Improved Image Based Protein Representations with Application to Membrane Protein Type Prediction

Juan D. Clares*, Victoria Sanchez*, Antonio M. Peinado*, Juan A. Morales-Cordovilla*,
Concepción Iribar† and José M. Peinado†
*Dpt. Teoría de la Señal, Telemática y Comunicaciones and CITIC-UGR
† Dpt. Bioquímica y Biología Molecular III e Inmunología
Universidad de Granada, Granada, Spain
Email: jdavidclares@correo.ugr.es, {victoria, amp, jamc, mciribar, jpeinado}@ugr.es

***Abstract*—With the explosion of protein sequences generated in the postgenomic era, there is a need for the development of computational methods to characterize and classify them as an alternative to the experimental methods that are expensive and time consuming. Although the amino acid chains that constitute proteins are originally symbolic chains they can be converted into numerical sequences and processed as signals. One recent approach represents a protein as a set of images derived from numerical representations of the protein based on the physicochemical properties of amino acids. Then a feature vector is extracted from texture descriptors of the set of images. In this paper we adopt the same approach of representing proteins as sets of images but we propose to generate the images from evolutionary or structural characterization of proteins instead of generating them from physicochemical properties. We also propose the use of an alternative texture descriptor that, in combination with the proposed approach, obtains a significant improvement of classification accuracy in a membrane protein type prediction task.**

***Keywords*—image representation; membrane protein; protein type prediction; support vector machine; texture descriptor;**

I. Introduction

Biomedical signal processing aims at extracting significant information from observations of physiological activities of organisms, ranging from gene and protein sequences, to neural and cardiac rhythms, to tissue and organ images [1]. In the case of protein sequences they consist of amino acid chains where each amino acid is represented by a letter from a 20-element alphabet. Thus, in order to use signal processing methods, the symbolic chain must first be converted into a numerical representation. One of the most common numerical representations for proteins is based on the physicochemical or biochemical properties of amino acids, i.e., amino acid indices [2], where each index assigns a numerical value to each amino acid, depending on the specific physicochemical or biochemical property that the index represents. Signal processing tools such as Fourier or wavelet transforms [3]–[5], digital filters [6], or spectral similarity measures [7] have been used in different protein sequence processing tasks.

Another very interesting approach is the one proposed by Nanni et al. in [8] where proteins are represented as images and feature vectors, built from texture descriptors, are obtained and used for classification tasks. In the two methods, named PR (physicochemical representation) and CW (continuous wavelet), that represent a protein as an image the protein is first represented as a numerical signal based on the physicochemical or biochemical properties of amino acids. Then, in the case of PR, an image of size $L \times L$ (where L is the protein length) is built for every physicochemical property where the value of pixel (i, j) is the sum of the value of that property for the amino acid in position i of the protein and the value of the same property for the amino acid in position j. In the case of CW the Meyer continuos wavelet is applied to that same numerical representation and the wavelet power spectrum is extracted by considering different decomposition scales and processed as an image.

In this paper we propose to generate PR and CW images as well but derived from evolutionary or structural characterization of proteins instead of physicochemical or biochemical properties of amino acids. We will show that the proposed approach in combination with an alternative texture descriptor to those proposed in [8] obtains significant improvements in a membrane protein type prediction task.

The paper is organized as follows. The original physicochemical based image representations are presented in Section 2. The proposed approach is described in Section 3. The experimental framework and the obtained results are discussed in Section 4. The last section is devoted to conclusions.

II. Physicochemical based image representations

In order to represent each protein as an image, it is necessary to first represent the amino acid chain that constitutes the protein in numerical form as indicated before and, as it was also stated, one common numerical representation is the one based on the physicochemical or biochemical properties of amino acids, i.e., the amino acid indices. There are 544 indices [2] where each index assigns a number to each amino acid, depending on the specific physicochemical or biochemical property that the index represents. Let $\mathbf{a} = [a(0), a(1), \ldots, a(L-1)]$ represent the symbolic amino acid chain in the protein that we are considering, where L is the number of amino acids in the protein. We would generate the

This research was supported by Project P12.TIC.1485 funded by Consejeria de Economia, Innovacion y Ciencia (Junta de Andalucía).

