Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 419485, 9 pages
http://dx.doi.org/10.1155/2013/419485

Research Article

On Classification of PDZ Domains: A Computational Study

Wasim Aftab,[1] Adnan Memic,[2] and Dumitru Baleanu[3,4,5]

[1] *Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia*
[2] *Centre of Nanotechnology, King Abdulaziz University, Jeddah, Saudi Arabia*
[3] *Department of Mathematics and Computer Sciences, Faculty of Arts and Sciences, Cankaya University, 06530 Ankara, Turkey*
[4] *Chemical and Materials Engineering Department, Faculty of Engineering, King Abdulaziz University, P.O. Box 80204, Jeddah 21589, Saudi Arabia*
[5] *Institute of Space Sciences, Magurele-Bucharest, 077125, Romania*

Correspondence should be addressed to Adnan Memic; amemic@kau.edu.sa and Dumitru Baleanu; dumitru@cankaya.edu.tr

Received 3 July 2013; Revised 31 July 2013; Accepted 31 July 2013

Academic Editor: J. A. Tenreiro Machado

Our goal in this present study is to introduce new wavelet based methods for differentiating and classifying Class I and Class II PDZ domains and compare the resulting signals. PDZ domains represent one of the most common protein homology regions playing key roles in several diseases. To perform the classification, we developed two methods. The first of our methods was comparable to the standard wavelet approaches while the second one surpasses it in recognition accuracy. Our models exhibited interesting results, and we anticipate that it can be used as a computational technique to screen out the misfit candidates and to reduce the search space, while achieving high classification and accuracy.

1. Introduction

PDZs are structural domains contained in many proteins. They have been shown to act as key players involved in numerous diseases mediated with the PDZ domain interactions they contain [1–3]. PDZs represent one of the most common protein domains found in the human genome that are made up of approximately 90 amino acids. They were initially identified based on the homology of three different proteins from which their name is derived. One of the main functions that PDZ domains have been recognized for is their scaffolding or mediator role [4–6] in the assembly of receptors at the cellular membrane interface. These reactions are governed by PDZ domain binding of C-termini of their ligand protein, more specifically the last four to six amino acid residues [7–9]. Overall, there are several PDZ domain classifications; however, the two most dominant recognition motifs are X-S/T-X-Φ for Class I PDZ domains and X-Φ-X-Φ for Class II PDZ domains [10]. Most PDZ domains share a similar 3D globular module [11] as shown by NMR and crystallography structures. In general, PDZs are composed of six beta strands and two alpha helices; however, predicting peptide-PDZ domain interaction can be difficult. Our goal was to generate a computational model that could aid in interaction classifications. Given that PDZ domains often exhibit strong binding towards certain peptide ligands, they can be classified based on the recognition sequence of interacting peptides. However, with more recent reports, it is clear that PDZ domains display significant promiscuity in binding [12], proving to be a challenge for their classification. By extrapolating the relevant sequence information and using several wavelet transformation, we intended to develop a method that would be time saving and offer an alternative to expensive experiments. Work on this topic by Kalyoncu et al. [13] was based on using a statistical learning model for automated prediction of PDZ domains. Chen et al. [14] used a statistical model to perform prediction studies based on domain-peptide interactions. Another interesting classification technique according to the two critical positions of PDZ domains has been introduced by Bezprozvanny and Maximov [15]. Chen et al. [14] used a multidomain selectivity model to predict PDZ domain-peptide interactions across

Table 1: Comparisons between our methods and MODWT.

Methods	Sensitivity (%)			Specificity (%)			Positive predictive value (%)			Negative predictive value (%)		
	WAD-1	WAD-2	MODWT	WAD-1	WAD-2	MODWT	WAD-1	WAD-2	MODWT	WAD-1	WAD-2	MODWT
Parametric classifier	81.82	91.67	81.82	66.67	100	55.56	75	100	69.23	75	90	71.43
K-nearest neighbors	72.73	75	63.64	33.33	77.78	88.89	57.14	81.82	87.5	50	70	66.67

Sensitivity gives true positive rate, or the recall rate of prediction algorithm. See (10).
Specificity gives true negative rate of prediction algorithm. See (11).
Positive predictive value confirms that a correct prediction is actually correct. See (12).
Negative predictive value confirms that a false prediction is actually false. See (13).
Parametric classifier and K-nearest neighbors are popular pattern classification algorithms. See [27].

the mouse proteome. However, in this study, we have converted the biological prediction problem into an engineering challenge guided by mapping protein domain sequences into signals based on the 7 key physiochemical properties of amino acids (hydrophobicity [16], electronic [17], isoelectric point [18], polarity [19], volume [19], composition [20], and molecular weight [20]).

In this paper, the problem statement is as follows: given any PDZ sequence S = GTRITLEEITLERA, predict to which class of PDZ (I or II) S belongs.

The core of our study is the feature extraction from the amino acid sequences. In order to extract features from our dataset, we have used wavelets-based signal processing methodologies because it retains more abundant information of sequence order in frequency domain and time domain [21], which is a key to our method. In this study, we were able to develop two novel methods for feature extraction and classification. Feature extraction in the first method named WAD-1 is achieved by first smoothing the signal incorporating empirical mode decomposition (EMD) [1] and second treating normalized signals with maximum overlap discrete wavelet transform (MODWT) [22].

