International Journal of Engineering &Technology, 8 (4) (2019) 478-489

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Combination between DE and SVM to enhance Protein Structure Prediction based on Secondary Structural information

Thair A. Kadhim [1] *, Mohammed Hasan Aldulaimi [2], Suhaila Zainudin [3], Azuraliza Abu Bakar [4]

[1,2] Ministry of Education, Babylon-Iraq
[3,4]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia
**Corresponding author E-mail: altaeeth@gmail.com*

Abstract

The effective selection of protein features and the accurate method for predicting protein structural class (PSP) is an important aspect in protein folding, especially for low-similarity sequences. Many promising approaches are proposed to solve this problem, mostly via computational intelligence methods. One of the main aspect of the prediction is the extraction of an excellent representation of a protein sequence. An integrated vector of dimensions 71 was extracted using secondary and hydropathy information in this study Using newly developed strategies for categorizing proteins into their respective main structures classes, which are all-α, all-β, α/β, and α+β. Support Vector Machine (SVM) and Differential Evolution (DE) were combined using the wrapper method to select the top N features based on the level of their respective importance. The classification can be made more accurate by tuning the kernel parameters for the SVM in the training phase. In this study, the mean of the classification rate from using the SVM classifier was used to evaluate the selected subset of features. This study was tested using two low - similarity data sets (D640 and ASTRAL). A comparison between the proposed (SVM + DE) based on DE feature selection approach and (SVM+DE) based on grid search (a traditional method to search for parameters) forms the core of this work. The proposed SVM+DE model is competitive and highly reliable in terms of time and performance accuracy compared with other reported methods in literature.

Keywords: *Feature Selection; Differential Evolution; Hydropathical Information; Secondary Structure; Computational Time*.

1. Introduction

The PSP acts an essential role in knowing the structures and the functions of proteins using the information amino acid sequence [1]. The SCOP is commonly used to classify the classes of the protein structures [2]. There are a total of 112,722 PDB inputs, with 50,825,784 scopes or proteins, whose their structure classes are known as SCOP1.75 C (as of 7 October 2015) [3], with 90% going to the four major categories of the protein (all -α , all -β , α / β and α + β).

The vast growing of proteomics, genomics and the sequences of protein generated a significant gap between the number of known structure and the sequence-known of proteins. Constraints for example cost and time of methods for protein structure determination such as (X-ray[4], NMR [5] and ESR [6])

Which means that they may not be sufficient to resolve each structural sequence of the protein. Thus, the development of a reliable computation method for determine the class of protein is very demanding. In general, three methods are used in pervious researches of PSP. The first method relates to the similarity alignment of the sequence; which is consider the common method used for PSP because of the large number of sequences discovered. It neglects when the protein sequences are related to each other but have different functions [7]. The second method used the tertiary structure [8]. The protein function is related to the tertiary structure of the protein, which is determined by the protein folding process. However, it has been confirmed that the similarities between protein structures are not always related to similarity of the catalysts [1]. The third method represented the primary structure by using physicochemical features such as information. Currently, many researches had been conducted based on features derived from the secondary structural information, which is helpful in structural class prediction [9-15]. This study will use the information of secondary structure to classify the protein's structure, as well as the hydropathical information of proteins.

In previous works, the amino acid (AA) sequence served as a platform of feature extraction [16-18]. Recently, the extraction of secondary structures sequences (SSS) features were assumed to improve accuracy prediction of low - similar sequences [13, 19], for example the smallest α-helices and β-strands length. The extracted features can be categorized into three types of methods: (1) distance-related features (2) content-related, and (2) order-related [20]. Despite success with protein structural class prediction methods, an excellent good prediction method for low-similarity structures remains elusive [21].

Determining the feature space via search strategy is also required, due to its usefulness for selecting features subset that are more accurate, [22]. Search algorithms have been developed based on two categories; optimal solution and computational cost. For example, by considering a problem with 5000 features for 90 samples, the total population matrix size will be (90 × 5000). If only 30 features subset

Copyright © Thair A. Kadhim et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is required, then the search space will be reduced to (90 × 30). This will lead to a lesser memory capacity requirement while also reducing computational time.

Two main steps in PSP: extracting features from the sequence of amino acid (in the arrangement of vector dimension), then, these features are fed into the classifier in order to classify them to a particular class [10].

Previous works on the first step exposed that the sequential features presentation can come in many forms, such as pseudo amino acids (PseAA) composition [16, 23], polypeptide composition [24] amino acids composition [25, 26], functional domain composition [17], amino acid sequence reverse encoding [27], position-specific score matrix (PSSM) [9] and predicted secondary structure information [19], [20], [27], [28]. The utilization of the reduced secondary structural class information as opposed to the standard twenty letters of amino acid results in enhanced performance [29]. It should also be pointed out that recently, this is because of the content of the secondary structural and spatial order are vital elements for the distribution of structural elements [19], [30], [31], the proposed method used the secondary structural information to enhance the PSP.

The earlier methods are capable of 90% accuracy when testing data set of high sequences similarities. But, they did not do well in low-similar data sets because they suffered from the over-fitting problem due to the high dimensions of the number of the features [32].The reporting accuracies vary between 50-70% [11]. However, this is manageable via predicted secondary structure and physicochemical properties. A set of classification systems were used protein structure prediction of low-similarity data sets , such as (SVM) [9], [21], [23], [29], [30], [33], neural network [34], fuzzy k-nearest neighbor [35], fuzzy clustering [36], Bayesian classification [37], Logistic regression [30] rough sets [38] and classifier fusion techniques [39], [40]. Among these classification techniques used to tackle this problem, SVM classifier has achieved the best results [41].

Support Vector Machine (SVM) [42], while efficient for PSP, it faces two important problems. The first is its selection of the best feature subset, and the second is the parameters optimization of the kernel. The first is significantly affected by the second, and vice versa. It is therefore prudent that choosing the best feature subset and the optimization of the parameters for the SVM be implemented concurrently. Incidentally, the grid feature selection method do take these factors into account [41], [43], [44], however, it is more time consuming [23]. The feature selection method could be implemented with the wrapper or filter approaches [45], [46]. However, wrapper based methods demonstrate better precision due to the usage of classification algorithm [47], and due to this reason, the wrapper technique is adopted in this work .

There are many feature selection methods that were reported for the SVM in literature [10, 39, 48-53] in the context of avoiding the over-fitting problem. However, most of the results reported by these methods are quite poor. We also took into account the fact that models with lesser features are less computationally demanding [54]. The meta-heuristic algorithms have been proven to be quite adept at solving such problems due to their capability in tackling large dimensional features in biology to select optimal features. For example, Genetic Algorithm (GA) have been successfully used in the past to accurately select biological data [23]. Recently, an algorithm called Differential Evolution (DE) was proven to be a very efficient optimizer [55-58]. Other feature selection methods utilizing DE were proposed in [51], [59-61]. However, most of them focused on selecting the best features without accounting for the time and computational costs of the parametric settings for the SVM classifier.