PR image of size $L \times L$ corresponding to amino acid index I_i, PRim_i, in the following way [8]

$$\text{PRim}_i(j,k) = I_i(a(j)) + I_i(a(k)) \quad j,k = 0,\ldots,L-1 \quad (1)$$

where the value of pixel (j,k) is obtained as the sum of the corresponding numerical values of property I_i for amino acid $a(j)$ and amino acid $a(k)$. If the image has a size larger than 250×250 it is resized to 250×250 [8].

In the case of the CW image, CWim_i, corresponding to amino acid index I_i we have to first generate the numerical representation of the amino acid chain corresponding to amino acid index I_i, $\mathbf{x}_i = [x_i(0), x_i(1), \ldots, x_i(L-1)]$, simply by substituting each amino acid in the chain by the corresponding numerical value of the property I_i for that amino acid, i.e., $\mathbf{x}_i = [I_i(a(0)), I_i(a(1)), \ldots, I_i(a(L-1))]$. Then the Meyer continuos wavelet considering 100 decomposition scales is applied to signal $\mathbf{x}_i$. The resulting magnitude scalogram constitute the CWim_i image (size $100 \times L$) corresponding to amino acid index I_i [8].

Next a feature vector is extracted from each image based on the calculation of texture descriptors. Two texture descriptors are used on the PR and CW images, local phase quantization (LPQ) [9] and local binary pattern histogram Fourier features (LBP-HF) [10]. The LPQ texture analysis method utilizes the Fourier phase information computed locally in a window for every image position. The phases of the four low-frequency coefficients are then decorrelated and uniformly quantized in an eight-dimensional space where a histogram of the resulting codewords is created and used as texture descriptor. The LBP-HF method is based on the computation of uniform local binary pattern histograms and then the extraction of features from the histograms using the discrete Fourier transform.

LPQ represents each image with a feature vector of size 256 and LBP-HF with a feature vector of size 176.

III. Proposed approach

In this paper we propose to alternatively generate PR and CW images from evolutionary or structural characterization of proteins instead of generating them from physicochemical or biochemical properties of amino acids.

Regarding the evolutionary features we propose to use Profile Hidden Markov Models (PHMM) [11] that build an HMM architecture for representing profiles of multiple sequence alignments. We have obtained PHMM profiles from HHblits+HHmake software [12] with default options on UniProt 2016 dataset. The PHMM profile of an L length protein sequence provides a $20 \times L$ matrix, $\mathbf{PH}$, containing the following conditional probabilities

$$PH(i,j) = P(i|j) \quad i = 1,..,20 \quad j = 0,\ldots,L-1 \quad (2)$$

where $P(i|j)$ represents the emission probability of amino acid i at position j where the twenty essential amino acids are numbered as shown in table I.

TABLE I. The 20 essential amino acids

1,A alanine	2,R arginine	3,N asparagine	4,D aspartic acid	5,C cysteine	6,Q glutamine	7,E glutamic acid	8,G glycine	9,H histidine	10,I isoleucine
11,L leucine	12,K lysine	13,M methionine	14,F phenylalanine	15,P proline	16,S serine	17,T threonine	18,W tryptophan	19,Y tyrosine	20,V valine

We then consider each one of the rows of matrix $\mathbf{PH}$ as a numerical representation of the protein, i.e., $\mathbf{x}_i = [PH(i,0), PH(i,1), \ldots, PH(i,L-1)]$ and generate the evolutionary image PREim_i corresponding to amino acid i as

$$\text{PREim}_i(j,k) = PH(i,j) + PH(i,k) \;\; j,k = 0,\ldots,L-1 \quad (3)$$

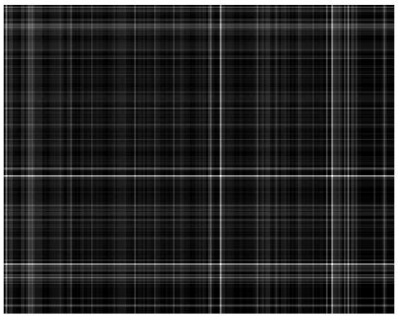

Fig. 1. PRE image corresponding to amino acid 1 (PREim_1) of membrane protein P79336.

