Computational Biology and Chemistry 92 (2021) 107450

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Review Article

Prediction of protein structural class based on symmetrical recurrence quantification analysis

Ines Abdennaji[a,b,*], Mourad Zaied[a,b], Jean-Marc Girault[a,b]

[a] Research Team in Intelligent Machines, National School of Engineers of Gabes, B.P. W, 6072 Gabes, Tunisia
[b] GSII ESEO – LAUM UMR CNRS 6613, 49000 Angers, France

ARTICLE INFO

Keywords:
Protein structural classes
Symmetry
Symmetrical recurrence quantification analysis
Recurrence plot
Machine learning
SVM
LDA
Random Forest

ABSTRACT

Protein structural class prediction for low similarity sequences is a significant challenge and one of the deeply explored subjects. This plays an important role in drug design, folding recognition of protein, functional analysis and several other biology applications. In this paper, we worked with two benchmark databases existing in the literature (1) 25PDB and (2) 1189 to apply our proposed method for predicting protein structural class. Initially, we transformed protein sequences into DNA sequences and then into binary sequences. Furthermore, we applied symmetrical recurrence quantification analysis (the new approach), where we got 8 features from each symmetry plot computation. Moreover, the machine learning algorithms such as Linear Discriminant Analysis (LDA), Random Forest (RF) and Support Vector Machine (SVM) are used. In addition, comparison was made to find the best classifier for protein structural class prediction. Results show that symmetrical recurrence quantification as feature extraction method with RF classifier outperformed existing methods with an overall accuracy of 100% without overfitting.

1. Introduction

Today, the structural classes in four levels (quaternary, ternary, secondary and primary) play a significant role in theoretical and experimental studies of protein science. The protein quaternary and the tertiary structures are determined via the process of protein folding. Protein secondary structure is the three-dimensional form of local segments of proteins whose amino acids linear sequence (in a peptide or protein) forms the protein primary structure. As mentioned by Chou et Zhang in 1995 (Chou and Zhang, 1995), it is important and helpful to predict higher proteinic classes from primary proteinic sequences for two reasons. Firstly, if the structural class of the protein under study is known then the searching scope of conformation can be reduced (Bahar et al., 1997). Secondly, the structural class is related to various protein properties (Nishikawa and Ooi, 1982). Since there is no simple and direct way for the protein tertiary structure prediction from its primary structure, four secondary structural classes of proteins based on the types and arrangement of their secondary structural class are proposed by Levitt and Chothia (1976). These classes are the α, the β and those with a mixture of α and β shapes called the α/β and the $\alpha+\beta$.

These four protein structural classes can be used to (1) implement a heuristic method for deciding tertiary structure (Carlacci et al., 1991), (2) reduce search space of probable conformations of tertiary structure (Chou, 1992; Bahar et al., 1997), (3) improve prediction of secondary structure accuracy and (4) predict function from amino acid sequence information. Protein structural class prediction plays an essential role in functional analysis, protein structures, drug designs and a lot of other similar applications in biology (Gromiha and Selvaraj, 1998).

For the last 10 years, prediction of protein structural class for low similarity sequences (Kurgan and Homaeian, 2006; Chou, 2005) is a tough challenge for the scientific community. Therefore, an automated and accurate protein structural class prediction for newly established proteins is required. In order to extract the feature sequences from protein, various feature extraction techniques are used in the recent studies which can be later useful for classification of the structural classes. Most of used techniques include amino acid composition (AAC) (Chou, 1995; Kim et al., 2006; Cai et al., 2002), average chemical shift (ACS) (Fias et al., 2008; Lin et al., 2012; Feng et al., 2016), pseudo amino acid (PSeAA) (Liang and Zhang, 2017), polypeptides composition (Sun and Huang, 2006), PsiBlast (Liu et al., 2010; Li et al., 2014), etc. These techniques do not facilitate to reach 70% of classification results individually therefore, extracted features from different feature extraction

[*] Corresponding author at: Research Team in Intelligent Machines, National School of Engineers of Gabes, B.P. W, 6072 Gabes, Tunisia.
E-mail address: abdennaji.ines.1992@ieee.org (I. Abdennaji).

https://doi.org/10.1016/j.compbiolchem.2021.107450
Received 8 December 2020; Accepted 3 February 2021
Available online 8 February 2021

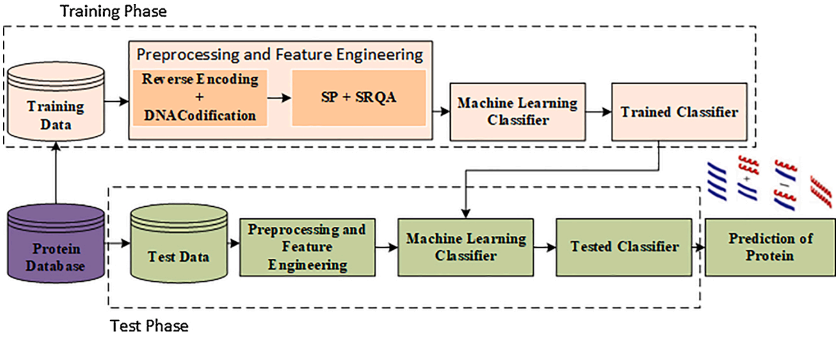
Training Phase
Training Data
Preprocessing and Feature Engineering
Reverse Encoding + DNA Codification
SP + SRQA
Machine Learning Classifier
Trained Classifier
Protein Database
Test Data
Preprocessing and Feature Engineering
Machine Learning Classifier
Tested Classifier
Prediction of Protein
Test Phase

Fig. 1. Framework diagram.

techniques are fused. Furthermore, to classify the structural classes, various classification methods are applied such as Fisher's Linear Discriminant Algorithm (LDA) (Yang et al., 2009), Support Vector machines (SVM) (Kim et al., 2006; Sudha et al., 2018; Anand et al., 2008; Zhu et al., 2019), Artificial Neural Network (ANN) (Bao et al., 2014) and Bayesian Classifier (Aydin et al., 2011).

