BIOINFORMATICS

Vol. 26 ECCB 2010, pages i467–i473
doi:10.1093/bioinformatics/btq371

Solenoid and non-solenoid protein recognition using stationary wavelet packet transform

An Vo[1,*,†], Nha Nguyen[2,†] and Heng Huang[3]

[1]The Feinstein Institute for Medical Research, North Shore LIJ Health System, NY, [2]Department of Electrical Engineering and [3]Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA

ABSTRACT

Motivation: Solenoid proteins are emerging as a protein class with properties intermediate between structured and intrinsically unstructured proteins. Containing repeating structural units, solenoid proteins are expected to share sequence similarities. However, in many cases, the sequence similarities are weak and non-detectable. Moreover, solenoids can be degenerated and widely vary in the number of units. So that it is difficult to detect them. Recently, several solenoid repeats detection methods have been proposed, such as self-alignment of the sequence, spectral analysis and discrete Fourier transform of sequence. Although these methods have shown good performance on certain data sets, they often fail to detect repeats with weak similarities. In this article, we propose a new approach to recognize solenoid repeats and non-solenoid proteins using stationary wavelet packet transform (SWPT). Our method associates with three advantages: (i) naturally representing five main factors of protein structure and properties by wavelet analysis technique; (ii) extracting novel wavelet features that can capture hidden components from solenoid sequence similarities and distinguish them from global proteins; (iii) obtaining statistics features that capture repeating motifs of solenoid proteins.

Results: Our method analyzes the characteristics of amino acid sequence in both spectral and temporal domains using SWPT. Both global and local information of proteins are captured by SWPT coefficients. We obtain and integrate wavelet-based features and statistics-based features of amino acid sequence to improve the classification task. Our proposed method is evaluated by comparing to state-of-the-art methods such as HHrepID and REPETITA. The experimental results show that our algorithm consistently outperforms them in areas under ROC curve. At the same false positive rate, the sensitivity of our WAVELET method is higher than other methods.

Availability: http://www.naaan.org/anvo/Software/Software.htm

Contact: anphuocnhu.vo@mavs.uta.edu

1 INTRODUCTION

With several interesting features of repeating sequences, significant progress has been made in the identification of the DNA and protein repeats, understanding the duplication mechanism and special features of the repeat evolution (Kajava, 2001). Repeats are usually found in non-coding genomic regions. However, repeating sequences are also found in about 14% of all proteins coded by all known genes with about 25% of all eukaryotic proteins (Marcotte *et al.*, 1999). The known protein structures can be classified by the length of their repeats, which can provide information about a possible 3D structure of the repetitive protein (Kajava, 2001). There are four main structural classes (Kajava, 2001): Class I, crystalline structures (1- to 2-residue repeats); Class II, fibrous proteins (3- to 4-residue repeats); Class III, solenoid proteins (5- to 42-residue repeats); and Class IV, domain-forming repeats (30 or more residues). Solenoid proteins contain a superhelical arrangement of repeating structure units (Kobe *et al.*, 2000). This arrangement contrasts the structure of most Class-IV proteins that fold into globular domains in more complex manners.

Repeats in Class I and Class II have only 1−4 residues, hence they have low sequence complexity and can be easily detected. Globular repeats in Class IV have their sufficient length to be detected by database search tools like PSI-BLAST (Altschul *et al.*, 1997). Solenoid proteins are built of repeated structural units. The repeating units of the solenoids consist of one to several segments of secondary structure, among which are α-helices (Kajava *et al.*, 2002), β-strands (Hennetin *et al.*, 2006) and 3_{10}-helices. The solenoid proteins have purely α-helices or β-strands or a mixture of the secondary structures (Kobe *et al.*, 2000). They are expected to share sequence similarities. However, in some cases, the sequence similarities are weak such as protein farnesyltransferase (FTase; Boguski *et al.*, 1992) and insulin-like growth factor-1 receptor (IGF-1R; Bajaja *et al.*, 1987), so that they are non-detectable (Kobe *et al.*, 2000). Therefore, database search tools like PSI-BLAST relying on clear conservation pattern are not good tools to detect solenoid repeats.

In recent years, several methods have been proposed to identify solenoid repeats. Some of them are based on self-alignment of the sequence such as REPRO (George *et al.*, 2000), RADAR (Heger *et al.*, 2000), TRUST (Szklarczyk *et al.*, 2004), HHrep (Soding *et al.*, 2006) and HHrepID (Biegert *et al.*, 2008). HHrep and HHrepID utilized hidden Markov model comparison (HMM–HMM), while the others used sequence–sequence comparison to find suboptimal self-alignments. Repeating parts of the sequence appear as off-diagonal regions of similarity. They allow the detection of basic repeating units and locations of units along the sequence. HHrepID has been reported to be the most sensitive self-alignment approach to detect repeats (Biegert *et al.*, 2008). However, HHrepID often cannot detect repeats with weak similarities.