Our second method named WAD-2 uses trigram frequencies of amino acids followed by MODWT for feature extraction. The idea of amino acid frequency has been used popularly in many past [23, 24] and recent [13, 25, 26] proteomics analysis with various mathematical and statistical models.

Finally, we used different classifiers, namely, parametric classifier [27] and K-nearest neighbors [27], to classify these features. The performance of current methods showed improvement compared with wavlets only method in Table 1.

2. Methods

2.1. Empirical Mode Decomposition (EMD). According to Huang et al. [28], the empirical mode decomposition (EMD) [28] is a technique used to decompose a given signal into a set of elementary signals called "intrinsic mode functions" (IMFs). The algorithm operates by removing IMFs in every iteration. Rato et al. [29] has proposed an improved EMD algorithm, which is as follows from their paper.

Presenting any signal, $\chi(t)$, the IMFs are generated by an iterative activity titled sifting operations, which consist of the following steps.

(i) Generate each and every localized maxima, M_x, $x = 1, 2 \ldots$, and so forth, and minima, m_y, $y = 1, 2 \ldots$ and so on, in $\chi(t)$.

(ii) Generate the interpolating signals for maxima as $M(t) = f_M(M_x, t)$ and for minima as $m(t) = f_m(m_y, t)$. The interpolating signals thus computed contributes to the upper and lower envelopes of the signal.

(iii) Consider $e(t) = (M(t) + m(t))/2$.

(iv) Remove $e(t)$ from the original signal $\chi(t)$, and update it as $\chi(t) = \chi(t) - e(t)$.

(v) Go back to step (i) and continue until $\chi(t)$ produced in step (iv) remains almost unvaried.

(vi) After obtaining an IMF $\varphi(t)$, subtract it from updated $\chi(t)$ as $\chi(t) = \chi(t) - \varphi(t)$, and jump to step (i) if multiple extremum (a maxima or minima) for $\chi(t)$ is noticed.

This IMF can be mathematically defined as [29]

$$\begin{aligned} \varphi(t) &= \mathrm{Re}\left\{|y(t)|\, e^{\mathrm{jarg}(y(t))}\right\}, \\ \varphi(t) &= |y(t)| \cos[\theta(t)], \\ \varphi(t) &= |y(t)| \cos[\arg(y(t))]. \end{aligned} \tag{1}$$

We also compute a ratio,

$$R = \frac{E(\chi(t))}{E(e(t))}. \tag{2}$$

Here, $E(\chi(t))$ and $E(e(t))$, respectively, indicate the energy of the original signal before sifting and the average energy of the upper and lower envelopes. However, if R crosses-over an allowed threshold of resolution τ, then the IMF computation is stopped.

2.2. Maximum Overlap Discrete Wavelet Transform (MODWT). MODWT was described by Percival and Walden [22]; it is a special case among the currently available wavelet-based techniques for the analysis of arbitrary-length discrete time series. MODWT is different from DWT in a sense that it is a circular shift-invariant transform [6]; the idea here is that, if a circular shift operation is applied to the real time

series data, then it generates identically shifted MODWT coefficients. The MODWT is well suited for any sequence length N, whereas for a complete decomposition of J levels the DWT requires N to be a multiple of 2^J.

Moreover, we can interpret the MODWT as a cyclic version of the DWT, and it is achieved by averaging over all nonredundant DWTs of shifted versions of the original series. However as this operation smoothens the original DWT signals, it also increases the running time of the computer program. If we apply DWT on an n-point time series, it requires $O(n)$ multiplications whereas MODWT for the same series requires $O(n \log 2n)$ multiplications [30].

It is well suited for our study because, once the protein sequence is translated to a numerical sequence, it becomes a time-series sequence $\{X_t \mid t = 0, 1, 2, \ldots, N-1\}$. In order to filter the signals at each level of $\{X_t\}$, MODWT treats the time series as a periodical. The MODWT coefficients are given by [31]

$$\widetilde{W}_{j,t} = \sum_{l=0}^{Lj-1} \widetilde{h}_{j,l} X_{t-l \bmod N}, \tag{3a}$$

$$\widetilde{V}_{j,t} = \sum_{l=0}^{Lj-1} \widetilde{g}_{j,l} X_{t-l \bmod N}. \tag{3b}$$

Here, $\widetilde{h}_{j,l}$, $\widetilde{g}_{j,l}$ are the high and low pass filters, respectively, of level j. It is also evident from the above equations that the unseen values $X_{-1}, X_{-2}, X_{-3}, \ldots, X_{-N+1}, X_{-N}$ are the same as the observed values $X_{N-1}, X_{N-2}, X_{N-3}, \ldots, X_1, X_0$, which indicates that MODWT induces "cyclic boundary conditions" during wavelet transformation.