SVM selects the kernel function, optimizes the kernel parameters, and subsequently decide on the soft margin for the penalty parameter (represented by a constant (C) [22]. The SVM classification can theoretically be enhanced by parametric optimization [55]. Parametric optimization should be carried out on the parameter of penalty $(C)$and the functional parameter of kernel, including the gamma (γ) under the radial basis function (RBF) kernel [29]. A conventional technique along the RBF function was utilized to calculate the best results, while $(\gamma)$is the grid search method. The method involves attempting of the pairs (C,γ), then one is chosen for his best accuracy for cross-validation. After the result of a "better" area on the grid, enhanced search for the grid will test the features to find a better area on the grid [22] (See Fig. 1).

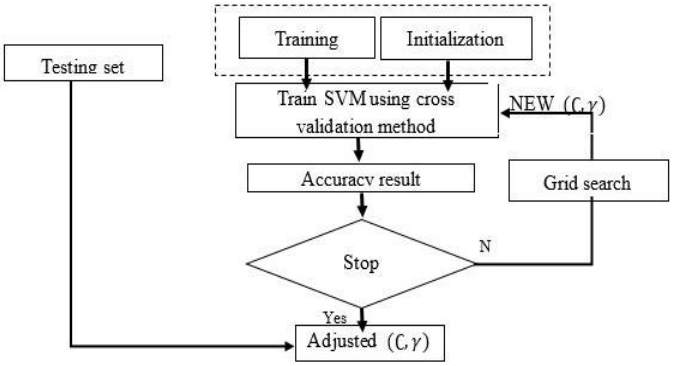


Fig. 1: The Process of Combining Grid Algorithm with SVM and Parameters Setting Based on Grid Search Algorithm.

This methodology involves adjusting grids in the space of the decision and assessing the values of the objective function at each grid point. The best value of the objective function will be regarded as the best solution. The main problem of this method is the exponential increase of the grid's point number that corresponds to the decision variables' number. This will increase the cost of the computation process [21], [39].

However, this method is inefficient and performs quite poorly [22]. One disadvantage of the grid search method is reported in W. Li et al [62], where he posited that the grid search method is time intensive, especially when additional parameters are needed and when it becomes difficult to estimate the parametric range and penalty factor. Therefore, we proposed an improved PSP method by adopting Differential Evolution (DE), which is able to optimally select the features' subset while improving the parameters of SVM. This methodology selects the features and adjusts the parameters in an evolutionary manner.

In this paper, we focused on improving the performance of the classifier, which is measured in terms of its accuracy, while also reducing the computational time that is required for parametric optimization of the classifier based on DE-based feature selection algorithm.

2. Related work

Many feature selection methods proposed for selecting best features for SVM [63-65]. However, the results reported by these methods are quite poor due to the high dimensionality of features. We also took into account the fact that models with lesser features are less computationally demanding. Other feature selection methods that utilizes DE were proposed in [51, 59-61, 66]. However, although the results promising, most of them focused on selecting the best features without accounting for what kind of features that have a highly impact on protein structural class prediction can be used, the time consuming for the training and parametric settings for the SVM classifier. Moreover, they didn't fully explore the combination between DE+SVM in protein structural class prediction. This research proposes a thorough investigation of the feature-based method within the wrapper method for the purpose of enhancing the performance of the classifier for the protein structural class prediction process, as well as reduce the time needed for tuning the SVM parameters.

3. Differential evolution: background

Differential Evolution (DE) is regarded as an evolutionary algorithm. Since 1995, DE is regarded to be a very effective optimization algorithm [55]. The development of the Genetic Annealing (GA) algorithm by Kenneth Price and Dr. Dobb [67] resulted in the DE algorithm. DE has features such as its simplicity, parallel and quick processing speeds, its ability to search directly, user friendliness, efficient convergence, and fast implementation process [67]. DE can provide the optimum features for both the subset and SVM parameters concurrently [49]. DE utilizes three main processes: mutation, crossover, and selection. These operators are utilized to evolve from generating initial population randomly until it reaches the final individual solution [50], [68]. Mutation and crossover are used to generate the experimental vector, then the selection is used to determine whether the new vector may survive until the next generation. The procedure of DE is as follows:

Step 1: parameters Initialization

Four main parameters must be identified by the user for DE algorithm prior to its start are: size of population (M), number of iterations (K), scale factor (F), and crossover rate (CR).

Step 2: initialize size of population (M) randomly

For a usual Evolutionary Algorithm (EA), DE initiates a solution of random population of candidate solutions for the optimization problem:

$$X_{i,j} = 1 \ldots NP \tag{1}$$

Where i is the population index, j is the generation of the population, and NP is the number of variables.

Step 3: Mutation

This step intends to enhance the exploration ability and increase the solution vectors for the DE algorithm [69]. The trial parameters are generated by the addition of weighted difference vector between the two population vectors ($X_{r1,g}$ and $X_{r2,g}$) and a third member $X_{r0,g}$. In Eq (2), the combination of three completely random vectors is presented, where the goal is to produce a mutant vector $V_{i,g}$ based on the current generation:

$$V_{i,g} = X_{r0,g} + F * (X_{r1,g} - X_{r2,g}) \tag{2}$$

Where $F \in (0, 1)$ is scale of the evolution rate of the population.

Step 4: Crossover

DE operates based on the uniform crossover, in other words, DE has a discrete recombination to produce trial vectors based on parameters obtained from two different vectors. Particularly, DE crosses each vector with a mutant vector:

$$U_{i,g} = \begin{cases} V_{i,g}, if\ (rand(0,1) \le Cr \text{ or } g = g_{rand}) \\ X_{i,g}, otherwise \end{cases} \tag{3}$$

The probability of crossover, $Cr \in [0,1]$, alters the parametric values that are adopted from the mutant. When a new vector has a higher objective function value (higher fitness) compared to the determined population individuals, then the new vector will replace the existing vector where the comparison takes place [60].

Step 5: Selection

During selection, the probability for fitter chromosomes to be selected for the recombination pool process is invariably higher. The selection uses the tournament selection or roulette wheel method, as shown in Fig. 2.

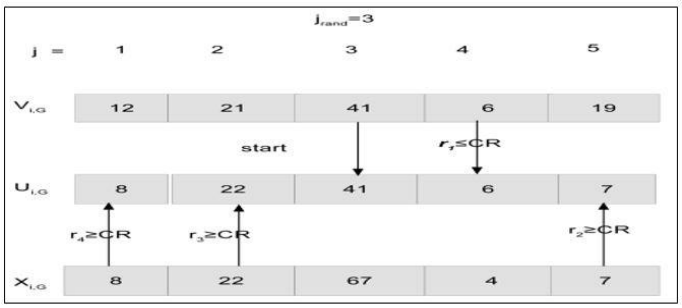


Fig. 2: Uniform Crossover: The Parameters That are Chosen in Random Mode Is Originated from the Mutant (E.G. $J_{Rand} = 3$) . Unselected Parameters are Filled with D-1 Independent Trials. If$Rand_J(0,1) \le CR$, The Origin of Parameter Values Is Mutant, Otherwise Its Donated from A Target.

