International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 11 (2017) pp. 2973-2980

Classification of Structural Protein Domain Based on Hidden Markov Model

Tarek AbuShanab[1] and Rami Al-Hmouz[2]

Department of Electrical and Computer Engineering, King Abdulaziz University, P.O. Box 80204, Jeddah 21589, Saudi Arabia.

[1]ORCID: 0000-0001-6331-4152, [2]ORCID: 0000-0001-8710-070

Abstract

PDZ domains are a standout amongst the most ordinarily discovered protein-protein connection domains in organisms from bacteria to people. These domains are classified into classes I, II and III, contingent upon their binding accomplices and the type of bonds that have been framed. PDZ domains comprise of around 80 to 90 amino acids and distinguished as a locale of taxonomy homology among a varied list of signaling proteins. They have been presented to act as key players that are required in various diseases intervened with the PDZ domain connections; they contain ranging from cystic fibrosis to cancer. In this investigation, we concentrated on the taxonomy of PDZ domains as Class I or II given the primary sequence of the PDZ domains. We utilized the Hidden Markov Model and in view of the domain's essential amino acid successions in helping PDZ domain taxonomy. We assemble our model utilizing a data set which contains 115 interesting human and mice PDZ domains. Three models of emission matrices are investigated utilizing a unigram, bigram, and trigram of amino acid. Our model achieves the high forecast precision of PDZ domain classes with an accuracy of (83.25%).

INTRODUCTION

Protein-protein interactions participate mainly in signal transduction, formation of functional protein complexes and protein modification. One of the most common protein interaction domains in the cell is PDZ domain [1-3]. PDZ is an acronym consolidating the main letters of three proteins post-synaptic density protein (PSD95), Drosophila disc large tumor suppressor (Dlg1), and Zonula occludens-1 protein (zo-1).The interactions of PDZ domain indicated that they involved in numerous diseases that include cystic fibrosis and cancer [4-6]. They have been presented to act as key players required in various diseases mediated with the PDZ domain connections they contain going from cystic fibrosis to cancer. The PDZ domains, among other about 70 particular acknowledgment domains, are essential on the grounds that they are included in the development of multi-cellular organisms by building cell polarity, coordination of the intercellular signaling framework and coordinating the specificity of signaling proteins. The PDZ domain is a typical structural domain of 80-90 amino-acids found in the signaling proteins of microscopic organisms, yeast, plants, viruses, and animals. Moreover, most PDZ domains have a conserved 3D fold made of six β strands and two α helices. PDZ domains can bind to the C-terminal peptides of various proteins and act as a glue, grouping distinctive protein complexes together, focusing on particular proteins and routing these proteins in signaling pathway. These domains are classified into classes I, II and III, contingent upon their binding accomplices and the nature of bonds framed.

A few endeavors [1-4] have been made to arrange PDZ domains as well as their ligands. The most generally utilized taxonomy framework depends on the successions of C-terminal peptide ligands. The last three or four amino acid residues are normally considered and two fundamental classes are characterized. This yields the signatures (S/T)xΦCOOH as class I and ΦxΦCOOH as class II. This PDZ domain characterization (class I and class II) depends on the interaction between ligand position - 2 and the residue situated at the N-terminal end of the αB-helix (position αB:1) of the PDZ domain: class I is controlled by a polar residue (for the most part histidine) while class II contains predominantly hydrophobic amino acid. Generally, PDZ domain taxonomy can be drawn approach from either a structure or taxonomy based forecasts. Despite the fact that, structural based techniques can give an in-depth comprehension of PDZ binding, these strategies can be plagued by low throughput and high cost. Moreover, they frequently give an indication into a solitary or few PDZ domains, giving less concentration to the role of point mutations in PDZ taxonomy, upstream and downstream successions with respect to the PDZ binding motif, and in addition the correct PDZ domain cutoffs that influence ligand affinity.