The cyclic boundary condition can be difficult to implement in case of nonperiodic signals that exhibit discontinuities between start and end times [31]; therefore, common adoption [31] is "reflection boundary conditions", in which the time series is extended to $2N$ instead of N. Due to reflection symmetry, the unseen values $X_{-1}, X_{-2}, X_{-3}, \ldots, X_{-N+1}, X_{-N}$ are assigned to the seen values $X_0, X_1, X_2, \ldots, X_{N-2}, X_{N-1}$. Therefore, (3a) and (3b) can be rewritten as in [31]:

$$\widetilde{W}_{j,t} = \sum_{l=0}^{Lj-1} \widetilde{h}_{j,l} \dot{X}_{t-l \bmod 2N}, \tag{4a}$$

$$\widetilde{V}_{j,t} = \sum_{l=0}^{Lj-1} \widetilde{h}_{j,l} \dot{X}_{t-l \bmod 2N}. \tag{4b}$$

Here, $\{\dot{X}_t\}$ is the extension of $\{X_t\}$.

3. Prediction with WAD-1

The improved EMD normalizes the signal to a unit power [29], which was not the case with original EMD [28]. Hence, improved EMD is more suitable for low amplitude biomedical signals [29] since it keeps all the components of a signal and does not decompose into different levels of resolution. The improved EMD algorithm also does not reduce the feature space. Feature space reduction is an important part of our job because high dimensional feature space increases the running time of prediction algorithms.

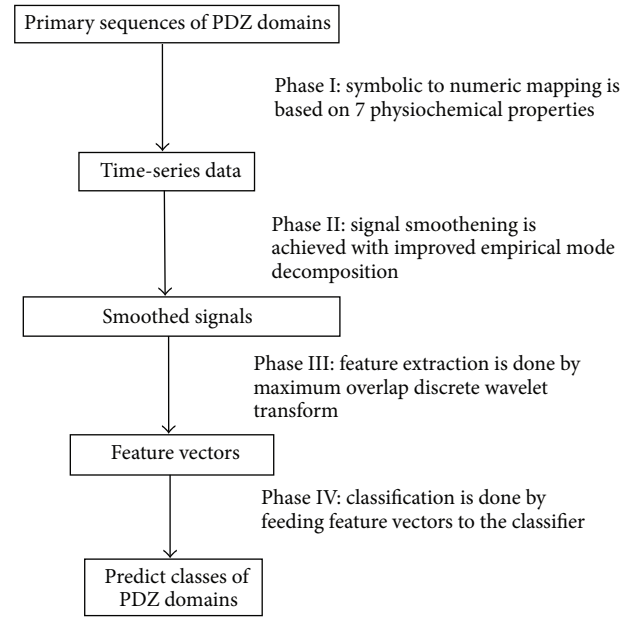


Figure 1: Block diagram depicting flowchart of method WAD-1.

Therefore, we used a combination of MODWT and improved EMD for our feature extraction part because improved EMD preprocesses a signal by interpolating all its maxima and minima, which in turn enhances the signal fit. The signal thus generated is fed to MODWT for feature extraction because it preserves more abundant information of sequence order in both frequency as well as time domains, and at the same time it reduces the feature space, which in turn improves the run time complexity of the classifiers. In essence, with our combination, we achieved reducing the feature space without losing sequence order information.

Our dataset is primarily based on the dataset of Tonikian et al. [32]. We have extracted only human PDZ domains for our work from their dataset.

Our first algorithm WAD-1 operates in four phases as shown in Figure 1. It can be seen from Figure 1 that, in the first phase, mapping from symbolic domain to numeric equivalent is performed. For WAD-1 method this mapping is based on the 7 physiochemical properties of amino acids; that is, every amino acid in each sequence is converted to a real number. With this operation, we obtain time-series data. We smoothen this signal by applying EMD. Let $\widehat{S} = \{\widehat{s}_0, \widehat{s}_1, \ldots, \widehat{s}_{N-1}\}$ indicate a smoothened signal, where $\widehat{s}_i$ indicates the smoothened signal value corresponding to the ith position in the sequence.

We then applied MODWT operation of level J on $\widehat{S}$ as follows from [33]:

$$W = \widetilde{W}\widehat{S}. \tag{5}$$

Here, W is a nonorthogonal real matrix of dimension $(J+1)N \times N$. The MODWT coefficient vector from (1) can be resolved into $(J+1)$ vector as in [33]:

$$\widetilde{W} = \left[\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_J, \widetilde{V}_J\right]. \quad (6)$$

Here, $\widetilde{W}_j$ $(j = 1, 2, \ldots, J)$ implies the length of $N/2^j$ vector of wavelet coefficients associated with the change on scale of length $L = 2^{j-1}$, and $\widetilde{V}_J$ implies the length of $N/2^J$ vector of scaling coefficients associated with averages on scale of length 2^j.

From [22, 34], it is evident that MODWT transformation preserves energy of the signal. The following equation proposed from Gupta et al. [33] is shown here:

$$\begin{aligned} \left\|\widehat{S}\right\|^2 &= \|W\|^2 \\ \left\|\widehat{S}\right\|^2 &= \sum_{j=1}^{J}\left\|\widetilde{W}_j\right\|^2 + \left\|\widetilde{V}_J\right\|^2. \end{aligned} \quad (7)$$

The variance of $\widehat{S}$ can be written similarly as in Gupta et al. [33]:

$$\sigma^2 = \frac{1}{N}\left\|\widehat{S}\right\|^2 - \overline{\widehat{S}}^2. \quad (8)$$

The feature vector generated through WAD-1 is influenced by the work of Gupta et al. with G protein coupled receptors (GPCRs) [33], and it is made up of the variances of several physiochemical properties of the amino acids in the PDZ sequence. For example, using (8), WAD-1 computes the part of the feature vector for a PDZ sequence based on EIIP values as

$$F_{\text{EIIP}} = \left[\sigma^2_{\text{EIIP}}(1), \sigma^2_{\text{EIIP}}(2), \ldots, \sigma^2_{\text{EIIP}}(3)\right]. \quad (9)$$

Here, J indicates upper limit of the level of decomposition. The complete feature vector is obtained by concatenating 7 such feature vectors computed for 7 physiochemical properties of amino acids.

