Efficient Dynamic Analysis of Low-similarity Proteins for Structural Class Prediction

M.A. Zervou[1,2], E. Doutsi[1], P. Pavlidis[1], P. Tsakalides[1,2]
[1]*Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece*
[2]*Department of Computer Science, University of Crete, Heraklion, Greece*
E-mails: zervou@ics.forth.gr, edoutsi@ics.forth.gr, pavlidis@ics.forth.gr, tsakalid@ics.forth.gr

***Abstract*—Prediction of protein structural classes from amino acid sequences is a challenging problem as it is profitable for analyzing protein function, interactions, and regulation. The majority of existing prediction methods for low-homology sequences utilize numerous amount of features and require an exhausting search for optimal parameter tuning. To address this problem, this work proposes a novel self-tuned architecture for feature extraction by modeling directly the inherent dynamics of the data in higher-dimensional phase space via chaos game representation (CGR) and generalized multidimensional recurrence quantification analysis (GmdRQA). Experimental evaluation on a real benchmark dataset demonstrates the superiority of the herein proposed architecture when compared against the state-of-the-art unidimensional RQA taking under consideration that our method achieves similar performance in a data-driven manner with a smaller computational cost.**

***Index Terms*—Protein structure prediction, chaos game representation, multidimensional recurrence quantification analysis, nonlinear time series analysis**

I. Introduction

Protein structure prediction is one of the most significant and challenging problems in bioinformatics, as it has a prominent role in understanding the function and evolution of proteins. Acquiring knowledge of the dynamics and properties that force each protein into a unique 3D structure is highly important and useful in medicine and biotechnology for drug design, enzymes composition, interpretation of disease-related phenotype, etc. [1]. In essence, the biological function of a protein is associated with its tertiary structure, which is determined by its amino acid sequence via the process of protein folding [2]. Along these lines, the tertiary protein structure can be classified into four structural categories based on the protein's folding patterns known as (i) all α-fold, in which are classes of structural domains that are mainly composed of α-helices and only a little amount of β-strands (a.k.a. β-sheets), (ii) all β-fold, that is mostly formed by β-strands and a few isolated α-helices, (iii) the $\alpha + \beta$-fold, forming helices and mostly β anti-parallel strands, and the (iv) α/β-fold, that consists of helices and almost all parallel strands [3].

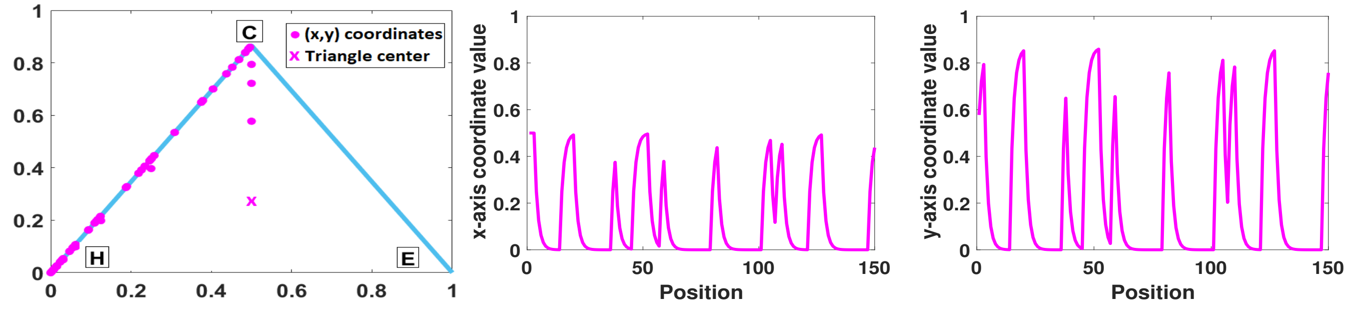
In the last decade, a plethora of machine learning based algorithms has been developed for the protein structural class prediction. Some widely used methods for extracting meaningful features to represent protein samples are the amino acid composition [4], the pseudo-amino acid composition [5], the dipeptide and tripeptide compositions [6], the PSI-BLAST profile [7] and the predicted secondary structure information [8], among others. However, the performance of these techniques is less significant when low-similarity proteins are encountered. Hence, a lot of effort has been made to improve the prediction accuracy for low-homology proteins [9]–[14] with numerous studies [15]–[18] demonstrating the potential of using chaos game representation (CGR) [19] along with recurrence quantification analysis (RQA) [20]. In particular, CGR is a method of converting a long unidimensional sequence into a graphical form where the x- and y-coordinates of each point on this graph are considered as two individual time series. The RQA is a powerful non-linear analysis tool, capable of dealing with non stationary time-series of varying lengths while extracting several features. Although the combination of the aforementioned methods leads to high-precision results, their enhanced performance comes at the expense of time and memory complexity. This is a result mainly from the fact that (i) a pair of coordinates has to be processed separably resulting in large number of features and (ii) the fine-tuning of the RQA parameters for each time-series is highly demanding.