Many recent approaches [18-23] group PDZ domains from several species into one category, or even worse, group amino acids into pseudo categories, thus the features that are responsible of PDZ domain classification have less concern. These approaches were built on this model as a pilot for their future work. However, they also focused less on multi species domain structure function relationships and machine learning

techniques. At last, their approach in utilizing these already accessible outcomes amplified the model into a position-specific scoring medium in light of the essential succession of both the 82 mouse PDZ domains and 93peptides encoded in the mouse proteome.

The motivation behind our work is for the most part centered around PDZ domain taxonomy gone for understanding the basic elements and taxonomies motifs in charge of correct taxonomy. We proposed a strategy for PDZ domain taxonomy by utilizing just the essential amino acid taxonomy data from 115 extraordinary human and mouse PDZ domains utilizing Hidden Morkov Model HMM. HMM has been used in numerous temporal recognition applications including speech, handwriting, gesture, and bioinformatics. HMM is the process of moving from one state then onto the next, the likelihood of each consequent state depends just on what was the past state:

Set of states: $\{s_1, s_2, \mathrm{K}, s_N\}$

$$P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1}) \quad (1)$$

In the context PDZ classification problem we have two classes that are formed from combinations of amino acids:

A G V I L F P Y M T S H N Q W R K D E C

The problem is now for a given the sequence of amino acids (protein) what is the probability of this sequence to be class I or class II. Our goal is to develop an approach based on computational methods that would allow PDZ domain classification problem to rely on natural amino acid sequences. To be more specific, HMM gives a solution for the PDZ domain classification problem. We will study pseudo-amino acid designation (two classes). We will also extend the study to cover the bigrams and trigrams of the 20 amino acids composing the PDZ domain primary sequences. The key elements of HMM are, the observed sequence (protein), transition matrix of the states and the emission matrix for unigram, bigram and trigram of the protein sequence.

The rest of the paper is structured as follows. Section 2 will cover the studies that are related to PDZ domain. The methodology of HMM will be discussed in section 3. In section 4, we will show the experimental results. The paper will end with conclusions in section 5.

LITERATURE REVIEW

PDZ domain classification problem can be solved from either a structure of the protein or the sequence of amino acids [7-10]. The drawbacks of the structure based methods are low throughput and high cost [11-14]. More recently, several groups have started work on primary sequence based machine learning approaches to classify the PDZ domains [18-23]. However, in many studies, several species were grouped into one category many of or group amino acids into pseudo categories, this will provide less interest on features that are responsible for PDZ domain classifications. Most works in the literatures are focused on predicting various peptide interactions, with different levels of classification rates. MacBeath lab [24-25] has produced the two published works that in this field. lthough the underlying correct forecast rate has been just 48%, and correct negative expectation for those domains that do not connect at 88%, they based on this model as an antecedent for their future work. Be that as it may, they also focused less on multi-species domain structure function associations and highlight extraction strategies. However, feature extraction techniques were not intensively investigated. Eventually, the position of primary sequence of amino acids approach extended the model using scoring matrix based on the primary sequence of both the 82 mouse PDZ domains and 93 peptides encoded in the mouse proteome. Kalyoncu et al [26] used a computational approach using primary sequence data for the classification and prediction of binding interactions of PDZ domains. They focused in feature extraction for the binary sequence and the built a feature vector that is based on bigram and trigram frequency from PDZ domain amino acid primary sequences. Then, they fed their classifier (random forest) with extracted features. They have achieved an impressive classification rate that is above 90% when using trigrams for classification. Similar scores in predicting PDZ domain interactions were achieved. Since the dataset which used in the experiment was imbalanced dataset, they used resampling with replacement. However, the role of amino acid which would be played in the classification was not clear in which that was the major drawback of their study. Specificity map of PDZ domain was presented by Tonikian et al [27], they found that PDZ domain can be categorized into 16 unique classes from a database of only 72 PDZ domain members. Also, the idea of domain mutational effects was introduced in the peptide binding prediction. In Shao et al [28], a classification of PDZ domain from primary sequence was presented using a regression based model. It showed that peptide PDZ binding can be computationally quantified. Instead of generating a simple prediction of binding. The achieved score that is related to area under curve (AUC) was 0.88 using 23 different PDZ domains. Likewise, Wiedemann et al [29] presented a quantitative model for three PDZ domains; they developed a model based on parameters for ligand affinity prediction within a rational design framework of high affinity binders. Similarly, proposed an approach for predicting binding affinity from complexes of PDZ domains and small peptides was proposed by Roberts et al [30]. In [21], Wavelet transform of amino acid properties was proposed and the results were obtained using neural network. The purpose of our work is mainly focused on PDZ domain classification aimed at understanding the critical features for the correct class. In this study, we propose HMM method for PDZ domain classification by using only the primary amino acid sequence information.

