

Wavelet images and Chou's pseudo amino acid composition for protein classification

Loris Nanni · Sheryl Brahnam · Alessandra Lumini

Received: 5 April 2010 / Accepted: 28 September 2011 / Published online: 13 October 2011
© Springer-Verlag 2011

Abstract The last decade has seen an explosion in the collection of protein data. To actualize the potential offered by this wealth of data, it is important to develop machine systems capable of classifying and extracting features from proteins. Reliable machine systems for protein classification offer many benefits, including the promise of finding novel drugs and vaccines. In developing our system, we analyze and compare several feature extraction methods used in protein classification that are based on the calculation of texture descriptors starting from a wavelet representation of the protein. We then feed these texture-based representations of the protein into an Adaboost ensemble of neural network or a support vector machine classifier. In addition, we perform experiments that combine our feature extraction methods with a standard method that is based on the Chou's pseudo amino acid composition. Using several datasets, we show that our best approach outperforms standard methods. The Matlab code of the proposed protein descriptors is available at <http://bias.csr.unibo.it/nanni/wave.rar>.

Keywords Proteins classification · Machine learning · Ensemble of classifiers · Support vector machines

Introduction

There are a number of areas where reliable methods for protein classification are of benefit. They are important in detecting remote sequence homologous (Saigo et al. 2004), in sub-cellular localization (Chou and Shen 2007), and in the protein–protein interaction problem (Nanni and Lumini 2006). They are also of value in designing novel drugs and vaccines, especially in the discovery phase, for many human diseases, such as those implicated in mitochondrial defects (Lowell and Shulman 2005; Lin and Beal 2006). To reduce costs and to speed up the development of protein research, it has become crucial to develop better computational methods for solving these protein classification problems.

Extracting features from proteins is an essential first step in building these systems. Various methods have been developed for extracting features from proteins, including pseudo amino acid composition (Chen and Li 2007; Chou 2001; Ding and Zhang 2008) and Markov chains model (Bulashevskaya and Eils 2006). The Chou's pseudo amino acid composition (PseAAC) (Chou and Shen 2007) is one of the most widely used feature extractors for peptides and proteins (Chen et al. 2009; Xiao et al. 2011a, b; Fang et al. 2008; Nanni and Lumini 2008). While maintaining much of the sequence order information, PseAAC represents a protein sequence using a discrete model that is composed of a set of more than 20 discrete factors. The first 20 factors represent the components of its conventional amino acid (AA) composition. Additional factors incorporate some sequence order information using various modes (e.g., a

L. Nanni (✉)
Department of Information Engineering, University of Padua,
Via Gradenigo, 6, 35131 Padova, Italy
e-mail: loris.nanni@unipd.it

S. Brahnam
Computer Information Systems, Missouri State University,
901 S. National, Springfield MO 65804, USA
e-mail: sbrahnam@missouristate.edu

A. Lumini
Department of Electronic, Informatics and Systems (DEIS),
Università di Bologna, Via Venezia 52, 47023 Cesena, Italy
e-mail: alessandra.lumini@unibo.it

series of rank-different correlation factors along a protein chain).

There are many variations of the PseAAC. In Du and Li (2006), feature vectors were constructed by combining PseAAC with dipeptide composition and the occurrence frequencies of different residues. In Du et al. (2009), PseAAC was created by computing the correlation function of the physicochemical properties of two residues for subchloroplast location prediction. In Zeng et al. (2009), a substitution model was constructed based on the augmented PseAAC that was composed of amino acid composition and auto covariance (AC) variables. In Nanni and Lumini (2008), genetic programming was applied to extract artificial features as PseAAC.

Another important class of feature extraction methods includes those that are based on kernels. One of the first kernels used to represent proteins was the Fisher kernel (Jaakkola et al. 1999). Another kernel method is the mismatch string kernel (Leslie et al. 2004), which measures the sequence similarity based on shared occurrences of subsequences. String kernels perform as well as the Fisher kernel but are lower in computational cost. A new class of kernels, proposed in Lei and Dai (2005) derives vectors from k-peptide vectors mapped by a matrix of high-scored pairs, measured using BLOSUM62 scores of k-peptides. Another interesting approach is the bio-basis function neural network (Yang and Thomson 2005), where sequences are not encoded within a feature space. The distances obtained by sequence alignment are used to train the neural network.

The aim of this work is to analyze and to compare several feature extraction methods that are based on the calculation of texture descriptors starting from a wavelet representation of the protein. Recent studies have proposed methods for extracting features from the proteins using wavelets (Li 2008; Qiu et al. 2009; Nanni and Lumini 2010; Shi et al. 2011). These approaches, described below, encode the protein sequence as a numerical sequence by substituting each amino acid with a value corresponding to a given physicochemical property.