Other approaches to recognize solenoid repeats use periodic patterns in proteins such as (Coward *et al.*, 1998), (Murray *et al.*, 2002), (Murray *et al.*, 2004), REPPER (Gruber *et al.*, 2005) and REPETITA (Marsella *et al.*, 2009). Repeating protein motifs, TIM barrels, propellor blades, coiled coils and leucine-rich repeating structures have been analyzed (Murray *et al.*, 2002) and used to detect repeats in known protein structure. The data utilized in Murray *et al.* (2002) are relative accessible surface area and simple hydrophobicity that provide information of the protein structure and

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

sequence. REPPER (Gruber *et al*., 2005) searches for periodicities of particular, user-defined types (hydrophobic, polar, positively charged) using the Fourier transform of sequence. However, it primarily aims at the analysis of fibrous proteins and does not allow insertions between repeating units. REPETITA (Marsella *et al*., 2009) utilizes sequence profile with the discrete Fourier transform to detect degenerated repeats (Lupas *et al*., 1997). It includes the five numeric scales proposed by Atchley (Atchley *et al*., 2005) to characterize the amino acid sequence. Compared to TRUST and RADAR, REPETITA has been reported to be more sensitive. However, similar to HHrepID, REPETITA also cannot detect some repeats with weak similarities.

Wavelet analysis has been widely applied to process biomedical signals (Unser *et al*., 1996). Some applications of wavelet transforms in genome sequence analysis and gene expression data analysis have been proposed (Li, 2003). A key advantage of wavelet transforms over the Fourier transform is their ability to simultaneously capture both spectral and temporal information within the signal (Daubechies, 1992, Mallat *et al*., 1989). In contrast, the Fourier transform does not give local information of proteins, because the Fourier coefficients only contain globally averaged information. Wavelet analysis has an improved ability to capture hidden components from biological data and is a good link between biological systems and the mathematics objects (Li, 2003). The wavelet transforms can be categorized as: continuous wavelet transform (CWT) and discrete wavelet transform (DWT). The CWT was used to detect and characterize repeating motifs in protein sequence and structure data (Murray *et al*., 2002). It maps a signal to a time-scale joint representation calculated by continuously shifting a continuously scalable function over a signal and calculating the correlation between them. The resulting wavelet coefficients are highly redundant. In molecular biology and genetics, we are more interested in discretely sampled rather than continuous functions. The DWT was used to classify protein subcellular location images (Chebira *et al*., 2007). Stationary wavelet packet transform (SWPT) is one of discrete wavelet analysis techniques. A main advantage of the stationary wavelets over the DWT is its shift invariant property (Coifman *et al*., 1995). The SWPT is suitable for many bioinformatics applications, such as DNA copy numbers smoothing and detection (Huang *et al*., 2008; Nguyen *et al*., 2010).

In this article, the SWPT technique is proposed to characterize proteins by five representation factors: polarity, secondary structure, molecular volume, codon diversity and electrostatic charge. We propose to extract new features from the SWPT of five factors and from statistics of amino acid sequence, and employ them to classify solenoid and non-solenoid proteins. Empirical studies on solenoid protein detection have been performed to compare proposed method to other related methods. Experimental results demonstrate the promising performance of our proposed approach.

2 METHODS

A flow diagram with all steps of our algorithm is given in Figure 1. The individual steps including wavelet analysis technique, feature extraction, feature selection, and classification will be described in following subsections.

2.1 The proposed method

The framework of proposed method is shown in Figure 1. At first, a protein sequence is translated into five numerical signals derived by Atchley (Atchley *et al*., 2005). These five signals represent polarity, secondary structure, molecular volume, codon diversity and electrostatic charge. All signals are normalized to zero mean, since the averages are not significant to detect repeats. Next, each signal is decomposed into 16 subbands by SWPT wavelet transform (see Section 2.2). Eighty subbands obtained from five signals are used to extract eighty wavelet features as in Equation (5). In addition to wavelet features, probability features of 20 amino acids are also computed for each protein sequence. Seventeen features are selected from total 100 features using the forward feature selection method. Finally, the quadratic discriminant classifier described in Section 2.5 is applied to classify protein sequences.

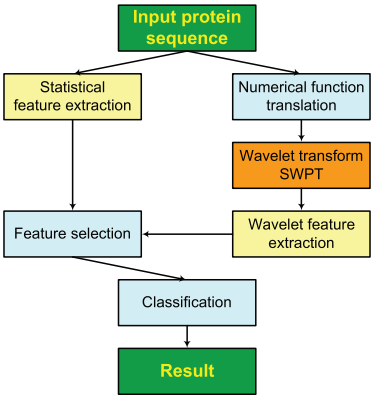

Fig. 1. The block diagram of WAVELET method for solenoid and non-solenoid classifications. Novel statistics-based and wavelet-based features are the key components of our proposed method.