METHODOLOGY

Hidden Markov model (HMM) is computational tool that model the sequence of observations into probability distributions. It has been used extensively for many applications like speech applications [31], face identification [32], lip and speech-reading [33], optical character recognition [34] and time DNA modeling [35]. The purpose behind this extensive variety of uses is the rich numerical structure the HMMs are based on, yielding ideal outcomes if utilized appropriately. The Hidden Markov model gets its name from two characterizing properties. In the first place, it expects that the perception at time *t* was created by some procedure whose state S_t is hidden from the observer. Second, it accept that the condition of this shrouded procedure fulfills the Markov property: that is, given the esteem of $S_{(t-1)}$, the present state S_t is free of the considerable number of states preceding $t - 1$. As it were, the detail in some time typifies all we have to think about the historical backdrop of the procedure so as to foresee the eventual fate of the procedure. The yields likewise fulfill a Markov property as for the state: given S_t, Y_t is free of the states and perceptions at all other time records. Taken together, these Markov properties implies that the joint probability of a succession of states and perceptions considered in the accompanying path:

$$P(S_{1:T},Y_{1:T}) = P(S_1)\,P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \quad (1)$$

The notation $X_{1:T}$ means $X_1 \ldots X_T$. The joint probability in (1) can be drawn graphically as shown in Figure1.

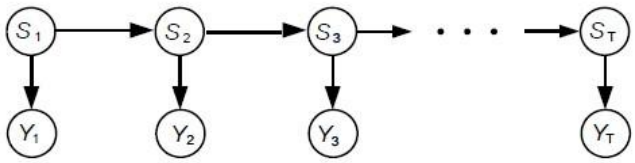

Figure 1. A Bayesian network specifying conditional independent relations for a Hidden Markov model.

Figure (1), known as a Bayesian network, belief network, probabilistic graphical model, or probabilistic independence network, demonstrates the conditions between the factors in the model. Every variable is spoken to by a node in the graph; every node gets immediate circular segments from nodes which it is restrictively on in the factorization of the joint conveyances. A third suspicion of the concealed Markov model is that the shrouded state variable is discrete: S_t can go up against K values which we will mean by the whole numbers $\{1 \ldots K\}$ that represent the available classes in PDZ domain problem.

The probability distribution over sequence of observations is specified by: (1) initial state $P(S_1)$, (2) the $K \times K$ state transition matrix $P(S_t|S_{t-1})$ and (3) the output model (emission matrix) $P(Y_t|S_t)$. HMM assumes that model is time invariant in which the state transition matrices and output models are independent of *t*. If the observations are discrete which can take one of *L* values, the output model can be obtained through a $K \times L$ observation (or emission) matrix. For the observation matrix, represented by $P(Y_t|S_t)$, can be modeled in different ways such as a Gaussian distribution, mixture of Gaussian, or a neural network. HMMs can be augmented to allow for input variables. An HMM, for the most part, comprises of the accompanying five segments where the initial two portray the structure of the model and the last three the parameters:

(a) A set of *S* states, $S_1, S_2, \ldots, S_T$

(b) A distinct observation symbols $D\ \{d_1, d_2, \ldots d_D\}$.