Using graphic approaches to study biological systems is highly intuitive and can provide useful insights in the analysis of complicated relations in these systems, as indicated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Zhou and Deng 1984; Andraos 2008; Chou 1989), protein-folding kinetics and folding rates (Chou 1990), inhibition of HIV-1 reverse transcriptase (Althaus et al. 1993), inhibition kinetics of processive nucleic acid polymerases and nucleases (Chou et al. 1994), and drug metabolism systems (Chou 2010). The graphical analysis via the “cellular automaton image” (Wolfram 1984) has also been applied to study hepatitis B viral infections (Xiao et al. 2006) and

HBV virus gene missense mutation (Xiao et al. 2005), as well as representing complicated biological sequences (Xiao et al. 2005) and providing assistance in the identification of various important protein attributes (Xiao et al. 2006, 2011; Xiao and Chou 2007). Recently, the Wenxiang diagram (see Chou et al. 1997 and the web-server at <http://icpr.jci.edu.cn/bioinfo/wenxiang>) has been used to study protein–protein interactions and provided very useful insights (Zhou 2011).

Although Adaboost-based approaches have been used to predict protein structural classes (see, e.g., Niu et al. 2006), we use an Adaboost ensemble of neural networks or a support vector machine classifiers to train our texture-based representations of the protein. In addition, we perform experiments that combine our feature extraction methods with a standard method that is based on the Chou’s pseudo amino acid composition. Moreover, we try to develop methods that succeed in handling a number of datasets. For this reason, we train our classifiers using several datasets that represent several different classification problems. This strengthens our approach so that it outperforms standard methods trained specifically on individual datasets.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the feature extraction methods analyzed in this work. In Sect. 3, we report the experimental results obtained on four classification problems. Finally, in Sect. 4, we draw some conclusions and indicate areas for future research.

Feature extraction

Recently, a number of studies have searched for a compact yet effective representation of proteins (see, e.g., Chou and Shen 2007; Nanni and Lumini 2008). One solution that works well in many problems is based on a fixed length encoding that is combined with a general-purpose classifier. In our study, the set of physicochemical properties we use are obtained from the amino acid index (Kawashima and Kanehisa 2000) database.¹ An amino acid index is a set of 20 numbers that represent different physicochemical properties. In this work, for each protein descriptor, a set of 50 physicochemical properties are used. A different classifier is trained for each physicochemical property, and then they are combined by sum rule. The different properties are selected using the training data by sequential forward floating selection as in Nanni and Lumini (2006).

For each feature extraction method employed in this study, we use an Adaboost of neural networks or a support

¹ Available at <http://www.genome.jp/dbget/aaindex.html>. We have not considered the properties where the amino acids have value 0 or 1.

vector machine² (Cristianini 2000) as our classifiers. Before the classification step, all the features used for training are linearly normalized to [0 1] considering the training data.

In addition to extracting the texture descriptors from the wavelet images, we also use Chou's pseudo amino acid composition (PseAAC) (Zeng et al. 2009). A set of pseudo amino acid-based features are extracted from a given protein. This is the concatenation of the 20 standard amino acid composition values and m (here $m = 20$) values that reflect the effect of the sequence order.

In our experiments, we combine 50 Chou's pseudo amino acid feature vectors using 50 different physicochemical properties extracted from the AAindex³ (Kawashima and Kanehisa 2000). Different classifiers are trained for each Chou's pseudo amino acid feature vector. The 50 classifiers are then combined using the sum rule.

Below, we discuss the feature extraction methods used in our experiments. In Sect. 2.1, we describe several methods for encoding the protein sequences as wavelets. In Sect. 2.2, we describe the texture features we extract from the resulting wavelet images.

Wavelet descriptor (WA)

Proposed methods for extracting features from the proteins using wavelets include Li (2008), Qiu et al. (2009), Nanni and Lumini (2010), Shi et al. (2011), and Wen et al. (2005). These methods encode the protein sequence as a numerical sequence by substituting each amino acid with a value corresponding to a given physicochemical property. Different decomposition scales have different results in analyzing protein sequences. A too high a decomposition scale would introduce redundancy in the decomposing process while a too low decomposition level would omit much detailed information (Wen et al. 2005).

In Li (2008), the Meyer continuous wavelet is then applied to this encoding, and the wavelet power spectrum is extracted by considering different decomposition scales. In this paper, we name this method for representing a protein as an image CW.