Inspired by wavelet analysis for the protein structure and sequence (Murray *et al*., 2002), we propose to use the SWPT to analyze characteristics of amino acid sequence. However, there are two main differences between Murray's method and ours. The first difference is that the data utilized in Murray *et al.* (2002) is relative accessible surface area and simple hydrophobicity, while the data utilized in our method is five patterns that summarize a large portion of the physio-chemical and biological properties of amino acids (Atchley *et al*., 2005; Marsella *et al*., 2009). To obtain these five factors, 495 amino acid indices which include general attributes, such as molecular volume or size, hydrophobicity and charge, as well as more specific measures, such as the amount of nonbonded energy per atom or side chain orientation angle are analyzed. This analysis simplifies high-dimensional data by generating a smaller number of factors that would summarize the entire constellation of amino acid physiochemical properties. The resultant factors are linear functions of the original data, are fewer in number than the original, and reflect clusters of covarying traits that describe the underlying or latent structure of the variables. The first resultant factor is a polarity index which is bipolar, large positive and negative factor coefficients and also reflects simultaneous covariance in portion of polarity versus nonpolarity, and hydrophobicity versus hydrophilicity. Therefore, the hydrophobicity feature used in the previous study is also involved in the first factor. The second difference is that Murray (Murray *et al*., 2002) used the continuous wavelets which are highly redundant, but we are more concerned with discrete sampling rather than continuous functions for protein sequence representation. Moreover, since some amino acids play important roles in structure of solenoids, we propose statistics features to capture repeating motifs of solenoids.

2.2 Stationary wavelet packet transform

The SWPT is based on filters H_1 and G_1 and on an up-sampling operator. The filter H_1 is a low-pass filter defined by a sequence $h_1(n)$ and the high-pass filter G_1 defined by a sequence $g_1(n)$. Given a signal $s(n)$ of length N_s, the first level of the SWPT produces two subbands: approximation subband $s_{11}(n)$

and detail subband $s_{12}(n)$. These subbands are obtained by convolving $s(n)$ with the low-pass filter $h_1(n)$, and with the high-pass filter $g_1(n)$ as follows:

$$s_{11}(n)=\sum_k h_1(n-k)s(k),\ s_{12}(n)=\sum_k g_1(n-k)s(k). \quad (1)$$

Sequences $h_1(n)$ and $g_1(n)$ are obtained by using MATLAB function *wfilters* where n is an integer. An example of $h_1(n)$ and $g_1(n)$ is given by $h_1(n)=\{-0.1294, 0.2241, 0.8365, 0.4830\}$ and $g_1(n)=\{-0.4830, 0.8365, -0.2241, -0.1294\}$ when *db2* wavelets are used. Values of k in the above summation are from 1 to N_s, where N_s is the length of $s(n)$. Two new subbands s_{11} and s_{12} have the same length as the original signal $s(n)$. The low-pass filter H_1 is assumed to satisfy the internal orthogonal relation as

$$\sum_n h_1(n)h_1(n+2i)=0, \quad (2)$$

for all integers $i\neq 0$ and $\sum_n h_1^2(n)=1$. The high-pass filter G_1 is defined by

$$g_1(n)=(-1)^n h_1(1-n). \quad (3)$$

The high-pass filter G_1 satisfies the same internal orthogonal relation as H_1 and the mutual orthogonal relation as

$$\sum_n h_1(n)g_1(n+2i)=0, \quad (4)$$

for all integers i. For further details of these filters, please see Daubechies (1992) and Mallat *et al*. (1989).

The filters are modified at each level by an up-sampling operator. The filter $h_l(n)$ at level l is obtained by inserting a zero between every adjacent pair of elements of the filter $h_{l-1}(n)$ at level $(l-1)$, and similarly for filter $g_l(n)$ (Nason *et al*., 1995). Each filter is an upsampled version of the previous one. The second level of the SWPT produces four subbands: $s_{21}(n), \cdots, s_{24}(n)$. These subbands are obtained by convolving $s_{11}(n)$ and $s_{12}(n)$ with the filters $h_2(n)$ and $g_2(n)$. The process is iterated until an expected level l is reached. The number of subbands at level l is equal to 2^l. For example, in order to obtain the SWPT with 16 subbands, we select $l=4$ because $2^4=16$. We calculate $h_2(n)$ and $g_2(n)$ by inserting one zero between samples of $h_1(n)$ and $g_1(n)$. Similarly, we compute $h_3(n)$, $g_3(n)$ and $h_4(n)$, $g_4(n)$ by inserting one zero between samples of $h_2(n)$, $g_2(n)$ and $h_3(n)$, $g_3(n)$.

The SWPT does not employ a decimator after filtering. The absence of a decimator leads to a full rate decomposition. Each subband contains the same number of samples as the input. The absence of a decimator makes the SWPT shift invariant. This shift invariant property provides the SWPT preferable for the usage in various signal processing applications such as smoothing (Huang *et al*., 2008; Nguyen *et al*., 2010) and classification because they capture more spatial information. It has been shown that many of the artifacts could be suppressed by a redundant representation of the signal (Coifman *et al*., 1995). The SWPT offers a richer range of possibilities for signal analysis. Thus, we propose to use the SWPT to analyze characteristics of proteins. In our experiments, a signal was decomposed into 16 subbands corresponding to with four-level SWPT. The numbers of subbands and filter type were chosen to give the best results in classification performance. We applied several l-level SWPTs with $l=2,3,4,5$ and several types of filters. The four-level SWPT and Coif5 filter produce the best results. From wavelet subbands how to generate features for solenoid classifier will be described in the next subsection.

