



**HAL**  
open science

## Prediction of protein structural class based on symmetrical recurrence quantification analysis

Ines Abdennaji, Mourad Zaied, Jean-Marc Girault

► **To cite this version:**

Ines Abdennaji, Mourad Zaied, Jean-Marc Girault. Prediction of protein structural class based on symmetrical recurrence quantification analysis. *Computational Biology and Chemistry*, 2021, 92, pp.107450. 10.1016/j.compbiolchem.2021.107450 . hal-03676706

**HAL Id: hal-03676706**

**<https://hal.science/hal-03676706v1>**

Submitted on 10 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Prediction of Protein Structural Class Based on Symmetrical Recurrence Quantification Analysis

Ines Abdennaji , Mourad Zaied , Jean-Marc Girault

*Research Team in Intelligent Machines, National School of Engineers of Gabes, B.P. W  
6072 Gabes, Tunisia*

*GSII ESEO -LAUM UMR CNRS 6613, 49000 Angers, France*

---

## Abstract

Protein structural class prediction for low similarity sequences is a significant challenge and one of the deeply explored subjects. This plays a 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 , Random Forest and Support Vector Machine 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.

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

---

## 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

*Preprint submitted to Neural Networks*

*May 15, 2021*

5 determined via the process of protein folding. Protein secondary structure is  
6 the three-dimensional form of local segments of proteins whose amino acids  
7 linear sequence (in a peptide or protein) forms the protein primary structure.  
8 As mentioned by Chou et Zhang in 1995[1], it is important and helpful to  
9 predict higher proteinic classes from primary proteinic sequences for two  
10 reasons. Firstly, if the structural class of the protein under study is known  
11 then the searching scope of conformation can be reduced [2]. Secondly, the  
12 structural class is related to various protein properties [3]. Since there is  
13 no simple and direct way for the protein tertiary structure prediction from  
14 its primary structure, four secondary structural classes of proteins based on  
15 the types and arrangement of their secondary structural class are proposed  
16 by Levitt and Chothi [4]. These classes are the  $\alpha$ , the  $\beta$  and those with a  
17 mixture of  $\alpha$  and  $\beta$  shapes called the  $\alpha/\beta$  and the  $\alpha + \beta$ .

18 These four protein structural classes can be used to 1) implement a heuristic  
19 method for deciding tertiary structure [5], 2) reduce search space of probable  
20 conformations of tertiary structure [6, 2], 3) improve prediction of secondary  
21 structure accuracy and 4) predict function from amino acid sequence  
22 information. Protein structural class prediction plays an essential role in  
23 functional analysis, protein structures, drug designs and a lot of other similar  
24 applications in biology [7].

25 For the last 10 years, prediction of protein structural class for low similarity  
26 sequences [8, 9] is a tough challenge for the scientific community. Therefore,  
27 an automated and accurate protein structural class prediction for newly  
28 established proteins is required. In order to extract the feature sequences  
29 from protein, various feature extraction techniques are used in the recent  
30 studies which can be later useful for classification of the structural classes.  
31 Most of used techniques include Amino Acid Composition(AAC) [10, 11, 12],  
32 Average Chemical Shift (ACS) [13, 14, 15], Pseudo Amino Acid (PSeAA) [16],  
33 polypeptides composition [17], PsiBlast [18, 19] and etc. These techniques  
34 do not facilitate to reach 70% of classification results individually therefore,  
35 extracted features from different feature extraction techniques are fused. Furthermore,  
36 to classify the structural classes, various classification methods are applied  
37 such as Fisher's Linear Discriminant Algorithm (LDA) [20], Support  
38 Vector machines (SVM) [11, 21, 22, 23], Artificial Neural Network (ANN)  
39 [24] and Bayesian Classifier [25].

40 From the studies presented above, it is noticed that there is a great disparity  
41 in the protein sequences encoding and feature extraction. Furthermore,  
42 classification performance can be improved by using the fused feature en-

43 gineering technique and machine learning methods. The need to introduce  
44 new simple methods with high performance is expected.

45 The proposed work is a continuation of the previously undertaken stud-  
46 ies [21, 26] on the use of recurrences and the recent work done by Girault  
47 [27] based on the link between recurrences and symmetries. In addition, the  
48 presence of symmetry in the tertiary structures of proteins [28] suggests that  
49 symmetry can be an important property which has to be explored. Conse-  
50 quently, it is appropriate to investigate the consideration of symmetries for  
51 the classification of proteins. The major contributions of this study is to  
52 present:

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

58 The remaining paper is arranged as follows. Material and methodology is  
59 presented in Section 2. Results illustrate in section 3 accompanied by dis-  
60 cussion in section 4. Finally, Section 5 concludes the paper.