To overcome these limitations, we propose (i) the generalized multidimensional recurrence quantification analysis (GmdRQA) [21] as a sophisticated, non-linear analysis of the multidimensional time series data that allows us to exploit both the intra- and inter-data correlations in an automated fashion and (ii) a data-driven fine-tuning (DDFT) of the GmdRQA parameters based on the Average Mutual Information (AMI) and the False Nearest Neighbors (FNN) methods. To the best of our knowledge, there is no prior work in the literature that employs multidimensional recurrence quantification analysis for the protein structural class prediction especially in an automatic parameter selection. The contributions of this paper are summarized below:

(i) The design of a novel protein structural class prediction scheme that combines CGR and a unidimensional data-driven fine-tuned RQA (DDFT-1DRQA).

(ii) The design of a protein structural class prediction scheme that combines CGR and a multidimensional grid search tuned RQA (GS-GmdRQA).

This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under HFRI faculty grant no. 1725, and by the Stavros Niarchos Foundation within the framework of the project ARCHERS.



(a) Chaos Game Representation

(b) Time series t_x

(c) Time series t_y

Fig. 1: (a) Chaos Game Representation of protein's 1ASH predicted secondary structure based on the three secondary structural elements, C (coil), E (strand) and H (helix). (b),(c) Time series representation of 1ASH protein based on CGR.

(iii) The design of a novel protein structural class prediction scheme that combines CGR and data-driven fine-tuned GmdRQA (DDFT-GmdRQA).

The rest of the paper is organized as follows: Section II is an overview of the dataset employed by our study. Section III is an introduction to CGR, GmdRQA and the data-driven parameter-tuning algorithms, namely AMI and FNN. Section IV evaluates the performance of the proposed schema to the state-of-the-art in terms of classification accuracy using support vector machines (SVM) and Fisher's linear discriminant algorithm (FDA). Finally, Section V draws the conclusion of this work and gives directions for future extensions.

II. Dataset description

This work employs the 25PDB dataset [22] that includes 1673 proteins of varying length with 25% sequence homology. The proteins are categorised based on their structural class with 443 of them belonging to α-fold, 443 to β-fold, 346 to α/β-fold and 441 to $\alpha+\beta$-fold. Each protein is a long chain called polypeptide or polymer, which is formed when several monomers, known as amino acids, are joined together. Only 20 amino acids are known as proteinogenic, meaning they participate in the synthesis of a protein primary structure. In the literature, there are several works [16]–[18] that instead of dealing with the protein primary structure they use the PSI-PRED tool [23] that predicts the role of each amino acid in the protein secondary structure. Particularly, PSI-PRED transforms the initial amino acid sequence to a sequence of equal length that now consists of only three states that describe its secondary structure, namely coils (C), strands (E) and helices (H). This simplification not only reduces the dimensionality of our data from 20 amino acids to three structural elements but also the overall computational complexity. Hence, in this work we have decided to use as input data the prediction of the protein secondary structure.

