



Article

Prediction of Protein Subcellular Localization Based on Fusion of Multi-view Features

Bo Li ¹, Lijun Cai ^{1,*}, Bo Liao ^{1,2,*}, Xiangzheng Fu ¹, Pingping Bing ³ and Jialiang Yang ^{2,*}

¹ College of Information Science and Engineering, Hunan University, Changsha 410082, China; hn.libo@163.com (B.L.); excelsior511@126.com (X.F.)

² School of Mathematics and Statistics, Hainan Normal University, Haikou 570100, China

³ Academics Working Station, Changsha Medical University, Changsha 410219, China; bpping@163.com

* Correspondence: ljcai@hnu.edu.cn (L.C.); dragonbw@163.com (B.L.); jialiang.yang@mssm.edu (J.Y.)

Academic Editor: Quan Zou

Received: 24 December 2018; Accepted: 28 February 2019; Published: 6 March 2019



Abstract: The prediction of protein subcellular localization is critical for inferring protein functions, gene regulations and protein-protein interactions. With the advances of high-throughput sequencing technologies and proteomic methods, the protein sequences of numerous yeasts have become publicly available, which enables us to computationally predict yeast protein subcellular localization. However, widely-used protein sequence representation techniques, such as amino acid composition and the Chou's pseudo amino acid composition (PseAAC), are difficult in extracting adequate information about the interactions between residues and position distribution of each residue. Therefore, it is still urgent to develop novel sequence representations. In this study, we have presented two novel protein sequence representation techniques including Generalized Chaos Game Representation (GCGR) based on the frequency and distributions of the residues in the protein primary sequence, and novel statistics and information theory (NSI) reflecting local position information of the sequence. In the GCGR + NSI representation, a protein primary sequence is simply represented by a 5-dimensional feature vector, while other popular methods like PseAAC and dipeptide adopt features of more than hundreds of dimensions. In practice, the feature representation is highly efficient in predicting protein subcellular localization. Even without using machine learning-based classifiers, a simple model based on the feature vector can achieve prediction accuracies of 0.8825 and 0.7736 respectively for the CL317 and ZW225 datasets. To further evaluate the effectiveness of the proposed encoding schemes, we introduce a multi-view features-based method to combine the two above-mentioned features with other well-known features including PseAAC and dipeptide composition, and use support vector machine as the classifier to predict protein subcellular localization. This novel model achieves prediction accuracies of 0.927 and 0.871 respectively for the CL317 and ZW225 datasets, better than other existing methods in the jackknife tests. The results suggest that the GCGR and NSI features are useful complements to popular protein sequence representations in predicting yeast protein subcellular localization. Finally, we validate a few newly predicted protein subcellular localizations by evidences from some published articles in authority journals and books.

Keywords: protein subcellular localization; protein primary sequence; generalized chaos game representation; statistical method; support vector machine; unitary distance

1. Introduction

Assigning subcellular localizations for a protein is a significant step to elucidate its interaction partners, functions and potential roles in the cellular machinery [1,2]. However, experimental methods to determine subcellular localization usually involve immunolabelling or tagging, which could be

laborious and time-consuming [1,3–5]. With the development of high-throughput genomic and proteomic sequencing techniques, there have been increasing number of protein sequences sequenced and cataloged in the protein data banks. So there is an urgent need for effective and efficient computational methods to predict protein subcellular localizations, especially for species like yeast.

Typical computational methods to predict protein subcellular localizations consist of two steps including: (1) protein sequence representation, in which each primary protein sequence was transformed into a numerical feature vector; and (2) protein classification, in which a classification model was then trained based on the feature vectors and labels of the training samples. Currently, there are generally three categories of sequence representation methods: (1) the amino acids composition based-methods, which calculate the occurrence frequencies of the 20 amino acids, but ignore the sequence-order information of each residue; (2) the Chou's Pseudo Amino Acid Composition (PseAAC)-based methods [6,7], which not only model the amino acid composition information but also incorporate the interactions among adjacent residues. The Chou's PseAAC based-methods achieved about an increase of 20 percent of predicting accuracy than amino acids composition-based methods; (3) the hybrid methods allowing for integrating features from multiple views, which usually increase prediction accuracy [8–10]. After the sequence feature was constructed, various classifiers including covariant discriminant (CDC) [10,11], nearest neighbor (NN) [12,13], support vector machine (SVM) [14], deep learning [15] and ensemble classifier [16,17] were adopted to predict protein subcellular localization.

During the past decades, significant progresses have been made on developing efficient protein sequence representations and subsequent classifiers. For example, Zhang et al. introduced several amino acid hydrophobic patterns and average power-spectral density to define a modified PseAAC. Based on these features, they predicted protein subcellular localization by employing the covariant discriminant predictor [3]. Liao et al. attempted to identify protein subcellular locations based on amino acid composition components and adjacent triune residues [6]. Chen et al. utilized the measure of diversity and increment of diversity on protein primary sequences [18]. Ding et al. represented the apoptosis protein sequences by a novel approximate entropy (ApEn)-based PseAAC and employed an ensemble classifier model as the prediction engine, of which the basic classifier is the fuzzy K-nearest neighbor [16]. Lin et al. refined the PseAAC based on the physico-chemical characteristics of the 20 amino acids, and adopted SVM to predict protein subcellular locations [19]. Zhang et al. introduced the concept of distance frequency to capture the positional distribution information of amino acids and also adopted SVM to classify proteins [2]. More recently, Yu et al. implemented the CELLO2GO (<http://cello.life.nctu.edu.tw/cello2go/>) web-based system server for providing protein subcellular location prediction service based on functional Gene Ontology Annotation [20]. Wan et al. introduced a multi-label subcellular-localization predictor named HybridGO-Loc that leverages not only the GO term occurrences but also the inter-term relationships [21]. Dehzangi et al. proposed two segmentation-based feature extraction methods to explore potential local evolutionary-based information for Gram-positive and Gram-negative subcellular localizations [22]. Finally, Shao et al. employed a deep model-based descriptor (DMD) to extract high-level features from protein images, which was proven to be useful for determining the subcellular localization of proteins [23]. However, due to the limitations of feature representation schemes and the relative low accuracy of classification algorithms, most current algorithms still cannot be widely employed in real applications.

