \le m\ and$$
$$B[0,j] = h + g.j\ , for\ 1 \le j \le n.$$

For array C;

$$C[i,0] = h + g.j\ , for\ 0 \le i \le m\ and$$
$$C[0,j] = -\infty, for\ 1 \le j \le n.$$

To get the final result, the optimal alignment is constructed by tracing back from the maximum among the matrices A[m, n], B[m, n] and C[m, n]. From (m, n) to (0, 0), at every step maximum among the possibilities was chosen and the alignment is formed. So at matrix filling process at the beginning the structure is augmented so that not only the current position, additionally which array that position belongs to and from which array we came from is also remembered.

III. RESULTS

A. Dataset

At the experiments, we have selected a representative set of proteins from SH3 and SH2 families. Gene sequence and structural information were obtained from SWISS-MODEL Repository and Protein Data Bank [8, 9, 10]. The selected proteins from SH2 family are:

- Human P56 Tyrosine Kinase [11]
- Phosphatidylinositol 3-Kinase Regulatory Alpha Subunit [12]
- P85 Alpha [13]

Proteins from SH3 family are:

- Rho guanine nucleotide exchange factor 7 [14]
- Proto-oncogene tyrosine-protein kinase Yes [15]
- FYN Tyrosine Kinase SH3 Domain [16]
- Hematopoietic Cell Kinase [17]

B. Results

We have made Multiple Sequence Alignment using our computer program and loaded the resulting alignment to Jalview, a multiple alignment editor [8].