Step 6: Selection

If the termination condition (maximum number of iterations K) is met, the process is completed. This is followed by the repetition of steps 3-5. It is decided that a common set of parametric values be used, specifically F = 0.8 and CR = 0.95 for this study. The following algorithm (See Fig. 3) shows the pseudo code of a typical strategy in DE (DE/rand/1) adopted in this work [70]:

P_t: Population at time t
TP Temporary population
$t \leftarrow 0$;
$InitializeP_t$;
Evaluate P_t;
While *not termination condition* **do**
$TP \leftarrow \phi$;
for $\forall \text{indiv}_i \in P_t$ do
$offspring \leftarrow TRIALVECTORGENERATION(indiv_i)$;
Evaluate Offspring;
if Offspring is fitter than Offspring indiv_i **then**
Put Offspring into TP;
else
Put indiv_i into TP;
end if
end for
$t = t + 1$;
$P_t \leftarrow$ TP;
end while

Fig. 3: Pseudo Code of DE/Rand/1 Source [70].

During each generation, each individual indiv_i undergoes a trial vector generation to generate offspring's (restored by the subroutine TRIALVECTORGENERATION). During this subroutine process, three individuals were selected from a trial vector randomly. The base vector is formed by one individual, while a different vector is formed by the difference between other individuals. Summing these two vectors resulted in a trial vector, which is then recombined with a parent (indiv_i) to form offspring's.

4. Materials and method

The parametric selection of SVM was realized via two methods; combining SVM with grid search algorithm and combining SVM with DE algorithm method in the tuning parameters process.

4.1. Data set

Two almost non-similar datasets ($ASTRAL$ and D640) are used to evaluate and design the proposed methodology. The selected ASTRAL had sequences similarities less than 20%, containing 6424 protein sequences [44].
In this study four important classes were selected in the study of protein form which are (all-α, all-β, α/β, α+ β) from a total of 7. The dataset with 5,626 sequences was divided into two equal subsets in a random manner, one was used as the training set ($ASTRAL_{training}$), while the other was utilized as a testing set ($ASTRAL_{testing}$) [10]. Both datasets are available at http://web. xidian.edu.cn/slzhang/paper.html.
Another commonly used low sequence identity benchmark dataset is D640, with 25% sequences similarity selected from [27] for the purpose of providing a fair and impartial comparison with the current prediction methods. Both datasets are commonly used as standard datasets in some state-of-the-art systems [10, 13, 19, 27, 30, 31, 71]. The detailed description of the used data sets is presented in Table 1.

Table 1: Data Sets Details

Dataset	all-α	all-β	α/β	α+β	Total
$\text{ASTRAL}_{\text{training}}$	640	662	748	763	2813
$\text{ASTRAL}_{\text{testing}}$	640	662	747	764	2813
640	138	154	177	171	640

4.2. Protein sequence representation

In this work, 71 features are extracted from the aforementioned datasets using the secondary structure and hydropathy information to build the model of the classifier. All of these features can be classified into:

4.2.1. Features extraction using secondary structural elements

The conversion of all amino acids of the protein into secondary structural form led to a more effective use of the proposed method. These elements are E (strand), H (helex), and C (coil), where the order of these elements are seen as a secondary structure sequence (SSS), which can be obtained from a database server predicting a structure of protein named PSIPRED [72].
Example 1
Amino Acid Sequence (AAS):
PVITLPDGSQRHYDHAVSPMDVALDIGPGLAKACIAGRVNGELVDACDLIEN
Predicted Secondary Structure Sequence:
CEEECCCCCEEECCCCCCHHHHHHHCCHHHHCCEEEEEECCEECCCCCCCCC
By using the above representation, a number of features can be generated using the formula:

$$p(i) = \frac{n_i}{N} \tag{4}$$

Where n_i is the amount of SSS number i along of the sequences, with i ∈ {E, H, C}, N is represent the length sequence of SSS.
After this process 2 features (F1and F2) will be created (See Table 2).

$$p(xy) = \frac{n_{xy}}{N} \tag{5}$$

Where n_{xy} is number of 2 repeated patterns, xy ∈ {EH, HE} After this process six features (F3… F8) will be created (see Table 2).

$$NcalcSeg(i) = \frac{CalcSeg(i)}{N} \tag{6}$$

$CalcSeg(i)$ is the H or E segments number, i ∈ {E, H}. This process will result in the generation of 2 features (F9 and F10) (See Table 2).

$$CMV_i = \frac{1}{N(N-1)} \sum_{j=1}^{n_i} x_{ij} \tag{7}$$

Where i ∈ {H, E} and N is the amino acids amount of the protein sequences, while x_{ij} is the index of the j^{th} location of the i^{th} structures, respectively.
This process results in the generation of 2 features (F11 and F12) (See Table 2).

$$NMaxSeg_i = \frac{MaxSegi}{N} \tag{8}$$

Where the i ∈ {E, H},$MaxSegi$ is the longest length of α-helices (β-strands) in the sequences of the protein. This process resulted in the generation of 2 features (F16 and F17) (see Table 2).
The (SCMV) "probability changing the State of Moment Vectors" are considered to detect changing state positions in SSS. These features are created for the purpose of helping in the distinction between α/β and α+β classes of the protein sequences.
The classes of sequences are exposed to the phenomenon of changing the state, but in another situations of the order. In SCMV, there are 6 changes states that can be calculated using the next formula:

$$SCMV_i = \frac{1}{N(N-1)} \sum_{j=1}^{n_i} x_{ij} \tag{9}$$

Where $i \in \{EC, CE, EH, HE, CH, HC\}$ and N is the changes number of state in the proteins sequences. This process resulted in the generation of six features (F18 …. F23) (see Table 2). The next section will involve the extraction of other types of features using hydropathy information.