From the studies presented above, it is noticed that there is a great disparity in the protein sequences encoding and feature extraction. Furthermore, classification performance can be improved by using the fused feature engineering technique and machine learning methods. The need to introduce new simple methods with high performance is expected.

The proposed work is a continuation of the previously undertaken studies (Sudha et al., 2018; Olyaee et al., 2016) on the use of recurrences and the recent work done by Girault (2015) based on the link between recurrences and symmetries. In addition, the presence of symmetry in the tertiary structures of proteins (Xu et al., 2005) suggests that symmetry can be an important property which has to be explored. Consequently, it is appropriate to investigate the consideration of symmetries for the classification of proteins. The major contributions of this study is to present:

1. a simpler protein sequences encoding;
2. an easy to use method;
3. new feature vectors based on symmetry concept and recurrence;
4. the best classifier by comparing different protein structural class prediction models such as SVM, LDA and RF.

The remaining paper is arranged as follows. Material and methodology is presented in Section 2. Results illustrate in Section 3 accompanied by discussion in Section 4. Finally, Section 5 concludes the paper.

2. Materials and methods

2.1. The framework

The framework diagram of this study is shown in Fig. 1. First, data set is split up into training and test sets with a ratio of 80:20. Second, the training and test sets are preprocessed through a coding phase. Then, symmetrical recurrence plots (SRP) are calculated and the feature extraction step is performed by applying symmetrical recurrence quantification analysis (SRQA). In total, three different features data sets are calculated: 8-SRQA-R, 8- SRQA-I, 16-SRQA, their definition will be presented in Section 2.4. Third, the machine learning models such as the RF, SVM and LDA are used to training data set for training. The model parameters iteratively tune to improve the performance in the training process. Lastly, test data set is used to evaluate the trained models.

2.2. Database

In this work, we used two benchmark databases containing low similarity proteins which are widely used for predicting protein structural classes: the database 25PDB includes 1673 protein sequences with 40% sequence homology and the database 1189 contains 1092 protein sequences with 25% sequence homology. Table 1 gives more details about the two databases (Kurgan and Homaeian, 2006) and the distribution of the four secondary structural classes.

Table 1
Structure of the two data sets used in our study.

Dataset	α	β	α/β	$\alpha+\beta$	Total
25PDB	443	443	346	441	1673
1189	223	294	334	241	1092

2.3. Reverse encoding & DNA codification

Each protein is formed with a linear sequence of amino acids (AAs). In addition, there are 20 standard genetic codes and multi-coded methods. So, each one protein could be expressed by different kinds of nucleotide sequences. The reverse encoding goes in inverse from protein to DNA sequence. As there is no uniqueness in the universal code of translating DNA into AAs, we used the codon (see in Table 2) as presented by Deschavanne and Tuffery (2008). In their study, the authors guarantee the balance in base composition to maximize the difference between the AAs codes.

There are a lot of representations of DNA sequences used in the biology field like: numerical representation (Kwan and Arniker, 2009), Chaos Game representation (Jeffrey, 1990), binary representation (Voss, 1992), etc. For the sake of simplicity, we used a unique DNA representation performed by Conte and Giuliani (2009) which is based on attributing:

- (+1) to the purine: Adenine (A) and Guanine (G);
- (−1) to the pyrimidine: Cytosine (C) and Thymine (T).

The simple reverse binary encoding (reverse encoding + binary DNA encoding) constitutes the first contribution of our proposed approach. It permits the transformation of one protein sequence into a binary sequence, one example is shown in Fig. 2. This will help to visualize, extract and identify characteristics from the sequences such as symmetries and recurrences.

2.4. Proposed approach

Our second contribution is an improvement of previously undertaken studies (Yang et al., 2009; Olyaee et al., 2016) that are based on the use of recurrences. The improvement extracts four kinds of symmetrical recurrences as proposed initially in the recent work done by Girault (2015). These extracted symmetrical recurrences have two advantages: they use symmetry properties that have not been used

Table 2
Reverse encoding.

A=GCT	C=TGC	D=GAC	E=GAG	F=TTC	G=GGT	H=CAC	I=ATT	K=AAG	L=CTA
M=ATG	N=AAC	P=CCA	Q=CAG	R=CGA	S=TCA	T=ACT	V=GTG	W=TGG	Y=TAC

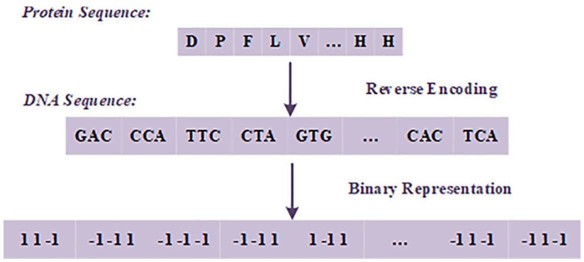

Fig. 2. Representation of protein 1A6M by a binary sequence.

currently and the symmetrical nature of recurrences does not require an embedding phase, therefore, making it much simpler. To further explore the symmetrical recurrences, we recalled the concept of standard recurrences plot in Appendix A.