(c) The transition probability matrix; where, for q_t representing the state visited at time t
$P = P_r(q_{t+1} = S_j | q_t = S_i)$

(d) The emission probabilities for each state S_i and d in D.
$r_{id} = P_r(S_i \text{ emits symbol } d)$

(e) An initial distribution vector $\pi = (\pi_i)$, $\pi_i = P_r(q_1 = S_i)$.

RESULTS AND DISCUSSION

As mentioned earlier an HMM consists of hidden states and emissions emitted from those states. The PDZ classification at a particular position in the protein sequence depends on the PDZ classification found at the previous state. In the experiment, we retrieved a total of 115 PDZ domains from human and/or mouse for classification and categorization. The beginning and end of PDZ domain designations were made in accordance with published UniProt annotations, or as they were reported in retrieved datasets [27,35,36]. The dataset consisted of 78 sequences of Class I, and 37 sequences Class II domains. The dataset was based on PDZ domains retrieved from PDZBase [37], and publications by Kalyoncu et al. [36], Tonikian et al.[27] and Stiffler et al. [33]. Also the dataset has been tested by algorithms in [20-21].

When constructing a statistical model such as HMM algorithm, it is important to have enough data. If the dataset used to train the algorithm is too small the problem of interest might not be described by the model since it does not contain all information needed. The test set is supposed to be independent of the model constructed. Thus the model's performance can be tested on the test set. The full database containing the PDZ domains, their reported classification. The data was split into training and testing sets. In the training phase, we used 90% of dataset, while the remaining 10% was considered for testing. The emission matrix was formed from the training set. The classification results were reported by running the previous experiment 100 times for random selections of training and test sets.

A HMM has two important matrices that hold its parameters. The first is the HMM transition matrix, which contains the probabilities of switching from one state to another. The second important matrix is the HMM emission matrix, which holds the probabilities of amino acid in each position of the protein. The classification results were obtained by considering the cases

according the number amino acid used in the emission matrix:

1- Single amino acid (unigram)
2- Two amino acid (bigram).
3- Three amino acid (trigram).

The probability of occurrences of each amino acid (case1) in both classes is represented by the emission matrix (see figures 2, 3). Here we have 20 elements for each class in the emission matrix. While Figures 3 and 4 shows the probabilities for case 2. There would be 400 combinations of amino acid that will be represented in the emission matrix, the brighter regions (refer to figure 4 and 5) are the more likely to appear in the sequences. In case 3, the emission matrix will composed from 8000 elements for each class.

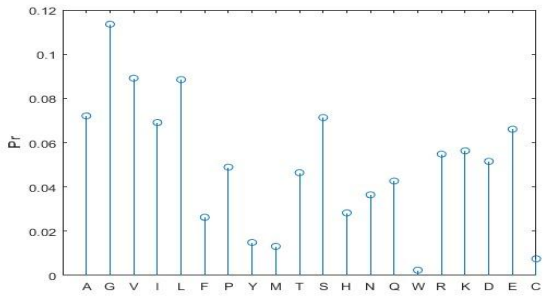

Figure 2. Amino acid probability in class I (case 1)

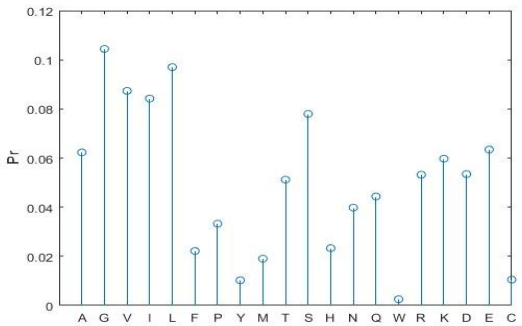

Figure 3. Amino acid probability in class II (case 1)

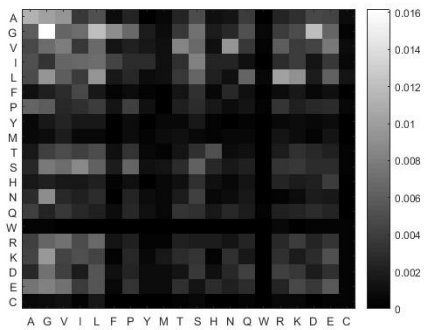

Figure 4. Bigram occurrences in class I (case 2)