## 61 **2. Materials and Methods**

### 62 *2.1. The Framework*

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

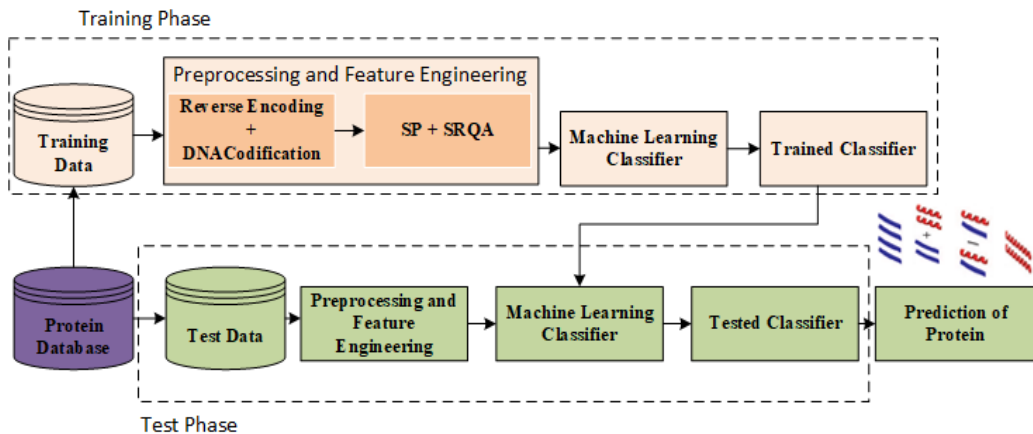


Figure 1: Framework diagram.

74 *2.2. Database*

75 In this work, we used two benchmark databases containing low similar-  
 76 ity proteins which are widely used for predicting protein structural classes:  
 77 the database 25PDB includes 1673 protein sequences with 40% sequence ho-  
 78 mology and the database 1189 contains 1092 protein sequences with 25%  
 79 sequence homology. Table 1 gives more details about the two databases [8]  
 80 and the distribution of the four secondary structural classes.

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

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

81 *2.3. Reverse Encoding & DNA Codification*

82 Each protein is formed with a linear sequence of Amino Acids (AAs). In  
 83 addition, there are 20 standard genetic codes and multi-coded methods. So,  
 84 each one protein could be expressed by different kinds of nucleotide sequences.  
 85 The reverse encoding goes in inverse from protein to DNA sequence. As there  
 86 is no uniqueness in the universal code of translating DNA into AAs, we used  
 87 the codon (see in Table 2) as presented by Deschavanne and Tuffery [29]. In  
 88 their study, the authors prove that this encoding gives the best results for

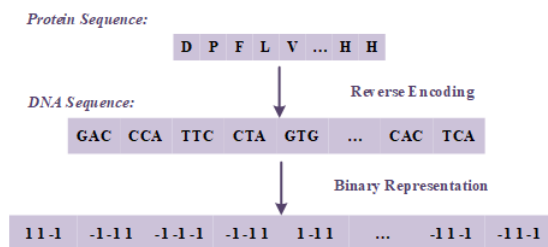


Figure 2: Representation of Protein 1A6M by a binary sequence.

89 the prediction of protein structural class. Furthermore, the authors guarantee  
 90 the balance in base composition to maximize the difference between the AAs  
 91 codes.

---

A=GCT	C=TG	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

---

Table 2: Reverse Encoding

92 There are a lot of representations of DNA sequences used in the biology  
 93 field like: numerical representation [30], Chaos Game representation[31], bi-  
 94 nary representation[32] and, Etc. For the sake of simplicity, we used a unique  
 95 DNA representation performed by Elio Conte et al. [33] which is based on  
 96 attributing:

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

99 The simple reverse binary encoding (reverse encoding + binary DNA encod-  
 100 ing) constitutes the first contribution of our proposed approach. It permits  
 101 the transformation of one protein sequence into a binary sequence, one ex-  
 102 ample is shown in Fig2. This will help to visualize, extract and identify  
 103 characteristics from the sequences such as symmetries and recurrences.