4.2.2. Features extraction using hydropathy information

In this study, the hydropathical features are used for protein representing on the hypothesis that it has a significant influence on the protein folding process. The hydropathical features describe the hydrophobic and hydrophilic nature of the protein sequence [73].
The 20 amino acids types are classified into three sets using their hydropathical features called (I) "Internal", (E)"External" and (A) "Ambivalent". The following rulings offered by Liu, & Wang [73]
This study was used to classify amino acids by using their hydropathy profile:

$$F(S(i)) = \begin{cases} I \; if \; S(i) = F, I, L, M, V \\ E \; if \; S(i) = D, E, H, K, N, Q, R \\ A \; if \; S(i) = S, T, Y, C, W, G, P, A \end{cases} \tag{10}$$

According to the above approaches, the 71 generated features are explained as the following table:

Table 2: Extracted features details

F#	Explanation	F#	Explanation
$f1$	Average helex (H)length.	$f37$	Pattern(CE_AE) probability position.
$f2$	Average strand (E)length strand.	$f38$	Pattern (EC_AE). probability position
$f3$	Average change of pattern(HC) position.	$f39$	Pattern (CE_EA) probability position.
$f4$	Average change of pattern (HE)position.	$f40$	Pattern (CE_EE) probability position.
$f5$	Average change of pattern (CE) position.	$f41$	Pattern (CE_IE) probability position.
$f6$	Average change of pattern (CH) position.	$f42$	Pattern (EC_IE) probability position.
$f7$	Average change of pattern (HE) position.	$f43$	Pattern (CH_IA) probability position.
$f8$	Average change of pattern (EC) position.	$f44$	Pattern (HC_IA) probability position.
$f9$	Helex segments (H) avarage.	$f45$	Pattern (CH_IE)probability position.
$f10$	Strand segments (E) avarage.	$f46$	Pattern (HC_IE) probability position.
$f11$	Helex (H) composition moment.	$f47$	Pattern(CH_AI) probability position.
$f12$	Strands (E) composition moment.	$f48$	Pattern (HC_AI) probability position.
$f13$	Average change of pattern (A) position.	$f49$	Pattern (CH_AA) probability position.
$f14$	Average hydropathical feature external(E) length.	$f50$	Pattern (CH_AA) probability position.
$f15$	The verage length of hydropatical feature of Internal (I).	$f51$	Pattern (CH_AE)probability position.
$f16$	The Average maximum length segments of helex(H).	$f52$	Pattern (HC_AE) probability position.
$f17$	The average maximum length strand segments (E).	$f53$	Pattern (CH_EA)probability position.
$f18$	Pattern (HC) probability.	$f54$	Pattern (HC_EA)probability position.
$f19$	Pattern (HE) probability.	$f55$	Pattern (CH_EE)probability position.
$f20$	Pattern (CH) probability.	$f56$	Pattern (HC_EE)probability position.
$f21$	Pattern (CE) probability.	$f57$	Pattern (HE_II)probability position.
$f22$	Pattern (EH) probability.	$f58$	Pattern (EH_II)probability position.
$f23$	Pattern (EC) probability.	$f59$	Pattern (HE_IE)probability position.
$f24$	Pattern (CE_II) probability position.	$f60$	Pattern (EH_IE)probability position.
$f25$	Pattern (HE_EE) probability position.	$f61$	Pattern (HE_IA)probability position.
$f26$	Pattern (EC_II) probability position.	$f62$	Pattern (EH_IA)probability position.
$f27$	Pattern (CH_II) probability position	$f63$	Pattern (HE_AA)probability position.
$f28$	Pattern (HC_II) probability position.	$f64$	Pattern (EH_AA)probability position.
$f29$	Pattern (CE_IE) probability position.	$f65$	Pattern (HE_AE)probability position.
$f30$	Pattern (EC_IE) probability position .	$f66$	Pattern (EH_AE)probability position.
$f31$	Pattern (CH_IA) probability position .	$f67$	Pattern (HE_EI)probability position.
$f32$	Pattern (HC_IA) probability position.	$f68$	Pattern (EH_EI)probability position.
$f33$	Pattern (CE_AI) probability position.	$f69$	Pattern (HE_EA)probability position.
$f34$	Pattern (EC_AI) probability position.	$f70$	Pattern (EH_EA)probability position.
$f35$	Pattern (CE_AA) probability position.	$f71$	Pattern (EH_EE)probability position.
$f36$	Pattern(EC_AA) probability position.		

In Eq (10), $S(i)$ represents the i^{th} amino acids in the main protein sequence, while $F(S(i))$ represents the replacement of constant according to its hydropathy nature. Such as, amino acids sequences is stated as:
(S=MDPFLVLLHSVSS represented by $F(S)$=IEAIIIIIEAIAAA).
By applying Formula (4) for each protein sequence, i ∈ {A, I, E}, three features extracted (F13, F14 and F15) which are mixed with the SSS feature (see Table 2). In order to detect the effect of the two sequential structures of the hydropathical information towards the two sequential structure of the secondary information state, in the same location, an extra new calculation formula of probability is needed. After that, 48 conditional probabilities feature, (F24, F71) are calculated for which revealed the effect of pair hydropathy state$h$. These states belong to the set:
$\{AE, EE, EI, II, IA, EA\ , IE, AA, AI, \}$ in the SSS form each pair s, belongs to the set { EC, CE, HE, EH, HC, CH} by using the following formula:

$$P(s\backslash h) = \frac{p(h,s)}{P(h)} \tag{11}$$

$P(s\backslash h\ )$ is the occurrence probability of the elements of secondary structural s. The occurrences should be in the similar positions with a given pair of hydropathy h in the proteins sequences probabilities, where the hydropathy features pair h happens with in the protein sequence. $P(h)$ is the occurrence probability of the elements of the hydropathical features with in the sequence, where $p(h,s)$ is the occurrence probability of s and h happening at same positions as $p(h,s)\ and\ P(h)$, which can be calculated by using the following formula:

$$p(h,s) = \frac{total_{s,h}}{N-1} \tag{12}$$

$$p(h) = \frac{total_h}{N-1} \tag{13}$$

Where $total_{s,h}$ is the entire total of times of occurring the pair s to , h is the hydropathy pair, N is the length of the sequences, and total_h is total of times of occurring pair h in the sequences.
The importance of membrane proteins in medicine and biology resulted in a few of them being available in the PDB. Despite their importance, only a few dozen membrane protein structures are available in the PDB. The method proposed in this paper is inapplicable for membrane proteins (constituting 3% of all genes), because these type of proteins possess large quaternary complexes that are associated with a complex lipid bilayer environment, which are particularly difficult to crystallize [74]. Therefore, we only focus on non-globular and non-membrane proteins.

5. The proposed DE-based approach architecture for feature selection

After the generation of the features, feature selection with parametric optimization based on DE can be designed via the following steps (depicted in Fig. 4):

5.1. Scaling

A demarcation that eliminates larger numeric ranges from dominating the smaller numeric ranges. The difficulty associated with calculating the numbers can be avoided [75]. The precision of SVM can be improved by scaling the feature value relative to the experimental results. In most cases, the features can be linearly scaled either in the range of [-1, +1] or [0, 1] by using the formula (14), where v is original value, v` is scaled value, max_a is upper bound of the feature value, and min_a is lower bound of the feature value.

$$\text{v`} = \frac{\text{v}-\text{min}_\text{a}}{\text{max}_\text{a}-\text{min}_\text{a}} \tag{14}$$

Once the DE operation is completed, the feature subset(s) can then be defined.