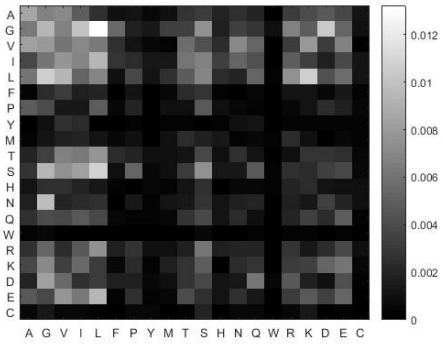

Figure 5. Bigram occurrences in class II (case 2)

In case 1, the probability of occurrences of amino acid in the protein sequence are almost the same for both classes. The most frequent amino acid for class 1 and 2 is G while the lease frequent is W for both classes. The statics for case 2 and case 3 are shown in Table 1 and 2 respectively.

Table 1: Statistic of case 2

Occurrences	**Class 1**	**Class 2**
Most frequent	GG	GL
Never appeared	WV,WI,WL NF, WF,CF WP,CP,YY HY,WY,CY WM,WT,WS WN,CN,WQ IW,FW,YW MW,SW,HW NW,WW,RW KW,DW,CW WK,MC,HC	YA, WV,WI YL,WL,FF YF,NF,WF CF,HP,WP FY,PY,YY MY,WY,WM CM,WT,YS HH,WH,KH CH,FN,WN AW,VW,IW FW,PW,YW TW,NW,QW WW,RW,KW EW,CW,MD WD,CD,VC WC,EC,CC

Occurred only in one class	CP,HY,CY WS,CN,WQ MW,SW,HW DW,WK,MC HC	YA,YL,FF YF,HP,FY PY,MY,CM YS,HH,WH KH,CH,FN AW,VW,PW TW,QW,EW MD,WD,CD VC,WC,EC CC

Table 2: Statistics of case 3

Occurrences	**Class 1**	**Class 2**
Most frequent	GGG	GGL
Occurred only in one class	CPP, HYY, CYY,WSS, CNN,WQQ MWW,SWW, HWW,DWW, WKK,MCC HCC	YAA,YLL FFF,YFF HPP,FYY PYY,MYY CMM,YSS HHH,WHH KHH,CHH FNN,AWW VWW,PWW QWW,EWW MDD,WDD CDD,VCC WCC,ECC CCC

We reported the classification rate for four different transition matrices: the transition between class 1 and class 2 are assumed to be a) 0.5, b) 0.2, c) 0.1 and d) 0.05.

In Table 3 we show the steps of HMM to calculate the probability of each stat for two Amin acids. The state probability was performed for case 1 and for a given transition matrix

		class I	class II
Transmission matrix (A) =	class I	0.8	0.2
	class II	0.2	0.8

Assuming that the initial probabilities of class I and class II are 0.5. This process will be repeated to all amino acid presented in the protein sequence. The sequence of (AG) is classified as class 1 since its probability (0.00564) is higher than class 2 (0.0028).

Table 3: HMM steps to calculate the state probability

X	**A**	**G**
***Sclass*1**	P(A |class1) = 0.075 P(class1) = 0.5 $Sclass1,1$= 0.075*0.5=**0.0375**	P(G|class1)=0.118 P(class1|class1)=0.8 0.0375*0.188*0.8=0.**00564** P(class1|class2)=0.2 0.031*0.118*0.2=0.**0007316** ***Sclass*1,2 =0.00564**
***Sclass*2**	P(A |class2) = 0.062 P(class2) = 0.5 $Sclass2,1$ =0.062*0.5=**0.031**	P(G|class2)=0.115 P(class2|class1)=0.2 0.2*0.0375*0.115=0.**0008625** P(class2|class2)=0.8 0.031*0.8*0.115=**0.002852** ***Sclass*2,2 =0.002852**

The averaged classification error for 100 iterations is shown in Table 4. The minimum classification error (0.35) was achieved in case 2 for transition =0.95. This error is quite considerable because the probabilities of occurrences of amino acid are almost the same for both classes.