#### 104 2.4. Proposed Approach

105 Our second contribution is an improvement of previously undertaken  
 106 studies [20, 26] that are based on the use of recurrences. The improvement  
 107 extracts four kinds of symmetrical recurrences as proposed initially in the  
 108 recent work done by Girault [27]. These extracted symmetrical recurrences

109 have two advantages: they use symmetry properties that have not been used  
 110 currently and the symmetrical nature of recurrences does not require an em-  
 111 bedding phase, therefore, making it much simpler. To further explore the  
 112 symmetrical recurrences and to make the paper autonomous, we recalled the  
 113 concept of standard recurrences plot in the appendix.

#### 114 2.4.1. Symmetrical Recurrence Plot

115 As proposed by Girault in [27], taking symmetrical properties of recur-  
 116 rences in consideration make processes understandable and detect invisible  
 117 transitions effectively. The present work is an application of this new concept  
 118 to biological discrete sequences. From the concept of symmetrical recurrence  
 119 plot, four novel recurrence matrices are proposed. In [27], it is seen that re-  
 120 spective matrices are sensitive to the occurrence of diverse symmetry types.  
 121 Four types of transformation are performed i.e. Translation, Reflection, In-  
 122 version and Glide (TRIG). Furthermore, corresponding components of the  
 123 two-dimensional matrix  $M_k$  (a new matrix) can be presented in the general-  
 124 ized framework as below:

$$M_k(j, i) = \Theta[\varepsilon - \| X(j) - G_k X(i) \|] \quad (1)$$

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

126 The theoretical framework proposed is similar to the one proposed in [34]:

$$\| X(j) - G_k X(i) \| \leq \varepsilon. \quad (2)$$

127 Four types of operations are considered:

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

135 An interesting properties of symmetrical recurrence plots are 1) not useful  
 136 to embed and 2) sojourn points are naturally removed. This means that  
 137 standard settings are fix to  $d = 1$  (embedding dimension),  $\tau = 0$  (time

138 delay). Also, the gauge is null ( $\varepsilon = 0$ ) since we are working on binary  
139 sequences. In the particular case of binary data,  $M_T = M_R$  and  $M_I = M_G$ .  
140 This is observed in Fig.3 where the four symmetrical recurrence plots (SRP)  
141 were computed using Equ.1 by considering a protein sequence. We clearly  
142 noticed that the Translation and Reflection presented the same plot. In  
143 addition, Glide and Inversion gave the identical plot. Owing to these two  
144 matching results, we will consider just the **Reflection** and the **Inversion**  
145 in the rest of the paper. Finally, The quantification step is very significant and  
146 useful to investigate the difference between local and global symmetries in  
147 the symmetrical recurrence analysis.

#### 148 2.4.2. Symmetry Recurrence Quantification analysis

149 In order to quantify the different types of recurrences, it is recommended  
150 to extend the current recurrence descriptors to other forms of recurrence such  
151 as symmetrical recurrences. Therefore, Symmetry Recurrence Quantification  
152 Analysis (SRQA) is proposed based on Recurrence Quantification Analysis  
153 (RQA) [35, 36, 37] .

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

- 160 • 8-SRQA-R ( $RR_R, DET_R, ENTR_R, LAM_R, Lmax_R, L_R, TT_R, TREND_R$ );
- 161 • 8-SRQA-I ( $RR_I, DET_I, ENTR_I, LAM_I, Lmax_I, L_I, TT_I, TREND_I$ );
- 162 • 16-SRQA ( $RR_R, DET_R, ENTR_R, LAM_R, Lmax_R, L_R, TT_R, TREND_R,$   
163  $RR_I, DET_I, ENTR_I, LAM_I, Lmax_I, L_I, TT_I, TREND_I$ ).

#### 164 2.5. Prediction Model and Performance Metrics

165 As discussed in section I, the purpose of the study is to predict the protein  
166 structural classes such as  $\alpha$ ,  $\beta$ ,  $\alpha/\beta$  and  $\alpha + \beta$ . The framework for classifi-  
167 cation is presented and described in Fig 1. In our study, 3 sets of features  
168 (8-SRQA-R, 8-SRQA-I, 16-SRQA) are fed into machine learning classifiers.  
169 Furthermore, machine learning classifiers such as SVM, LDA are used as  
170 suggested in [18, 19, 20] to predict the protein structural class. Besides, the