To address this problem, we first introduced two novel feature representations based on Generalized Chaos Game Representation (GCGR) and novel statistics and information theory (NSI), respectively. Using the two types of features, we developed a predicting model based on unitary distances. Our experiments indicate that the model can quickly and accurately predict the subcellular localizations for yeast even without classifiers. To further evaluate the effectiveness of the proposed new features, we proposed a multi-view feature by combining these features with well-known features like PseAAC and dipeptide composition, which was fed into a SVM classification system. We then

tested the performance of the proposed features and models on two yeast benchmark datasets and compared them with a few popular methods using the jackknife test.

2. Results and Discussions

We listed in Tables 1 and 2 the predicting results of the proposed model and other existing models for the jackknife test on CL317 and ZW225 respectively. As can be seen, our model achieved overall prediction accuracies of 0.8825 and 0.7736 respectively on CL317 and ZW225. The performance on CL317 outperforms some existing methods, such as Wei et al. [15] (with accuracy 0.827) and Zhang et al. [24] (with accuracy 0.88). The improvement is important considering that we only used a 2-D GCGR feature and a 3-D NSI feature, while other methods combined features like amino acid composition of 20-D and dipeptide of 400-D. We further tested the performance of combining GCGR and NSI with other widely-recognized features including pseudo-amino acid composition (PwAAC) and dipeptide composition (Dipeptide). Specifically, we applied three models including: (1) PwAAC alone, (2) fusion of features PwAAC and Dipeptide, and (3) fusion of features PwAAC, Dipeptide, GCGR and NSI into the CL317 dataset. Their prediction results for the jackknife test were summarized in Figure 1.

Table 1. The prediction results of dataset CL317 using unitary distance based on GCGR + NSI features in the jackknife test.

	Cy	Me	Nu	En	Mi	Se
Sn (%)	91.8	85.3	86.5	93.7	83.3	74.5
Sp (%)	86.2	99.6	91.9	86.2	91.5	90.9
MCC	0.83	0.83	0.86	0.88	0.86	0.81
Acc	0.8825					

Table 2. The prediction results of dataset ZW225 using unitary distance based on GCGR + NSI features in the jackknife test.

	Me	Cy	Nu	Mi
Sn (%)	0.6617	0.8286	0.88	0.8439
Sp (%)	0.7792	0.7432	0.9383	0.8841
MCC	0.6863	0.6789	0.9115	0.7745
Acc	0.7736			

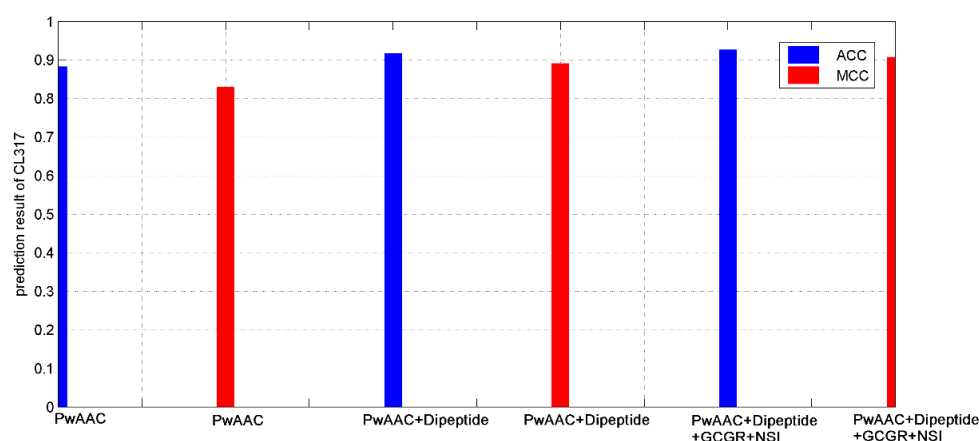


Figure 1. The prediction results based on CL317 using the support vector machine algorithm with different combination of features.

As Figure 1 depicts, the model combined all features achieved much higher prediction accuracy than others, indicating that: (1) feature fusion techniques are promising to improve the prediction accuracy since single-view feature can only reflect part of the information of a protein sequence; (2) the

two features GCGR and NSI can be served as a helpful complementary to features like PwAAC and Dipeptide, revealing the effectiveness of the two novel feature representation techniques as well.

To further evaluate the efficiency of the feature fusion technique and improve protein subcellular location prediction accuracy, we introduced the final multiple-views based model, in which the feature vector for each protein was represented by concatenating numerical vectors from GCGR, NSI, PwAAC and Dipeptide. In addition, SVM was selected as the classifier. Comparison with other existing models using the jackknife test on CL317 and ZW225 were shown in Tables 3 and 4, respectively.

Table 3. Comparison of prediction performance for CL317 in the jackknife test.

Predictor	MCC						Acc
	Cy	Me	Nu	En	Mi	Se	
[15]	0.80	0.77	0.73	0.90	0.74	0.68	0.827
[23]	0.87	0.90	0.86	0.95	0.86	0.80	0.909
[24]	0.84	0.85	0.84	0.91	0.77	0.80	0.88
[25]	0.89	0.88	0.87	0.95	0.88	0.78	0.911
[6]	0.946	0.909	0.885	0.957	0.882	0.706	0.912
This paper	0.896	0.913	0.929	0.892	0.853	0.905	0.921

Table 4. Comparison of prediction performance for ZW225 in the jackknife test.

Predictor	MCC				Acc
	Me	Cy	Nu	Mi	
[3]	0.933	0.90	0.634	0.60	0.831
[15]	0.91	0.929	0.732	0.68	0.858
[24]	0.921	0.871	0.732	0.64	0.84
[6]	0.91	0.871	0.756	0.72	0.849
This paper	0.909	0.892	0.867	0.778	0.889

As can be seen, our integration model achieved the highest overall accuracies, that is, 0.921 and 0.889 on CL317 and ZW225, respectively. There are two indications: (1) our proposed protein sequence feature representations including GCGR and NSI both contain some valuable information such as concentrated local information, which were not covered by previous features; (2) The integration of multiple informative features may improve prediction performance. In addition, our model achieved the highest MCCs for most of subcellular location classes on CL317 except for Cy and Me. Moreover, the MCCs and the Accs of the class Nu and Mi for our model are much better than other existing methods on ZW225.