2.3 Feature extractions

2.3.1 Wavelet feature extraction A solenoid protein comprises repeating structural units that arrange to let the polypeptide chain form a continuous superhelix (Kobe *et al*., 2000). It is expected to share sequence similarities. However, in some cases, the sequence similarities are weak, so that they are non-detectable. The knowledge of the structure can help the detections of weak sequence similarities and patterns among the structure units. The set of five factors derived by Atchley was used in analysis directed toward understanding the structural and functional aspects of protein variability. Factor II presents the secondary structure. There is an inverse relationship of relative propensity for various amino acids in various secondary structural configurations, such as a coil, a turn, or a bend versus the frequency in an α-helix (Atchley *et al*., 2005).

Instead of using the discrete Fourier transform (DFT; Marsella *et al*., 2009) for five factors combined with sequence profiles, we use the SWPT to analyze these five factors. It should be noted that the Fourier transform result is a global measure. Therefore, the period obtained from DFT method is a global period which could be invisible in many solenoids. In contrast, the SWPT analysis can capture hidden components from sequence similarities.

Each signal (factor) is decomposed into 16 subbands by SWPT. The energy of each subband is calculated by

$$E_k(s)=\frac{1}{N_s}\sum_{i=1}^{N_s}|s_k(i)|, \quad (5)$$

where $k=1,2,\cdots,16, s_k(i)$ are the k-th subband coefficients and N_s is the number of coefficients.

Each sequence is presented by five factors. As a result, we obtain 80 wavelet-based features for each protein sequence. The wavelet-based features of factor II extract the knowledge of the structure that can help weak sequence similarities detection. The wavelet-based features of other factors extract physio-chemical and biological properties of amino-acid sequence. Besides wavelet features, we also propose to use statistical information of amino acids in protein sequences for generating protein features.

2.3.2 Statistical feature extraction Solenoid proteins contain a superhelical arrangement of repeating structural units. The most common arrangement of the solenoid proteins is a single-stranded superhelix which is formed by one or several elements of secondary structure (such as α-helices, β-strands and 3_{10}-helices) winding along the superhelical axis. Each repeat contains at least one turn with an irregular conformation introduced between the secondary structures. The minimal structural unit corresponding to one repeat has one secondary structure element and one turn. Several representative structural repeating motifs are summarized below, where the asterisk denotes a polar residue, o denotes nonpolar residues, and x is any residue.

- A-(N/D)-L-*-x: the pentapeptide repeat protein (Bateman *et al*., 1998); S-x-(V/I)-x-G: the pentapeptide repeat of anti-freeze protein (Graether *et al*., 2000) contains one β-strand and one turn. Pentapeptide repeat proteins (PRPs) are found primarily in bacteria, especially cyanobacteria. Although the structure of the cyanobacterial pentapeptide proteins is not yet available, PRPs have important biochemical and physiological functions in cyanobacteria. The newly uncovered structural features may help scientists discover the biological role of pentapeptide repeat proteins within the cell.
- SxIGxx: the hexapeptide repeat of LpxA (Raetz *et al*., 1995) contains one β-strand and one turn. In hexapeptide, the regular superhelical surface of the β-helices might prefer to satisfy the protein-interacting demand by forming homo-oligomers (Kobe *et al*., 2000).
- DxLxGGxGx: the nine-residue repeat of serralysin (Baumann *et al*., 1993) contains one β-strand and one turn shown in Figure 2c.
- xxxGoxxLxxoLxxxxxLxxLxLxxNxL: the LRR of ribonuclease inhibitor (RI); LxxLxLxxN/CxL: the 11-residue repeat of the LRRs contains a β-strand and an α-helix (Kajava, 1998) as shown in Figure 2a. LRR proteins appear in different proteins with diverse functions. They bind a number of globular proteins by their concave surface. In the RI molecule, structural units consisting of a β-strand and an α-helix are arranged so that all the β-strands and the helices are parallel to a common axis, resulting in a nonglobular, horseshoe-shaped molecule with a curved parallel β-sheet lining the inner circumference of the horseshoe and the helices flanking its outer circumference (Kajava, 2001).

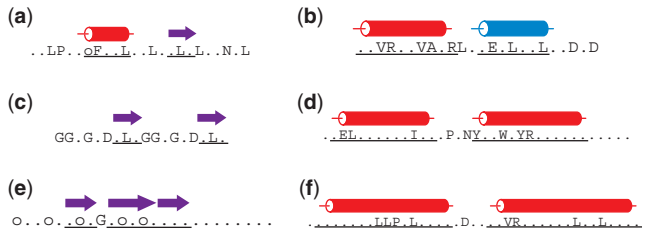

Fig. 2. Examples of consensus motifs in solenoid proteins. (**a**) Leucine-rich repeat (LRR), (**b**) Leucine-rich repeat variant (LRV), (**c**) Serralysin, (**d**) Protein FTase, (**e**) Insulin-like growth factor-1 receptor L domain (IGF-1R L) and (**f**) HEAT. Positions of α-helices (red cylinders), 3_{10}-helices (blue cylinders) and β-strands (magenta arrows) are shown in the above underlined portions of the sequences, and o denotes hydrophobic amino acid (Kobe *et al*., 2000).