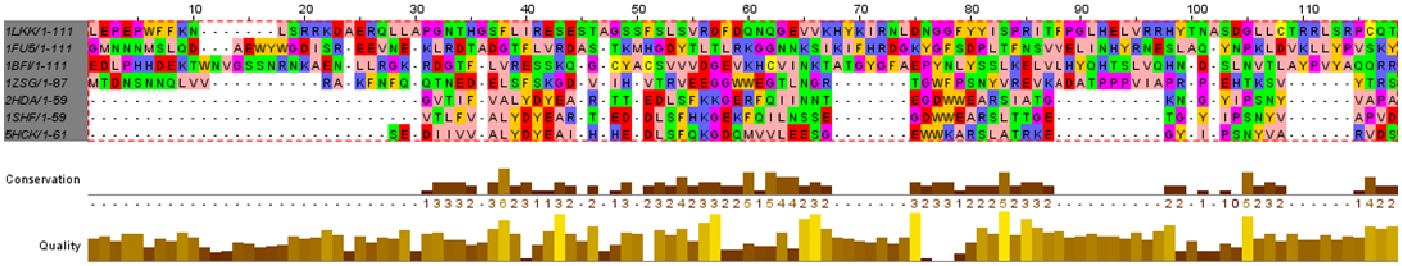

Figure 1. Multiple Sequence Alignment based on sequential information

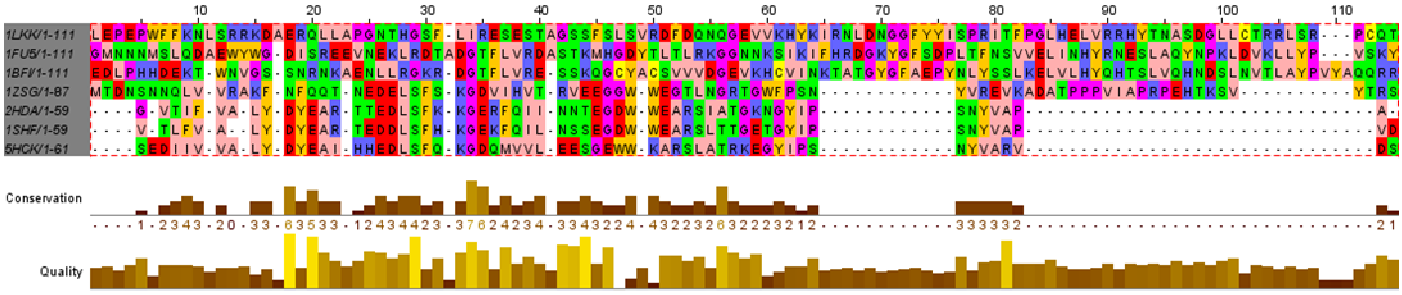

Figure 2. Multiple Sequence Alignment based on structural information

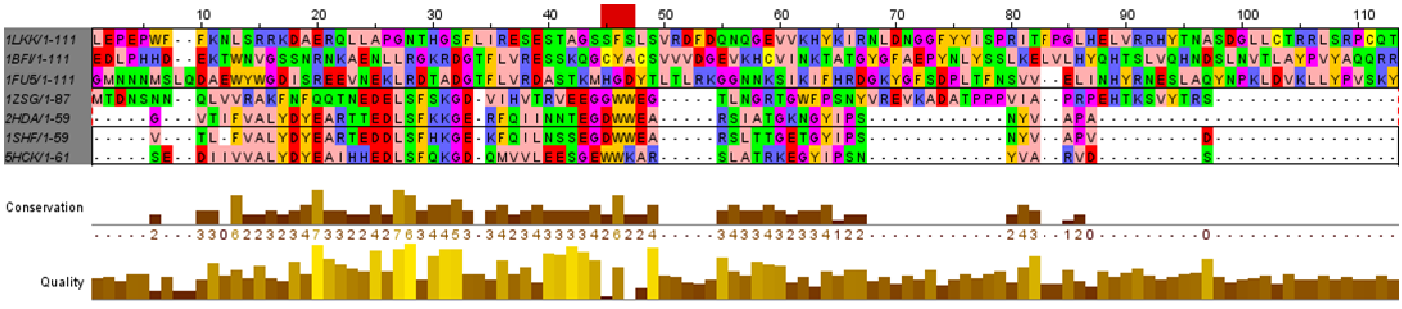

Figure 3. Multiple Sequence Alignment based on both sequential and structural information

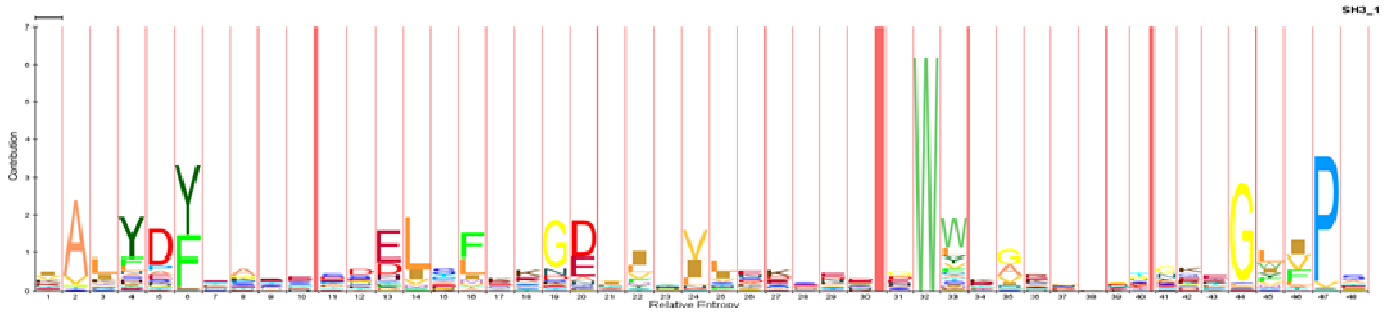

Figure 4. HMM Logo for SH3 Family

In order to emphasize on the advantage of our multiple sequence alignment program, we've tried to obtain 3 different alignments; using sequential information, using structural information and finally combined with both structural and sequential information. After the tests, we obtained the alignments presented in Figures 1, 2 and 3.

When the alignments, which have been made using 7 angstrom graph cutoff and atom type of C_β at results part, and HMM Logo at Figure 4 are taken into consideration, it can be seen that there are significant domains. These domains, such as "WW" at 45th position and "VAL" at 12th position, are conserved between proteins from different species at structural alignment and combined alignment (both structural an sequential information included), giving user information that those proteins are related with each other, whereas it is not the case for sequential alignment. Generally, it can be concluded from the Figures that, proteins from different families have different common domains as proven by their color patterns.

IV. DISCUSSION AND CONCLUSION

In this paper, we proposed a multiple sequence alignment method that uses structural information. By the experiments, we showed that such methods are viable and can give acceptable results even though the sequence and the actual 3D structural information is filtered out and only the graph properties are used. Additionally, as we also know that similar looking sequences should be related with each other, adding structural information made us converge to better alignment, which is implied by the graphs at results part.

An important advantage of our method is the ability to obtain alignments that capture the similarities between the folds and domains of multiple sequences even in the absence of sequence similarity. It is known that structural information is mostly conserved through the evolutionary processes, or reach to the similar structures due to convergent evolution. This results in a variety of protein families that have very low sequence similarity but high structural similarity. The converse is also true; there are proteins with very high sequence identity but are structurally different. As in those cases, multiple alignment based on sequence information can miss potential motifs, or may create false positives. Structural alignment of multiple proteins is also problematic in the case of proteins that are dissimilar in overall shape but carry a common similar domain. By using the structural information to create sequence alignment, these problems can be solved.

