Journal of Theoretical and Applied Information Technology
15th February 2018. Vol.96. No 3
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION OF HUMAN MEMBRANE PROTEIN TYPES USING OPTIMAL LOCAL DISCRIMINANT BASES FEATURE EXTRACTION METHOD

[1]NOR ASHIKIN MOHAMAD KAMAL, [2]AZURALIZA ABU BAKAR, [3]SUHAILA ZAINUDIN

[1]Department of Computer Science, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA
[2,3]Data Mining and Optimization Research Group, Center of Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, 53600 Bangi, Selangor, MALAYSIA
E-mail: [1]nor_ashikin@tmsk.uitm.edu.my, [2]azuraliza@ukm.edu.my

ABSTRACT

This paper presents a method of membrane protein feature extraction using a combination of the local discriminant bases (LDB) and three different classifiers. This method has adopted two dissimilarity measures of normalized energy difference and relative entropy to identify a set of orthogonal subspaces in optimal wavelet packets. The energy will be derived from the calculation of the two dissimilarity measures that have overlapping subspaces. This feature, in turn, serves as an input to support vector machine (SVM), decision tree and naïve Bayes classifiers. The proposed model yields the highest accuracy of 78.6%, 76.25%, 76.72% for dataset S1, S2, and S3 respectively by using SVM. This technique outperformed other feature extraction method for membrane protein type classification for dataset S2 and S3.

Keywords: *Membrane Proteins, Feature Extraction, Local Discriminant Bases, Wavelet, SVM*

1. INTRODUCTION

Membrane cell plays important roles in various biological processes, such as transporting molecules into or out of the cell, relay signals between the cell's internal and external environments, and provide the skeleton for the lipid bilayer membranes [1,2]. There are generally six types of membrane proteins, namely peripheral, single-pass type I, GPI-anchor, multi-pass, lipid-anchor, and single-pass type I [3]. Since the function of membrane proteins are related to their type, it is important to produce a classification model that can predict membrane protein types correctly. During the past few years, amino acid composition (AAC) [4] and pseudo amino acid composition (PseAAC) [5-7] have been the popular tools for protein feature extraction. AAC considers only the frequency of occurrence of each amino acid in the protein sequence. PseAAC was introduced by [18] to overcome the limitation of AAC. PseAAC complements AAC by reflecting the order of amino acid in the protein sequence. Nevertheless, feature space of PseAAC is redundant [3]. The integrated method that combines BLAST, protein-protein interaction and shortest distance methods was introduced by [3] as the feature extraction method for membrane protein type classification. Despite the numerous spatial feature extraction methods mentioned above, there has been a small number of studies that used frequency domain as the feature extraction method. These include the Fourier transform (FT) [9] and the discrete wavelet transform (DWT) [10]. However, FT is not a suitable technique to represent non-stationary signal since it cannot provide simultaneous time and frequency localization. On the other hand, DWT has good spatial and frequency domain localization properties make wavelet a powerful tool for characterizing signal. Wavelet packet transform (WPT) is a generalization of DWT whereby for each decomposition level, the approximation signal, as well as the detail signals, are filtered to obtain another low and high frequency signal. Although there are many ways ($> 2^L$) to analyze wavelet packet subbands using L-level decomposition, most of the researchers used the wavelet packet coefficient in the last decomposition to extract the signal features [11]. However, the last decomposition level coefficient may not be the best feature representation for a signal. Therefore, it is necessary to optimize the decomposition. In this paper, the combination of Wavelet Packet Transform (WPT) and Local Discriminant Basis (LDB) algorithms have not been investigated for membrane protein feature extraction. The algorithm is introduced and implemented by [11]. It utilized normalize energy difference and relative entropy dissimilarity measure to choose the optimal set of orthogonal subspace derived from WPT. Then, the energy features generated from the optimal wavelet packet subspaces are used as the input into support vector machine (SVM), decision tree and naïve Bayes classifiers for classification. This research shows the suitability of using WPT and improved LDB algorithms as the membrane protein feature extraction method. It shows that the WPT and improved LDB algorithm can improve the classification accuracy. This paper extends the work of [11] by using protein datasets and additional three classifiers. The remaining parts of the paper are organized as follows: Section 2 expresses materials and methods; Section 3 describes evaluations; Section 4 presents results and discussion, and the conclusion is provided in Section 5.