In this study, we chose the least asymmetric filter (LA8) for the wavelet transformation phase because it has a filter width short enough that any impact caused by boundary conditions stays within the tolerance limit.

We have used ANOVA in the feature extraction phase as this analysis is useful in comparing multiple variables for statistical significance. Gupta et al. [33] have used this method to extract features from amino acid sequences in the classification study of GPCRs.

This feature extraction is the most important phase in the overall classification and analysis since it yields the feature vector (FV), whose dimension is function of J, the number of levels chosen for decomposing the signal. For $J = 5$, (as in our case) the dimension of the feature vector is only $7 * 5 = 35$. This is a great reduction that improves the run time of the classification program. In the final phase, these FVs are fed to the classifier for prediction of PDZ domain.

For MODWT calculations in the 3rd phase, we have used the WMTSA wavelet toolkit [31] from MATLAB. With WAD-1 we are able to reduce the dimensionality of the feature space and at the same time retain the positional information of the amino acids in the original sequence. Table 1 clearly demonstrates the consistency of new WAD-1 when compared with popular MODWT.

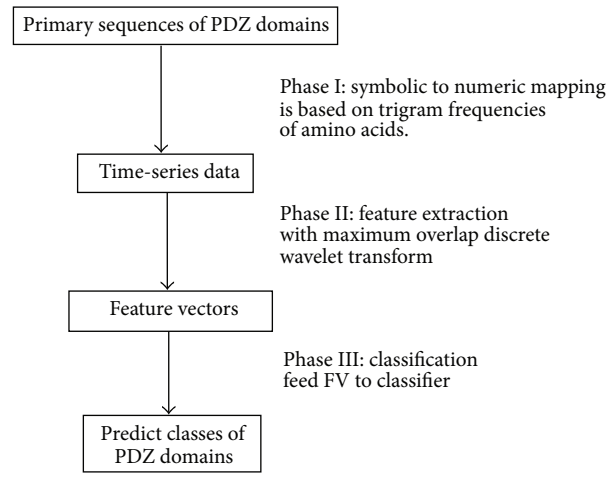


Figure 2: Block diagram depicting flowchart of method WAD-2.

4. Improved Prediction with WAD-2

In this method, we extract the trigram frequencies from the amino acids as previously reported [13]. For an amino acid sequence such as EITLERG, it has the following trigrams: EIT, ITL, TLE, LER, and ERG.

We calculate trigram frequency of amino acids of every PDZ sequence in the following way.

(i) First, count the number of times a trigram appears in the sequence.

(ii) Then, divide this number by the total number of trigrams in the sequence; in fact, it is found to be $(L-2)$, where L denotes length of a PDZ sequence.

(iii) Repeat steps (i) and (ii) for every possible trigram for the PDZ sequence.

We reduce the dimensions of the features by applying the idea introduced by Kalyoncu et al. [13], where 20 amino acids are grouped into 7 distinct classes based on dipoles and volumes.

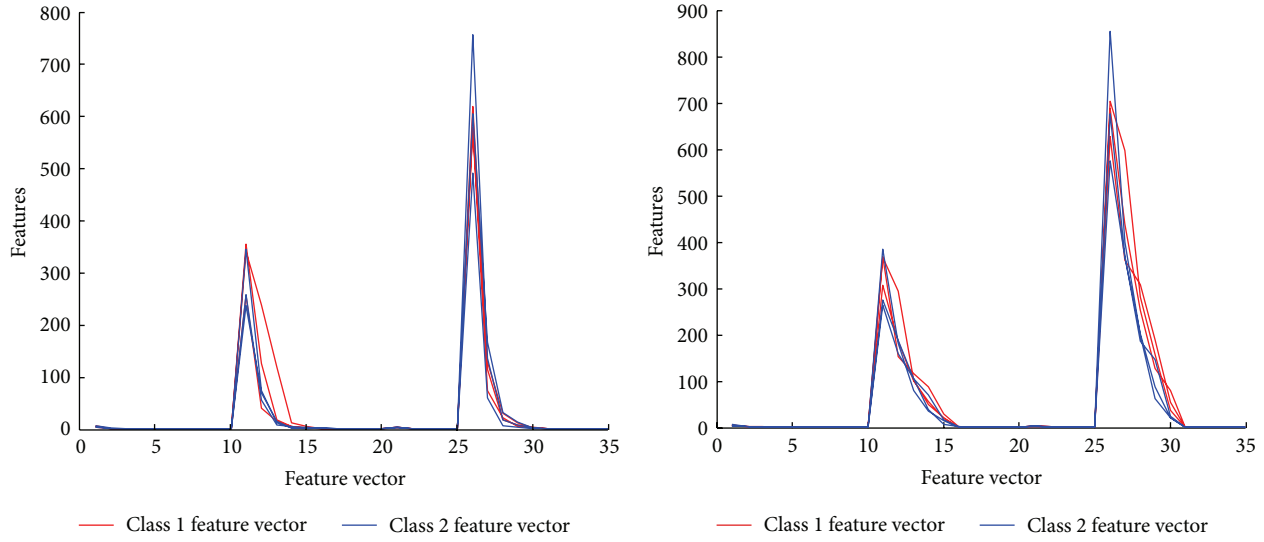
The block diagram of this method is shown in Figure 2, and we call it WAD-2. It operates in 3 phases.