- xxxxxxxxLLPxLxxxxxDxxxxVRxxxxxxLxxLxxxx: the HEAT repeat of PR65/A subunit (Groves *et al*., 1999). The HEAT repeats (Andrade *et al*., 1995) have two anti-parallel α-helices and make one complete coil shown in Figure 2f. They contain nuclear import factor importin-β used when conformational adjustment to the binding partners are needed (Kobe *et al*., 2000).
- Leucine-rich repeat variant (LRV) (Peters *et al*., 1996) contains an α-helix and 3_10-helix as shown in Figure 2b.
- Protein FTase α-subunit (Boguski *et al*., 1992) is shown in Figure 2d.
- Insulin-like growth factor-1 receptor L domain (IGF-1R L) (Bajaja *et al*., 1987) contains β-strands as shown in Figure 2e. The sequence similarities in IFG-1R and FTase are weak.

From above structural repeating motifs, some amino acids play key roles in structure of solenoids. Therefore, we propose using amino acid compositions as additional features to capture repeating motifs of solenoid proteins as follows

$$p(X) = \frac{N_X}{N_s}, \tag{6}$$

where X = [A, C, D, E, F, ⋯, V, W, Y] is the one-letter code corresponding to each of the 20 amino acids, and N_X is the number of amino acids X in a sequence, and N_s is the length of that sequence. The 80 wavelet-based features and 20 statistics-based features are combined, resulting a total of 100 features for each protein sequence. We can use all 100 features as inputs of a classifier. However, because the number of features is larger than the number of classes, one of the ways to select features is using forward selection.

2.4 Feature selection

The forward feature selection procedure (Kohavia *et al*., 1997) starts by evaluating all feature subsets, which consist of only one feature. We obtain 100 models at this point and the model that performs the best is chosen. The feature subset now contains one feature. Next, forward selection finds the best subset consisting of two features. Ninenty nine feature subsets are made by pairing this chosen feature with all the remaining 99 features, one by one. For 99 models, their statistics are compared to select the best performing model. The feature subset now contains two features. These iterations are continued until the best classification accuracy is obtained.

As a result, 17 features which include eight wavelet-based features and nine statistics-based features are selected from a total of 100 features using the forward feature selection method. Tables 1 and 2 show more details of selected features. All five factors which describe structure and properties of protein are present in selected feature subset at different subbands as in Table 1. The first row shows feature number (1–100). The first 80 features are from wavelet-based features and the last 20 features are from amino acid compositions. The second and third rows present factor (I–IV) and subband number (1–16). Statistical features of amino acids in sequence are also selected in Table 2. The selected feature corresponding to the occurrence frequency of amino acid *L* in sequences makes more sense with structural repeating motifs summarized in the previous subsection. After selecting a set of good features, we use them as inputs of a classifier described in the next subsection.

Table 1. Selected wavelet features from a total of 100 features using the forward feature selection method

Feature	35	49	27	41	10	68	46	51
Factor	V	IV	II	I	V	III	I	I
Subband	7	9	6	9	2	14	10	11

All five factors from I to IV at different subbands corresponding to solenoid properties and structure are selected.

Table 2. Selected statistics features from a total of 100 features using the forward feature selection method

Feature	90	92	99	94	82	89	93	97	88
Amino acid	L	N	W	Q	C	K	P	T	I

The occurrence frequency of amino acid *L* in sequences is the first choice to capture structural repeating motifs.

2.5 Classification

In our experiment, we use a quadratic discriminant analysis classifier (QDA; Krzanowski, 1988), which is a standard supervised classification approach. QDA models the likelihood of each class as a Gaussian distribution. The posterior distributions are used to estimate the class for a given test point. We can use maximum likelihood estimation algorithm to estimate the Gaussian parameters for each class from training points. A MATLAB function, *classify.m* in Statistics toolbox is used to perform QDA classification. A setting for this function is *quadratic* type. Feature vectors are normalized to have a unit variance before classification step. Some other complex classification methods may also be applied to further improve the classification results. However, this is beyond the scope of the paper and might smear the effect of the proposed features.

3 EXPERIMENTS AND DISCUSSIONS

3.1 Data sets

Two groups of data are studied: (i) 105 solenoid repeat proteins; (ii) 247 globular proteins (non-solenoid) without structural repeat. These data are downloaded from the website (http://protein.bio.unipd.it/repetita/) of the previous study (Marsella *et al*., 2009). Marsella took an initial set of 32 proteins with solenoid repeats and used the TESE server (Sirocco, 2008) to find more protein domains belonging to the same solenoid folds as the initial set. By limiting the maximal residual structural similarity according to the CATH classification (Pearl *et al*., 2003), TESE allows the user to generate ad hoc non-redundant sets of proteins with known structure. The final set of 105 solenoid domains was yielded by choosing representatives with at most 35% pairwise sequence identity (i.e. CATH 'S' level). Marsella also generated the set of 247 non-solenoid protein domains with TESE by randomly choosing X-ray structures with different topologies and no detectable sequence similarity (i.e. CATH 'T' level). A training set of 50 solenoid proteins and 119 non-solenoid, and a testing set

of 55 solenoid proteins and 128 non-solenoid ones are selected as in Marsella *et al*. (2009). The solenoid proteins contain the main repeat classes such as all α, all β, and α/β with available structure information or they have their structures and evolutions related to these major folds.