2 MATERIALS AND METHODS

2.1 Dataset

To evaluate the performance of the proposed method, we have selected 3 human membrane protein datasets from [12], which is SI, S2 and S3. Every dataset consists of six types of membrane proteins. S1 contains 2,876 protein sequences, that can be divided into 6 families: 1,414 multipass, 140 are lipid-anchored, 545 are peripheral, 546 single-pass type I, 161 single-pass type II and 70 GPI anchor membrane protein sequences. On the other hand, S2 comprises 2,073 membrane protein sequences classified into 879 multipass, 84 lipid-anchored, 470 peripheral, 436 single-pass type I, 144 single-pass type II, and 60 GPI-anchor membrane protein sequences. Dataset S3 comprises of 1,463 membrane protein sequences having 521 multipass, 60 lipid-anchored, 405 peripheral, 329 single pass type I, 103 single pass type II, and 45 GPI-anchor membrane protein sequences. The same dataset has also been used in [3]. The distribution of types of membrane proteins on dataset S1, S2 and S3 are shown in Figure 1, 2 and 3 respectively.

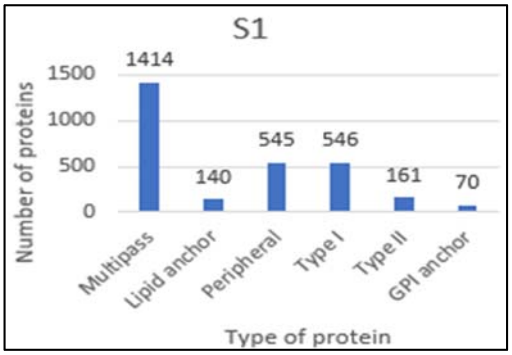

Figure 1: Distribution of Types of Membrane Proteins on Dataset S1

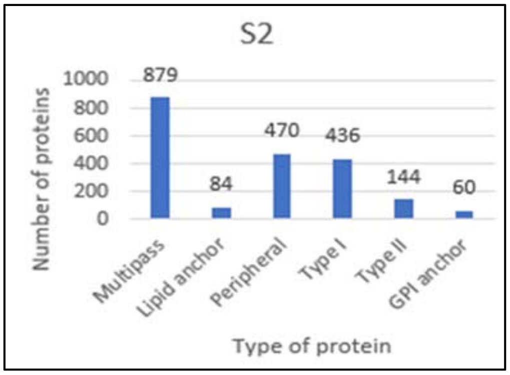

Figure 2: Distribution of Types of Membrane Proteins on Dataset S2

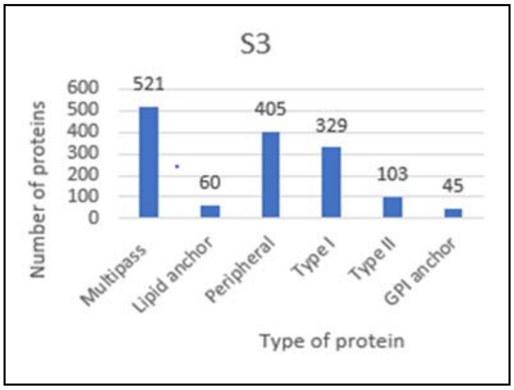

Figure 3: Distribution of Types of Membrane Proteins on Dataset S3

2.2 Feature Extraction Strategies

In this study, pseudo amino acid composition is used to extract features of membrane protein sequences. Subsequently, the extracted features are transformed into wavelet packets in order to obtain optimum subspaces via the local discriminant bases method.