2.4.1. Symmetrical recurrence plot

As proposed by Girault (2015), taking symmetrical properties of recurrences in consideration make processes understandable and detect invisible transitions effectively. The present work is an application of this new concept to biological discrete sequences. From the concept of symmetrical recurrence plot, four novel recurrence matrices are proposed. In Girault (2015), it is seen that respective matrices are sensitive to the occurrence of diverse symmetry types. Four types of transformation are performed, i.e. translation, reflection, inversion and glide (TRIG). Furthermore, corresponding components of the two-dimensional matrix M_k (a new matrix) can be presented in the generalized framework as below:

$$M_k(j,i) = \ominus[\varepsilon- \parallel X(j) - G_k X(i) \parallel] \tag{1}$$

with ε a gauge and $k \in \{T, R, I, G\}$.

The theoretical framework proposed is similar to the one proposed in Girault and Humeau-Heurtier (2018):

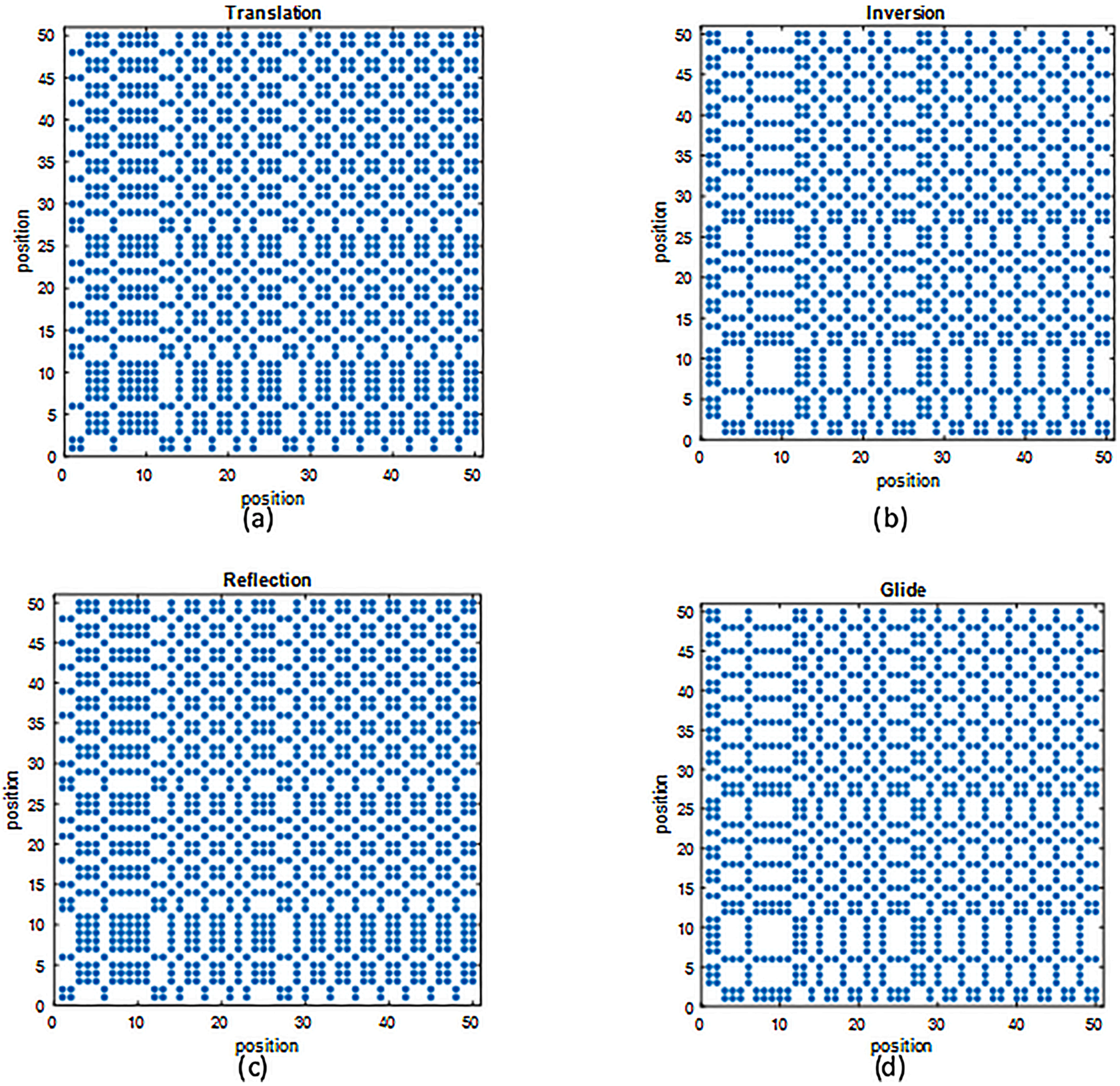

Fig. 3. (a) Translation recurrence plot, (b) reflection recurrence plot, (c) inversion recurrence plot, (d) glide recurrence plot, for the time series derived from protein 1A6M. The parameters used: $\varepsilon = 0$, $d = 1$, $\tau = 0$.

$$\| X(j) - G_k X(i) \| \leq \varepsilon. \tag{2}$$

Four types of operations are considered:

- $G_T[X(j)] = X(j+n)$ represents a translation of n samples, $k = T$;
- $G_R[X(j)] = X(-j+n)$ represents a reflection at the position n, $k = R$;
- $G_I[X(j)] = -X(-j+n)$ represents an inversion at the position n, $k = I$;
- $G_G[X(j)] = -X(j+n)$ represents a glide reflection of n samples, $k = G$.