As for the evolutionary CW image in order to form it we just apply the Meyer continuous wavelet in the same way as described in the section above but now to signal $\mathbf{x}_i = [PH(i,0), PH(i,1), \ldots, PH(i,L-1)]$. The resulting magnitude scalogram constitute the CWEim_i image (size $100 \times L$) corresponding to amino acid i.

In the case of the structural features we propose to explore the secondary structure of a protein. The sequence of amino acid residues that form a protein is called the primary structure. Secondary structure refers to the arrangement of the primary amino acid sequence into motifs (local conformations) such as α-helices, β-sheets and coils. α-helices are corkscrew-shaped conformations where the amino acids are packed tightly together, β-sheets are made up of two or more adjacent strands of the molecule, extended so that the amino acids are stretched out and the rest of conformations that are not associated with a regular secondary structure are refered to as coils. Given a protein sequence (primary structure) $\mathbf{a} = [a(0), a(1), \ldots, a(L-1)]$, the secondary structure prediction problem is to predict whether each amino acid $a(j)$, $j = 0,\ldots,L-1$ belongs to an α-helix (H), a β-sheet (E), or a coil (C). There is a number of secondary structure predictors available [13] but PSIPRED (version 3) is one of the most accurate [14]. PSIPRED incorporates two feed-forward neural

networks and generates, for an input protein, a $3 \times L$ matrix $\mathbf{SS}$ that contains the predicted likelihood of each one of the three secondary structures (H,E,C) at amino acid residue position $j,\ j = 0, \ldots, L-1$.

As with matrix $\mathbf{PH}$ we then consider each row of matrix $\mathbf{SS}$ as a numerical representation of the corresponding protein, i.e., $\mathbf{x}_i = [SS(i,0), SS(i,1), \ldots, SS(i,L-1)]$ and generate the structural PRSim$_i$ image corresponding to structure i as

$$\text{PRSim}_i(j,k) = SS(i,j) + SS(i,k) \quad j,k = 0, \ldots, L-1 \quad (4)$$

The structural CW image is generated as well by just applying the Meyer continuous wavelet to signal $\mathbf{x}_i = [SS(i,0), SS(i,1), \ldots, SS(i,L-1)]$. The resulting magnitude scalogram constitute the CWSim$_i$ image (size $100 \times L$) corresponding to secondary structure i.

Additionally, we propose to explore the use of a recently developed texture descriptor [15] named Binary Gabor Pattern (BGP). This texture descriptor constitutes a multi-resolution approach where the image is first convolved with J Gabor filters with different orientations. The obtained responses are then binarized obtaining J bits at each location and, by the use of some rule (consisting in obtaining the maximum after circular bitwise right shifts of the binarized response), a unique integer named BGPri (rotation invariant binary Gabor pattern) is assigned to each location. The descriptor is formed by the histogram of its BGPris at multiple scales. BGP represents each image with a feature vector of size 216 [15].

IV. Experimental framework and results

The experimental framework will be focused on the task of predicting membrane protein types. Although a lipid bilayer provides the basic structure of all cell membranes and serves as a permeability barrier to the molecules on either side of it, most membrane functions are carried out by membrane proteins [16]. There are different membrane protein types and the function of a membrane protein is closely correlated with the type it belongs to. Some membrane proteins transport particular nutrients, metabolites, and ions across the lipid bilayer. Others anchor the membrane to macromolecules on either side or function as receptors that detect chemical signals in the cells environment and relay them to the cell interior. Additionally others work as enzymes to catalyze specific reactions. With the explosion of protein sequences generated in the postgenomic era, the determination of membrane protein types by experimental methods is expensive and time consuming. However knowing the type of uncharacterized membrane proteins can be useful for both basic research and drug discovery. It is therefore important to develop computational methods to determine the types of membrane proteins.

The database used for this task is that described in [17] and used in Nanni *et al.* [8] (MEM database). This dataset contains 7582 membrane proteins classified into eight different types: (1) type-I single-pass transmembrane, (2) type-II single-pass transmembrane, (3) type-III single-pass transmembrane, (4) type-IV single-pass transmembrane, (5) multipass transmembrane, (6) lipid-chain anchored membrane, (7) GPI-anchored membrane and (8) peripheral membrane. Types (1)-(4) span the membrane only once, type (5) proteins span the membrane more than once, type (6) proteins are attached to lipids embedded within the cell membrane, type (7) proteins are anchored to the membrane by covalent linkage to glycosylphosphatidylinositol (GPI) and type (8) proteins adhere only temporarily to the biological membrane with which they are associated. The dataset is originally divided [17] into a training subset (3249 proteins) and a test subset (4333 proteins).