2.2.1 Pseudo Amino Acid Composition (PseAAC)

To overcome the limitation of amino acid composition—which only uses the frequency of occurrence of each amino acid—PseAAC [1] was proposed to preserve the protein sequence order and protein sequence length information. PseAAC has been used for predicting GPCRs and their types, using different physiochemical properties [13]. In PseAAC, the protein *P* can be expressed as follows:

$$PseAAC = P_1, P_2, ..., P_{20}, ..., P_{\wedge} \qquad (2.1)$$

where,

$$\wedge = 20 + n\lambda \qquad (2.2)$$

The first 20 elements from P_1 to P_{20} in Equation (2.1) represent the frequency of amino acid occurrence. λ is the number of tiers used in PseAAC, λ =1,…,m. These tiers obtain information from the correlation factors. The correlation factors are determined by the physiochemical properties. This paper follows the physicochemical characters of amino acid suggested in [13]. They are hydrophobicity, hydrophilicity, side chain mass, pK of the $\alpha - NH_3^+$ group, pK of the $\alpha - COOH$ group, and pI at $25^{\circ}C$ group. Type II PseAAC is used to represent proteins and set λ =25 is the optimal number of tiers that are able to lead to a higher prediction accuracy. Therefore, the number of features generated is $\wedge$ =20+6*25, which is 83. In this paper, PseAAC was generated using PseAAC-Builder [14]. The next process involved the transformation PseAAC into transform domain.

2.2.2 Wavelet Packet Transform (WPT)

The wavelet packet method is a generalization of wavelet decomposition that provides a wider range of signal analysis. For decomposition, the PseAAC is divided into approximation and detail components as shown in Figure 4, where *h(k)* is the low-pass filter and *g(k)* is the high-pass filters. At every decomposition levels, WPT enables all nodes in the tree structure to divide into approximation and detail coefficients at every decomposition levels.

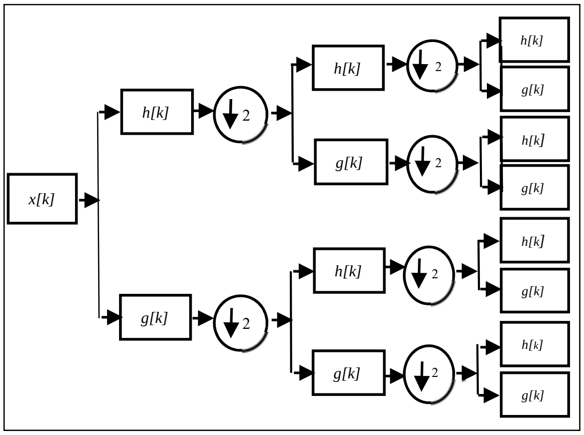

Figure 4: Wavelet Packet Transform

The equations of WPT filtering operations are described as follows:

$$a_i(k) = \sum_n h(n-2k)a_{i+1}(n) \qquad (2.3)$$

$$d_i(k) = \sum_n g(n-2k)a_{i+1}(n) \qquad (2.4)$$

where $a_i(k)$ and $d_i(k)$ represents approximation and detail coefficients of the wavelet packet decomposition, respectively.

An extensive search for the optimal decomposition is not feasible since the number of decompositions may be very large. Since WPT concentrates the energy of the signal into parts of trees, it is important to find an optimal node by using the basis selection algorithms [11]. In addition to feature extraction, the basis selection algorithms also enable feature selection, which is best for classification. Basis selection algorithms include Single Level Basis Selection (SLBS) [17], Best Basis Selection (BBS) [15], Local Discriminant Basis (LDB) [16], and Multi Level Basis Selection (MLBS) [17].

2.2.3 Local Discriminant Bases (LDB)

Local discriminant bases (LDB) were first introduced by Saito and Coifman [16]. The LDB algorithm is an extension of the 'best-basis' algorithm for the selection of a suitable orthogonal basis for the purpose of signal and image classification. Although the best-basis algorithm [15] is the first wavelet-based algorithm to reduce the feature dimensions, it is more appropriate to select a redundant basis of orthogonal functions to compress the signal. The best-basis algorithm chooses a basis by maximizing the entropy of orthogonal bases. Whereas LDB maximizes certain discriminant measures among classes.