Once the symbolic to numeric mapping is done, we applied MODWT as in WAD-1 for feature extraction.

We have used "PCP: a program for supervised classification of gene expression profiles" [27] in our classification phases for both WAD-1 and WAD-2.

5. Results and Discussions

In this paper, we show proof-of-principle application for our algorithms. We evaluated our approaches by 2 different classifiers and also compared them to a wavelet method,



(a) Feature vectors of both classes of PDZ computed using WAD-1 (EMD smoothening followed by MODWT)

(b) Feature vectors of both classes of PDZ computed using MODWT alone

Figure 3: Comparison of Signals generated after Feature Extraction with WAD-1 and MODWT for classes I and II of PDZ domains. WAD-1 clearly smoothens the signal more when compared to MODWT.

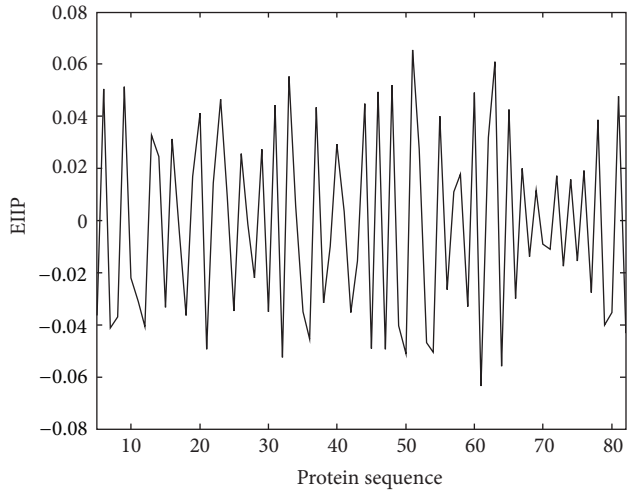


Figure 4: Mapped protein sequence (organism: human, PDZ name: DLG1, and length: 84 amino acids) using EIIP scale. This is the first step for WAD-1; it involves the transformation of amino acid sequences into a numerical sequence.

MODWT. We observed that our results are comparable with the wavelet method, and while our second method WAD-2 has shown promising results, the first method WAD-1 is able to produce consistent results (see Table 1).

Figure 4 shows the signal based on EIIP values. We show the resulting signals decomposed up to 5 levels by our first method in Figure 5(a). It shows a cumulative abstraction of the variations in the data over regions progressive to the wavelets scales with coefficients at higher levels; therefore, it depicts features at broader time intervals with the increase in the values of j.

We show the feature vector obtained for a PDZ sequence by our first method based on EIIP, isoelectric, composition, polarity, volume, and molecular weight values in Figures 5(b)–5(g), respectively. It is evident from these figures that wavelet variance value diminishes to almost zero for higher levels (after 3rd level) for almost all the physiochemical values. This clearly indicates that this wavelet variance vector has 3 principal components. The comparison between two feature extractions is depicted in Figure 3, from which it is clear that our signals are smoother with sharper peaks when compared to MODWT signals.

Table 1 shows the comparative predictive performance of WAD-1, WAD-2, and MODWT in terms of four statistical measures: sensitivity, specificity, negative predictive value, and positive predictive value.

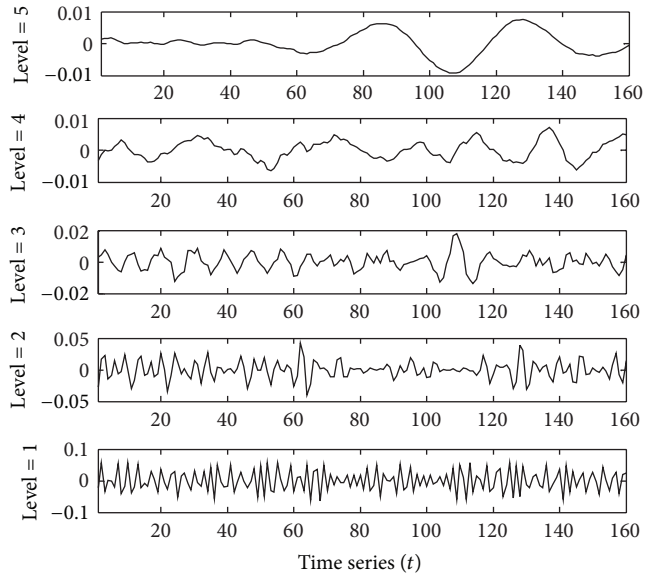
Sensitivity and specificity [35] are statistical measures, which help to evaluate the performance of a binary classification method. In order to understand the significance of sensitivity and specificity in our work, we need to comprehend a few more terms like true positive, false positive, true negative, false negative, which are defined as