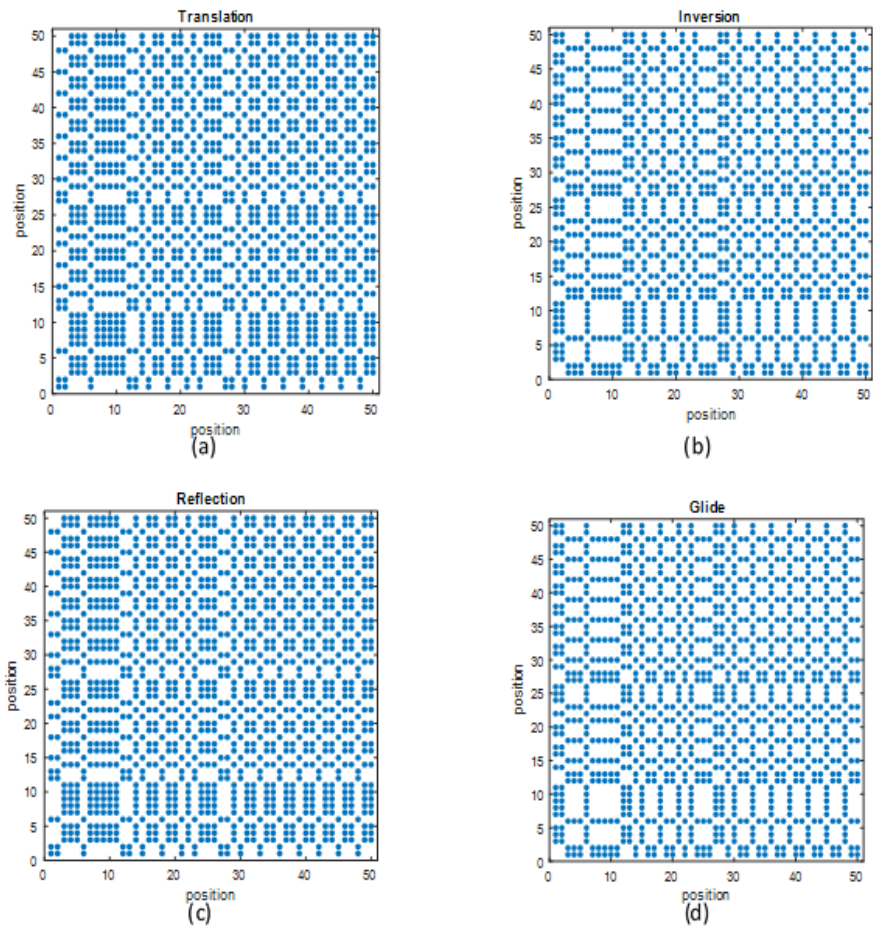


Figure 3: (a) Translation Recurrence Plot, (b) Reflection Recurrence Plot, (c) Inversion Recurrence Plot, (d) Glide Recurrence Plot, for the Time series derived from protein 1A6M. The parameters used:  $\varepsilon=0$ ,  $d=1$ ,  $\tau=0$ .

171 ensemble technique such as RF is also considered. In order to compare each  
 172 classifier and validate the accuracy of classification models, performance met-  
 173 rics are utilized. We decide to use the performance metrics in line with the  
 174 recent studies such as overall accuracy and sensitivity. These measures are  
 175 calculated as below:

$$OA = \frac{TN + TP}{TP + FP + TN + FN} \quad (3)$$

176

$$sensitivity = \frac{TP}{FN + TP} \quad (4)$$

177 where TP and TN are # True Positive and # True Negative respec-  
 178 tively. In addition, FP and FN are # False Positive and # False Negatives  
 179 accordingly.

### 180 3. Result

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

#### 186 3.1. Support Vector Machine(SVM) Classifier

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

Dataset	Scenarios	Sensitivity				OA
		All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$	
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 3: Sensitivity (%) of our method using SVM on the two benchmark datasets. Scenarios correspond to the two best feature sets.

191 In Table 3, the best result is obtained with SVM[16-SRQA] in the both  
 192 benchmark datasets for example All- $\alpha$ : 100%, All- $\beta$ : 81%,  $\alpha/\beta$ : 82% and  
 193  $\alpha+\beta$ : 79% and with 86% overall accuracy for the database 25PDB, and All- $\alpha$ :  
 194 62%, All- $\beta$ : 100%,  $\alpha/\beta$ : 94% and  $\alpha+\beta$ : 57% and with 80.4% overall accuracy  
 195 for the database 1189. According to Table 3 SVM classifier performs better  
 196 with the database 25PDB as compared to the database 1189 considering  
 197 sensitivity.