Finally, we searched authoritative journals and publications for further validation of the predicted subcellular location of some proteins, and found that some of them have already been validated by experiments. For example, we predicted that the protein YHR196W belongs to nucleolar, which have been reported by more than 20 publications such as Eswara et al. [26] and Polymenis et al. [27]. We also predicted Sec17p to be localized in cytoplasm and the endoplasmic reticulum, consistent with Aouida et al. [28]. Thus, our model is effective in screening out potential protein subcellular locations for further experimental validation.

To summarize: first, the new simple unitary distance-based method is comparable to many methods in prediction accuracy; second, the proposed new perspectives (GCGR and NSI) truly contain some valuable information from protein primary sequence, and can be served as a complement to the existing feature representations; third, the multi-feature based model can improve the prediction accuracy notably, thus can be used to help biologist determine protein subcellular location.

However, we are fully aware that there are several limitations in this study. First of all, we only used the average of the x- and y-axis of the points in the GCGR plot, which may retrieve only partial information of the plot. An immediate option is to try other statistics of the GCGR plot such as median

and percentiles. Second, the biological interpretation under the effectiveness of the features is not fully clear. Third, the current version of the software is not very user-friendly. In the future, we will devote to offer an online web service such that more biologists can use the software. We will also try to use some parallel algorithms for dealing with large scale eukaryote species including human data.

3. Materials and Methods

3.1. Datasets

In the paper, two yeast datasets CL317 and ZW225 are used for comparing different predicting models. The CL317 dataset was collected by Chen and Li [18]. The original 846 proteins explicitly annotated to one subcellular were derived from SWISSPROT (version 49.0) by European Bioinformatics Institute, Hinxton Cambridge, United Kingdom (www.ebi.ac.uk/swissprot) [25]. Since short sequences are more like to be homologous and it is also difficult to extract enough information from them, we removed the proteins with less than 80 residues similar to Chen and Li [18]. The remaining dataset contains 317 apoptosis proteins belonging to six subcellular locations including cytoplasmic (Cy), membrane (Me), nuclear (Nu), endoplasmic reticulum (En), mitochondrial (Mi) and secreted (Se) with 112, 55, 52, 47, 34 and 17 proteins, respectively. ZW225 was curated by Zhang and Wang [24], including 225 proteins in four subcellular locations with 89 membrane proteins, 70 cytoplasmic proteins, 41 nuclear proteins and 25 mitochondrial proteins. The proteins were extracted from SWISSPROT (version 50.3) by European Bioinformatics Institute, Hinxton Cambridge, United Kingdom using the same rules as CL317.

3.2. Generalized Chaos Game Representation (GCGR) of Protein Primary Sequences

The chaos game representation (CGR) was initially introduced to visualize DNA sequences [29] and later for protein sequences as well [30]. Here, we further developed a generalized chaos game representation (GCGR) to represent a protein sequence by a 2-dimensional numerical feature vector describing the frequency of 20 amino acids and their neighbor information in the sequence. The construction of GCGR consists of three steps:

3.2.1. Step 1: Convert a Protein Sequence into a Sequence on an Alphabet of Size 6

We converted the 20 amino acids into six groups (Table 5). Specifically, Proline (P), Glycine (G) and Cysteine (C) formed three separate groups because of their unique backbone properties. The remaining 17 amino acids were classified into the other three groups according to their hydropathy scale including strongly hydrophilic (denoted by H), strongly hydrophobic (L), and weakly hydrophilic or weakly hydrophobic (S) [31]. As a result, each primary protein sequence could be uniquely represented by a string on the alphabet {H, L, S, P, G, C}. For example, the protein sequence “YAMQESHFTCI” can be represented by “SLLHSHLSCL” according to Table 5.

Table 5. The six classes of the 20 amino acids.

Classification	Abbreviation	Amino Acids
Strongly hydrophilic or polar	H	H, R, D, E, N, Q, K
Strongly hydrophobic	L	L, I, V, A, M, F
Weakly hydrophilic or weakly hydrophobic (ambiguous)	S	S, T, Y, W
Proline	P	P
Glycine	G	G
Cysteine	C	C

3.2.2. Step 2: Construct the GCCR Plot

Firstly, we drew a regular hexagon, in which each vertex is associated with a distinct label of H, L, S, P, G and C, and each edge is of unit length. Then, for each encoded primary sequence in the

first step, we plotted its letters sequentially as vertices inside the hexagon as follows: the first vertex, corresponding to the first letter of the primary sequence, was placed in the center of the hexagon; and the i -th vertex, corresponding to the i -th letter, was placed in the middle of the first $(i-1)$ -th vertices and the vertex representing the i -th letter in the hexagon. After that, a plot named the GCGR of the primary sequence was drawn. As examples, we plotted in Figure 2 the GCGRs for six representative proteins with each belonging to a different subcellular location. From the six GCGR figures, we can directly retrieve some valuable information: for proteins in the Cy and Nu classes, the plotted points are close to vertices H and L; the protein in the Me class are uniformly distributed around all the vertices except for C; proteins in the Nu and En classes have fewer points around vertices G, C and P, G, respectively; proteins in the last two classes are almost uniformly distributed. In a word, the proteins in different subcellular locations distributed differently in the GCGR plots.