III. Methods

A. Chaos Game Representation

In order to transform a unidimensional sequence of secondary structural elements into a two-dimensional time series, Chaos Game Representation (CGR) is employed. More specifically, the sequence of a protein secondary structure is represented in a unit equilateral triangle with its vertices referring to the three secondary structure elements H, C, and E, with xy-plane coordinates (0,0), (0.5,$\sqrt{3}/2$) and (1,0), respectively. Then, the triangle centroid is defined at position (0.5,$\sqrt{3}/6$) while the first element of the sequence is calculated as the halfway distance point between the centre of the triangle and the vertex representing this element. Accordingly, the remaining consecutive elements in the sequence are plotted as the midpoint between the previous plotted point and the vertex representing the element being plotted as given in Eq.1,

$$(x_i, y_i) = \left(\frac{x_{i-1}+v_{ix}}{2}, \frac{y_{i-1}+v_{iy}}{2}\right) \qquad (1)$$

where v_{ix} and v_{iy} are respectively the x and y coordinates of the vertices corresponding to the secondary structure element at position i in the sequence. The resulted graph is given in Fig. 1(a). Finally, as depicted in Fig.1 (b) and (c), the CGR graph is decomposed into two time series that consist accordingly of the x and y coordinates that CGR algorithm produced so that $t_x = (x_1, x_2, \ldots, x_n)$ and $t_y = (y_1, y_2, \ldots, y_n)$, where n is the number of elements in the sequence.

B. Generalized Multidimensional Recurrence Quantification Analysis

Multidimensional recurrence quantification analysis extracts in general the underlying dynamics of the system by mapping the time series in a higher-dimensional phase space of trajectories by constructing state vectors via time delay embedding. The GmdRQA framework introduced in [21], transforms state vectors into state matrices in order to represent the time-delay embedding, since state matrices are considered more appropriate for describing multidimensional signals from a mathematical perspective, enabling them to model the correlations not only within a signal but also between different signals. Their experimental evaluation on real data revealed the superiority of GmdRQA framework, when compared against its vector-based counterpart, in terms of classification accuracy and robustness to noise.

More specifically, given a *multidimensional time series* r of length N the corresponding phase space representation is reconstructed as follows,

$$\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_{N_s} \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{1,1} & \mathbf{w}_{2,1} & \cdots & \mathbf{w}_{D,1} \\ \mathbf{w}_{1,2} & \mathbf{w}_{2,2} & \cdots & \mathbf{w}_{D,2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{w}_{1,N_s} & \mathbf{w}_{2,N_s} & \cdots & \mathbf{w}_{D,N_s} \end{pmatrix}, \quad (2)$$

where D is the number of dimensions of the data, $\mathbf{w}_{i,j} = (r_j^{(i)}, r_{j+\tau}^{(i)}, \ldots, r_{j+(m-1)\tau}^{(i)})$, $i = 1, \ldots, D$, $j = 1, \ldots, N_s$, with m being the embedding dimension, τ the delay and $N_s = N - (m-1)\tau$ the number of states. The state vectors u_j can be transformed into state matrices of the form $\mathbf{X}_j = (\mathbf{w}_{1,j}, \mathbf{w}_{2,j}, \ldots, \mathbf{w}_{D,j})^T$.

In the herein work, the dimensionality of the data is two, based on the CGR analysis mentioned in Section III-A. Thus, GmdRQA is directly applied on the respected two dimensional time series.