In our experiments, we explore another method for representing a protein as an image. The algorithm is as follows. Given a physicochemical property, the value of the element (i,j) of the image that describes a given protein is the sum of the value of the physicochemical property of the amino acid in position i of the protein and the value of the physicochemical property of the amino-acid in position j . Then, if the image has a size larger than 250×250 , it is

resized to 250×250 . We named this method for representing a protein as an image PR, some examples are reported in Fig. 1.

Finally, we use a method of applying wavelets that is derived from that proposed in Qiu et al. (2009), Nanni and Lumini (2010), and Shi et al. (2011), where a biorthogonal discrete wavelet is used to describe a protein. In this method, wavelet coefficients, using different scales, are extracted from the maximum, minimum, mean and standard deviation values. We also extract (as in Nanni and Lumini 2010) the first five discrete cosine coefficients from the approximation coefficients and maximum, minimum, mean and standard deviation values from both detail and approximation coefficients of the wavelet decomposition (four scales are used). This is because the high-frequency components are more noisy and hence only the low-frequency components are more important, just like the case of protein internal motions where the low-frequency components are functionally more important, as elucidated in Chou (1985, 1988, 1989). We have named this method as DW.

Texture descriptors

In our experiments, the following texture descriptors are extracted from the wavelet images: Dominant Local Ternary Patterns (DLTP), Local Phase Quantization (LPQ), Local Binary Pattern Histogram Fourier (LBP-HF), and the Multiscale Autoconvolution Histogram (MSAhist). Each of these texture descriptors is described below.

Dominant Local Ternary Patterns (DLBP)

DLPT combines Dominant Local Binary Patterns (DLBP) with Local Ternary Patterns (LTP). DLBP was proposed in Liao et al. (2009) for selecting the rotation invariant patterns to be selected in LBP. Rather than selecting uniform patterns, patterns were chosen in Liao et al. (2009) that represent 80% of the whole pattern occurrences in the training data. The LBP operator is calculated by evaluating

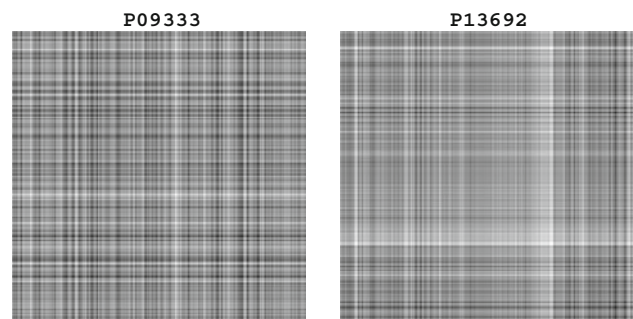


Fig. 1 PR representation of the proteins P09333 and P13692

² <http://sourceforge.net/projects/svm/>.

³ The IDs of the properties are available at <http://bias.csr.unibo.it/nanniVDw.docx>.

the binary differences between the gray value of a pixel \mathbf{x} and the gray values of P neighboring pixels on a circle of radius R around \mathbf{x} . The LBP operator is made rotation invariant by selecting the smallest value of $P - 1$ bitwise shift operations on the binary pattern. A pattern is considered uniform if the number of transitions in the sequence between 0 and 1 is less than or equal to two.

In LTP (Tan and Triggs 2007), the difference between a pixel \mathbf{x} and its neighbor \mathbf{u} is encoded by three values according to a threshold τ : 1 if $\mathbf{u} \geq \mathbf{x} + \tau$; -1 if $\mathbf{u} \leq \mathbf{x} - \tau$; else 0. The ternary pattern is then split into two binary patterns by considering its positive and negative components. Finally, the histograms that are computed from the binary patterns are concatenated to form the feature vector. Here $\tau = 0.15$; $P = 16$; $R = 2$.

Local Phase Quantization (LPQ)⁴

The LPQ operator as a texture descriptor was originally proposed by Ojansivu and Heikkilä (2008). LPQ is based on the blur invariance property of the Fourier phase spectrum. In this method, local phase information extracted using the 2D short-term Fourier transform (STFT) is computed over a rectangular neighborhood at each pixel position of the image. Only four complex coefficients are considered in LPQ that correspond to the 2D frequencies. Our experiments use the same code as was used in Ojansivu and Heikkilä (2008). For the mathematical details, refer to Ojansivu and Heikkilä (2008).

Local Binary Pattern Histogram Fourier (LBP-HF)⁵

LBP-HF, first proposed by Ahonen et al. (2009), is a rotation invariant image descriptor that is computed globally from the discrete Fourier transforms of LBP histograms. The LBP-HF descriptor computes a non-invariant LBP histogram over the entire region and then constructs rotationally invariant features from the histogram using Discrete Fourier Transform. The features are invariant to cyclic shifts in the input vector. The overhead of LBP-HF, compared with the LBP histogram, is low since only $P - 1$ Fast Fourier Transforms of P points needs to be computed from the LBP histogram to construct LBP-HF. Here ($P = 16$; $R = 2$) and ($P = 8$; $R = 1$).