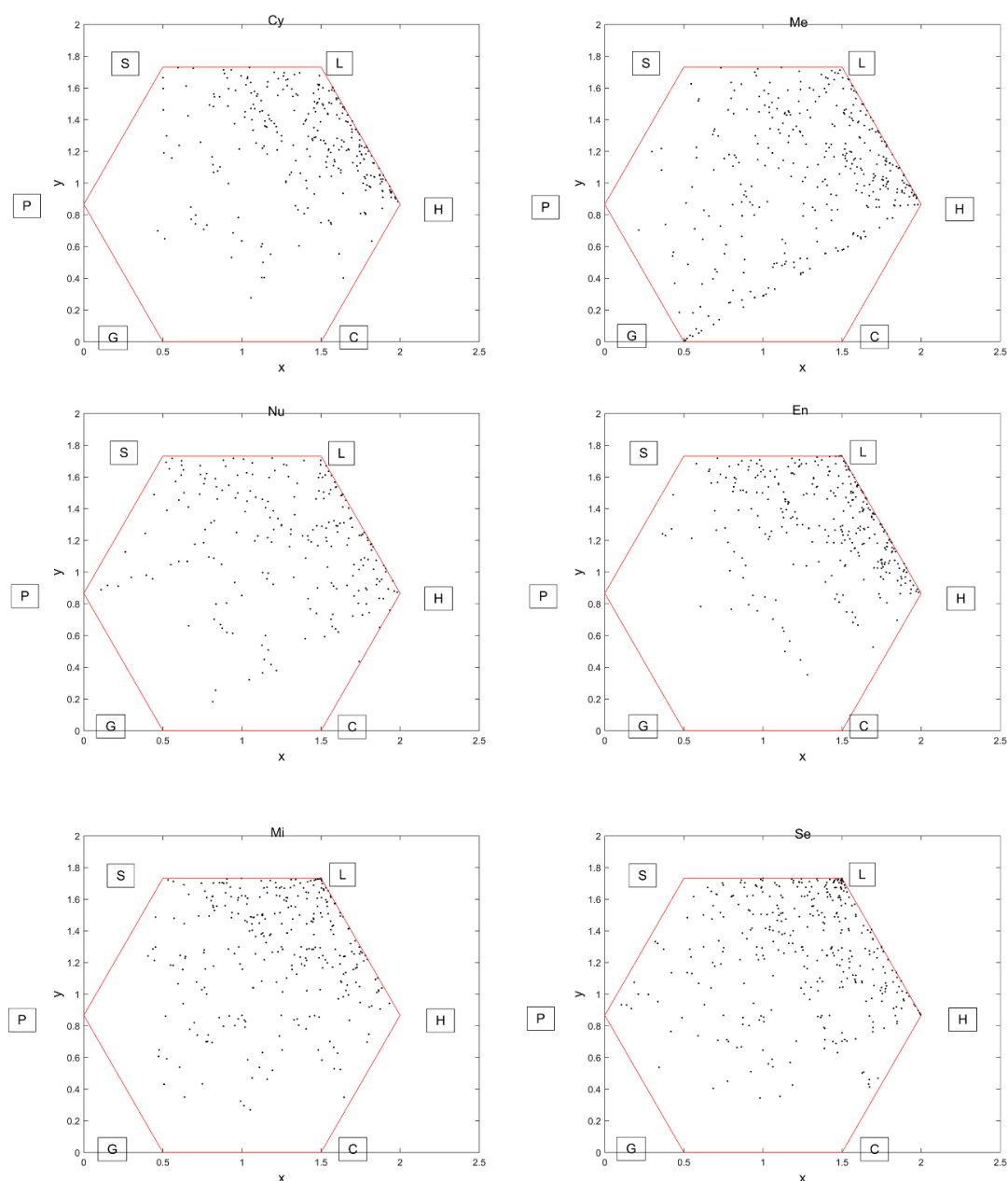


Figure 2. The GCGRs of primary sequence for proteins from six subcellular locations GCGR: Generalized Chaos Game Representation.

3.2.3. Step 3: Convert Each Protein Sequence into a 2-D Vector according to its GCGR Plot

As can be seen from Figure 2, each letter in the protein sequence corresponds to a (x, y) -coordinate in the GCGR plot. We then modelled the GCGR plot as a combination of two series: one is composed of the x -coordinates and the other is composed of the y -coordinates, which were named x -series and y -series, respectively. As can be seen from Figures 3 and 4, there are many useful observations: (1) The average values of the x -series and y -series for proteins in the En-class, denoted as \bar{x} and \bar{y} respectively, tend to be greater than those for proteins in the other classes; (2) Proteins in the first class Cy also have a large \bar{x} , but do not have a large \bar{y} ; (3) Proteins in the last two classes Mi and Se have moderate \bar{x} and \bar{y} , respectively.