Sequences are composed of long strings of alphabetic letters rather than arrays of numerical values. A metric for comparing such alphabetic data is sophisticated. Therefore, a method proposed by Atchley (Atchley *et al*., 2005) to quantify alphabetic information inherent to biological sequences was applied. Five patterns that summarize a large portion of the physio-chemical and biological properties of amino acids were obtained. These five factors represent polarity, secondary structure, molecular volume, codon diversity and electrostatic charge.

3.2 Comparisons to existing methods

We compare the classification performance of our WAVELET method against four exiting methods in protein repeats detection: RADAR (Heger *et al*., 2000), TRUST (Szklarczyk *et al*., 2004), HHrepID (Biegert *et al*., 2008) and REPETITA (Marsella *et al*., 2009). Sensitivity, specificity and accuracy of all methods are computed for training set, testing set and overall data.

RADAR and TRUST detect internal sequence symmetries by comparing the protein sequence itself and utilize sequence–sequence comparison to find suboptimal. RADAR builds a repeat profile to determine exact borders and extract a multiple alignment of repeats self-alignments, and TRUST explicitly makes use of consistency that has led to improvements in multiple sequence alignment. In TRUST and RADAR methods, predictions are considered when at least two repeat units are detected.

HHrepID utilizes hidden Markov model comparison. The maximum expected accuracy algorithm that maximizes the sum of posterior probabilities in the alignment and a probabilistic approach to consistency through a merging procedure based on posterior probabilities are also applied. HHrepID has been reported to be most sensitive to date (Biegert *et al*., 2008). We use default settings for HHrepID method. The MAC threshold is set to 0.5 and the *P*-value threshold for suboptimal alignment is set to 0.1.

REPETITA detects solenoid repeats and discriminates them from globular proteins using information from sequence profiles together with the discrete Fourier transform, based on the assumption that few characteristics of sequence repeats uniquely identify structural repeats.

3.3 Results and discussions

We summarize experimental results including sensitivity, specificity and accuracy in Table 3. Data sets used for evaluations are the training set, testing set and overall data. We evaluate both training set and testing set to verify that our predictive model does not overfit the training data. Because our model fits both training set and test set well with the accuracy of 94.1% and 91.3%, respectively, non-overfitting has taken place. Table 3 shows the optimal results of each method. The sensitivity and accuracy of our WAVELET method are consistently higher than the HHrepID and REPETITA methods for all three metrics. In overall data, the WAVELET method yields the best sensitivity of 93.3%, while the sensitivity of HHrepID and REPETITA methods are 66.7% and 70.0%, respectively.

Table 3. Comparisons of HHrepID, REPETITA and WAVELET methods for solenoid detection

	Method	Training (%)	Test (%)	Overall (%)
Sensitivity	HHrepID	70.0	63.6	66.7
	REPETITA	70.0	69.0	70.0
	WAVELET	**96.0**	**90.9**	**93.3**
Specificity	HHrepID	93.3	89.8	91.5
	REPETITA	85.0	83.0	84.0
	WAVELET	**93.3**	**91.4**	**92.3**
Accuracy	HHrepID	86.4	82.0	84.1
	REPETITA	80.5	78.7	79.6
	WAVELET	**94.1**	**91.3**	**92.6**

Sensitivity, specificity, and accuracy of each method using training set, testing set and overall data are computed. Both training and testing data are evaluated to verify that our predictive model does not overfit the training data. Bold values represent the proposed method.

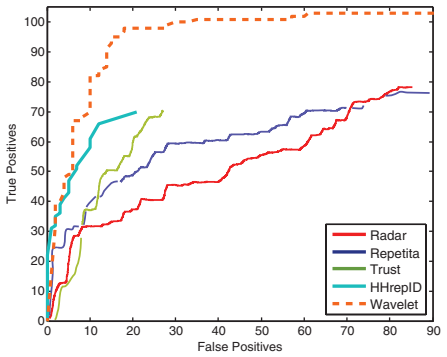

Fig. 3. Comparisons of RADAR, TRUST, HHrepID, REPETITA and WAVELET methods for solenoid detection: the number of true positive solenoid proteins against the number of false positive proteins detected above a threshold significance when overall data are used.

We evaluate all methods using ROC curve as shown in Figure 3. The number of true positive solenoid proteins against the number of false positive proteins detected above a threshold significance is computed. In the case of HHrepID method, we used the total repeat *P*-value for the threshold significance. The signed distance from the optimal line is used for significance measure in REPETITA method. For RADAR and TRUST methods, the number of repeat units is used. Predictions are considered where at least two repeat units have been detected. In our method, we use the posterior probabilities obtained from the QDA classifier for threshold significance. The performances of all methods in terms of areas under the ROC curve are shown in Table 4 when the false positive ranges from 0 to 20. The performance of WAVELET is better than other methods from 21% to 46% in areas under the ROC curve.