An interesting properties of symmetrical recurrence plots are (1) not useful to embed and (2) sojourn points are naturally removed. This means that standard settings are fix to $d = 1$ (embedding dimension), $\tau = 0$ (time delay). Also, the gauge is null ($\varepsilon = 0$) since we are working on binary sequences. In the particular case of binary data, $M_T = M_R$ and $M_I = M_G$. This is observed in Fig. 3 where the four symmetrical recurrence plots (SRP) were computed using Eq. (1) by considering a protein sequence. We clearly noticed that the Translation and Reflection presented the same plot. In addition, Glide and Inversion gave the identical plot. Owing to these two matching results, we will consider just the **Reflection** and the **Inversion** in the rest of the paper. Finally, the quantification step is very significant and useful to investigate the difference between local and global symmetries in the symmetrical recurrence analysis.

2.4.2. *Symmetry recurrence quantification analysis*

In order to quantify the different types of recurrences, it is recommended to extend the current recurrence descriptors to other forms of recurrence such as symmetrical recurrences. Therefore, symmetry recurrence quantification analysis (SRQA) is proposed based on recurrence quantification analysis (RQA) (Webber and Zbilut, 1994; Marwan et al., 2002; Trulla et al., 1996).

Eight descriptors are calculated for each recurrence matrix $M_R(j, i)$ and $M_I(j, i)$. Therefore, a total of sixteen descriptors were calculated with $k \in \{R, I\}$ (see Eqs. (A.3) to (A.11) in Appendix A): Recurrence Rate (RR_k), Determinism (DET_k), Entropy ($ENTR_k$), Laminarity (LAM_k), Maxline ($Lmax_k$), Meanline (L_k), Trapping Time (TT_k) and Trend ($TREND_k$). Finally, we can define 3 sets of features as input's classifiers simply:

- 8-SRQA-R (RR_R, DET_R, $ENTR_R$, LAM_R, $Lmax_R$, L_R, TT_R, $TREND_R$);
- 8-SRQA-I (RR_I, DET_I, $ENTR_I$, LAM_I, $Lmax_I$, L_I, TT_I, $TREND_I$);
- 16-SRQA (RR_R, DET_R, $ENTR_R$, LAM_R, $Lmax_R$, L_R, TT_R, $TREND_R$, RR_I, DET_I, $ENTR_I$, LAM_I, $Lmax_I$, L_I, TT_I, $TREND_I$).

2.5. *Prediction model and performance metrics*

As discussed in Section 1, the purpose of the study is to predict the protein structural classes such as α, β, α/β and $\alpha+\beta$. The framework for classification is presented and described in Fig. 1. In our study, 3 sets of features (8-SRQA-R, 8-SRQA-I, 16-SRQA) are fed into machine learning classifiers. Furthermore, machine learning classifiers such as SVM, LDA are used as suggested in Liu et al. (2010), Li et al. (2014) and Yang et al. (2009) to predict the protein structural class. Besides, the ensemble technique such as RF is also considered. In order to compare each classifier and validate the accuracy of classification models, performance metrics are utilized. We decide to use the performance metrics in line with the recent studies such as overall accuracy and sensitivity. These measures are calculated as below:

$$\mathrm{OA} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{3}$$

$$\mathrm{sensitivity} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}} \tag{4}$$

where TP and TN are # True Positive and # True Negative respectively. In addition, FP and FN are # False Positive and # False Negatives accordingly.

Table 3
Sensitivity (%) of our method using SVM on the two benchmark datasets. Scenarios correspond to the two best feature sets.

Dataset	Scenarios	Sensitivity				
		All-α	All-β	α/β	$\alpha+\beta$	OA
25PDB	8-SRQA-I	100	82	70	71	81.2
	16-SRQA	100	81	82	79	86.0
1189	8-SRQA-I	47	90	100	48	74
	16-SRQA	62	100	94	57	80.4

Table 4
Sensitivity (%) of our method using LDA on the two benchmark datasets. Scenarios correspond to the two best feature sets.

Dataset	Scenarios	Sensitivity				
		All-α	All-β	α/β	$\alpha+\beta$	OA
25PDB	8-SRQA-R	98	92	100	96	96.4
	16-SRQA	99	95	99	97	97
1189	8-SRQA-I	97	95	98	100	98
	16-SRQA	99	97	96	100	98.2

Table 5
Sensitivity (%) of our method using RF on the two benchmark data sets. Scenarios correspond to the two best feature sets.

Dataset	Scenarios	Sensitivity				
		All-α	All-β	α/β	$\alpha+\beta$	OA
25PDB	8-SRQA-I	100	100	100	100	100
	16-SRQA	78	81	75	94	82
1189	8-SRQA-I	100	100	100	100	100
	16-SRQA	91	93	97	87	92.2

3. Result

Sensitivity (%) and Overall Accuracy (%) were calculated considering two benchmark datasets (25PDB and 1189). For the sake of clarity, a synthesis of results obtained with the three classifiers (SVM, LDA, RF) is presented below in Tables 3–5. More details are presented in the appendix in Tables A.7–A.9 .