5.2. Fitness evaluation

The training dataset is applied to train the SVM classifier, while the classification accuracy is computed using the testing dataset. After producing the classification accuracy, the chromosomes are examined via fitness function _ formula (16).

5.3. Termination criteria

After achieving the termination criteria, the process terminates, otherwise, it will proceed to the next generation.

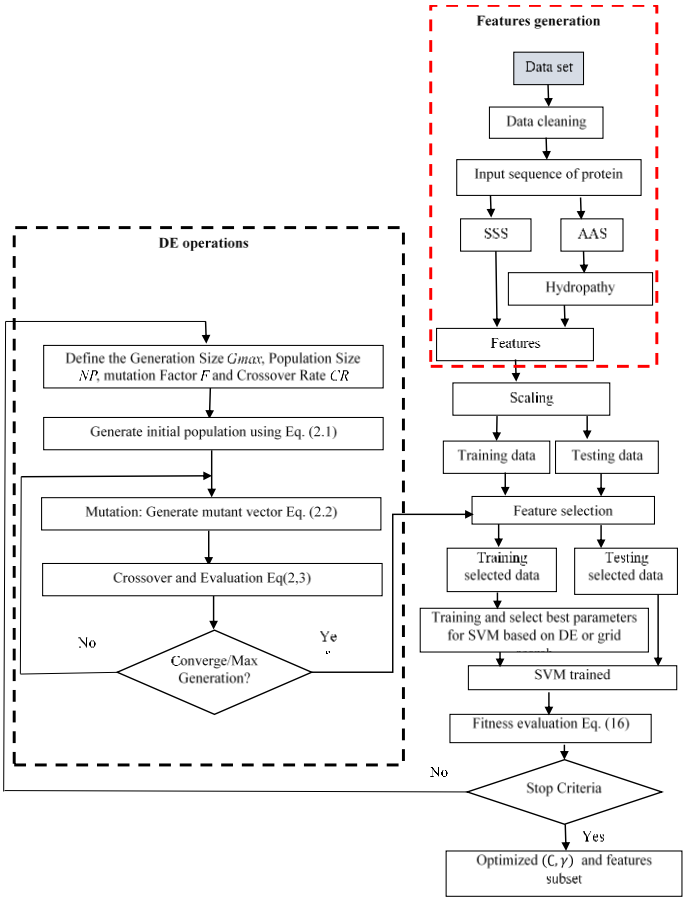


Fig. 4: The Proposed Methodology Architecture.

There are four types of kernel functions that are specific to predictions, which are polynomial, linear, sigmoid, and radial basis function (RBF). Experimental studies demonstrated that RBF performed better than the other kernel functions. The parameters of the regularization, (C), and the parameter of the kernel (γ), are both adjusted based on the 10-Fold Cross-Validation across two experimental datasets. Finally, the best value for the parameters is adjusted when C = 4 and gamma = 0.9, which are selected using the grid search approach that is available in the LIBSVM software.

6. The DE-based features selection and optimizing of parameters

The trail vector design, fitness function, and system architecture of the proposed DE-based feature selection and optimization of the parameters are described as follows:

6.1. Trail vector design

The DE individuals denotes a probable solution of created solutions; therefore, it is essential to discover a good method for representing all individuals. Amongst the 71 features that are existing in the datasets, many features are measured using each dataset performance.

6.2. Fitness function

Fitness function can be defined as how the resolution resolves the problem. In the proposed methodology, the objective was to maximize the PSP performance. So that, the grid search and the cross-validation method were utilized to get the maximization result. The RBF is also used in this study for the library of Libsvm. The fitness function result is the same result of SVM accuracy classifier of DE proposed by [47], as expressed in the following:

$$\text{fitness}(x) = \text{Accuracy}(x) \quad (15)$$

Where Accurecy (x) is the accuracy of the testing data x in the SVM classifier, which is built with the feature subset selection of training data. The classification accuracy of SVM is given by:

$$\mathrm{Acc(x)} = \left({}^{\mathrm{C}}/_{\mathrm{T}}\right) \times 100 \tag{16}$$

C is the samples number that are classified correctly in the testing data of SVM, while T is the total samples number in the testing data.

7. Performance measures

The 10-fold cross-validation procedure test is commonly used to ensure the statistical validity of a classifier [19]. The Fold Cross method can also be used to calculate efficiency in this work. For evaluation purposes, the individual sensitivity (sensitivity), with the overall prediction accuracy (OA) over the entire dataset is reported. We utilize it here to determine the consistency and strength of the proposed methodology. Four standard performance measures were assumed to evaluate the performance of the proposed methodology i.e., sensitivity (Sens, or accuracy), specificity (Spec), overall accuracy (OA) and Matthew's Correlation Coefficient (MCC). They were defined as per the following formulas [31].

$$\mathrm{Sens_j} = \mathrm{TP_j}/\left(\mathrm{TP_j} + \mathrm{FN_j}\right) = \mathrm{TP_j}/|\mathrm{C_j}| \tag{17}$$

$$\mathrm{Spec_j} = \mathrm{TN_j}/\left(\mathrm{FP_j} + \mathrm{TN_j}\right) = \mathrm{TN_j}/\sum_{k \neq j} |\mathrm{C_j}| \tag{18}$$

$$\mathrm{MCC_j} = \frac{(\mathrm{TP_j} * \mathrm{TN_j} - \mathrm{FP_j} * \mathrm{FN_j})}{\sqrt{(\mathrm{FP_j}+\mathrm{TP_j})(\mathrm{TP_j}+\mathrm{FN_j})(\mathrm{TN_j}+\mathrm{FP_j})(\mathrm{TN_j}+\mathrm{FN_j})}} \tag{19}$$

$$\mathrm{OA}=(\textstyle\sum_j \mathrm{TP_j})/(\sum_j |\mathrm{C_j}|) \tag{20}$$

Where $\mathrm{TP_j}$ is the target class that be classified correctly, $\mathrm{TN_j}$ is the number of opposed classes, $\mathrm{FP_j}$ is the number of target class that be classified as opposed classes and $\mathrm{FN_j}$ is the number of opposed classes that be classified as target class, On the other hand, $|\mathrm{C_j}|$ is the number of structural classes of protein $\mathrm{C_j}$ (all-α, all-β, α/β, and α+β). For example, for the all-α, $\mathrm{TP_j}$ (true positive) is the number of all-α protein sequences that were correctly classified. $\mathrm{TN_j}$ (true negative) is the number of classified protein sequences belongs to other three classes that are incorrectly classified as all-α .$\mathrm{FP_j}$ (false positive) is the number of proteins incorrectly classified as all-α, and $\mathrm{FN_j}$ (false negative) is the number of protein sequences classified as all-α but wrongfully classified as all-β, α/β, and α+β. $\mathrm{Sens_j}$(Sensitivity) is the accuracy which was the rate of protein sequences that were actually classified, and $\mathrm{Spec_j}$(Specificity) represents the reliability for prediction in procedure.