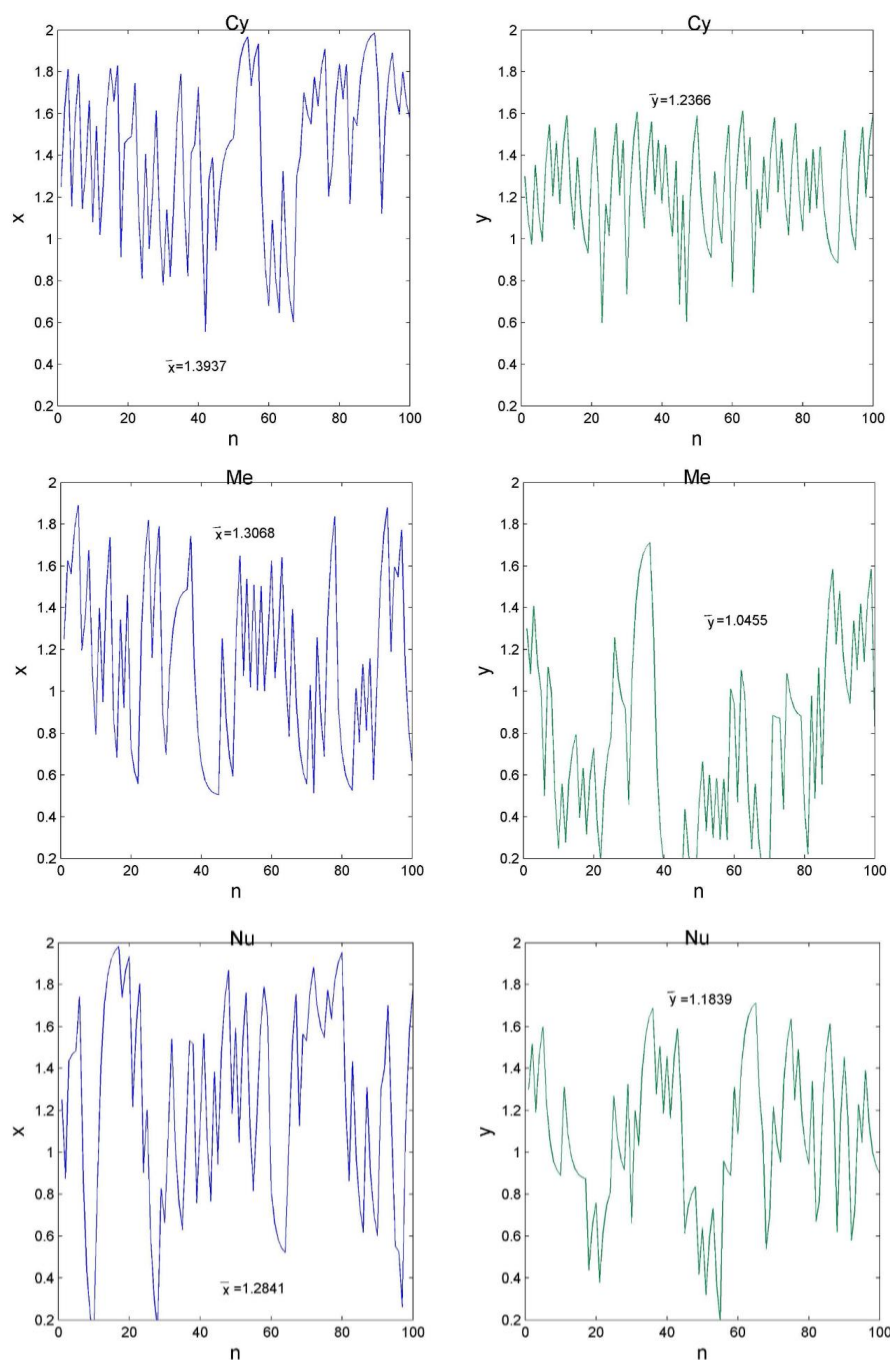


Figure 3. Six time series that represent the first three GCGRs in Figure 2. Each panel in Figure 2 gives rise to two time series.

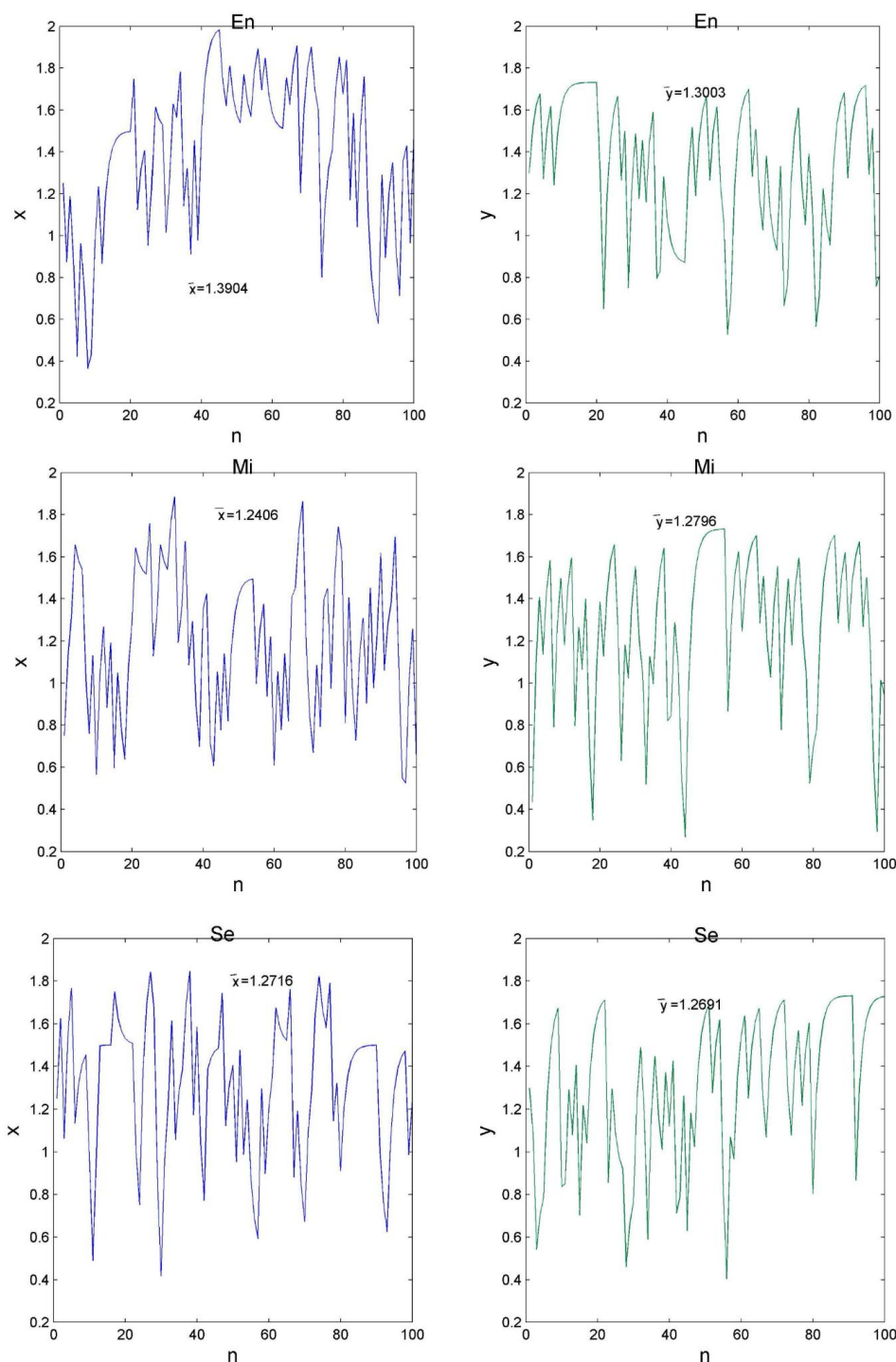


Figure 4. Six time series that represent the last three GCGRs in Figure 2. Each panel in Figure 2 gives rise to two time series.