3.1. *Support Vector Machine (SVM) classifier*

During the training process, three hyper-parameters were tuned such as the kernel coefficient gamma (auto mode), the polynomial kernel function degree (set to 3) and on/off probability estimates (set to TRUE). Finally, a test set was used to evaluate the model.

In Table 3, the best result is obtained with SVM[16-SRQA] in the both benchmark datasets for example All-α: 100%, All-β: 81%, α/β: 82% and $\alpha+\beta$: 79% and with 86% overall accuracy for the database 25PDB, and All-α: 62%, All-β: 100%, α/β: 94% and $\alpha+\beta$: 57% and with 80.4% overall accuracy for the database 1189. According to Table 3 SVM classifier performs better with the database 25PDB as compared to the database 1189 considering sensitivity.

3.2. *Linear Discriminant Analysis (LDA) classifier*

During the training process, default hyper-parameters were used with a dimensionality reduction. Finally, a test set was used to evaluate the model.

In Table 4, the best result is obtained with LDA[16-SRQA] in the both

Table 6
Comparison of our method (8-SRQA-I) with other studies.

Dataset	Methods	Classifier	Sensitivity				OA
			All-α	All-β	α/β	$\alpha+\beta$	
25PDB	AAD-CGR (Yang et al., 2009)	LDA	64.3	65	61.7	65	64
	SCPRED (Kurgan et al., 2008)	SVM	92.6	80.1	74	71	79.7
	Zhang et al. (2013)	SVM	96.7	80.8	82.4	75.5	83.7
	Olyaee et al. (2016)	LDA	95.6	89.5	88.1	87	90
	WD PseAAC (Yu et al., 2017)	SVM	95.7	97.7	94.8	84.4	93.1
	Wang and Wang (2019)	KNN	98	98.9	98	97.5	98.1
	Our method	LDA	99	95	99	97	97
		SVM	100	81	82	79	86
		RF	100	100	100	100	100
1189	AAD-CGR (Yang et al., 2009)	LDA	62.3	67.7	63.1	66.5	65.2
	SCPRED (Kurgan et al., 2008)	SVM	89.1	86.7	89.6	53.8	80.6
	Zhang et al. (2013)	SVM	92.4	84.4	84.4	73.4	83.6
	Olyaee et al. (2016)	LDA	92.3	90.1	86.5	75.2	-
	WD PseAAC (Yu et al., 2017)	SVM	98.7	99	94	68.9	90.8
	Wang and Wang (2019)	KNN	98.2	99.3	99.1	91.3	97.3
	Our method	LDA	99	97	96	100	98.2
		SVM	62	100	94	57	80.4
		RF	100	100	100	100	100

benchmark datasets for example All-α: 99%, All-β: 95%, α/β: 99% and $\alpha+\beta$: 97% and with 97% overall accuracy for the database 25PDB, and All-α: 99%, All-β: 97%, α/β: 96% and $\alpha+\beta$: 100% and with 98.2% overall accuracy for the database 1189. According to Table 4, LDA classifier performs better with the database 1189 as compared to the database 25PDB considering sensitivity.

3.3. Random Forest (RF) classifier

During the training process, hyper-parameter such as the number of estimators and the maximum depth were tuned. These parameters were selected as 9 (for the number of estimators) and 6 (for the maximum depth). Finally, a test set was used to evaluate the model.

In Table 5, the best result is obtained with RF [8-SQRA-I] in both benchmark datasets for example All-α: 100%, All-β: 100%, α/β: 100% and $\alpha+\beta$: 100% and with 100%. According to Table 5, RF classifiers performs in a similar way to whatever the dataset based on sensitivity.

3.4. Classifier comparison

From Tables 3–5, it can be claimed that the best combination between classifier input features and the classifier is RF[8-SRQA-I] with on overall of 100% without overfitting on both benchmark data sets with the same encoding. The second best combination is obtained with LDA [16-SRQA] with an overall of 97%. The worst combination is obtained with SVM[16-SRQA] with an overall of 80.5%. Consequently, we recommend using RF[8-SRQA-I].

4. Discussion

In this study, we showed the possibility to classify the 4 protein structural classes: All-α, All-β, α/β, $\alpha+\beta$ without error by considering: (1) a binary encoding of protein sequences, (2) the calculation of symmetrical recurrences and its 8 associated descriptors/features and (3) a classifier. In our study, the best combination of classifiers and their inputs is RF [8-SRQA-I]. From our point of view, the joint use of (1) a simple encoding, (2) taking into account descriptors based on symmetrical recurrences and (3) use the ensemble classifier is proved very significant for better results.

In Table 6 a comparison is made between our method (8-SRQA-I) and existing methods (RQA) obtained in Yang et al. (2009) and Olyaee et al. (2016) on the same data sets. In Yang et al. (2009), the protein sequences are encoded in two time series via the Chaos-Game-Representation (CGR) approach, 8-RQA and a LDA classifier were applied. In Yang et al. (2009), the data are embedded in a space with $d=8$ dimensions and with a delay $\tau=2$ and $\varepsilon=0.3$. The results obtained having sensitivity % 64.3(All-α), 65(All- β), 61.7(α/β) and 65 ($\alpha+\beta$) with an overall prediction accuracy of 64% for 25PDB dataset. Similar behavior is seen in Olyaee et al. (2016) with overall prediction accuracy 90% and LDA was applied.