### 198 3.2. Linear Discriminant Analysis (LDA) classifier

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

201 In Table 4, the best result is obtained with LDA[16-SRQA] in the both  
 202 benchmark datasets for example All- $\alpha$ : 99%, All- $\beta$ : 95%,  $\alpha/\beta$ : 99% and  
 203  $\alpha+\beta$ : 97% and with 97% overall accuracy for the database 25PDB, and All- $\alpha$ :  
 204 99%, All- $\beta$ : 97%,  $\alpha/\beta$ : 96% and  $\alpha+\beta$ : 100% and with 98.2% overall accuracy  
 205 for the database 1189. According to Table 4, LDA classifier performs better  
 206 with the database 1189 as compared to the database 25PDB considering  
 207 sensitivity.

Dataset	Scenarios	Sensitivity				OA
		All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$	
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 4: Sensitivity (%) of our method using LDA on the two benchmark datasets. Scenarios correspond to the two best feature sets.

### 208 3.3. Random Forest (RF) classifier

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

213 In Table 5, the best result is obtained with RF [8-SQRA-I] in both bench-  
 214 mark datasets for example All- $\alpha$ : 100%, All- $\beta$ : 100%,  $\alpha/\beta$ : 100% and  $\alpha + \beta$ :  
 215 100% and with 100%. According to Table 5, RF classifiers performs in a  
 216 similar way to whatever the dataset based on sensitivity.

217 *3.4. Classifier Comparison:*

218 From Tables 3, 4 and 5, it can be claimed that the best combination  
 219 between classifier input features and the classifier is RF[8-SRQA-I] with  
 220 on overall of 100% without overfitting on both benchmark data sets with  
 221 the same encoding. The second best combination is obtained with LDA[16-  
 222 SRQA] with an overall of 97%. The worst combination is obtained with  
 223 SVM[16-SRQA] with an overall of 80.5%. Consequently, we recommend us-  
 224 ing RF[8-SRQA-I].

Dataset	Scenarios	Sensitivity				
		All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$	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

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

225 **4. Discussion:**

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

234 In Table 6 a comparison is made between our method (8-SRQA-I) and  
 235 existing methods (RQA) obtained in [20] and [26] on the same data sets.  
 236 In [20], the protein sequences are encoded in two time series via the Chaos-  
 237 Game-Representation (CGR) approach, 8-RQA and a LDA classifier were  
 238 applied. In [20], the data are embedded in a space with  $d = 8$  dimensions  
 239 and with a delay  $\tau=2$  and  $\varepsilon=0.3$ . The results obtained having sensitivity %  
 240 64.3(All- $\alpha$ ), 65(All-  $\beta$ ), 61.7( $\alpha/\beta$ ) and 65( $\alpha + \beta$ ) with an overall prediction  
 241 accuracy of 64% for 25PDB dataset. Similar behavior is seen in [26] with  
 242 overall prediction accuracy 90% and LDA was applied.

243 In Table 6 a comparison is made between our best results obtained with  
244 RF[8-SRQA-I] and other existing methods [20, 38, 26, 39, 40, 41]. From  
245 Table 6, it is clearly shown that our best configuration i.e. RF[8-SRQA-I ]  
246 outperformed the recent results of Wang et al [39]. In addition, from Ta-  
247 bles A.7-A.9 we see that RF[ 8-SRQA-I], RF[8-SRQA-R] overpass the recent  
248 results of Wang et al [39] for any type of symmetry. Moreover, the percent-  
249 age accuracy and sensitivity on training and testing are the almost same.  
250 Besides, the RF model is tuned with 6 as maximum depth and results ap-  
251 proached to 100%. Cross-validation shows the same results. Therefore, our  
252 classification model is not overfitting.

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

## 260 5. Conclusion

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

## 273 Appendix A. Appendix

### 274 *Appendix A.1. Recurrence Plot*

275 Eckmann et al. [42] proposed the concept of recurrence plot initially to  
276 identify the presence of identical neighboring points in a time series such as

Dataset	Methods	Classifier	Sensitivity				OA
			All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$	
25PDB	AAD-CGR [20]	LDA	64.3	65	61.7	65	64
	SCPRED[40]	SVM	92.6	80.1	74	71	79.7
	Zhang et Al.[41]	SVM	96.7	80.8	82.4	75.5	83.7
	H. Olyaei[26]	LDA	95.6	89.5	88.1	87	90
	WD PseAAC[38]	SVM	95.7	97.7	94.8	84.4	93.1
	Wang [39]	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 [20]	LDA	62.3	67.7	63.1	66.5	65.2
	SCPRED [40]	SVM	89.1	86.7	89.6	53.8	80.6
	Zhang et Al[41]	SVM	92.4	84.4	84.4	73.4	83.6
	H. Olyaei[26]	LDA	92.3	90.1	86.5	75.2	-
	WD PseAAC[38]	SVM	98.7	99	94	68.9	90.8
	Wang et Al[39]	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	