However, unlike proteins in the Se class, those in the Mi class have a greater \bar{y} than its \bar{x} ; (4) \bar{y} of proteins in the second class Me is the smallest among all classes. In summary, \bar{x} and \bar{y} are two effective numerical features to identify subcellular location of proteins. It is not surprising since \bar{x} and \bar{y} contain not only the information about amino acids frequencies, but also their order in a protein sequence. For a better view, we also drew in Figure 5 the boxplots of \bar{x} and \bar{y} for proteins in each of the six classes in CL317.

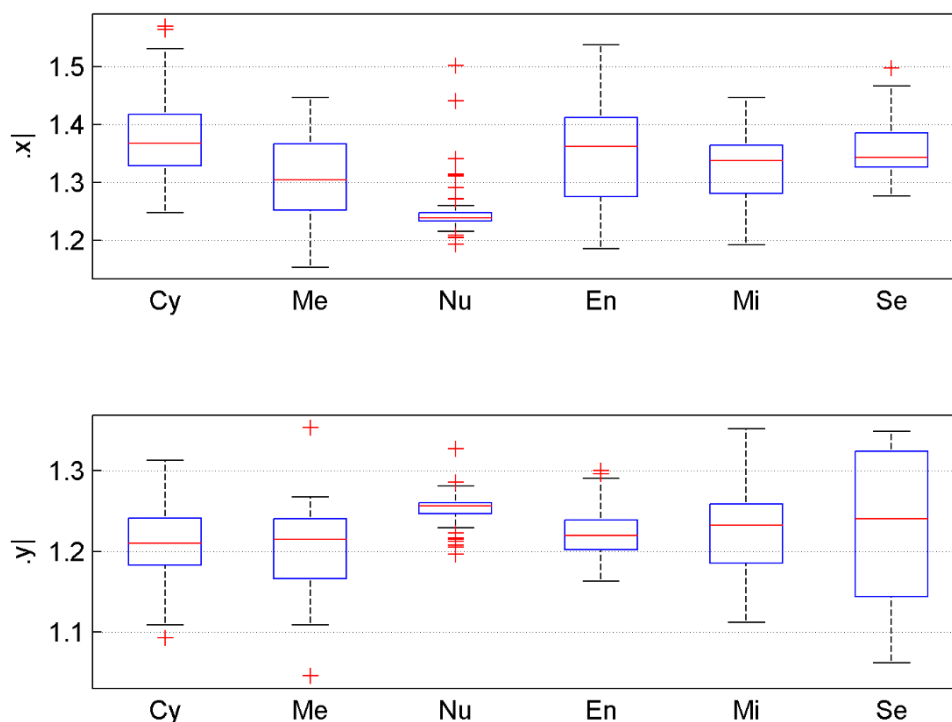


Figure 5. The boxplots for the \bar{x} and \bar{y} of all the proteins in dataset CL317 grouped into the six subcellular locations.

Theoretically, a class with narrow variation scope and less outliers can be discriminated more robustly. As can be seen, the proteins in the Nu class have substantially narrower variation with \bar{x} ranges approximately from 1.22 to 1.26 and \bar{y} ranges approximately from 1.23 to 1.28. For En, though \bar{x} is widely distributed, \bar{y} is more centralized, which can be used to differentiate this class. Similarly for Se, though \bar{y} is widely distributed, \bar{x} is more centralized. Finally, it is of note that all six classes has different medians and relatively differentiable variation scopes for both \bar{x} and \bar{y} . Therefore, by combining \bar{x} and the \bar{y} , it is possible to predict the localization of most proteins.

3.3. Novel Statistics and Information Theory (NSI) of Protein Primary Sequences

In order to acquire more position information of the primary sequence, we presented a novel statistics and information theory based method to extract features from the protein primary sequence. Different from the previous section, we just classified the 20 amino acids into three groups in view of their hydropathy profiles [21]: (1) internal group, in which the residues tend to appear in the inner side of the protein spatial structure, (2) external group, in which the residues tend to occur at the surface, and (3) ambivalent group, in which the residues do not have fixed common positions. Then, a protein sequence can be transformed into a 3-letter string according to the following rule:

$$F(P(j)) = \begin{cases} F & \text{if } P(j) = F, I, L, M, V \\ D & \text{if } P(j) = D, E, H, K, N, Q, R \\ S & \text{if } P(j) = S, T, Y, C, W, G, P, A \end{cases} \quad (1)$$

where $P(j)$ represents the j th letter in the protein primary sequence P , and $F(P(j))$ presents the encoded letter for $P(j)$. For example, given a protein sequence $P = \text{YAMQESHFTCI}$, its encoded sequence is $F(P) = \text{SSFDDSDFFSS}$.

After that, we calculated the position features of the encoded sequence to represent its local information. Specifically, let $W(k) (k \in \{F, D, S\})$ be the position sequence of a given amino acid k . We calculated the intervals between two consequent positions in $W(k)$, which formed a new

numerical distance sequence denoted by $N(k)$. Obviously, $N(k)$ contains the positional and distribution information of the given amino acid k in the primary sequences. For instance, for the encoded amino acid sequence, the position sequences for the reduced amino acids F, D, and S are: $W(F) = (3, 8, 11)$, $W(D) = (4, 5, 7)$, $W(S) = (1, 2, 6, 9, 10)$. Then the symbolic sequences are cyclic and their numerical sequences $N(k)$ are: $N(F) = (5, 3)$, $N(D) = (1, 2)$, $N(S) = (1, 4, 3, 1)$. The numerical sequence $N(k)$ provides a new profile to characterize correlation residues of the given sequence. In fact, the interval distance between two occurrences of k can be denoted by a random variable x . We calculated the probability $p_k(x)$ of the Matthew's correlation coefficient (MCC) of the variable x and obtain its distribution function. Based on the probability theory, we can further calculate the mean value $E_k(x)$ and the variance $D_k(x)$ by:

$$E_{(k)}(x) = \sum_x x \times p_{(k)}(x) \quad (2)$$

$$D_{(k)}(x) = E_{(k)}(x^2) - [E_{(k)}(x)]^2 \quad (3)$$

Then, we defined the positional information $I_{(k)}(x)$ as follows:

$$I_{(k)}(x) = E_{(k)}(x) / \sqrt{D_{(k)}(x)} \quad (4)$$

where $I_{(k)}(x)$ is a pivotal statistic for comparing the degree of variation from one data to another. Finally, $I_{(k)}(x)$ ($k \in \{F, D, S\}$) appropriately characterize the positional information of three encoding letters and thus form a novel feature vector of a protein primary sequence.