In Table 6 a comparison is made between our best results obtained with RF[8-SRQA-I] and other existing methods (Yang et al., 2009; Yu et al., 2017; Olyaee et al., 2016; Wang and Wang, 2019; Kurgan et al., 2008; Zhang et al., 2013). From Table 6, it is clearly shown that our best configuration, i.e. RF[8-SRQA-I] outperformed the recent results of Wang and Wang (2019). In addition, from Tables A.7–A.9 we see that RF [8-SRQA-I], RF[8-SRQA-R] overpass the recent results of Wang and Wang (2019) for any type of symmetry. Moreover, the percentage accuracy and sensitivity on training and testing are the almost same. Besides, the RF model is tuned with 6 as maximum depth and results approached to 100%. Cross-validation shows the same results. Therefore, our classification model is not overfitting.

It is often tricky to find the right parameters. In our approach, the data are binary coded therefore, $\varepsilon=0$. In order to choose the embedded dimension, a value greater than 1 eliminates false recurrences (sojourn point). With the combined use of binary data and symmetric recurrences, there are no more sojourn points. Therefore, search for an embedding dimension or a delay is not required. Consequently, the right parameters are $d=1$, $\tau=0$ and $\varepsilon=0$.

5. Conclusion

In this paper, we have shown that the judicious combination of (1) a simple reverse encoding followed by a binary coding, (2) the calculation of symmetrical recurrences features and, (3) a classifier like RF, provide improved classification of 4 structural classes of proteins such as All-α, All-β, α/β and $\alpha+\beta$ without overfitting. The simple recurrences settings ($d=1$, $\tau=0$ and $\varepsilon=0$) are proved useful to calculate the recurrences. The consideration of symmetry suggested by the presence of symmetric tertiary structures of proteins results in 100% classification without error. This proposed classification method can be used for other applications having binary or quaternary data. Furthermore, our proposed method will help in improving the drug design, folding recognition of protein, functional analysis and several other biology applications.

Conflict of interest

None declared.

Appendix A

A.1 Recurrence plot

Eckmann et al. (1995) proposed the concept of recurrence plot initially to identify the presence of identical neighboring points in a time series such as $x(n) = x_1, x_2, \ldots, x_N$. This time series is embedded into a phase space with an embedding dimension d and a time delay τ. Two points such as $X(i)$ and $X(j)$ in the d-dimensional space are considered recurrent if they satisfy the following test (Eckmann et al., 1995):

$$\| X(i) - X(j) \| \leq \varepsilon. \tag{A.1}$$

A two-dimensional matrix $N \times N$, M (recurrence Matrix) can be calculated as followed:

$$M(i,j) = \ominus[\varepsilon - \| X(i) - X(j) \|]. \tag{A.2}$$

The recurrence matrix M is a binary matrix composed of zeros and ones where zero components present the same state and non-zero components exhibit different states. Besides, the right selection of d is very important, as incorrect selection will lead towards recurrences contamination (Webber and Zbilut, 1994; March et al., 2005) (false recurrences/sojourn points). In addition, The embedding dimension is usually set with $d \geq 2$ to avoid the presence of sojourn points. The increase in dimensionality usually reduces the number of false recurrences; however, other approaches are proposed by Zaylaa et al. (2015).

A.2 Recurrence quantification analysis

The recurrence quantification analysis (RQA) extracts quantitative features (descriptors) from the binary matrix M. It permits to measure differently appearing recurrence plots (RPs) with the help of small-scale structures present inside it. A main advantage of RQA is to give effective information for non-stationary and short data while other techniques fail to do so. It may be applied to versatile types of data. Moreover, to quantify the complexity, different measures of RQA introduced heuristically in Webber and Zbilut (1994), Marwan et al. (2002) and Trulla et al. (1996) as described below.

Recurrence rate (RR): it measures of the density of recurrent points present in the matrix M. RR ranges between 0% to 100% where 100% reflect that all the points are recurrent:

$$\mathrm{RR} = \frac{1}{N^2} \sum_{i,j-1}^{N} M(i,j) \quad \forall i \neq j. \tag{A.3}$$

Determinism (DET): it measures the presence of temporal correlation and appears through the presence of diagonal/anti diagonal. DET is the percentage of recurrence points assembled to build diagonal lines:

$$\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{N} 1\ \mathrm{p}(l)}{N^2(\mathrm{RR})}, \tag{A.4}$$

where $p(l)$ represents the probability of finding diagonal line/anti-diagonal of l. $l_{m}in$ is the segment which is shortest and considered often as 2.

Entropy: it measures the deterministic structureś complexity within the system. It depends on the bin-number sensitively.

$$\mathrm{ENTR} = -\sum_{l=l_{\min}}^{N} p(l) ln(p(l)), \tag{A.5}$$

where $p(l)$ represents the chances of occurrence is that diagonal segment is of exact length (l) which is calculated based on frequency distribution $P(l)$:

$$p(l) = \frac{p(l)}{\sum_{l=l_{\min}}^{N} p(l)}. \tag{A.6}$$

Laminarity: it measures of the total number of recurrence points which combine to form a vertical line:

$$\mathrm{LAM} = \frac{\sum_{v=v_{\min}}^{N} \mathrm{v}\ \mathrm{p}(v)}{N^2(\mathrm{RR})}, \tag{A.7}$$

where $p(v)$ represents the probability of finding vertical lines of v which has at least v_{m}in as length.