Recurrence plots (RPs) [20] have been proposed as an advanced graphical technique of visual non-linear data analysis, which reveals all the times of recurrences, that is, when the phase space trajectory of the dynamical system visits roughly the same area in the phase space. Accordingly, the generalized multidimensional recurrence plot (GmdRP) is defined by

$$\mathbf{GmdR}_{i,j} = \Theta\left(\varepsilon - ||\mathbf{X}_l - \mathbf{X}_k||_F\right) \ , \quad (3)$$

where $k, l = 1, \ldots, N_s$, ε is a threshold, $||\cdot||_F$ is the Frobenius norm between state matrices, and $\Theta(\cdot)$ is the Heaviside step function whose discrete form is defined by,

$$\Theta(n) = \begin{cases} 1, & \text{if } n \geq 0 \\ 0, & \text{if } n < 0 \end{cases} , \ n \in \mathbb{R} \ . \quad (4)$$

The visual interpretation of RPs, which is often difficult and subjective, is enhanced by means of several numerical measures for the quantification of the structure and complexity of RPs. In this work the following 10 measures are utilized in order to later form our feature matrix (ref. [24] for the mathematical definitions):

- **Recurrence rate:** Measures the density of points in the RP or in other words, the probability that a similar state recurs to its neighbourhood in phase space.
- **Determinism:** The ratio of the number of recurrence points forming diagonal structures to the total number of recurrence points is regarded as determinism or predictability of the system.
- **Average diagonal length:** Refers to the mean time that the next recurrence of states can predicted from the state that is now observed. These lines indicate how different trajectories diverge during the evolution of the system and as time passes by.
- **Length of longest diagonal/vertical line:** Refers to the maximum length of the diagonal/vertical lines in the recurrence plot that represent the maximum time that the system evolves or remains in a certain state respectively.
- **Entropy of diagonal/vertical length:** Indicates the complexity of the recurrence plot in respect of the diagonal/vertical lines. The entropy of vertical lines reflects the distribution of time-periods for which the system abides in laminar phases.
- **Laminarity:** Provides information about the occurrence of the laminar states in the system. However, it does not describe the length of the laminar states.
- **Trapping time:** The average length of vertical lines is called trapping time and it is related with laminarity time. This value contains information about the frequency and the length of laminar states.
- **Size of the phase space:** Refers to the number of states created via time delay embedding.

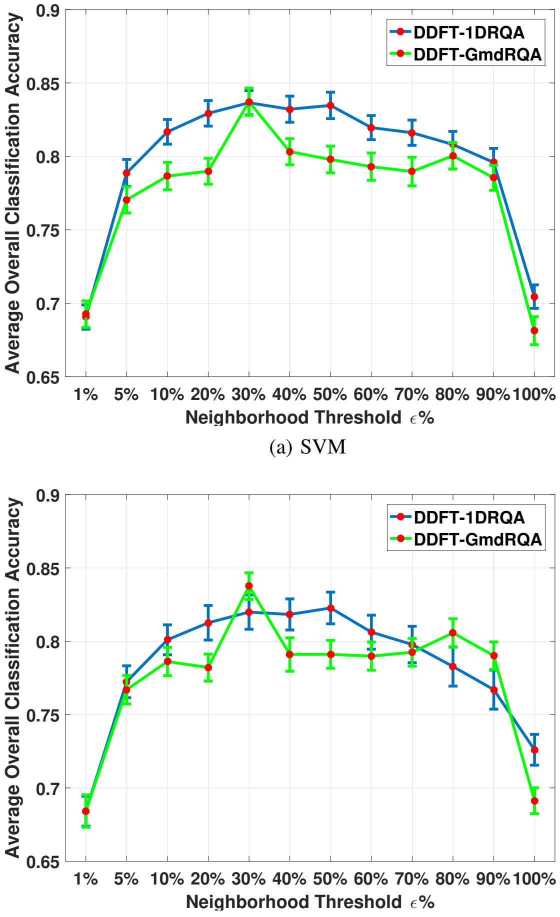
IV. Performance Evaluation

The herein work proposes three novel architectures for feature extraction and classification that combine CGR and RQA. Namely, a unidimensional RQA scheme with data-driven parameter selection, a multidimensional RQA scheme with a grid search of parameters and finally a multidimensional protein structural class prediction scheme with a data-driven parameter selection. Moreover, the case of parameter selection with GmdRQA is also evaluated in a grid search manner (GS-GmdRQA). The overall architecture of the proposed protein structure prediction frameworks that are implemented in MATLAB, on a desktop computer equipped with a CPU processor (Intel Core i5-4590) clocked at 3.30GHz, and a 8 GB RAM.