Multiscale Autoconvolution Histogram (MSAhist)

The MSAhist, proposed by Rahtu et al. (2005), is a derivation of the multiscale autoconvolution (MSA) transform (Rahtu et al. 2005). The MSAhist is a new way of

constructing affine invariant histograms from images. It uses the same affine invariant random variable as MSA but estimates the complete distribution rather than an expected value and has about the same computational complexity as MSA. MSAhist can be considered a generalization of the gray scale histogram except that it also encodes spatial information. For the mathematical details, refer to Rahtu et al. (2005).

Datasets

In this section, we briefly describe the datasets and the related protein problems that are used in our experiments:

- *GPCR* (Xiao et al. 2009) is a dataset that contains 365 G protein-coupled receptors (GPCR) and 365 non-GPCR. None of the proteins included in this dataset has $\geq 40\%$ pairwise sequences that are identical to any other in the same subset.
- *Virulent dataset* (Garg and Gupta 2008) contains bacterial virulent protein sequences that were retrieved from two databases, the SWISS-PROT (Bairoch and Apweiler 2000) and VFDB (Chen et al. 2005), it contains 1,025 virulent and 1,030 non-virulent bacterial sequences. In the original protocol (Garg and Gupta 2008), the Virulent dataset is used as training set. The ADHESINS dataset (Garg and Gupta 2008) was used as the testing set, it contains 469 adhesins and 703 non-adhesins proteins, including several archaeobacterial, viral, and yeast non-virulent proteins. None of the proteins included in this dataset has $\geq 40\%$ pairwise sequences that are identical to any other in the same subset.
- *HUM dataset* (Bock and Gough 2003) examines human protein–protein interaction. HUM contains a total of 1,882 human interacting/non-interacting protein pairs. None of the proteins included in this dataset has $\geq 40\%$ pairwise sequences that are identical to any other in the same subset.
- *HEL dataset* (Bock and Gough 2003) examines helicobacter protein–protein interaction. HEL contains a total of 2,916 helicobacter interacting/non-interacting protein pairs. None of the proteins included in this dataset has $\geq 40\%$ pairwise sequences that are identical to any other in the same subset.
- *Membrane sub-cellular (MEM)* (Chou and Shen 2007) dataset contains 7,582 membrane proteins that belong to eight membrane types. None of the proteins has 80% sequence identity to any other in the same subset (subcellular location).
- *GRAM* (Shen and Chou 2007) dataset contains 452 Gram-positive proteins that belong to five subcellular

⁴ <http://www.cse.oulu.fi/Downloads/LPQMatlab>.

⁵ <http://www.cse.oulu.fi/MVG/Downloads/LBPMatlab>.

location sites. To eliminate redundancy and homology bias, only those proteins that have <25% sequence identity to any other in a same subcellular location were allowed to be included in the benchmark datasets.

- *RAW* (Shen and Chou 2007) dataset contains 413 proteins that belong to four subchloroplast locations, including stroma, thylakoid lumen, thylakoid membrane, and envelope. To eliminate redundancy and homology bias, only those proteins that have <60% sequence identity to any other in a same subcellular location were allowed to be included in the benchmark datasets.

Experiments

In this section, we evaluate the protein descriptors described above by performing experiments using the descriptors on several datasets. In the experiments reported in this section, we use the area under the ROC curve (AU)⁶ (Fawcett 2004; Qin 2006) as the performance indicator. AU is a scalar measure, which can be interpreted as the probability that the classifier will assign a lower score to a randomly picked positive pattern than to a randomly picked negative pattern. The area under the ROC is widely considered one of the most reliable performance indicators (Qin 2006). It is based on both sensitivity and specificity.

Instead, when a multiclass dataset is used, the accuracy is employed as performance indicator. Among the independent dataset tests [sub-sampling (e.g., two or fivefold cross-validation) test and the jackknife test, which are often used for examining the accuracy of a statistical prediction method (Chou and Zhang 1995)], the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, as elucidated in (Chou and Shen 2010) and demonstrated by Eqs. 28–32 of Chou (2011). Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (see, e.g., Chen et al. 2009; Zeng et al. 2009; Li 2008; Lin 2008; Zhou et al. 2007; Lin et al. 2008; Ding et al. 2009; Kandaswamy et al. 2011; Masso and Vaisman 2010; Hayat and Khan 2011; Jiang et al. 2008; Hu et al. 2011; Chou and Shen 2010; Mohabatkar 2010). However, to reduce the computational time, we adopted the twofold cross-validation in this study,⁷ as done by many investigators with complex machine learning classifiers as the prediction engine. The exceptions to this protocol are with the Virulent, GRAM, and MEM datasets. Since those datasets are originally

divided in training set and independent dataset (used as test set). For details, see the papers where the datasets are proposed.