Table 4: error rate using HMM

	error rate			
	a	b	C	D
case 1	0.52	0.49	0.43	0.44
case 2	0.40	0.41	0.40	0.35
case 3	0.45	0.41	0.42	0.36

The decision of classification problem was made at the last state T. We tested the decision at each state (1 to T) and reported the final decision based on voting process. The decision will be class 1 if the decision of class 1 in all states is more than class 2 and viscera for class 2. The classification error was improved in all cases (see table 5).

Table 5: error rate (voting process)

	error rate			
	a	b	c	D
case 1	0.4842	0.3367	0.3317	0.2983
case 2	0.2517	0.1625	0.1650	0.1683
case 3	0.2683	0.1617	0.1642	0.1692

We also reported the classification error in which the decision is made base on the multiplication of all state probabilities assuming that no transition among classes. Here, the decision of class 1 is made if the multiplication of state probabilities of class 1 is greater than class 2 otherwise the decision will be class 2. This process mimic Naive Bayes classifier and then combine (fuse) all state probabilities. Some improvement was achieved over the previous results as shown in Table 6.

Table 6: error rate (fusion of state probabilities)

	error rate
case 1	0.2850
case 2	0.1658
case 3	0.1675

Table7 show a comparison for reported classification errors for different approaches that have been testes on the same dataset used in this study. The proposed method showed less error rate over other methods.

Table 7: error rate (fusion of state probabilities)

method	Classification error (%)
[21]	18.59
[37]	28.49
[20]	25.90
Our method	16.75

CONCLUSION

This study aims to classify PDZ domain as Class I or II based on a given sequence of amino acid. Hidden markov model was constructed by using interaction dataset (consists of 115 unique human and mouse PDZ domain). We predicted the classes of PDZ domain with accuracies of (83.25%) when no transition among states. With this highly enquiring result, this study could be an important step in the automated prediction of PDZ domain classes.

REFERENCES

[1] Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci 114: 3219–3231. pmid:11591811.

[2] Dev KK (2007) PDZ domain protein-protein interactions: A case study with PICK1. Current Topics in Medicinal Chemistry 2007, 7(1):3-20.

[3] Nourry C, Grant SG, Borg JP (2003) PDZ domain proteins: plug and play! Sci STKE 2003, 2003(179):RE7.

[4] K. K. Dev (2004) "Making protein interactions druggable: targeting PDZ domains," *Nature Reviews Drug Discovery*, vol. 3, pp. 1047-1056.

[5] J. Doorbar, (2006) "Molecular biology of human papillomavirus infection and cervical cancer," *Clinical science*, vol. 110, pp. 525-541, 2006

[6] B. D. Moyer, J. Denton, K. H. Karlson, D. Reynolds, S. Wang, J. E. Mickle, (1999)., "A PDZ-interacting domain in CFTR is an apical membrane polarization signal," *Journal of Clinical Investigation*, vol. 104, p. 1353.

[7] Gujral TS, Karp ES, Chan M, Chang BH, MacBeath G (2013) Family-wide investigation of PDZ domain-mediated protein-protein interactions implicates beta-catenin in maintaining the integrity of tight junctions. Chem Biol 20: 816-827.

[8] Ernst A, Appleton BA, Ivarsson Y, Zhang Y, Gfeller D, et al. (2014) A structural portrait of the PDZ domain family. J Mol Biol 426: 3509-3519.

[9] Ernst A, Sazinsky SL, Hui S, Currell B, Dharsee M, et al. (2009) rapid evolution of functional complexity in a

domain family. Sci Signal 2: ra50

[10] Ye F, Zhang M (2013) Structures and target recognition modes of PDZ domains: recurring themes and emerging pictures. Biochem J 455: 1-14.