In the original PR and CW methods [8], 50 physicochemical properties are used and consequently 50 images are generated by each method for every protein, one image for each physicochemical property considered. Regarding the proposed methods, in the case of PRE and CWE, 20 images are generated (one image for each one of the 20 essential amino acids) and, in the case of PRS and CWS, 3 images are generated (one image for each one of the 3 possible secondary structures H, E, C). Then each image is represented by a texture descriptor (LPQ, LBP-HF or the proposed BGP) feature vector and support vector machines (SVM) used as classifiers. In all the cases an RBF kernel is used with parameter selection by 3-fold crossvalidation and grid search.

In table II we show the performance in terms of accuracy (percentage of correctly classified proteins to all the proteins in the test set) of the different approaches for the membrane protein type multi-class classification task described above.

TABLE II. Classification accuracy obtained by the different image based approaches for the membrane protein type prediction task (the best performances are in bold face).

	LPQ	**LBP-HF**	**BGP**
PR [8]	76.62	-	-
CW [8]	-	**82.18**	-
PR (48 properties from [18])	76.64	74.20	85.57
CW (48 properties from [18])	85.67	86.40	**87.28**
PRE	82.53	88.69	**92.20**
CWE	89.45	89.66	90.30
PRS	80.64	81.70	**85.28**
CWS	81.91	82.69	83.24
PRE+PRS	85.21	89.96	**92.75**
CWE+CWS	89.06	91.14	90.81

The first two rows correspond to the results presented in [8] where 50 physicochemical properties are used. In [8] it is not indicated which particular 50 physicochemical indices were the ones selected from the amino acid index database [2]. As we were interested in testing the behaviour of the original PR and CW approaches with the new BGP texture descriptor we used instead the 48 physicochemical properties selected by Gromiha *et al.* and listed in [18]. The results obtained by PR and CW with these 48 physicochemical properties in combination with the three types of texture descriptors are shown in the next two rows. The rest of the table entries correspond to the results of the proposed approaches PRE,CWE and PRS,CWS with the three texture descriptors and to the combined approaches PRS+PRE and CWS+CWE. In these classifiers we have built a unique feature vector for

each protein that is formed by the concatenation of the texture descriptor vectors corresponding to the images. In table III we show the feature vector length for all the techniques and the different texture descriptors. In the combined approaches PRE+PRS and CWE+CWS we represent each protein by the images generated by both PRE and PRS or both CWE and CWS. In all the cases an SVM has been trained for each class and a one-versus-the-rest approach adopted.

TABLE III. FEATURE VECTOR SIZE FOR EACH ONE OF THE DIFFERENT IMAGE BASED APPROACHES

	LPQ	**LBP-HF**	**BGP**
PR [8] CW [8]	50×256	50×176	50×216
PR(48 properties from [18]) CW(48 properties from [18])	48×256	48×176	48×216
PRE CWE	20×256	20×176	20×216
PRS CWS	3×256	3×176	3×216
PRE+PRS CWE+CWS	23×256	23×176	23×216

As we can observe the best results are obtained by the proposed schemes PRE and CWE in combination with the BGP texture descriptor where we are obtaining 10 points of improvement in accuracy with respect to the results in [8] and 5 points of improvement with respect to the physicochemical based PR and CW schemes in [8] when combined with the BGP descriptor. Regarding the PRS and CWS schemes both also improve the results in [8] when combined with BGP but are two points below the 48 physicochemical based PR and CW schemes with BGP. It is however worth while noticing that in the case of PRS and CWS only three images represent one protein in comparison with the PR and CW schemes where 50 or 48 images are generated for each protein. The best overall result is obtained by the PRE+PRS combined scheme with BGP as texture descriptor where each protein is represented by 23 images. It is interesting to point out that in the proposed schemes the PR images (PRE or PRS) combined with the BGP descriptor always obtain better results than the corresponding CW images (CWE or CWS) but it is just the opposite in the physicochemical based PR and CW schemes where the CW images always obtain better results. This is an advantage as the PR images are much faster to generate as they do not imply any wavelet transform as it is the case with the CW images. We can then conclude that in the evolutionary and structural based representations the BGP texture descriptor is particularly well suited to the PR type images.