Experimental results

In Table 1, we compare the wavelet-based descriptors using as classifier the AdaBoost of neural networks. In these tests, only the two-class datasets are used since AdaBoost is a two-class classifier⁸ and for reducing the computation time.

The method named WAVE is the weighted fusion, by weighted sum rule, among CWT+LBP-HF, with a weight of 1. PR-LPQ also has a weight of 1, and DW has a weight of 2.

The following fusion approaches⁹ are reported in Table 2:

- CHOU, the Chou's pseudo amino acid composition descriptor;
- WAVE+X*CHOU, the weighted sum rule where the weight of WAVE is 1 and the weight of CHOU is X.

Now in Tables 3 and 4, we report other results obtained using the support vector machine (SVM) as classifier.¹⁰ Only the methods that belong to WAVE are reported (to reduce the computation time).

Since SVM works better than AdaBoost of neural network, we have tested only SVM in the multi-class datasets (to reduce the computation time).

It would be interesting to consider as baseline CHOU since it is widely used for protein classification. The following conclusions can be drawn from the results reported in this section (Tables 5, 6):

- our experiments, show that there is not a “best” stand-alone “wavelet-based” method that clearly performs better than others in all the case studies, it is interesting to note that in the VIR1 dataset only DW works well (among the wavelet-based approaches), while in HEL dataset the method DW works poorly.
- better performance stability among different test sets is obtained by combining different methods, see the good performance of WAVE and the impressive results obtained combining CHOU and WAVE. This an experimental confirm that texture descriptors and the protein

⁶ Implemented as in DDtool 0.95 Matlab Toolbox.

⁷ It is performed 10 times and the average results are reported.

⁸ For a multi-class classification with a two-class classifiers the one-versus-one or one-versus-all approach should be used (Cristianini 2000).

⁹ Before the fusion the scores of each method are normalized to mean 0 and standard deviation 1.

¹⁰ We have tested both linear and Gaussian kernels, the parameters are estimated using a grid search in the training set.

Table 1 Comparison among the tested wavelet based descriptors using AdaBoost as classifier

	Datasets			
	HUM	HEL	GPCR	VIR1
<i>CW</i>				
DLTP	61.2	68.3	95.9	59.4
LBP-HF	61.3	88.1	96.0	65.6
LPQ	60.2	83.6	96.7	59.4
MSAhist	62.3	79.8	93.7	56.7
<i>PR</i>				
DLTP	64.0	77.6	93.1	64.3
LBP-HF	60.7	89.3	84.0	59.7
LPQ	62.5	90.6	86.3	57.2
MSAhist	61.5	70.9	93.7	74.5
<i>DW</i>				
DW	60.0	77.0	98.7	82.7
<i>WAVE</i>				
WAVE	63.1	86.4	98.8	75.5

The bold number represents the best result in a given dataset

Table 2 AU obtained by different methods in the different datasets

	Datasets			
	HUM	HEL	GPCR	VIR1
<i>ENSEMBLE</i>				
WAVE	63.1	86.4	98.8	75.5
WAVE+CHOU	67.9	91.1	99.6	77.7
WAVE+2*CHOU	70.7	93.3	99.7	78.6
WAVE+3*CHOU	72.0	94.0	99.8	78.9
WAVE+4*CHOU	72.7	94.3	99.8	79.0
WAVE+5*CHOU	72.9	94.3	99.8	78.9
<i>CHOU</i>				
CHOU	70.1	90.3	98.8	75.8

The bold number represents the best result in a given dataset

Table 3 Comparison among the tested wavelet based descriptors using SVM as classifier

	Datasets			
	HUM	HEL	GPCR	VIR1
<i>CW+LBP-HF</i>				
CW+LBP-HF	66.1	92.2	96.6	67.1
<i>PR+LPQ</i>				
PR+LPQ	62.8	90.8	87.8	57.5
<i>DW</i>				
DW	67.4	87.0	98.7	82.1
<i>WAVE</i>				
WAVE	67.5	91.6	99.0	75.5

The bold number represents the best result in a given dataset

descriptor based on amino acid sequence bring complementary information.