[11] Jin R, Ma Y, Qin L, Ni Z (2013) Structure-based prediction of domain-peptide binding affinity by dissecting residue interaction profile at complex interface: a case study on CAL PDZ domain. Protein Pept Lett 20: 1018-1028.

[12] Hui S, Xing X, Bader GD (2013) Predicting PDZ domain mediated protein interactions from structure. BMC Bioinformatics 14: 27.

[13] Ernst A, Gfeller D, Kan Z, Seshagiri S, Kim PM, et al. (2010) Coevolution of PDZ domain-ligand interactions analyzed by highthroughput phage display and deep sequencing. Mol Biosyst 6: 1782 1790.

[14] Mu Y, Cai P, Hu S, Ma S, Gao Y (2014) Characterization of diverse internal binding specificities of PDZ domains by yeast two-hybrid screening of a special peptide library. PLoS One 9: e88286.

[15] McLaughlin RN, Jr., Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. Nature 491: 138-142.

[16] Gianni S, Haq SR, Montemiglio LC, Jurgens MC, Engstrom A, et al. (2011) Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. J Biol Chem 286: 27167-27175.

[17] Luck K, Charbonnier S, Trave G (2012) The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett 586: 2648-2661.

[18] Hui S, Xing X, Bader GD (2013) Predicting PDZ domain mediated protein interactions from structure. BMC Bioinformatics 14: 27.

[19] Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H: (2205) PDZBase: a protein-protein interaction database for PDZdomains. Bioinformatics, 21(6):827-828

[20] Nakariyakul S, Liu ZP, Chen L (2014) A sequence-based computational approach to predicting PDZ domain-peptide interactions. Biochim Biophys Acta 1844: 165–170. doi: 10.1016/j.bbapap.2013.04.008. pmid:23608946

[21] Khaled Daqrouq, Rami Alhmouz, Ahmed Balamesh, Adnan Memic (2015) Application of Wavelet Transform for PDZ Domain Classification, Plos one,10(4), 2015.

[22] Hui S, Bader GD (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. BMC Bioinformatics 11: 507.

[23] Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26: 1041-1045.

[24] Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences.

[25] Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome.

[26] Kalyoncu S, Keskin O, Gursoy A (2010) Interaction prediction and classification of PDZ domains. BMC Bioinformatics 11: 357.

[27] Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6: e239. doi: 10.1371/journal.pbio.0060239. pmid:18828675.

[28] Shao X, Tan CS, Voss C, Li SS, Deng N, et al. (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. Bioinformatics 27: 383-390.

[29] Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides.

[30] Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR (2012) Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity.

[31] Mark Gales and Steve Young (2008), "The Application of Hidden Markov Models in Speech Recognition", Foundations and Trends® in Signal Processing: Vol. 1: No. 3, pp 195-304

[32] Ara Nefian ,A Hidden Markov Model-Based Approach For Face Detection And Recognition, PhD thesis, Georgia Institute of Technology. 1999.

[33] P. Sujatha and M. R. Krishnan, (201 2) "Lip feature extraction for visual speech recognition using Hidden Markov Model," International Conference on Computing, Communication and Applications, Dindigul, Tamilnadu, , pp. 1-5.

[34] Karishma Tyagi, Vedant Rastogi (2014) ,Implementation of Character Recognition using Hidden Markov Model, International Journal of Engineering Research & Technology, Vol. 3 - Issue 2.

[35] John Henderson,Steven Salazberg,Kenneth H. Fasman (2009),Finding Genes in DNA with a Hidden Markov Model, Journal of Computational Biology. , 4(2): 127-141.

[36] Kalyoncu S, Keskin O, Gursoy A (2010) Interaction prediction and classification of PDZ domains. BMC Bioinformatics 11: 357.

[37] Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H (2005) PDZBase: a protein–protein

interaction database for PDZdomains. Bioinformatics 21: 827–828. pmid:1551399

[38] Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, Allen JE, Zaslavskaia LA, MacBeath G (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364-369.