In the first step of LDB, the results of PseAAC transformation into WPT in section 2.2.2 are used to calculate the time-frequency energy maps, C_l. This is done for $l=1,\dots,L$ to wavelet packet coefficient using (2.5).

$$C_L(j,k,m) \equiv \frac{\sum_{i=1}^{N_l} (\omega_{j,k,m}^T x_i^{(l)})^2}{\sum_{i-1}^{N_l} \left\| x_i^{(l)} \right\|^2} \qquad (2.5)$$

Suppose $A_{j,k} = B_{j,k}$ and $\Delta_{j,k} = D(\{C_l(j,k,.)\})_{l=1}^{L}$ whereby this array contains dissimilarity measures for nodes (j,k), for $k = 0,\dots,2^{J-1}$. The best subspaces $A_{j,k}$ are obtained by the condition when $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$ that is if the dissimilarity measure for the parent node is greater than the cumulative dissimilarity of the children nodes, then $A_{j,k} = B_{j,k}$ else $A_{j,k} \geq A_{j+1,2k} \oplus A_{j+1,2k+1}$ and $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$. When all sets of orthogonal subspaces have been found, each basis function is ranked from high to low based on their discrimination power. Subsequently, t (value less than n) most discriminant basis functions can be used for construction classifier.

The selection of LDB subspaces is used to differentiate every class [11]. It will determine the classification accuracy that will be obtained. By using only one dissimilarity measure, probably the characteristics for certain classes are unable to be recognized [2]. Therefore, this research used the same dissimilarity measures as in [11], which are normalized energy difference and relative entropy in order to gain high accuracy.

Normalized energy difference, D_1 is calculated based on the following equation:

$$D_1^{(1,2)} = \left| E_{j,k}^1 - E_{j,k}^2 \right| \qquad (2.6)$$

where $E_{j,k}^1$ and $E_{j,k}^2$ are the normalized energy of the corresponding wavelet packet nodes (j,k) that can be calculated using the following formula:

$$E_{j,k} = \frac{\sum_{m=0}^{m=2^{n0}-1} \alpha_{j,k,m}^2}{E_{x_i}} \qquad (2.7)$$

where $j=0,1,\dots,J, k=0,1,..2^j-1, n_0 = \log_2 n \geq J$. n is the signal size and n_0 is the maximum level of signal decomposition. $\alpha_{j,k,m}$ denotes the wavelet packet coefficient for nodes (j,k) at position m. $E_{j,k}$ is the total energy of signals. The second dissimilarity measure is relative entropy. It is expressed as follows:

$$D_2^{(1,2)} = \sum_{i=1}^{n} p_i^{(1)} \log \frac{p_i^{(1)}}{p_i^2} \qquad (2.8)$$

where $n = 2^{n_0-j} - 1, \sum_i p_i^{(1)} = \sum_i p_i^{(2)} = 1$ and $p_m(j,k) = \alpha_{j,k,m}^2 / \sum_{i=1}^{n} \left|\alpha_{j,k,i}\right|^2$ is the energy proportion for wavelet coefficient $\alpha_{j,k,m}$ that produces the total energy for nodes (j,k). The normalized energy difference and relative entropy for more than two class problems can be defined as follows:

$$D_1 = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D_1^{(i,j)} \qquad (2.9)$$

$$D_2 = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D_2^{(i,j)} \qquad (2.10)$$

The process to obtain LDBs using two discriminative measures can be described as follows. The PseAAC features are first decomposed by WPT. Next, the calculation for normalized energy difference as well as relative entropy for each subspace is done for every class using Equations (2.9) and (2.10). Subsequently, the wavelet packet tree is pruned based on the following rules, from top to bottom: If the discriminative measure for the parent node is greater than the cumulative discriminative measure of the children nodes, the parent node