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

277  $x(n) = x_1, x_2, \dots, x_N$ . This time series is embedded into a phase space with  
278 an embedding dimension  $d$  and a time delay  $\tau$ . Two points such as  $X(i)$   
279 and  $X(j)$  in the  $d$ -dimensional space are considered recurrent if they satisfy  
280 the following test [42]:

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

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

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

283 The recurrence matrix  $M$  is a binary matrix composed of zeros and ones  
284 where zero components present the same state and non-zero components  
285 exhibit different states. Besides, the right selection of  $d$  is very important,  
286 as incorrect selection will lead towards recurrences contamination [35, 43]

287 (false recurrences/sojourn points). In addition, The embedding dimension is  
 288 usually set with  $d \geq 2$  to avoid the presence of sojourn points. The increase  
 289 in dimensionality usually reduces the number of false recurrences; however,  
 290 other approaches are proposed by Zaylaa et al.[44].

291 *Appendix A.2. Recurrence Quantification Analysis*

292 The Recurrence Quantification Analysis (RQA) extracts quantitative fea-  
 293 tures (descriptors) from the binary matrix M. It permits to measure differ-  
 294 ently appearing recurrence plots (RPs) with the help of small-scale structures  
 295 present inside it. A main advantage of RQA is to give effective information  
 296 for non-stationary and short data while other techniques fail to do so. It  
 297 may be applied to versatile types of data. Moreover, to quantify the com-  
 298 plexity, different measures of RQA introduced heuristically in [35, 36, 37] as  
 299 described below.

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

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

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

$$DET = \frac{\sum_{l=l_{min}}^N lp(l)}{N^2(RR)} \quad (\text{A.4})$$

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

308 Entropy: it measures the deterministic structures' complexity within the  
 309 system. It depends on the bin-number sensitively.

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

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

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

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

$$LAM = \frac{\sum_{v=v_{min}}^N vp(v)}{N^2(RR)} \quad (\text{A.7})$$

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

316 Maxline is the longest length of the diagonal line.

$$L_{MAX} = \max(l_i; i = 1, \dots, N_l) \quad (\text{A.8})$$

317 Meanline: the vertical and diagonal line's length can be measured. There-  
 318 fore, the average diagonal line length is called the meanline which is associ-  
 319 ated with the predictability interval of the dynamic system.

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

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

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

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

$$TREND = \frac{\sum_{i=1}^{\bar{N}} (i - \frac{\bar{N}}{2})(RR - \langle RR \rangle)}{\sum_{i=1}^{\bar{N}} (i - \frac{\bar{N}}{2})^2} \quad (\text{A.11})$$

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



328 *Appendix A.3. Tables*

329 Tables A.7-A.9 present sensitivity (%) obtained with the three different  
 330 classifiers (SVM, LDA and RF). In each table, the two benchmark data sets  
 331 25PDB and 1189 are analyzed. For each row, All- $\alpha$ , All- $\beta$ ,  $\alpha/\beta$  and  $\alpha + \beta$   
 332 are tested.

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

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

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

Dataset	Methods	Sensitivity			
		All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$
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.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.

344 **References**

- 345 [1] K.-C. Chou, C. Zhang, Prediction of protein structural classes, *Critical reviews in biochemistry and molecular biology* 30 (1995) 275–349.  
 346 doi:10.3109/10409239509083488.  
 347
- 348 [2] I. Bahar, A. R. Atilgan, R. L. Jernigan, B. Erman, Understanding  
 349 the recognition of protein structural classes by amino acid composition,  
 350 *Proteins: Structure, Function, and Bioinformatics* 29 (1997) 172–185.

Dataset	Methods	Sensitivity			
		All- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$
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.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- $\alpha$	All- $\beta$	$\alpha/\beta$	$\alpha + \beta$
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

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.