A. Estimation of RQA parameters

As stated in [15]–[18] parameters such as τ and m are chosen through an exhausting grid search manner. In this study grid search of parameters is also employed in terms of fair comparison to the literature, however another more efficient approach is suggested. Characteristically, the optimal parameter estimation is performed in an automated fashion for each protein sequence separately. Specifically, the optimal time delay τ is estimated as the first minimum of the AMI function averaged over all the dimensions in the data [25]. In the implementation of AMI max lag is set to 10 and the number of bins for calculating the histogram is 10. Concerning the embedding dimension m, a minimal sufficient value is estimated using FNN [26]. In particular, the optimal m is reached when one of the following conditions is satisfied: *i)* FNN drops to 0, *ii)* sequential embeddings have the same number of false neighbors, *iii)* the point before which the number of FNNs starts to increase again, *iv)* the point where the difference between the number false neighbors is less than a threshold T_{tol}. Based on the dataset employed in this work, optimal $T_{\text{tol}} = 2\%$. For proteins of length lower than 45 time delay embedding parameters are set to $\tau = 1$ and $m = 2$, respectively.

Concerning the neighborhood threshold ε, following the empirical rule proposed in [27] it is given as a percentage of the distribution of all the pairwise distances that describe the phase space trajectory. Specifically, the average classification accuracy as a function of the neighborhood threshold ε for both DDFT-1DRQA and DDFT-GmdRQA using SVM and FDA classifier is depicted in Fig. 2. Both approaches achieve higher average overall accuracy when the percentage lies between 20%-50% for SVM and FDA as well as. The variation is roughly the same. Hence, we select the percentage of ε to be 30% for DDFT-1DRQA and DDFT-GmdRQA. This rule also applies in the case of time delay embedding parameters estimation using a grid search approach.



(a) SVM

(b) FDA

Fig. 2: Mean classification accuracy as function of the percentage of the neighborhood threshold ε for DDFT-1DRQA and DDFT-GmdRQA with SVM and FDA.

B. Classification

In this analysis, data are split into 70%-30% for training and testing, respectively. Each experiment is repeated 150 times, and the average overall accuracy is reported. Initially, the feature matrix is z-score normalized. Then, the performance of each architecture is evaluated by employing two well known classifiers independently. Particularly, a Gaussian-kernel support vector machine (SVM) as well as Fisher's Linear Discriminant Analysis (FDA) are applied separately on the normalized feature matrix for discriminating between the four structural classes α-fold, β-fold, α/β-fold and $\alpha+\beta$-fold. The regularization parameter C and kernel width parameter γ can take all positive values log-scaled in the range $[10^{-3}, 10^{3}]$.

C. Evaluation results

As proposed in the state-of-the-art, when DDFT-GmdRQA is performed separately in each unidimensional time series results in the DDFT-1DRQA architecture. The optimal set of parameters is evaluated in an automated fashion, using FNN and AMI, in each of the two time series and 20 features (10 features×2 dimensions) are extracted in total. Table I indicates the performance of the proposed DDFT-1DRQA against the state-of-the-art in terms of average overall classification accuracy. As shown, the automated selection of optimal parameters for each one of the proteins slightly increases the overall classification accuracy for both SVM and FDA classifiers when compared to its state-of-the-art counterparts [15], [16]. Furthermore, its self-tune manner reduces by far the time complexity. Along these lines, in order to be consistent with the literature, this work also investigates the performance of the proposed unidimensional framework with a grid search of parameters (GS-1DRQA). The range of embedding dimension m and time delay τ is 1 to 8. In the case where both m and τ are equal to 8, the phase space can not be constructed for small proteins, hence no results are reported and the grid search procedure is terminated. The optimal parameters are $m = 7$ and $\tau = 1$, however, the overall accuracy remains roughly unaltered for all the other combinations for both SVM and FDA. As indicated in Table I the parameter selection based on grid search achieves the highest accuracy that is though an improvement on a quit small scale. However, it is extremely time consuming and requires the utilization of the maximum number of features.