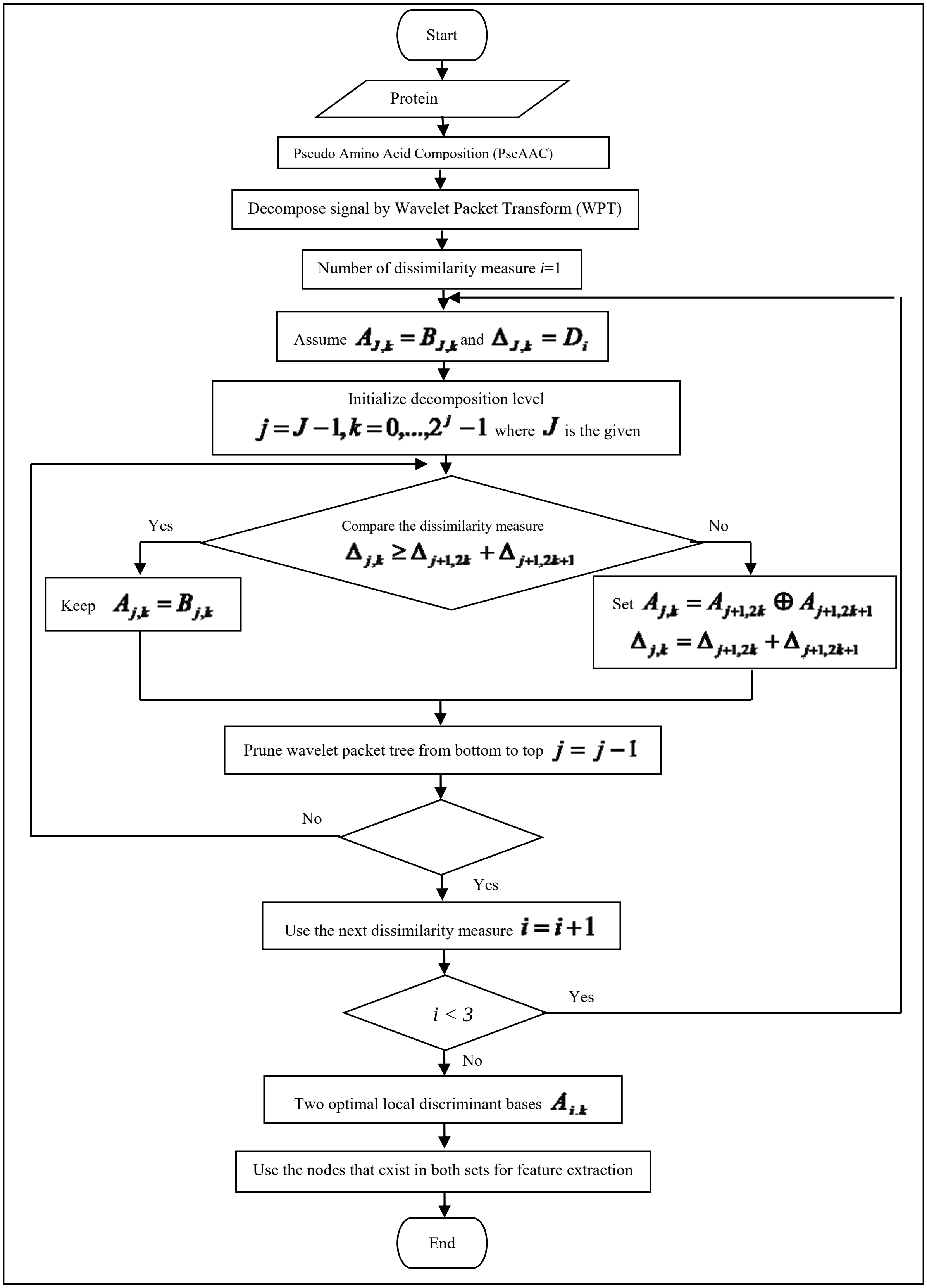
Start
Protein
Pseudo Amino Acid Composition (PseAAC)
Decompose signal by Wavelet Packet Transform (WPT)
Number of dissimilarity measure i=1
Assume $A_{J,k} = B_{J,k}$ and $\Delta_{J,k} = D_i$
Initialize decomposition level
$j = J-1, k = 0,...,2^j - 1$ where J is the given
Yes
Compare the dissimilarity measure
$\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$
No
Keep $A_{j,k} = B_{j,k}$
Set $A_{j,k} = A_{j+1,2k} \oplus A_{j+1,2k+1}$
$\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$
Prune wavelet packet tree from bottom to top $j = j-1$
No
Yes
Use the next dissimilarity measure $i = i+1$
$i < 3$
Yes
No
Two optimal local discriminant bases $A_{i,k}$
Use the nodes that exist in both sets for feature extraction
End

Figure 2: Flow Chart Of Protein Feature Extraction Using LDB Method