Table 4: SVM-DE Results Based on 10- Fold Cross

Dataset	Class	Sensitivity (%)	Specificity (%)	MCC (%)
	all-α	93.00	98.9	91.2
	all-β	84.00	96.00	81.3
$\mathrm{ASTRAL_{training}}$	α/β	86.2	90.00	75.00
	α + β	69.07	90.3	60.00
	OA	82.6		
	all-α	96.3	98.6	94.3
	all-β	84.3	97.07	83.42
$\mathrm{ASTRAL_{testing}}$	α/β	87.1	91.3	77.2
	α + β	74.5	90.5	65.2
	OA	85.07		
	all-α	93.03	98.01	89.7
	all-β	91.6	97.5	89.2
D640	α/β	92.1	94.8	86.08
	α + β	91.9	97.2	89.95
	OA	91.9		

As shown in Table 3 it can be conclude that although the values of specificity were high for all classes, the values of sensitivity were varying. This shows that $\mathrm{FP_j}$ (false positive) in comparison with true negative was very small. Therefore $\mathrm{TN_j}$ (true negative) controls the results. Thus, in this study it can be seen the sensitivity increased by including the features of the proposed method.

8. Results and discussion

We executed our algorithm in Matlab R2012a development environment by extending Libsvm, which was originally designed by Lin and Chang [76]. The experimental evaluation was implemented on a Pentium® Dual-Core CPU with the speed of 2.3 GHz and 3.00 GB of RAM. The 10 fold-CV test is executed upon all of the benchmark datasets to assess and compare the proposed classification method to seven previous methods. As pointed out previously, the aim of the proposed method is to enhance the accuracy of the predictions and reduce the computational time of the prediction. Therefore, the experiments are performed using the 10- fold cross validation test with selected features, and the resulting outcomes are presented in Tables 2 and 3.

8.1. Analysis and comparison with other prediction methods experimental results and comparison in terms of performance prediction

The following best parametric setting for the DE algorithm is a size of population of 100, the probability of crossover of 0.5, minimum mutation of 0.5, and the maximum mutation of 0.8 [58]. When the generation reached 100, or the fitness function fails to further enhance via the last 10 generations, the termination criteria will take over, which will culminate in the best chromosome.

Table: 3: Performances Comparison Across Several Methods for the Two Datasets.

Dataset	Reference	Sensitivity (%) (SVM+DE) based on grid search feature selection					Sensitivity (%)(SVM+DE) based on DE feature selection				
		all-α	all-β	α/β	α + β	OA	all-α	all-β	α/β	α + β	OA
$ASTRAL_{train}$	Experimental study	93.00	84.00	86.2	69.07	**82.60**	92.2	83.7	85.4	69.7	**82.30**
$ASTRAL_{testi}$	Experimental study	**96.3**	**84.3**	87.1	**74.5**	**85.07**	95.9	84.00	86.7	74.5	**84.9**
	[64]	93.13	78.33	83.38	64.27	79.14	93.13	78.33	83.38	64.27	79.14
	[2]	94.53	77.49	87.28	71.47	82.33	94.53	77.49	87.28	71.47	82.33
	[63]	95.16	80.7	83.94	72.51	82.69	95.16	80.7	83.94	72.51	82.69
D640	Experimental study	93.03	**91.6**	**92.1**	**91.9**	**91.9**	90.6	90.3	90.1	91.8	**90.9**
	[3]	90.60	81.8	85.9	66.7	80.80	90.60	81.8	85.9	66.7	80.80
	[15]	89.10	85.10	88.10	71.40	83.10	89.10	85.10	88.10	71.40	83.10
	[12]	94.93	76.62	89.27	74.27	83.44	94.93	76.62	89.27	74.27	83.44
	[46]	92.75	81.82	89.27	74.27	84.22	92.75	81.82	89.27	74.27	84.22

As per Table 4, this method has been compared with previous studies, such as SCPRED [11] and MODAS [13], which are frequently used as a standard. We also compared the method with other PSP systems [10, 44]. The overall highest accuracy is obtained from the proposed method in $ASTRAL_{testing}$ and D640 datasets were (85.07%, and 91.9%), respectively, which were improvements of up to 2.38% and 7.68%, respectively, compared to the previous best-performance results. As for $ASTRAL_{testing}$, the all-α, all-β, α/β and α + β accuracies were 1.14%, 3.6%, 3.16%, and 1.9%, respectively, which are higher than [44]. For the D640 the all-α, all-β, α/β, and α + β, the accuracies were 0.55%, 9.78%, 2.83%, and 17.63%, respectively, which are more than [44].

Although the improvements seemed insignificant, the overall mean values are quite significant. For example, approximately only 112,722, as of October 7, 2015, PDB entries with 50,825,784 domains or proteins had structural class labels in SCOP as of October 7, 2015 [3], while there were more than 39 million non-redundant proteins sequences in the PDB at the NCBI. Hence, a 0.1% enhancement in accuracy will help to determine the accurate structural classes labels for ~39,000 proteins. After analyzing the results and based on the aforementioned example for the $ASTRAL_{testing}$ dataset with 2813 sequences of proteins, the improvement accuracy is up to 2.38%, and the improvement of the proposed method will help determine the accurate structure of ~67 proteins. For the D640 dataset with 640 sequences of proteins, the improvement accuracy is up to 7.68%, which will help determine the accurate structure for ~49 proteins. Therefore, the total number of determined protein is 116 using the proposed method, which can be added to the PDB structure (See Fig. 5).

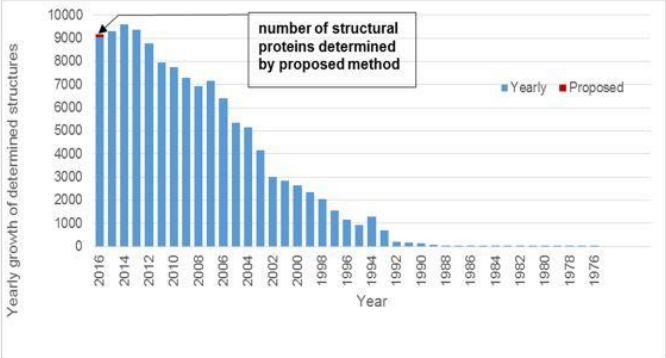


Figure 5: Illustrates the annual growth of the total structure in the PDB database. The blue bars represent (the number of known protein structures in Protein Data Bank) [3], while the red bars represent the total number of protein sequences determined by the proposed method.