At a false positive rate (FPR) of 8% (20/247), WAVELET method is able to detect about 60% more solenoid proteins than RADAR and about 50% more than REPETITA. TRUST and HHrepID can detect better than RADAR and REPETITA. However, WAVELET method performs about 37% better than TRUST and 29% better than HHrepID in sensitivity at FPR of 8%. When the FPR is from

Table 4. Performance of all methods in terms of areas under ROC curve in overall data when false positives are from 0 to 20

Method	RADAR	TRUST	REPETITA	HHrepID	**WAVELET**
Area	798	1009	1057	1168	**1475**

Bold values represent the proposed method.

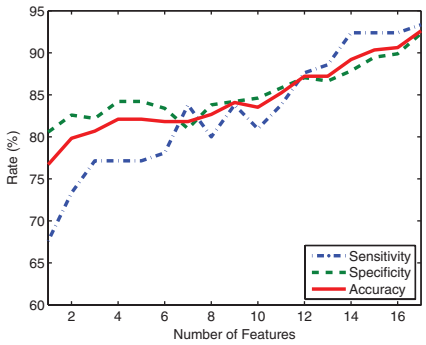

Fig. 4. Sensitivity, specificity and accuracy of WAVELET method for solenoid and non-solenoid classification against the number of selected features. Overall data are used.

3% to 8%, WAVELET also outperforms the others. The solenoid proteins detected using WAVELET is close to using HHrepID, when FPR is 1–3%. But they are much more than RADAR, TRUST and REPETITA. Compared to RADAR and TRUST in detecting solenoids, HHrepID is an identifier with higher sensitivity. This agrees with results reported in Biegert *et al.* (2008) that HHrepID is most sensitive to date. When FPR is <1%, the sensitivity of HHrepID method is higher than the others. However, the highest sensitivity of HHrepID in this simulation is about 67% for overall data, while that of WAVELET method can reach 93% at the same FDR of 8%. Because the structure information using wavelet feature is combined with statistical information in our WAVELET method, a higher classification accuracy was achieved.

Figure 4 shows the overall performances of WAVELET when the number of features changes from 1 to 17. The WAVELET's performance almost increases when the number of features increases. The accuracy of the WAVELET ranges from 77% to 93% corresponding to 1 to 17 features. In REPETITA method, only two features are used. A small number of features often cannot capture enough biological information for detection. Therefore, there are only 70% solenoid data detected by REPETITA method, while WAVELET method with 17 features can detect 93% solenoid data.

We also show some examples of RADAR, TRUST, REPETITA, HHrepID and WAVELET methods for solenoid protein detection in Table 5. These solenoid proteins are plotted in Figure 5. Three of six sample solenoids are non-detectable by RADAR and TRUST methods. REPETITA can detect four solenoids, WAVELET can detects all of six solenoids, while HHrepID cannot find any repeats in these solenoid sequences. All above simulation results illustrate that WAVELET method outperforms the others.

Table 5. Examples of solenoid detection using RADAR, TRUST, REPETITA, HHrepID and WAVELET methods

CATH Domain	RADAR	TRUST	REPETITA	HHrepID	WAVELET
1p5qA02	True	False	True	False	True
1xat000	False	True	False	False	True
1ee6A00	False	True	False	False	True
1ho8A01	True	False	True	False	True
2a4zA03	False	True	True	False	True
1tdtA02	True	False	True	False	True

All sequences shown in this table are solenoids. 'True' is a correct detection and 'False' is a wrong detection.

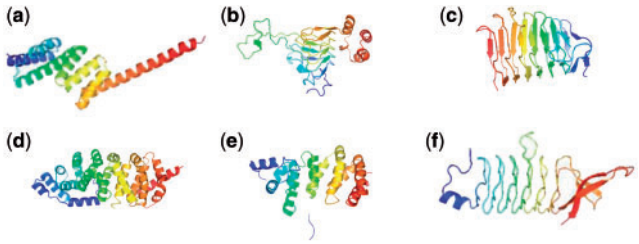

Fig. 5. Examples of solenoid proteins: (**a**) 1p5qA02, (**b**) 1xat000, (**c**) 1ee6A00, (**d**) 1ho8A01, (**e**) 2a4zA03 and (**f**) 1tdtA02. Detection results of these solenoids using RADAR, TRUST, REPETITA, HHrepID and WAVELET methods are shown in Table 5. Rainbow coloring from blue to red shows the topology from *N* to *C* terminus. Cartoon representations of these sample solenoid proteins are available in previous study (Marsella *et al*., 2009).

4 CONCLUSION

In this article, we proposed a new WAVELET method to recognize solenoid proteins and global proteins using SWPT and statistical features of amino acid sequences. In order to detect solenoid repeats with weak similarities, we took advantages of the integration of wavelet-statistics features and the SWPT analysis of five factors representing protein structure and properties. Our new features can capture structure, properties of solenoid proteins and hidden components from sequence similarities, to distinguish them from global proteins. The proposed approach was validated by comparing to other state-of-the-art methods in solenoid proteins detection experiments. In all results, our new scheme improved the solenoid protein recognition in all statistical metrics, including sensitivity, specificity and accuracy. The WAVELET method is a promising approach for solenoid protein classification. Based on different types of training data, the WAVELET method can be applied to classify different kinds of solenoids or different kinds of protein structures in future work.