Multidimensional RQA is employed as it processes the data concurrently, exploits the intra- and inter-data correlations but also produces half the number of features when compared to the proposed unidimensional RQA scheme. It is remarkable to notice that in comparison with the minimum number of features reported in the state-of-the-art, namely 16, GmdRQA utilizes only 10 features. The performance of GmdRQA is initially evaluated using a grid search estimation of time delay embedding parameters where m and τ are set equal for every single protein. The optimal set of parameters is $m = 5$ and $\tau = 7$, although, as mentioned previously for the GS-1DRQA, the performance for both SVM and FDA classifiers is consistent for all different pairs. The grid search setup remains the same as in the GS-1DRQA scenario. The results indicate that the efficiency of GS-GmdRQA remains unaffected as time delay embedding parameter combinations differentiate. Thereafter, DDFT-GmdRQA is examined. As given in Table I, the overall classification accuracy does not provide particular variation among most of the approaches. It is though worth mentioning that DDFT-GmdRQA is the most efficient scheme as it yields a high overall accuracy, utilizing the minimum number of features that has been reported so far in the state-of-the-art. Therefore, DDFT-GmdRQA not only reduces the

	Methods	Number of Features	Overall Accuracy SVM	Overall Accuracy FDA	Computational Complexity (sec)
Yang *et al.* [15]	CGR,RQA	16	-	0.6508	-
Yang *et al.* [16]	PSIPRED, CGR, RQA	16	-	0.8140	-
This work	PSIPRED, CGR, GS-1DRQA	20	0.8503	0.8489	25671.78
This work	PSIPRED, CGR, GS-GmdRQA	10	0.8371	0.8409	13968.37
This work	PSIPRED, CGR, DDFT-1DRQA	20	0.8369	0.8218	491.11
This work	PSIPRED, CGR, DDFT-GmdRQA	10	0.8388	0.8360	262.89

TABLE I: Summarized predicted quality results for 25PDB dataset.

time complexity due to its automated parameter estimation, but the memory complexity as well.

V. Conclusions and future work

In this work, we designed and implemented three novel self-tuned protein structure prediction architectures based on the representation of secondary structure data in higher-dimensional phase spaces using unidimensional and multi-dimensional RQA for capturing the underlying dynamics of the data. The experimental evaluation on real data revealed the superiority of the DDFT-GmdRQA-based framework in extracting and exploiting the underlying temporal dynamics of the data generating processes, resulting in lower time and memory complexity, when compared against the state-of-the-art unidimensional RQA and the herein proposed DDFT-1DRQA approach.

An extension of this work will consider to combine the proposed DDFT-GmdRQA feature extraction framework with other feature extraction schemes and evaluate the overall prediction accuracy utilizing several classification algorithms as well as feature selection techniques. Later, this work can be further expanded by directly processing on the primary amino acid protein sequence without employing intermediate tools such as PSI-PRED and CGR.

References

[1] M. E. Noble, J. A. Endicott, and L. N. Johnson, "Protein kinase inhibitors: insights into drug design from structure," *Science*, vol. 303, no. 5665, pp. 1800–1805, 2004.

[2] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[3] M. Levitt and C. Chothia, "Structural patterns in globular proteins," *Nature*, vol. 261, no. 5561, pp. 552–558, 1976.