In terms of time consumption of the proposed method, for the (SVM+DE) based on the grid search feature selection approach, its overall sensitivity is 82.6%, 85.07%, 91.9% for ASTRALtraining, ASTRALtesting, and 640 benchmarks, respectively, and the average number of features is 45, with all features for C and γ. Note that although the overall sensitivity performance in (SVM+DE) based on DE feature selection approach is higher than the previous best performing results and slightly less than the overall accuracy performance in (SVM+DE) based on the grid search feature selection approach for all experimented data sets (see Table 4). Table 4 shows the results for the overall accuracy performance for the two datasets using two approaches. (SVM+DE) based on DE feature selection approach generated a small optimized feature subsets C and γ, while (SVM+DE) based on grid search feature selection approach uses all of the features. As shown in Table 3, the overall highly performing accuracy is obtained through the proposed methodologies. Some results were reported to be lower compared to the best values of the methods being analyzed.

The results in terms of CPU time are tabulated in Table 5. Since both methods were implemented using similar software and hardware, it is clear that the values of computational cost in terms of time belongs to (SVM+DE) based on DE feature selection method decrease significantly compared to (SVM+DE) based on grid search feature selection method in tuning the parameters of the LIBSVM process.

Table 5: Comparing the CPU Time (in Seconds) between Two Proposed Methods

Data sets	(SVM+DE) based on grid search feature selection	(SVM+DE) based on DE feature selection
$ASTRAL_{training}$	7532.82	261.31
$ASTRAL_{testing}$	6821.21	215.09
D640	1534.85	19.33

When investigating the results in Table 5, the (SVM+DE) based on grid search feature selection method is not as fast as the (SVM+DE) based on DE feature selection. The former is time consuming because it searches though a high dimensionality search space and does not perform well [22], while the latter is a parallel, direct search for best parameters, with excellent convergence and quick implementation [61]. Comparing these two methods, we can surmise that the proposed methods enhanced the performance of the classifier, especially its accuracy. However, the second method reduced the computational time at a level that is more significant compared to the first. The first method require more computational time as it searches in a high dimensional space[22], while the second method is a parallel, direct search for best parameters, having excellent convergence and quick execution properties [61]. However, as seen in Table 4, in spite of the performance prediction of this method, it is slightly lower than the former ((SVM+DE) based on grid search feature selection method, while the latter managed to decrease the computational time. Comparatively, the (SVM+DE) based on grid search feature selection method) is faster by 30 epochs in seconds.

9. Conclusion

The search procedure, as part of the feature selection method, operates based on Differential Evolution (DE) optimization technique as its engine, because the DE engine provides the most optimal and quickest solution. There were two methods proposed in this work for tuning SVM parameters, and comparing them both, it was shown that both possess the capability to increase the performance of SVM, while the (SVM+DE) based on DE feature selection method is quicker that the (SVM+DE) based on grid search feature selection method in terms of CPU time. The SVM+DE combined the natures of the deferential evolution algorithm with the great capability of SVMs within the search process that resulted in the best set of features. The validation of the proposed methodology was carried out with two data sets reported in literature. The results demonstrated that the proposed method herein is capable of producing a promising accurate classification outcome. By comparing our results with previously proposed methods, it can be concluded that this new method is an accurate classifier for future data mining tasks, especially for the case of low similarities of sequences in protein. It was not only able to find the best classification model, but it is also able to minimize the number of features for the SVM classifier, and consequently decrease the time needed to detect rates of PSP.

Acknowledgement

The research was sponsored by UKM “University Kebangsaan Malaysia” under Grant (FRGS/1/2016/ICT02/UKM/01/2).

References

[1] Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17:455-60. https://doi.org/10.1093/bioinformatics/17.5.455.
[2] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology 1995;247:536-40. https://doi.org/10.1016/S0022-2836(05)80134-2.
[3] Rashid MA, Khatib F, Sattar A. Protein preliminaries and structure prediction fundamentals for computer scientists. arXiv preprint arXiv:1510.02775 2015.
[4] Rupp B. Biomolecular crystallography: principles, practice, and application to structural biology. Garland Science, 2009.
[5] Protein N. Spectroscopy: Principles and Practice. Palmer AG III 2007.
[6] Rhodes CJ. Electron spin resonance. Part one: a diagnostic method in the biomedical sciences. Science progress 2011;94:16-96. https://doi.org/10.3184/003685011X12982218769939.
[7] Li L, Cui X, Yu S, Zhang Y, Luo Z, Yang H, et al. PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations. PLoS One 2014;9:e92863. https://doi.org/10.1371/journal.pone.0092863.
[8] Ghanty P, Pal NR. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE transactions on nanobioscience 2009;8:100-10. https://doi.org/10.1109/TNB.2009.2016488.
[9] Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. Enhancing protein fold prediction accuracy using evolutionary and structural features. IAPR International Conference on Pattern Recognition in Bioinformatics 2013:196-207. https://doi.org/10.1007/978-3-642-39159-0_18.
[10] Ding S, Zhang S, Li Y, Wang T. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie 2012;94:1166-71. https://doi.org/10.1016/j.biochi.2012.01.022.
[11] Kurgan L, Cios K, Chen K. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC bioinformatics 2008;9:1. https://doi.org/10.1186/1471-2105-9-226.
[12] Liu T, Jia C. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. Journal of theoretical biology 2010;267:272-5. https://doi.org/10.1016/j.jtbi.2010.09.007.
[13] Mizianty MJ, Kurgan L. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC bioinformatics 2009;10:1. https://doi.org/10.1186/1471-2105-10-414.
[14] Zhang F, Wang D. An effective feature selection approach for network intrusion detection. Networking, Architecture and Storage (NAS), 2013 IEEE Eighth International Conference on 2013:307-11. https://doi.org/10.1109/NAS.2013.49.
[15] Zhang J, Niu Q, Li K, Irwin GW. Model Selection in SVMs using Differential Evolution. IFAC Proceedings Volumes 2011;44:14717-22. https://doi.org/10.3182/20110828-6-IT-1002.00584.

[16] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics 2001;43:246-55. https://doi.org/10.1002/prot.1035.

[17] Chou K-C, Cai Y-D. Predicting protein structural class by functional domain composition. Biochemical and biophysical research communications 2004;321:1007-9. https://doi.org/10.1016/j.bbrc.2004.07.059.

[18] Wu J, Li M-L, Yu L-Z, Wang C. An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. The protein journal 2010;29:62-7. https://doi.org/10.1007/s10930-009-9222-z.

[19] Kong L, Zhang L. Novel structure-driven features for accurate prediction of protein structural class. Genomics 2014;103:292-7. https://doi.org/10.1016/j.ygeno.2014.04.002.

[20] Kong L, Zhang L, Lv J. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology 2014;344:12-8. https://doi.org/10.1016/j.jtbi.2013.11.021.

[21] Paliwal KK, Sharma A, Lyons J, Dehzangi A. Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information. BMC bioinformatics 2014;15:S12. https://doi.org/10.1186/1471-2105-15-S16-S12.