Maxline is the longest length of the diagonal line:

$$L_{\mathrm{MAX}} = \max(l_i; i = 1, \ldots, N_l) \tag{A.8}$$

Meanline: the vertical and diagonal line's length can be measured. Therefore, the average diagonal line length is called the meanline which is associated with the predictability interval of the dynamic system:

$$L = \frac{\sum_{l=l_{\min}}^{N} l(p(l))}{\sum_{l=l_{\min}}^{N} p(l)}. \tag{A.9}$$

Trapping time (TT): it measures the average length of the vertical lines, which is directly connected to the laminarity interval of the dynamic system, i.e. how long the dynamic system will remain in some specific state.

$$\mathrm{TT} = \frac{\sum_{v=v_{\min}}^{N} v(p(v))}{\sum_{v=v_{\min}}^{N} p(v)}. \tag{A.10}$$

Table A.7
Sensitivity of our proposed method using SVM on the two benchmark datasets. Classifier input features are 8-SRQA-R (Reflection), 8-SRQA-I (Inversion) and 16-SRQA.

Dataset	Methods	Sensitivity			
		All-α	All-β	α/β	$\alpha+\beta$
25PDB	8-SRQA-R	100	76	66	79
	8-SRQA-I	100	82	70	71
	16-SRQA	100	81	82	79
1189	8-SRQA-R	44	100	100	98
	8-SRQA-I	47	90	100	48
	16-SRQA	62	100	94	57

Table A.8
Sensitivity of our method using LDA on the two benchmark datasets. Classifier input features are 8-SRQA-R (Reflection), 8-SRQA-I (Inversion) and 16-SRQA.

Dataset	Methods	Sensitivity			
		All-α	All-β	α/β	$\alpha+\beta$
25PDB	8-SRQA-R	98	92	100	96
	8-SRQA-I	99	92	99	95
	16-SRQA	99	95	99	97
1189	8-SRQA-R	100	93	98	100
	8-SRQA-I	97	95	98	100
	16-SRQA	99	97	96	100

Table A.9
sensitivity of our method using RF on the two benchmark datasets. Classifier input features are 8-SRQA-R (Reflection), 8-SRQA-I (Inversion) and 16-SRQA.

Dataset	Methods	Sensitivity			
		All-α	All-β	α/β	$\alpha+\beta$
25PDB	8-SRQA-R	99	99	100	100
	8-SRQA-I	100	100	100	100
	16-SRQA	78	81	75	94
1189	8-SRQA-R	100	100	100	100
	8-SRQA-I	100	100	100	100
	16-SRQA	91	93	97	87

Trend (TREND): it is the regression coefficient of the linear association among the density of recurrence points in a line parallel to the line of Identity and its distance to the line of Identity. In addition, the trend gives significant information about the system's stationarity:

$$\mathrm{TREND} = \frac{\sum_{l=1}^{\overline{N}}(i-\frac{\overline{N}}{2})(\mathrm{RR}-<\mathrm{RR}>)}{\sum_{i=1}^{\overline{N}}(i-\frac{\overline{N}}{2})^2}, \tag{A.11}$$

where $\overline{N}$ is the Maximal number of diagonals parallel to the LOI.

A.3 Tables

Tables A.7–A.9 present sensitivity (%) obtained with the three different classifiers (SVM, LDA and RF). In each table, the two benchmark data sets 25PDB and 1189 are analyzed. For each row, All-α, All-β, α/β and $\alpha+\beta$ are tested.

Firstly, considering the SVM classifier, it is observed in Table A.7 that the best performances are globally obtained with 16-SRQA (Fusion) in 25PDB benchmark data set. However, with data set 1189, the best outcome comes from 8-SRQA-R.

Secondly, considering the LDA classifier, it is noticed in Table A.8 that the best performances are globally obtained with 16-SRQA (Fusion) in both benchmark data sets.

Thirdly, considering the RF classifier, it can be seen in Table A.9 that the best performances are globally obtained with 8-SQRA-I in both benchmark data sets. Note that outcomes obtained with 8-SRQA-R and 16-SRQA are fairly close to those obtained with 8-SRQA-I.

References

Anand, A., Pugalenthi, G., Suganthan, P., 2008. Predicting protein structural class by svm with class-wise optimized features and decision probabilities. J. Theoret. Biol. 253, 375–380.

Aydin, Z., Singh, A., Bilmes, J., Noble, W.S., 2011. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC Bioinformatics 12, 154.

Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B., 1997. Understanding the recognition of protein structural classes by amino acid composition. Proteins: Struct. Funct. Bioinformatics 29, 172–185.

Bao, W., Chen, Y., Wang, D., 2014. Prediction of protein structure classes with flexible neural tree. Biomed. Mater. Eng. 24, 3797–3806.

Cai, Y.-D., Hu, J., Liu, X., Chou, K.-C., 2002. Prediction of protein structural classes by neural network method. J. Mol. Des. 1, 332–338.

Carlacci, L., Chou, K.C., Maggiora, G.M., 1991. A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry 30, 4389–4398.

Chou, K.-C., Zhang, C., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349. https://doi.org/10.3109/10409239509083488.

Chou, K.-C., 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223, 509–517.

Chou, K.-C., 1995. A novel approach to predicting protein structural classes in a (20-1)-d amino acid composition space. Proteins: Struct. Funct. Bioinformatics 21, 319–344.

Chou, K.-C., 2005. Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein and Peptide Sci. 6, 423–436.