[4] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *The Journal of Biochemistry*, vol. 99, no. 1, pp. 153–162, 1986.

[5] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[6] R.-y. Luo, Z.-p. Feng, and J.-k. Liu, "Prediction of protein structural class by amino acid and polypeptide composition," *European Journal of Biochemistry*, vol. 269, no. 17, pp. 4219–4225, 2002.

[7] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and psi-blast profile," *Biochimie*, vol. 92, no. 10, pp. 1330–1334, 2010.

[8] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," *Journal of theoretical biology*, vol. 267, no. 3, pp. 272–275, 2010.

[9] J.-Y. Yang, Z.-G. Yu, and V. Anh, "Protein structure classification based on chaos game representation and multifractal analysis," in *2008 Fourth International Conf. on Natural Computation*, vol. 4. IEEE, 2008, pp. 665–669.

[10] S. Zhang, F. Ye, and X. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via pssm," *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 6, pp. 1138–1146, 2012.

[11] Y. Liang, S. Liu, and S. Zhang, "Prediction of protein structural classes for low-similarity sequences based on consensus sequence and segmented pssm," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[12] J. Wang, C. Wang, J. Cao, X. Liu, Y. Yao, and Q. Dai, "Prediction of protein structural classes for low-similarity sequences using reduced pssm and position-based secondary structural features," *Gene*, vol. 554, no. 2, pp. 241–248, 2015.

[13] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, and B. Tian, "Prediction of protein structural class for low-similarity sequences using chou's pseudo amino acid composition and wavelet denoising," *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 260–273, 2017.

[14] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowledge-Based Systems*, vol. 163, pp. 787–793, 2019.

[15] J.-Y. Yang, Z.-L. Peng, Z.-G. Yu, R.-J. Zhang, V. Anh, and D. Wang, "Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation," *Journal of Theoretical Biology*, vol. 257, no. 4, pp. 618–626, 2009.

[16] J.-Y. Yang, Z.-L. Peng, and X. Chen, "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure," *BMC bioinformatics*, vol. 11, no. 1, p. S9, 2010.

[17] M. H. Olyaee, A. Yaghoubi, and M. Yaghoobi, "Predicting protein structural classes based on complex networks and recurrence analysis," *Journal of theoretical biology*, vol. 404, pp. 375–382, 2016.

[18] H. Jiang, A. Zhang, Z. Zhang, Q. Meng, and Y. Li, "Protein tertiary structure prediction based on multiscale recurrence quantification analysis and horizontal visibility graph," in *International Symposium on Neural Networks*. Springer, 2019, pp. 531–539.

[19] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic acids research*, vol. 18, no. 8, pp. 2163–2170, 1990.

[20] J. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europh. Lett.*, vol. 4, no. 9, p. 973, 1987.

[21] M. A. Zervou, G. Tzagkarakis, and P. Tsakalides, "Automated screening of dyslexia via dynamical recurrence analysis of wearable sensor data," in *2019 IEEE 19th International Conf. on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2019, pp. 770–774.

[22] http://biomine.cs.vcu.edu/datasets/SCPRED/SCPRED.html.

[23] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of molecular biology*, vol. 292, no. 2, pp. 195–202, 1999.

[24] J. Zbilut and C. Webber, "Embeddings and delays as derived from quantification of recurrence plots," *Phys. Lett. A*, vol. 171, pp. 199–203, 1992.

[25] I. Vlachos and D. Kugiumtzis, "State space reconstruction from multiple time series," in *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conf.* World Scientific, 2009, pp. 378–387.

[26] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical review A*, vol. 45, no. 6, p. 3403, 1992.

[27] K. H. Krämer, R. V. Donner, J. Heitzig, and N. Marwan, "Dimension-scalable recurrence threshold estimation," *arXiv preprint arXiv:1802.01605*, 2018.