(a) true positive: the sample belongs to Class 1, and the algorithm also recognizes it,

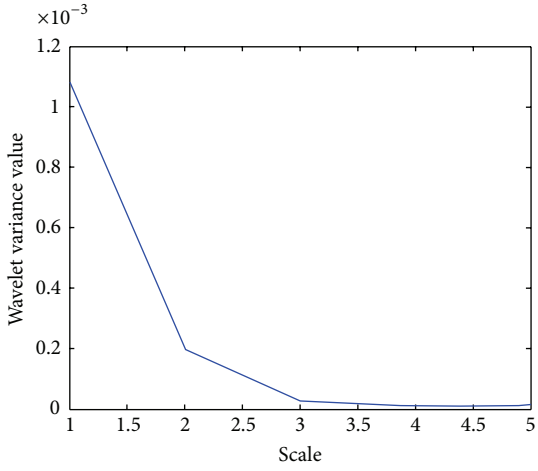
(b) false positive: the sample belongs to Class 2, but the algorithm recognizes it otherwise,

(c) true negative: the sample belongs to Class 2, and the algorithm also recognizes it,

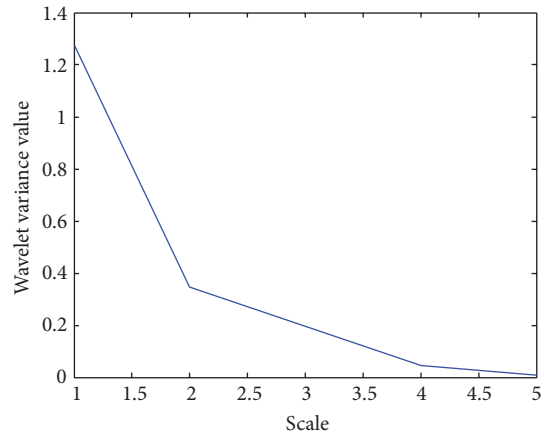
(d) false negative: the sample belongs to Class 1, but the algorithm recognizes it otherwise.



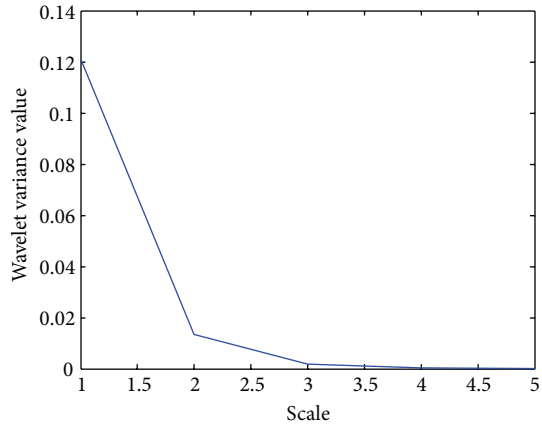
(a) Wavelet coefficients of the mapped protein sequence up to level 5 for EIIP value



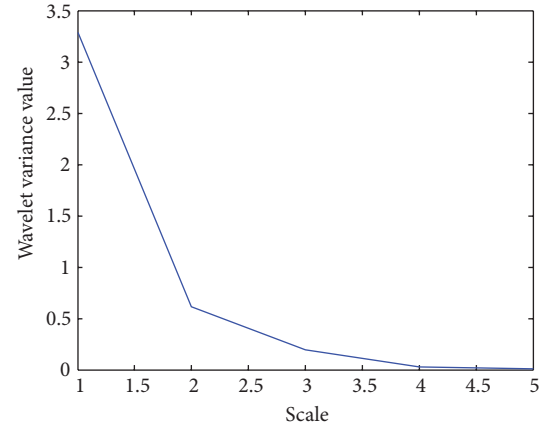
(b) Wavelet variance vector, V_{EIIP} for EIIP value, obtained for the input protein sequence



(c) Wavelet variance vector, $V_{\text{ISOELECTRIC}}$ for isoelectric value, obtained for the input protein sequence

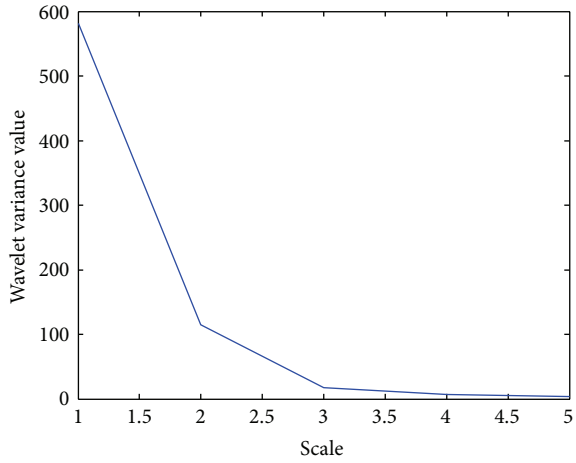


(d) Wavelet variance vector, $V_{\text{COMPOSITION}}$ for composition value, obtained for the input protein sequence

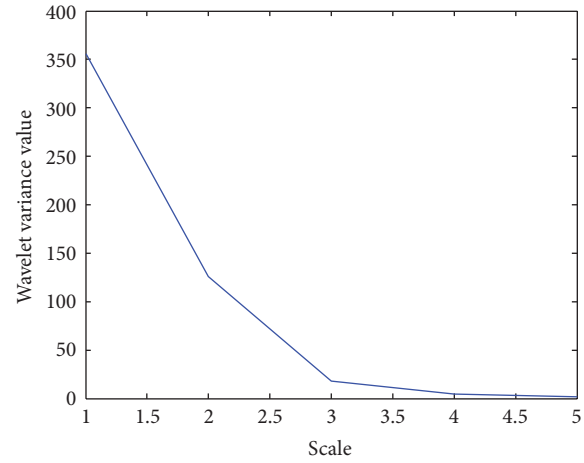


(e) Wavelet variance vector, V_{POLARITY} for polarity value, obtained for the input protein sequence

Figure 5: Continued.