In order to give a statistical validation to the hypothesis that the fusion between CHOU and WAVE is the best approach the Wilcoxon Signed-Rank test (as suggested by

Table 4 Comparison among the tested wavelet based descriptors and CHOU

	Datasets			
	HUM	HEL	GPCR	VIR1
<i>ENSEMBLE</i>				
WAVE	67.5	91.6	99.0	76.6
WAVE+CHOU	71.2	94.5	99.7	77.6
WAVE+2*CHOU	73.2	95.6	99.7	78.5
WAVE+3*CHOU	74.3	95.8	99.7	78.8
WAVE+4*CHOU	74.4	95.6	99.7	78.8
WAVE+5*CHOU	74.6	95.6	99.7	78.8
<i>CHOU</i>				
CHOU	70.4	90.1	99.2	76.2

The bold number represents the best result in a given dataset

Table 5 Comparison, in the multi-class datasets, among the tested wavelet-based descriptors

	Datasets		
	GRAM	RAW	MEM
<i>CW+LBP-HF</i>			
CW+LBP-HF	78.45	84.47	82.18
<i>PR+LPQ</i>			
PR+LPQ	54.31	75.24	76.62
<i>DW</i>			
DW	81.47	86.11	86.17
<i>WAVE</i>			
WAVE	83.19	84.47	85.39

The bold number represents the best result in a given dataset

Table 6 Comparison, in the multi-class datasets, among the tested wavelet-based descriptors and CHOU

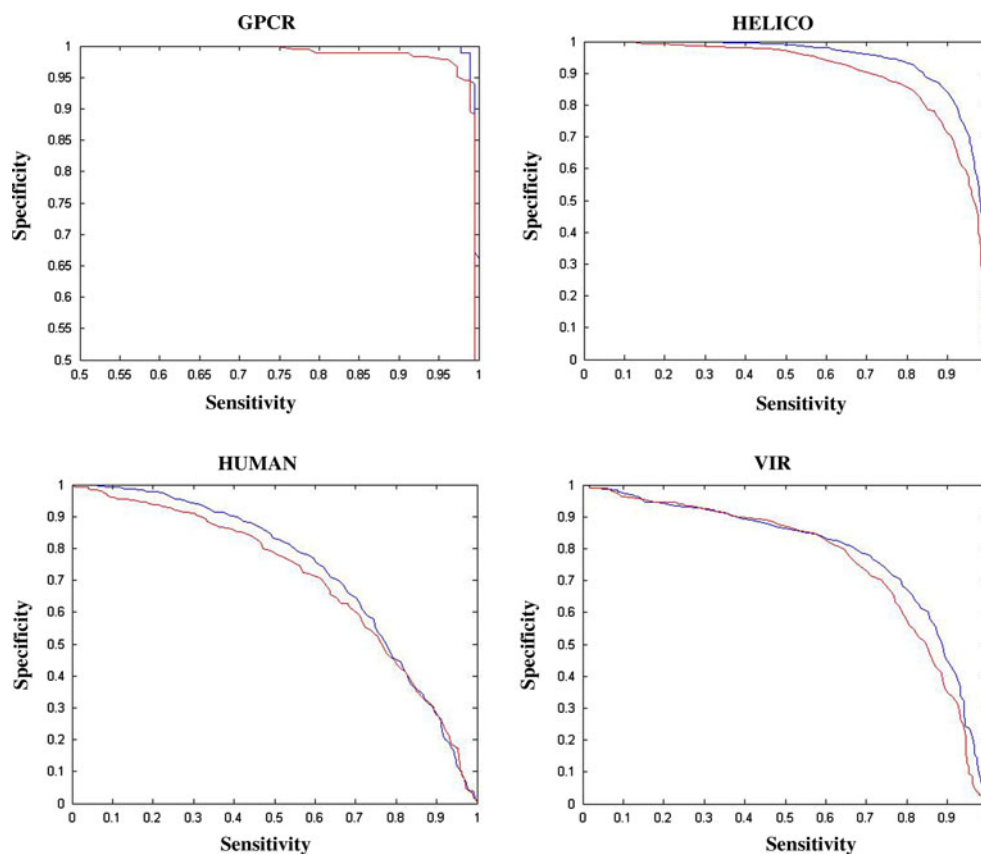
	Datasets		
	GRAM	RAW	MEM
<i>ENSEMBLE</i>			
WAVE	83.19	84.47	85.39
WAVE+CHOU	84.91	87.38	88.09
WAVE+2*CHOU	84.91	90.78	90.05
WAVE+3*CHOU	84.91	90.29	90.81
WAVE+4*CHOU	84.05	90.29	91.37
WAVE+5*CHOU	84.05	90.29	91.65
<i>CHOU</i>			
CHOU	81.47	88.83	89.08

The bold number represents the best result in a given dataset

Demsar 2006) is used to compare *WAVE+5*CHOU* with CHOU. The null hypothesis is that there is no difference between the accuracies of the two methods. The results are that the two approaches are statistically different (reject the null hypothesis, level of significance 0.05).