3.4. Unitary Distance

In this article, rather than the serial combination method which combines different feature vectors into a super-vector, the parallel combination method combines two feature vectors by a complex vector [32], which was defined by

$$z = u + vi \quad (5)$$

where i is an imaginary unit. Here, we defined the parallel combined feature space on \mathfrak{R} as $C = \{u + vi | u \in A, v \in B\}$. Thus C is an m -dimensional complex vector space, where $m = \max(\dim A, \dim B)$. The inner product of two vectors in the complex space is given by $(a, b) = a^H b$, where $a, b \in C$, and H is the denotation of conjugate transpose. The complex vector space defined by the above inner product is usually called unitary space. The norm in unitary space is given by:

$$|z| = \sqrt{z^H z} = \sqrt{\sum_{k=1}^n (a_k^2 + b_k^2)} \quad (6)$$

where $z = (a_1 + i \cdot b_1, \dots, a_n + i \cdot b_n)^T$.

Then, the unitary distance between two complex vectors z_1 and z_2 is calculated by:

$$|z_1 - z_2| = \sqrt{(z_1 - z_2)^H (z_1 - z_2)} \quad (7)$$

3.5. Performance Assessment

For evaluating the effectiveness of two proposed features GCGR and NSI, we first introduced an easy model for fast predicting protein subcellular location, which is described as follows: for a given protein, its numeric features from GCGR and NSI were firstly extracted. Then, these two features are concatenated in parallel and then classified by a classifier free prediction model. As well-acknowledged, all the prediction models in the real number space could be extended to the complex number space by using different similarity measures. For example, the Euclidian distance is a commonly-used similarity metric in the real space, while the unitary distance is often used in the complex space, which was

adopted in this paper. Note that the dimensionalities of u and v in the complex space must be equal, pad the lower-dimensional one with \bar{x}/\bar{y} until its dimensionality is equal to the higher-dimensional one before vectors combination.

In order to measure the predictive capability of the algorithm, we adopted the following commonly used measures:

$$\text{Sensitivity}(S_n) = TP / (TP + FN) \quad (8)$$

$$\text{Specificity}(S_p) = TN / (TN + FP) \quad (9)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \quad (10)$$

$$\text{Overall prediction accuracy}(A_c) = \frac{\sum TP_i}{N} \quad (i = 1, 2, 3, \dots, n) \quad (11)$$

where True Positive (TP) represents the number of true positives in its subcellular location, True Negative (TN) represents the number of true negatives in its subcellular location, False Positive (FP) denotes the number of false positives and False Negative (FN) denotes the number of false negatives in its subcellular location. N is the total number of the protein sequences. Sensitivity means the rate of correct prediction. Specificity means the reliability level for predictive model. The Matthew's correlation coefficient (MCC) shows the comprehensive performance of the prediction algorithm.

In this paper, all the experiments were completed by MatLab and a Library for Support Vector Machines (LIBSVM). In the experiments, we chose the jackknife test for validating the performance of each model. The jackknife test is done by dropping in turn each sample from the data set as the test sample and fitting the model for the remaining set of observations as the training samples. The predicting accuracy can be obtained by the right classified samples divided by the total number of samples. It worth's noticing that though there is no parameter in our feature construction process, there are two model parameters for LIBSVM, namely c and g . We adopted a grid search to find the best c and g in the jackknife test. Specifically, c and g both varied from 2^{-5} to 2^5 with a multiple 2, and the best c is 8 and g is 0.0625 for the dataset CL317, while the numbers are 16 and 0.125 respectively for the dataset ZW225.

4. Conclusions

In this study, we first proposed a novel and quick method for predicting yeast subcellular locations based on generalized chaos game representation of the protein primary sequence and the statistics and information theory to uncover the residues distribution among the sequence. Implementation on two benchmark yeast datasets suggests that this model achieves comparable classification performance as those of machine learning-based classifiers. In addition, a fusion model incorporating GCGR and NSI with some known features including PwAAC and Dipeptide were presented, which gains the highest overall accuracy and MCC on the two benchmark datasets. The results also indicate that the new features extracted contain some useful information, which is not mined in previous methods.

Author Contributions: B.L. (Bo Liao), L.C. and J.Y. conceived the concept of the work. B.L. (Bo Li), X.F and P.B. performed the experiments. B.L. (Bo Li), and J.Y. wrote the paper. All authors read and approved the final manuscript.

Funding: This study is supported by the Program for National Nature Science Foundation of China (Grant Nos. 61863010, 61873076, 61370171, 61300128, 61472127, 61572178, 61672214 and 61772192), and the Natural Science Foundation of Hunan, China (Grant Nos. 2018JJ2461, 2018JJ3570).

Acknowledgments: The authors would like to thank Dengyi Long for kind suggestions and discussions that have helped improve the presentation of this paper.

Conflicts of Interest: The authors confirm that this article content has no conflict of interest.