- 351 [3] K. Nishikawa, T. Ooi, Correlation of the amino acid composition of  
352 a protein to its structural and biological characters, The Journal of  
353 Biochemistry 91 (1982) 1821–1824.
- 354 [4] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature  
355 261 (1976) 552–558.
- 356 [5] L. Carlacci, K. C. Chou, G. M. Maggiora, A heuristic approach to  
357 predicting the tertiary structure of bovine somatotropin, Biochemistry  
358 30 (1991) 4389–4398.
- 359 [6] K.-C. Chou, Energy-optimized structure of antifreeze protein and its  
360 binding mechanism, Journal of molecular biology 223 (1992) 509–517.
- 361 [7] M. M. Gromiha, S. Selvaraj, Protein secondary structure prediction in  
362 different structural classes., Protein engineering 11 (1998) 249–251.

- 363 [8] L. A. Kurgan, L. Homaeian, Prediction of structural classes for pro-  
364 tein sequences and domains—impact of prediction algorithms, sequence  
365 representation and homology, and test procedures on accuracy, *Pattern*  
366 *Recognition* 39 (2006) 2323–2343.
- 367 [9] K.-C. Chou, Progress in protein structural class prediction and its im-  
368 pact to bioinformatics and proteomics, *Current Protein and Peptide*  
369 *Science* 6 (2005) 423–436.
- 370 [10] K.-C. Chou, A novel approach to predicting protein structural classes in  
371 a (20–1)-d amino acid composition space, *Proteins: Structure, Function,*  
372 *and Bioinformatics* 21 (1995) 319–344.
- 373 [11] J. K. Kim, S.-Y. Bang, S. Choi, Sequence-driven features for prediction  
374 of subcellular localization of proteins, *Pattern recognition* 39 (2006)  
375 2301–2311.
- 376 [12] Y.-D. Cai, J. Hu, X. Liu, K.-C. Chou, Prediction of protein structural  
377 classes by neural network method, *J Mol Des* 1 (2002) 332–338.
- 378 [13] S. Fias, S. Van Damme, P. Bultinck, Multidimensionality of delocal-  
379 ization indices and nucleus independent chemical shifts in polycyclic  
380 aromatic hydrocarbons, *Journal of computational chemistry* 29 (2008)  
381 358–366.
- 382 [14] H. Lin, C. Ding, Q. Song, P. Yang, H. Ding, K.-J. Deng, W. Chen,  
383 The prediction of protein structural class using averaged chemical shifts,  
384 *Journal of Biomolecular Structure and Dynamics* 29 (2012) 1147–1153.
- 385 [15] Z. Feng, X. Hu, Z. Jiang, H. Song, M. A. Ashraf, The recognition of  
386 multi-class protein folds by adding average chemical shifts of secondary  
387 structure elements, *Saudi Journal of Biological Sciences* 23 (2016) 189–  
388 197.
- 389 [16] Y. Liang, S. Zhang, Predict protein structural class by incorporating two  
390 different modes of evolutionary information into chou’s general pseudo  
391 amino acid composition, *Journal of Molecular Graphics and Modelling*  
392 78 (2017) 110–117.
- 393 [17] X.-D. Sun, R.-B. Huang, Prediction of protein structural classes using  
394 support vector machines, *Amino acids* 30 (2006) 469–475.

- 395 [18] T. Liu, X. Zheng, J. Wang, Prediction of protein structural class for low-  
396 similarity sequences using support vector machine and psi-blast profile,  
397 *Biochimie* 92 (2010) 1330–1334.
- 398 [19] L. Li, X. Cui, S. Yu, Y. Zhang, Z. Luo, H. Yang, Y. Zhou, X. Zheng,  
399 Pssp-rfe: accurate prediction of protein structural class by recursive  
400 feature extraction from psi-blast profile, physical-chemical property and  
401 functional annotations, *PLoS One* 9 (2014) e92863.
- 402 [20] J.-Y. Yang, Z.-L. Peng, Z.-G. Yu, R.-J. Zhang, V. Anh, D. Wang, Pre-  
403 diction of protein structural classes by recurrence quantification analysis  
404 based on chaos game representation, *Journal of Theoretical Biology* 257  
405 (2009) 618–626.
- 406 [21] P. Sudha, D. Ramyachitra, P. Manikandan, Enhanced artificial neural  
407 network for protein fold recognition and structural class prediction, *Gene*  
408 *Reports* 12 (2018) 261–275.
- 409 [22] A. Anand, G. Pugalenth, P. Suganthan, Predicting protein structural  
410 class by svm with class-wise optimized features and decision probabili-  
411 ties, *Journal of theoretical biology* 253 (2008) 375–380.
- 412 [23] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, L. Hao, Predicting protein  
413 structural classes for low-similarity sequences by evaluating different fea-  
414 tures, *Knowledge-Based Systems* 163 (2019) 787–793.
- 415 [24] W. Bao, Y. Chen, D. Wang, Prediction of protein structure classes with  
416 flexible neural tree, *Bio-medical materials and engineering* 24 (2014)  
417 3797–3806.
- 418 [25] Z. Aydin, A. Singh, J. Bilmes, W. S. Noble, Learning sparse models for  
419 a dynamic bayesian network classifier of protein secondary structure,  
420 *BMC bioinformatics* 12 (2011) 154.
- 421 [26] M. H. Olyae, A. Yaghoubi, M. Yaghoobi, Predicting protein structural  
422 classes based on complex networks and recurrence analysis, *Journal of*  
423 *theoretical biology* 404 (2016) 375–382.
- 424 [27] J.-M. Girault, Recurrence and symmetry of time series: Application to  
425 transition detection, *Chaos, Solitons & Fractals* 77 (2015) 11–28.