Conte, S., Giuliani, A., 2009. Identification of Possible Differences in Coding and Non-Coding Fragments of DNA Sequences by Using the Method of the Recurrence Quantification Analysis. arXiv:0910.3516.

Deschavanne, P., Tuffery, P., 2008. Exploring an alignment free approach for protein classification and structural class prediction. Biochimie 90, 615–625.

Eckmann, J., Kamphorst, S.O., Ruelle, D., et al., 1995. Recurrence plots of dynamical systems. World Sci. Ser. Nonlinear Sci. Ser. A 16, 441–446.

Feng, Z., Hu, X., Jiang, Z., Song, H., Ashraf, M.A., 2016. The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi J. Biol. Sci. 23, 189–197.

Fias, S., Van Damme, S., Bultinck, P., 2008. Multidimensionality of delocalization indices and nucleus independent chemical shifts in polycyclic aromatic hydrocarbons. J. Comput. Chem. 29, 358–366.

Girault, J.-M., Humeau-Heurtier, A., 2018. Centered and averaged fuzzy entropy to improve fuzzy entropy precision. Entropy 20, 287.

Girault, J.-M., 2015. Recurrence and symmetry of time series: application to transition detection. Chaos Solitons Fract. 77, 11–28.

Gromiha, M.M., Selvaraj, S., 1998. Protein secondary structure prediction in different structural classes. Protein Eng. 11, 249–251.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.

Kim, J.K., Bang, S.-Y., Choi, S., 2006. Sequence-driven features for prediction of subcellular localization of proteins. Pattern Recogn. 39, 2301–2311.

Kurgan, L.A., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn. 39, 2323–2343.

Kurgan, L., Cios, K., Chen, K., 2008. Scpred: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinformatics 9, 226.

Kwan, H.K., Arniker, S.B., 2009. Numerical representation of dna sequences. 2009 IEEE International Conference on Electro/Information Technology 307–310.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.

Li, L., Cui, X., Yu, S., Zhang, Y., Luo, Z., Yang, H., Zhou, Y., Zheng, X., 2014. Pssp-rfe: accurate prediction of protein structural class by recursive feature extraction from psi-blast profile, physical-chemical property and functional annotations. PLOS ONE 9, e92863.

Liang, Y., Zhang, S., 2017. Predict protein structural class by incorporating two different modes of evolutionary information into Chou's general pseudo amino acid composition. J. Mol. Graph. Modell. 78, 110–117.

Lin, H., Ding, C., Song, Q., Yang, P., Ding, H., Deng, K.-J., Chen, W., 2012. The prediction of protein structural class using averaged chemical shifts. J. Biomol. Struct. Dyn. 29, 1147–1153.

Liu, T., Zheng, X., Wang, J., 2010. Prediction of protein structural class for low-similarity sequences using support vector machine and psi-blast profile. Biochimie 92, 1330–1334.

March, T., Chapman, S., Dendy, R., 2005. Recurrence plot statistics and the effect of embedding. Physica D: Nonlinear Phenom. 200, 171–184.

Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., Kurths, J., 2002. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. Phys. Rev. E 66, 026702.

Nishikawa, K., Ooi, T., 1982. Correlation of the amino acid composition of a protein to its structural and biological characters. J. Biochem. 91, 1821–1824.

Olyaee, M.H., Yaghoubi, A., Yaghoobi, M., 2016. Predicting protein structural classes based on complex networks and recurrence analysis. J. Theoret. Biol. 404, 375–382.

Sudha, P., Ramyachitra, D., Manikandan, P., 2018. Enhanced artificial neural network for protein fold recognition and structural class prediction. Gene Rep. 12, 261–275.

Sun, X.-D., Huang, R.-B., 2006. Prediction of protein structural classes using support vector machines. Amino acids 30, 469–475.

Trulla, L., Giuliani, A., Zbilut, J., Webber Jr., C., 1996. Recurrence quantification analysis of the logistic equation with transients. Phys. Lett. A 223, 255–260.

Voss, R.F., 1992. Evolution of long-range fractal correlations and 1/f noise in dna base sequences. Rev. Lett. 68, 3805–3808.

Wang, S., Wang, X., 2019. Prediction of protein structural classes by different feature expressions based on 2-d wavelet denoising and fusion. BMC Bioinformatics 20, 701.

Webber Jr., C.L., Zbilut, J.P., 1994. Dynamical assessment of physiological systems and states using recurrence plot strategies. J. Appl. Physiol. 76, 965–973.

Xu, R., Li, M., Chen, H., Huang, Y., Xiao, Y., 2005. A symmetry-related sequence-structure relation of proteins. Chin. Sci. Bull. 50, 536.

Yang, J.-Y., Peng, Z.-L., Yu, Z.-G., Zhang, R.-J., Anh, V., Wang, D., 2009. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J. Theoret. Biol. 257, 618–626.

Yu, B., Lou, L., Li, S., Zhang, Y., Qiu, W., Wu, X., Wang, M., Tian, B., 2017. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J. Mol. Graph. Modell. 76, 260–273.

Zaylaa, A., Charara, J., Girault, J.-M., 2015. Reducing sojourn points from recurrence plots to improve transition detection: application to fetal heart rate transitions. Comput. Biol. Med. 63, 251–260.

Zhang, L., Zhao, X., Kong, L., 2013. A protein structural class prediction method based on novel features. Biochimie 95, 1741–1744.

Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., Hao, L., 2019. Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl.-Based Syst. 163, 787–793.