Funding: NSF-CCF 0830780; NSF-CCF 0939187; NSF-CCF 0917274; NSF-DMS 0915228; NSF-CNS 0923494; University of Texas Arlington Research Enhancement Program.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,M.A. *et al*. (1995) HEAT repeats in the Huntington's disease protein. *Nat. Genet.*, **11**, 115—116.

Atchley,W. *et al*. (2005) Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA*, **102**, 6395–6400.

Bajaja,M. *et al*. (1987) On the tertiary structure of the extracellular domains of the epidermal growth factor and insulin receptors. *Biochim. Biophys. Acta.*, **916**, 220–226.

Bateman,A. *et al*. (1998) Structure and distribution of pentapeptide repeats in bacteria. *Protein Sci.*, **7**, 1477—1480.

Baumann,U. *et al*. (1993) Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif. *EMBO J.*, **12**, 3357–3364.

Biegert,A. *et al*. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.

Boguski,M. *et al*. (1992) Novel repetitive sequence motifs in the alpha and beta subunits of prenyl-protein transferases and homology of the alpha subunit to the MAD2 gene product in yeast. *New Biol.*, **4**, 408–411.

Chebira,A. *et al*. (2007) A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, **8**, 1–10.

Coifman,R.R. *et al*. (1995) Translation invariant de-noising. *Lecture Notes Stat.*, **103**, 125–150.

Coward,E. *et al*. (1998) Detecting periodic patterns in biological sequences. *Bioinformatics*, **14**, 498–507.

Daubechies,I. (1992) Chapter 3. In *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Society for Industrial and Applied Mathematics (SIAM), pp. 53–106.

George,R.A. *et al*. (2000) The REPRO server: finding protein internal sequence repeats through the web. *Trends Biochem. Sci.*, **25**, 515–517.

Graether,S.P. *et al*. (2000) Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature*, **406**, 325–328.

Groves,M.R. *et al*. (1999) Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.*, **9**, 383–389.

Gruber,M. *et al*. (2005) REPPER-repeats and their periodicities in fibrous proteins. *Nucleic Acids Res.*, **33**, W239–W243.

Heger,A. *et al*. (2000) Rapid automatic detection and alignment of repeats in protein sequence. *Proteins Struct. Funct. Genet.*, **41**, 224–237.

Hennetin,J. *et al*. (2006) Standard conformations of beta-arches in beta-solenoid protein. *J. Mol. Biol.*, **358**, 1094–1105.

Huang,H. *et al*. (2008) Array CGH data modeling and smoothing in stationary wavelet packet transform domain. *BMC Genomics*, **9**, S2–S17.

Kajava,A. *et al*. (2002) What curve alpha-solenoid? *J. Biol. Chem.*, **227**, 49791–49798.

Kajava,A.V. (1998) Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.*, **277**, 519–527.

Kajava,A.V. (2001) Review: proteins with repeated sequence-structural prediction and modeling. *J. Struct. Biol.*, **134**, 132–144 .

Kobe,B. *et al*. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends in Biochemical. Sci.*, **25**, 509–515.

Kohavia,R. *et al*. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273–324.

Krzanowski,W.J. (1988) Chapter 12. In *Principles of Multivariate Analysis: A User's Perspective (Oxford Statistical Science Series)*. Oxford University Press, p. 340.

Li,P. (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, **19**, 2–9.

Lupas,A. *et al*. (1997) Self-compartmentalizing proteases. *Trends Biochem. Sci.*, **22**, 399–404.

Mallat,S. *et al*. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Pattern Anal. Mach. Intell.*, **11**, 674–693.

Marcotte,E.M. *et al*. (1999) Census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.

Marsella,L. *et al*. (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by Fourier transform. *Bioinformatics*, **25**, i289–i295.

Murray,K.B. *et al*. (2002) Wavelet transforms for the characterization and detection of repeating motifs. *J. Mol. Biol.*, **316**, 341–363.

Murray,K.B. *et al*. (2004) Toward the detection and validation of repeats in protein structure. *Proteins*, **57**, 365–380.

Nason,G.P. *et al*. (1995) The stationary wavelet transform and some statistical applications. *Lecture Notes Stat.*, **103**, 281–299.

Nguyen,N. *et al*. (2010) Stationary wavelet packet transform and dependent Laplacian bivariate shrinkage estimator for array-CGH data smoothing. *J. Comput. Biol.*, **17**, 139–152.

Pearl,F.M. *et al*. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.

Peters,J.W. *et al*. (1996) A leucine-rich repeat variant with a novel repetitive protein structural motif. *Struct. Biol.*, **3**, 1181–1187.

Raetz,C.R. *et al*. (1995) A left-handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase. *Science*, **270**, 997–1000.

Sirocco,F. *et al*. (2008) TESE: generating specific protein structure test set ensembles. *Bioinformatics*, **24**, 2632–2633.

Soding,J. *et al*. (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.*, **34**, W137–W142.

Szklarczyk,R. *et al*. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**, i311–i317.

Unser,M. *et al*. (1996) A review of wavelets in biomedical applications. *Proc. IEEE*, **84**, 626-638.