While our method can overcome such limitations, it is also affected by the high parameter count that is also problematic in many multiple sequence alignment algorithms. Due to the parameter dependency, it requires optimization for precise alignments.

Multiple Sequence Alignment based on structural properties also facilitates the detection of the presence of structural domains and motifs, which can be later used in function detection and structural studies.

As future work, parameters can be optimized, emphasizing important graph properties by weighting the effect of a property on the final score. Thus, optimization is very important.

Sequence weight bias also should be constructed correctly, in other words similar looking sequences don't affect the alignment of other sequences. At this process, by using guide tree it will be tried to assign weights in inverse ratio with respect to their similarities to each other.

V. ACKNOWLEDGMENTS

B.O, thanks to Ahmet Gunec for his support on preparing the dataset. B.O, thanks to Ahmet Sinan Yavuz for reviewing the manuscript.

VI. REFERENCES

[1] F.Corpet, "Multiple sequence alignment with hierarchical clustering," /Nucleic Acids Research, /vol. 16, July 1988.

[2] J. D. Thompson , D. G. Higgins, T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," /Nucleic Acids Research/, vol. 22, pp.4673-4680, July 1994

[3] N. Leibowitz, Z. Y. Fligelman, R. Nussinov, and H. J. Wolfson, "Automated Multiple Structure Alignment and Detection of a Common Substructural Motif", */PROTEINS: Structure, Function, and Genetic/*, vol. 43, pp. 235-245, 2001

[4] A. Kucukural, O. U. Sezerman, "Structural Alignment of Proteins Using Graph Theoretical Properties," Istanbul, Turkey, 2009

[5] N. Gupta, N. Mangal, S. Biswas, "Evolution and similarity evaluation of protein structures in contact map space," in vol. 59, pp. 196-204, 2005

[6] A. R. Atilgan, P. Akan, C. Baysal; "Small-World Communication of Residues and Significance for Protein Dynamics," //Biophysical Journal/, /vol 86, pp. 85-91, January 2004

[7] Setubal, J. and Meidanis, J. (1997). Introduction to computational molecular biology. PWS Publishing Company.pp. 64-66

Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M., Barton, G.J (2009), "Jalview version 2: A Multiple Sequence Alignment and

Analysis Workbench, "*Bioinformatics* doi: 10.1093/bioinformatics/btp033

[8] Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201

[9] Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.

[10] Schwede T, Kopp J, Guex N, and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31: 3381-3385.

[11] PDB ID: 1LKK

Tong, L., Warren, T.C., King, J., Betageri, R., Rose, J., Jakes, S., Crystal structures of the human p56lck SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 A and 1.8 A resolution., J.Mol.Biol. 256: 601-610, 1996

[12] PDB ID: 1FU5

Weber, T., Schaffhausen, B., Liu, Y., Gunther, U.L., NMR structure of the N-SH2 of the p85 subunit of phosphoinositide 3-kinase complexed to a doubly phosphorylated peptide reveals a second phosphotyrosine binding site. , Biochemistry 39: 15860-15869, 2000

[13] PDB ID: 1BFI

Siegal, G., Davis, B., Kristensen, S.M., Sankar, A., Linacre, J., Stein, R.C., Panayotou, G., Waterfield, M.D., Driscoll, P.C Solution structure of the C-terminal SH2 domain of the p85 alpha regulatory subunit of phosphoinositide 3-kinase., J.Mol.Biol. 276: 461-478, 1998

[14] PDB ID: 1ZSG

Mott, H.R., Nietlispach, D., Evetts, K.A., Owen, D., Structural Analysis of the SH3 Domain of beta-PIX and Its Interaction with alpha-p21 Activated Kinase (PAK), Biochemistry 44: 10977-10983 , 2005

[15] PDB ID: 2HDA

Martin-Garcia, J.M., Luque, I., Mateo, P.L., Ruiz-Sanz, J., Camara-Artigas, A Crystallographic structure of the SH3 domain of the human c-Yes tyrosine kinase: Loop flexibility and amyloid aggregation. , Febs Lett. 581: 1701-1706 , 2007

[16] PDB ID: SHF

Noble, M.E., Musacchio, A., Saraste, M., Courtneidge, S.A., Wierenga, R.K., Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. , EMBO J. 12: 2617-2624, 1993

[17] PDB ID: 5HCK

Horita, D.A., Baldisseri, D.M., Zhang, W., Altieri, A.S., Smithgall, T.E., Gmeiner, W.H., Byrd, R.A. , Solution structure of the human Hck SH3 domain and identification of its ligand binding site. , J.Mol.Biol. 278: 253-265, 1998