References

1. Yu, D.; Wu, X.; Shen, H.; Yang, J.; Tang, Z.; Qi, Y.; Yang, J. Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features. *IEEE Trans. Nanobiosci.* **2012**, *11*, 375–385. [[CrossRef](#)] [[PubMed](#)]
2. Kuo-Chen, C.; Yu-Dong, C. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765–45769.
3. Ernst, J.; Bar-Joseph, Z. STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinform.* **2006**, *7*, 191. [[CrossRef](#)] [[PubMed](#)]
4. Mei, S.; Fei, W.; Zhou, S. Gene ontology based transfer learning for protein subcellular localization. *BMC Bioinform.* **2011**, *12*, 44. [[CrossRef](#)] [[PubMed](#)]
5. Wang, Z.; Zou, Q.; Jiang, Y.; Ju, Y.; Zeng, X. Review of Protein Subcellular Localization Prediction. *Curr. Bioinform.* **2014**, *9*, 331–342. [[CrossRef](#)]
6. Liao, B.; Jiang, J.; Zeng, Q.; Zhu, W. Predicting Apoptosis Protein Subcellular Location with PseAAC by Incorporating Tripeptide Composition. *Protein Pept. Lett.* **2011**, *18*, 1086–1092. [[CrossRef](#)] [[PubMed](#)]
7. Wang, Z.; Jiang, L.; Li, M.; Sun, L.; Lin, R. Fast Fourier Transform-based Support Vector Machine for Subcellular Localization Prediction Using Different Substitution Models. *Acta Biochim. Biophys. Sin.* **2007**, *39*, 715–721. [[CrossRef](#)] [[PubMed](#)]
8. Qiu, J.; Luo, S.; Huang, J.; Sun, X.; Liang, R. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. *Amino Acids* **2010**, *38*, 1201–1208. [[CrossRef](#)] [[PubMed](#)]
9. Gao, Q.; Jin, Z.; Wu, C.; Sun, Y.; He, J.; He, X. Feature Extraction Techniques for Protein Subcellular Localization Prediction. *Curr. Bioinform.* **2009**, *4*, 120–128. [[CrossRef](#)]
10. Chou, K. Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477–483. [[CrossRef](#)] [[PubMed](#)]
11. Yu-Xi, P.; Zhi-Zhou, Z.; Zong-Ming, G.; Guo-Yin, F.; Zhen-De, H.; Lin, H. Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. *J. Protein Chem.* **2003**, *22*, 395–402.
12. Jia, P.; Qian, Z.; Zeng, Z.; Cai, Y.; Li, Y. Prediction of subcellular protein localization based on functional domain composition. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 366–370. [[CrossRef](#)] [[PubMed](#)]
13. Khan, A.; Khan, M.F.; Choi, T. Proximity based GPCRs prediction in transform domain. *Biochem. Biophys. Res. Commun.* **2008**, *371*, 411–415. [[CrossRef](#)] [[PubMed](#)]
14. Shen, Y.Q.; Burger, G. TESTLoc: Protein subcellular localization prediction from EST data. *BMC Bioinform.* **2010**, *11*, 563. [[CrossRef](#)] [[PubMed](#)]
15. Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* **2018**, *117*, 212–217. [[CrossRef](#)]
16. Ding, Y.-S.; Zhang, T.-L. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognit. Lett.* **2008**, *29*, 1887–1892. [[CrossRef](#)]
17. Wan, S.; Duan, Y.; Zou, Q. HPSLPred: An Ensemble Multi-label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source. *Proteomics* **2017**, *17*, 1700262. [[CrossRef](#)] [[PubMed](#)]
18. Chen, Y.; Li, Q. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol.* **2007**, *245*, 775–783. [[CrossRef](#)] [[PubMed](#)]
19. Lin, H.; Wang, H.; Ding, H.; Chen, Y.; Li, Q. Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta Biotheor.* **2009**, *57*, 321–330. [[CrossRef](#)] [[PubMed](#)]
20. Yu, C.; Cheng, C.; Su, W.; Chang, K.; Huang, S.; Hwang, J.; Lu, C. CELLO2GO: A web server for protein subCELLular LOcalization prediction with functional gene ontology annotation. *PLoS ONE* **2014**, *9*, e99368. [[CrossRef](#)] [[PubMed](#)]
21. Wan, S.; Mak, M.; Kung, S. HybridGO-Loc: Mining Hybrid Features on Gene Ontology for Predicting Subcellular Localization of Multi-Location Proteins. *PLoS ONE* **2014**, *9*, e89545. [[CrossRef](#)] [[PubMed](#)]
22. Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.* **2015**, *364*, 284–294. [[CrossRef](#)] [[PubMed](#)]

23. Shao, W.; Ding, Y.; Shen, H.; Zhang, D. Deep model-based feature extraction for predicting protein subcellular localizations from bio-images. *Front. Comput. Sci. China* **2017**, *11*, 243–252. [[CrossRef](#)]
24. Zhang, Z.; Wang, Z.; Zhang, Z.; Wang, Y. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* **2006**, *580*, 6169–6174. [[CrossRef](#)] [[PubMed](#)]
25. Boeckmann, B.; Bairoch, A.M.; Apweiler, R.; Blatter, M.; Estreicher, A.; Gasteiger, E.; Martin, M.; Michoud, K.; Odonovan, C.; Phan, I. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365–370. [[CrossRef](#)] [[PubMed](#)]
26. Eswara, M.B.; McGuire, A.T.; Pierce, J.B.; Mangroo, D. Utp9p facilitates Msn5p-mediated nuclear reexport of retrograded tRNAs in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **2009**, *20*, 5007–5025. [[CrossRef](#)] [[PubMed](#)]
27. Polymenis, M.; Aramayo, R. Translate to divide: Control of the cell cycle by protein synthesis. *Microb. Cell* **2015**, *2*, 94–104. [[CrossRef](#)] [[PubMed](#)]
28. Aouida, M.; Ramotar, D. Identification of essential yeast genes involved in polyamine resistance. *Gene* **2018**, *677*, 361–369. [[CrossRef](#)] [[PubMed](#)]
29. Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.* **1990**, *18*, 2163–2170. [[CrossRef](#)] [[PubMed](#)]
30. Yang, J.; Peng, Z.; Yu, Z.; Zhang, R.; Anh, V.; Wang, D. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *J. Theor. Biol.* **2009**, *257*, 618–626. [[CrossRef](#)] [[PubMed](#)]
31. Panek, J.; Eidhammer, I.; Aasland, R. A new method for identification of protein (sub)families in a set of proteins based on hydrophathy distribution in proteins. *Proteins* **2005**, *58*, 923–934. [[CrossRef](#)] [[PubMed](#)]
32. Yang, J.; Yang, J.; Zhang, D.; Lu, J. Feature fusion: Parallel strategy vs. serial strategy. *Pattern Recognit.* **2003**, *36*, 1369–1381. [[CrossRef](#)]

Sample Availability: Samples of the compounds are not available from the authors.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).