In Fig. 2, we report the specificity/sensitivity curve (for the 2-class datasets) obtained by CHOU (the red line) and our proposed method, *WAVE+5*CHOU* (the blue line).

Fig. 2 Comparison among the tested wavelet based descriptors



These plots confirm our previous conclusions that texture descriptors generalize well across different protein classification problems.

Conclusions

In this paper, we have presented an empirical study where different feature extraction approaches for representing proteins are compared and combined. We obtain a number of statistically robust observations regarding the behavior of the different texture descriptors here tested, the main conclusions that can be drawn from the results are:

- in each dataset, the best “wavelet approach” is different;
- combining different wavelet descriptors, it is possible to obtain an approach that works well in all the four datasets;
- it is clear that the texture descriptors and the protein descriptor based on amino acid sequence bring complementary information, this is experimentally confirmed by the fusion between the two approaches that outperform all the other methods.

Another way for further improving the performance of system used in our experiments is to use information from

different sources. For example, features can be extracted directly from an analysis of the protein's spatial structure (Daras et al. 2006). The position-specific scoring matrix (PSSM) (Ben-Gal et al. 2005) could also be considered. Finally, since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models or predictors (Chou and Shen 2009), we are currently working on providing a web-server for the method presented in this paper.

Acknowledgments We wish to thank Ojansivu and Heikkila for sharing their LPQ code; Rahtu, Salo and Heikkila for sharing their MSAhist code; Ahonen, Matas, He and Pietikäinen for sharing their LBP-HF code.

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ahonen T et al (2009) Rotation invariant image description with local binary pattern histogram Fourier features, *Image Analysis, SCIA 2009. Lect Notes Comp Sci* 5575:61–70
- Althaus IW et al (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J Biol Chem* 268:6119–6124
- Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without

- rate laws: new methods based on directed graphs. *Can J Chem* 86:342–357
- Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucl Acids Res* 28:45–48
- Ben-Gal I et al (2005) Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics* 21(11):2657–2666
- Bock J, Gough D (2003) Whole-proteome interaction mining. *Bioinformatics* 19:125–135
- Bulashevskaya A, Eils R (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinform* 7:298
- Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. *J Theor Biol* 248:377–381
- Chen L et al (2005) VFDB: a reference database for bacterial virulence factors. *Nucl Acids Res* 33:D325–D328
- Chen C et al (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Peptide Lett* 16:27–31
- Chou KC (1985) Low-frequency motions in protein molecules: beta-sheet and beta-barrel. *Biophys J* 48:289–297
- Chou KC (1988) Review: low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem* 30:3–48
- Chou KC (1989a) Graphic rules in steady and non-steady enzyme kinetics. *J Biol Chem* 264:12074–12079
- Chou KC (1989b) Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem Sci* 14:212
- Chou KC (1990) Review: applications of graph theory to enzyme kinetics and protein folding kinetics: steady and non-steady state systems. *Biophys Chem* 35:1–24
- Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct Funct Genet* 43:246–255
- Chou KC (2010) Graphic rule for drug metabolism systems. *Curr Drug Metab* 11:369–378
- Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J Theor Biol* 273:236–247
- Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. *Anal Biochem* 370:1–16
- Chou KC, Shen HB (2007b) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 360:339–345
- Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. *Nat Sci* 2:63–92. (openly accessible at <http://www.scirp.org/journal/NS/>)
- Chou KC, Shen HB (2010a) Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Sci* 2:1090–1103
- Chou KC, Shen HB (2010b) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE* 5:e11335
- Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. *Crit Rev Biochem Mol Biol* 30:275–349
- Chou KC, Kezdy FJ, Reusser F (1994) Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal Biochem* 221:217–230
- Chou KC, Zhang CT, Maggiora GM (1997) Disposition of amphiphilic helices in heteropolar environments. *Proteins Struct Funct Genet* 28:99–108
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
- Daras P et al (2006) Three-dimensional shape-structure comparison method for protein classification. *IEEE Trans Comput Biol Bioinform* 3(3):193–207
- Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognit Lett* 29:1887–1892
- Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Peptide Lett* 16:351–355
- Du PF, Li YD (2006) Prediction of protein submitochondria locations by hybridizing pseudoamino acid composition with various physicochemical. *BMC Bioinform* 7:518
- Du PF, Cao SJ, Li YD (2009a) SubChlo: predicting protein subchloroplast locations with pseudo- amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. *J Theor Biol* 261:330–335
- Du P, Cao S, Li Y (2009b) SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. *J Theor Biol* 261(2):330–335
- Fang Y et al (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. *Amino Acids* 34(1):103–109
- Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. HP Laboratories, Palo Alto
- Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinform* 9:62. doi:10.1186/1471-2105-9-62
- Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J Theor Biol* 271:10–17
- Hu L et al (2011) Predicting functions of proteins in mouse based on weighted protein–protein interaction network and protein hybrid properties. *PLoS ONE* 6:e14556
- Jaakkola T, Diekhans M, Haussler D (1999) Using the Fisher kernel method to detect remote protein homologies. In: Seventh international conference on intelligent systems for molecular biology. AAAI Press, Menlo Park, pp 149–158
- Jiang X et al (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Peptide Lett* 15:392–396
- Kandaswamy KK et al (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. *J Theor Biol* 270:56–62
- Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucl Acids Res* 20:1
- Lei Z, Dai Y (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinform* 6:291
- Leslie CS et al (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics* 20:467–476
- Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Peptide Lett* 15:612–616
- Liao S, Law MWK, Chung ACS (2009) Dominant local binary patterns for texture classification. *IEEE Trans Image Process* 18(5):1107–1118
- Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J Theor Biol* 252:350–356
- Lin MT, Beal MF (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* 443:787–795