(f) Wavelet variance vector, V_{VOLUME} for volume value, obtained for the input protein sequence



(g) Wavelet variance vector, $V_{\text{MOL-WT}}$ for molecular weight value obtained for the input protein sequence

FIGURE 5: (a) Wavelet coefficients and (b)–(g): wavelet variance vectors for different physiochemical values. Wavelet variance vector is computed by concatenating the wavelet variances for all the levels from $j = 1$ to J. It is computed for all the 7 physiochemical values of amino acids. And the final feature vector for any PDZ sequence is a concatenation of such wavelet variance vectors. From Figures 5(b)–5(g), it is evident that, for most of the physiochemical properties, the wavelet variance vector has 3 principal components.

Therefore, mathematically, sensitivity and specificity are defined as

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}, \quad (10)$$

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false negatives}}. \quad (11)$$

According to Akobeng [36], sensitivity and specificity are important measures of the uninflected statement of an assay but cannot be used to categorize the future developments of the sickness in an individual. But other measures like positive and negative predictive values unveil the prospects regarding the probability of a prediction being correct.

Positive predictive value [37] can be described mathematically as

$$\text{positive predictive value (PPV)} = \frac{\alpha}{\alpha + \alpha'}. \quad (12)$$

Here, α = number of samples that actually belongs to Class 1 and α' = number of samples that does not actually belong to Class 1.

Similarly, negative predictive value [37] is defined mathematically as

$$\text{negative predictive value (NPV)} = \frac{\beta}{\beta + \beta'}. \quad (13)$$

Here, β = number of samples that actually belongs to Class 2 and β' = number of samples that does not actually belong to Class 2.

Both NPV and PPV are crucial estimations of the performance of a prediction algorithm because PPV estimates the probability that a true prediction is actually true, and NPV confirms that a false prediction is actually false.

From the context of medical diagnostics, if we consider Class 1 samples as positives and Class 2 samples as negatives, then a high negative predictive value means that the method very rarely recognizes a negative sample (Class 2) as a healthy one (Class 1). In this regard, our second method performs very well (see Table 1).

However, NPV does not consider cases when a positive sample is identified as a negative one, or in our case it, does not give any idea on the chances of misclassification of sample from Class 1 to Class 2.

Therefore, we calculated PPV which complements this drawback of NPV. Similar arguments also apply in case of predictive drawbacks of PPV.

From Table 1, we observe that WAD-1 performed considerably well and WAD-2 surprised us in recognition accuracy. Kalyoncu et al. [13] have done a sophisticated computational study on interaction prediction and classification of PDZ domains, where they achieved an excellent maximum sensitivity of 90.7%. With our WAD-2 method, we are able to achieve similar results with the highest sensitivity of 91.67%.

6. Conclusions

In this work, we have successfully classified PDZ domains of Classes I and II only. We introduced a new method for feature extraction: an EMD smoothing of signals followed by a MODWT transform. We further introduced WAD-2 method based on the trigram frequency of amino acids and found that the results were improved in terms of sensitivity and accuracy. We note that our second WAD-2 method performed better in recognition accuracy than the WAD-1 method. As mentioned earlier, our work in this paper is meant as a proof-of-principle application for our algorithms. We are enthusiastic that further improvement of our algorithm can lead to even

better accuracy and predicting of the promiscuity of PDZ domains.

Abbreviations

EMD: Empirical mode decomposition
MODWT: Maximum overlap discrete wavelet transform
IMFs: Intrinsic mode functions
GPRC: G protein coupled receptors
FV: Feature vector
ANOVA: Analysis of variance
PPV: Positive predictive value
NPV: Negative predictive value.

Acknowledgments

This study was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia under Grant no. (903-004-D1434). The authors thank the DSR for the technical and financial support.

References

[1] V. Raghuram, D. O. D. Mak, and J. K. Foskett, "Regulation of cystic fibrosis transmembrane conductance regulator single-channel gating by bivalent PDZ-domain-mediated interaction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 1300–1305, 2001.

[2] C. P. Ponting, C. Phillips, K. E. Davies, and D. J. Blake, "PDZ domains: targeting signalling molecules to sub-membranous sites," *BioEssays*, vol. 19, no. 6, pp. 469–479, 1997.

[3] K. K. Dev, "Making protein interactions druggable: targeting PDZ domains," *Nature Reviews Drug Discovery*, vol. 3, no. 12, pp. 1047–1056, 2004.

[4] B. Z. Harris and W. A. Lim, "Mechanisma and role of PDZ domains in signaling complex assembly," *Journal of Cell Science*, vol. 114, no. 18, pp. 3219–3231, 2001.