- 426 [28] R. Xu, M. Li, H. Chen, Y. Huang, Y. Xiao, A symmetry-related  
427 sequence-structure relation of proteins, *Chinese Science Bulletin* 50  
428 (2005) 536.
- 429 [29] P. Deschavanne, P. Tuffery, Exploring an alignment free approach for  
430 protein classification and structural class prediction, *Biochimie* 90 (2008)  
431 615–625.
- 432 [30] H. K. Kwan, S. B. Arniker, Numerical representation of dna sequences,  
433 in: 2009 IEEE International Conference on Electro/Information Tech-  
434 nology, IEEE, 2009, pp. 307–310.
- 435 [31] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic acids*  
436 *research* 18 (1990) 2163–2170.
- 437 [32] V. R. F, Evolution of long-range fractal correlations and  $1/f$  noise in  
438 dna base sequences, *Rev. Lett* 68 (1992) 3805–8.
- 439 [33] S. Conte, A. Giuliani, Identification of possible differences in coding  
440 and non-coding fragments of dna sequences by using the method of  
441 the recurrence quantification analysis, arXiv preprint arXiv:0910.3516  
442 (2009).
- 443 [34] J.-M. Girault, A. Humeau-Heurtier, Centered and averaged fuzzy en-  
444 tropy to improve fuzzy entropy precision, *Entropy* 20 (2018) 287.
- 445 [35] C. L. Webber Jr, J. P. Zbilut, Dynamical assessment of physiological  
446 systems and states using recurrence plot strategies, *Journal of applied*  
447 *physiology* 76 (1994) 965–973.
- 448 [36] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, J. Kurths,  
449 Recurrence-plot-based measures of complexity and their application to  
450 heart-rate-variability data, *Physical review E* 66 (2002) 026702.
- 451 [37] L. Trulla, A. Giuliani, J. Zbilut, C. Webber Jr, Recurrence quantification  
452 analysis of the logistic equation with transients, *Physics Letters A* 223  
453 (1996) 255–260.
- 454 [38] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, B. Tian,  
455 Prediction of protein structural class for low-similarity sequences using  
456 chou’s pseudo amino acid composition and wavelet denoising, *Journal*  
457 *of Molecular Graphics and Modelling* 76 (2017) 260–273.

- 458 [39] S. Wang, X. Wang, Prediction of protein structural classes by different  
459 feature expressions based on 2-d wavelet denoising and fusion, *BMC*  
460 *bioinformatics* 20 (2019) 701.
- 461 [40] L. Kurgan, K. Cios, K. Chen, Scpred: accurate prediction of protein  
462 structural class for sequences of twilight-zone similarity with predicting  
463 sequences, *BMC bioinformatics* 9 (2008) 226.
- 464 [41] L. Zhang, X. Zhao, L. Kong, A protein structural class prediction  
465 method based on novel features, *Biochimie* 95 (2013) 1741–1744.
- 466 [42] J. Eckmann, S. O. Kamphorst, D. Ruelle, et al., Recurrence plots of  
467 dynamical systems, *World Scientific Series on Nonlinear Science Series*  
468 *A* 16 (1995) 441–446.
- 469 [43] T. March, S. Chapman, R. Dendy, Recurrence plot statistics and the  
470 effect of embedding, *Physica D: Nonlinear Phenomena* 200 (2005) 171–  
471 184.
- 472 [44] A. Zaylaa, J. Charara, J.-M. Girault, Reducing sojourn points from  
473 recurrence plots to improve transition detection: Application to fetal  
474 heart rate transitions, *Computers in biology and medicine* 63 (2015)  
475 251–260.