- Lin H et al (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Peptide Lett* 15:739–744
- Lowell BB, Shulman GI (2005) Mitochondrial dysfunction and type 2 diabetes. *Science* 307:384–387
- Masso M, Vaisman II (2010) Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. *J Theor Biol* 266:560–568
- Mohabatar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Peptide Lett* 17:1207–1214
- Nanni L, Lumini A (2006) An ensemble of K-local hyperplane for predicting protein–protein interactions. *Bioinformatics* 22(10):1207–1210
- Nanni L, Lumini A (2008a) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* 34(4):653–660
- Nanni L, Lumini A (2008b) Genetic programming for creating Chou's pseudoamino acid based features for submitochondria localization. *Amino Acids* 34(4):653–660
- Nanni L, Lumini A (2010) A high performance set of descriptors extracted from the amino acid sequence for protein classification. *J Theor Biol* 266(1):1–10
- Niu B et al (2006) Predicting protein structural class with AdaBoost learner. *Protein Peptide Lett* 13:489–492
- Ojansivu V, Heikkilä J (2008) Blur insensitive texture classification using local phase quantization. In: *ICISP*
- Qin ZC (2006) ROC analysis for predictions made by probabilistic classifiers. In: *Fourth international conference on machine learning and cybernetics*, pp 3119–3124
- Qiu JD et al (2009) Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Anal Biochem* 390:68–73
- Rahtu E, Salo M, Heikkilä J (2005) Affine invariant pattern recognition using multi-scale autoconvolution. *IEEE Trans Pattern Anal Machine Intell* 27(6):908–918
- Saigo H et al (2004) Protein homology detection using string alignment kernels. *Bioinformatics* 20(11):1682–1689
- Shen H-B, Chou K-C (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng Design Select* 20:39–46
- Shi SP et al (2011) Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction. *Biochim Biophys Acta* 1813:424–430
- Tan X, Triggs B (2007) Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Analysis and modelling of faces and gestures*. In: *LNCS*, vol 4778, pp 168–182
- Wen ZN, Wang KL, Li ML, Nie FS, Yang Y (2005) Analyzing functional similarity of protein sequences with discrete wavelet transform. *Comput Biol Chem* 29:220–228
- Wolfram S (1984) Cellular automata as models of complexity. *Nature* 311:419–424
- Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. *Protein Peptide Lett* 14:871–875
- Xiao X et al (2005a) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. *J Theor Biol* 235:555–565
- Xiao X et al (2005b) Using cellular automata to generate Image representation for biological sequences. *Amino Acids* 28:29–35
- Xiao X, Shao SH, Chou KC (2006a) A probability cellular automaton model for hepatitis B viral infections. *Biochem Biophys Res Commun* 342:605–610
- Xiao X et al (2006b) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* 30:49–54
- Xiao X, Wang P, Chou KC (2009) GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J Comput Chem* 30(9):1414–1423
- Xiao X, Wang P, Chou KC (2011a) Quat-2L: a web-server for predicting protein quaternary structural attributes. *Mol Divers* 15:149–155
- Xiao X, Wang P, Chou KC (2011b) GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol Biosyst* 7:911–919
- Yang ZR, Thomson R (2005) Bio-basis function neural network for prediction of protease cleavage sites in proteins. *IEEE Trans Neural Netw* 16:263–274
- Zeng YH et al (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J Theor Biol* 259(2):366–372
- Zhou GP (2011) The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. *J Theor Biol* 284:142–148
- Zhou GP, Deng MH (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem J* 222:169–176
- Zhou XB et al (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J Theor Biol* 248:546–551