[5] A. S. Fanning and J. M. Anderson, "PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane," *Journal of Clinical Investigation*, vol. 103, no. 6, pp. 767–772, 1999.

[6] S. Tsunoda, J. Sierralta, Y. Sun et al., "A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade," *Nature*, vol. 388, no. 6639, pp. 243–249, 1997.

[7] D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon, "Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ," *Cell*, vol. 85, no. 7, pp. 1067–1076, 1996.

[8] Z. Songyang, A. S. Fanning, C. Fu et al., "Recognition of unique carboxyl-terminal motifs by distinct PDZ domains," *Science*, vol. 275, no. 5296, pp. 73–77, 1997.

[9] M. Sheng and C. Sala, "PDZ domains and the organization of supramolecular complexes," *Annual Review of Neuroscience*, vol. 24, pp. 1–29, 2001.

[10] U. Wiedemann, P. Boisguerin, R. Leben et al., "Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides," *Journal of Molecular Biology*, vol. 343, no. 3, pp. 703–718, 2004.

[11] H. Tochio, F. Hung, M. Li, D. S. Bredt, and M. Zhang, "Solution structure and backbone dynamics of the second PDZ domain of postsynaptic density-95," *Journal of Molecular Biology*, vol. 295, no. 2, pp. 225–237, 2000.

[12] H. J. Lee and J. J. Zheng, "PDZ domains and their binding partners: structure, specificity, and modification," *Cell Communication and Signaling*, vol. 8, article 8, 2010.

[13] S. Kalyoncu, O. Keskin, and A. Gursoy, "Interaction prediction and classification of PDZ domains," *BMC Bioinformatics*, vol. 11, article 357, 2010.

[14] J. R. Chen, B. H. Chang, J. E. Allen, M. A. Stiffler, and G. MacBeath, "Predicting PDZ domain-peptide interactions from primary sequences," *Nature Biotechnology*, vol. 26, no. 9, pp. 1041–1045, 2008.

[15] I. Bezprozvanny and A. Maximov, "Classification of PDZ domains," *FEBS Letters*, vol. 509, no. 3, pp. 457–462, 2001.

[16] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[17] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 12, pp. 1101–1114, 1994.

[18] J. M. Zimmerman, N. Eliezer, and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of Theoretical Biology*, vol. 21, no. 2, pp. 170–201, 1968.

[19] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[20] http://pages.pomona.edu/~ac044747/aroc/Genetic_Code.swf.

[21] Z. C. Li, X. B. Zhou, Z. Dai, and X. Y. Zou, "Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis," *Amino Acids*, vol. 37, no. 2, pp. 415–425, 2009.

[22] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge, UK, 2002.

[23] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

[24] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Research*, vol. 11, no. 5, pp. 863–874, 2001.

[25] J. A. Tenreiro Machado, "Shannon information and power law analysis of the chromosome code," *Abstract and Applied Analysis*, vol. 2012, Article ID 439089, 13 pages, 2012.

[26] J. A. Tenreiro Machado, A. C. Costa, and M. D. Quelhas, "Can power laws help us understand gene and proteome information?" *Advances in Mathematical Physics*, vol. 2013, Article ID 917153, 10 pages, 2013.

[27] L. J. Buturović, "PCP: a program for supervised classification of gene expression profiles," *Bioinformatics*, vol. 22, no. 2, pp. 245–247, 2006.

[28] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, no. 1971, pp. 903–995, 1998.

[29] R. T. Rato, M. D. Ortigueira, and A. G. Batista, "On the HHT, its problems, and some solutions," *Mechanical Systems and Signal Processing*, vol. 22, no. 6, pp. 1374–1394, 2008.

[30] "Nonlinear denoising via wavelet shrinkage," http://rss.acs.unt.edu/Rdoc/library/wmtsa/html/wavShrink.html.

[31] C. R. Cornish, C. S. Bretherton, and D. B. Percival, "Maximal overlap wavelet statistical analysis with application to atmospheric turbulence," *Boundary-Layer Meteorology*, vol. 119, no. 2, pp. 339–374, 2006.

[32] R. Tonikian, Y. Zhang, S. L. Sazinsky et al., "A specificity map for the PDZ domain family," *PLoS Biology*, vol. 6, no. 9, article e239, 2008.

[33] R. Gupta, A. Mittal, K. Singh, V. Narang, and S. Roy, "Time-series approach to protein classification problem," *IEEE Engineering in Medicine and Biology Magazine*, vol. 28, no. 4, pp. 32–37, 2009.

[34] D. P. Percival, "On estimation of the wavelet variance," *Biometrika*, vol. 82, no. 3, pp. 619–631, 1995.

[35] D. G. Altman and J. M. Bland, "Diagnostic tests 1: sensitivity and specificity," *British Medical Journal*, vol. 308, no. 6943, article 1552, 1994.

[36] A. K. Akobeng, "Understanding diagnostic tests 1: sensitivity, specificity and predictive values," *Acta Paediatrica*, vol. 96, no. 3, pp. 338–341, 2007.

[37] D. G. Altman and J. M. Bland, "Diagnostic tests 2: predictive values," *British Medical Journal*, vol. 309, no. 6947, article 102, 1994.