V. CONCLUSION

In this paper we have proposed alternative image based representations of proteins and applied them to a membrane protein type classification task. We have shown how to generate PR and CW images from evolutionary or structural characterization of proteins instead of generating them from physicochemical or biochemical properties of amino acids as it was originally proposed. In the case of the evolutionary based images (PRE,CWE), we have built them using PHMM profiles from HHblits and, in the case of the structural based images (PRS,CWS), we have used the secondary structure prediction obtained from PSIPRED. Then, texture descriptors have been used as feature vectors for an SVM based membrane protein type classification task. We have proposed to use the recently developed BGP texture descriptor and obtained that the evolutionary based image representations when combined with BGP significantly improve classificaction accuracy. The best results are finally obtained by the combined evolutionary-structural PRE+PRS approach with BGP, concluding that the BGP texture descriptor seems to be particularly well suited to the PRE and PRS type images.

REFERENCES

[1] H.-H. Chang, J.M.F. Moura, "Biomedical signal processing," in *Biomedical Engineering and Design Handbook*, 2nd ed. vol. 1, M. Kutz, Ed.: McGraw-Hill, 2010, pp. 559–579.

[2] S. Kawashima et al., "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, D202-D205, 2008.

[3] S. Zhang and T. Wang, "Feature analysis of protein structure by using discrete Fourier transform and continuous wavelet transform," *J. Math. Chem.*, vol. 46, pp. 562–568, 2009.

[4] V. Sanchez, A.M. Peinado, J.L. Peréz-Córdoba, A.M. Gómez, "A new signal characterization and signal-based Chou's PseAAC representation of protein sequences," *Journal of Bioinformatics and Computational Biology*, vol. 13, no. 5, pp. 1–24, Oct. 2015.

[5] J. Jia, Z. Liu, X. Xiao, "iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC," *Journal of Theoretical Biology*, vol. 377, pp. 47–56, 2015.

[6] S. Sekhar and G. Panda, "Efficient localization of hot spots in proteins using a novel s-transform based filtering approach," *IEEE-ACM Trans. Comput. Biol. Bioinform.*, vol. 8, pp. 1235–1246, 2011.

[7] K. Gupta et al., "Detailed protein sequence alignment based on Spectral Similarity Score (SSS)," *BMC Bioinformatics*, vol. 6:105, pp. 1–16, 2005.

[8] L. Nanni, S. Brahnam, A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, pp. 657–665, 2012.

[9] V. Ojansivu, J. Heikkil, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *LNCS:International Conference on Image and Signal Processing(ICISP08)*, France, 2008, pp. 239–243.

[10] T. Ahonen, J. Matas, C. He, M. Pietikainen, "Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features," in *LNCS:Scandinavian Conference on Image Analysis(SCIA09)*, Oslo, 2008, pp. 61–70.

[11] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, Oct. 1998.

[12] M. Remmert, A. Biegert, A. Hauser, J. Soding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012.

[13] F. G. Ledda, "Protein Secondary Structure Prediction: Novel Methods and Software Architectures,"Ph.D. dissertation, Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliary, Italy, 2011.

[14] D. W. A. Buchan et al., "Scalable web services for the PSIPRED Protein Analysis Workbench," *Nucleic Acids Research*, vol. 41, pp. W340–W348, Jul. 2013.

[15] L. Zhang, Z. Zhou, and H. Li, "Binary Gabor pattern: An efficient and robust descriptor for texture classification," in *Proc. 19th IEEE International Conference on Image Processing (ICIP2012)*, Orlando, FL, 2012, pp. 81–84.

[16] B. Alberts et al., *Essential Cell Biology. Third Edition*. Garland Science, New York, USA, 2010.

[17] K.C. Chou, H. B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochemical and Biophysical Research Communications*, vol. 360, no. 2, pp. 339–345, Aug. 2007.

[18] M.M. Gromiha, M. Oobatake1, A. Sarai, "Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins," *Biophysical Chemistry*, vol. 82, no. 1, pp. 51–67, Nov. 1999.