[22] Huang C-L, Wang C-J. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Systems with applications 2006;31:231-40. https://doi.org/10.1016/j.eswa.2005.09.024.

[23] Li Z-C, Zhou X-B, Lin Y-R, Zou X-Y. Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 2008;35:581-90. https://doi.org/10.1007/s00726-008-0084-z.

[24] Sun X-D, Huang R-B. Prediction of protein structural classes using support vector machines. Amino acids 2006;30:469-75. https://doi.org/10.1007/s00726-005-0239-0.

[25] Chou K-C. A key driving force in determination of protein structural classes. Biochemical and biophysical research communications 1999;264:216-24. https://doi.org/10.1006/bbrc.1999.1325.

[26] Zhou G-P. An intriguing controversy over protein structural class prediction. Journal of protein chemistry 1998;17:729-38. https://doi.org/10.1023/A:1020713915365.

[27] Yang J-Y, Peng Z-L, Chen X. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC bioinformatics 2010;11:1. https://doi.org/10.1186/1471-2105-11-S1-S9.

[28] Zhang L, Zhao X, Kong L. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou ׳s pseudo amino acid composition. Journal of theoretical biology 2014;355:105-10. https://doi.org/10.1016/j.jtbi.2014.04.008.

[29] Mohammad TAS, Nagarajaram HA. Svm-based method for protein structural class prediction using secondary structural content and structural information of amino acids. Journal of Bioinformatics and Computational biology 2011;9:489-502. https://doi.org/10.1142/S0219720011005422.

[30] Liu T, Zheng X, Wang J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. Biochimie 2010;92:1330-4. https://doi.org/10.1016/j.biochi.2010.06.013.

[31] Zhang S, Ding S, Wang T. High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. Biochimie 2011;93:710-4. https://doi.org/10.1016/j.biochi.2011.01.001.

[32] Ahmadi Adl A, Nowzari-Dalini A, Xue B, Uversky VN, Qian X. Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences. Journal of Biomolecular Structure and Dynamics 2012;29:1127-37. https://doi.org/10.1080/07391102.2011.672626.

[33] Wang L, Xu Y, Li L. Parameter identification of chaotic systems by hybrid Nelder–Mead simplex search and differential evolution algorithm. Expert Systems with Applications 2011;38:3238-45. https://doi.org/10.1016/j.eswa.2010.08.110.

[34] Cai Y-D, Zhou G-P. Prediction of protein structural classes by neural network. Biochimie 2000;82:783-5. https://doi.org/10.1016/S0300-9084(00)01161-5.

[35] Lyons J, Biswas N, Sharma A, Dehzangi A, Paliwal KK. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. Journal of theoretical biology 2014;354:137-45. https://doi.org/10.1016/j.jtbi.2014.03.033.

[36] Shen H-B, Yang J, Liu X-J, Chou K-C. Using supervised fuzzy clustering to predict protein structural classes. Biochemical and Biophysical Research Communications 2005;334:577-81. https://doi.org/10.1016/j.bbrc.2005.06.128.

[37] Wang ZX, Yuan Z. How good is prediction of protein structural class by the component-coupled method? Proteins: Structure, Function, and Bioinformatics 2000;38:165-75. https://doi.org/10.1002/(SICI)1097-0134(20000201)38:2<165::AID-PROT5>3.0.CO;2-V.

[38] Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K. Prediction of protein structural class with Rough Sets. BMC bioinformatics 2006;7:1. https://doi.org/10.1186/1471-2105-7-20.

[39] Chen Y-W, Lin C-J. Combining SVMs with various feature selection strategies. Feature extraction. Springer, 2006, 315-24. https://doi.org/10.1007/978-3-540-35488-8_13.

[40] Li X, Liu T, Tao P, Wang C, Chen L. A highly accurate protein structural class prediction approach using auto cross covariance transformation and recursive feature elimination. Computational Biology and Chemistry 2015;59:95-100. https://doi.org/10.1016/j.compbiolchem.2015.08.012.

[41] Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC genomics 2014;15:1. https://doi.org/10.1186/1471-2164-15-S1-S2.

[42] Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20:273-97. https://doi.org/10.1007/BF00994018.

[43] Leijôto LF, Rodrigues TADO, Záratey LE, Nobre CN. A Genetic algorithm for the selection of features used in the prediction of protein function. Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on 2014:168-74. https://doi.org/10.1109/BIBE.2014.42.

[44] Zhang L, Zhao X, Kong L, Liu S. A novel predictor for protein structural class based on integrated information of the secondary structure sequence. Biochimie 2014;103:131-6. https://doi.org/10.1016/j.biochi.2014.05.008.

[45] Chuang L-Y, Ke C-H, Yang C-H. A hybrid both filter and wrapper feature selection method for microarray classification2008.

[46] Uncu Ö, Türkşen I. A novel feature selection approach: combining feature wrappers and filters. Information Sciences 2007;177:449-66. https://doi.org/10.1016/j.ins.2006.03.022.

[47] Tan KC, Teoh EJ, Yu Q, Goh K. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications 2009;36:8616-30. https://doi.org/10.1016/j.eswa.2008.10.013.

[48] Bradley PS, Mangasarian OL. Feature selection via concave minimization and support vector machines. ICML 1998;98:82-90.

[49] Garcia-Nieto J, Alba E, Apolloni J. Hybrid DE-SVM approach for feature selection: application to gene expression datasets. 2009 2nd International Symposium on Logistics and Industrial Informatics 2009:1-6. https://doi.org/10.1109/LINDI.2009.5258761.

[50] Guyon I. Practical feature selection: from correlation to causality. NATO Science for Peace and Security 2008;19:27-43.

[51] He X, Zhang Q, Sun N, Dong Y. Feature selection with discrete binary differential evolution. Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on 2009;4:327-30. https://doi.org/10.1109/AICI.2009.438.

[52] Pal M, Foody GM. Feature selection for classification of hyperspectral data by SVM. IEEE Transactions on Geoscience and Remote Sensing 2010;48:2297-307. https://doi.org/10.1109/TGRS.2009.2039484.

[53] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs2000.

[54] Korn F, Pagel B-U, Faloutsos C. On the “dimensionality curse” and the “self-similarity blessing”. IEEE Transactions on Knowledge and Data Engineering 2001;13:96-111. https://doi.org/10.1109/69.908983.

[55] dos Santos GS, Luvizotto LGJ, Mariani VC, dos Santos Coelho L. Least squares support vector machines with tuning based on chaotic differential evolution approach applied to the identification of a thermal process. Expert Systems with Applications 2012;39:4805-12. https://doi.org/10.1016/j.eswa.2011.09.137.

[56] Koloseni D, Luukka P. Differential Evolution Based Nearest Prototype Classifier with Optimized Distance Measures and GOWA. Intelligent Systems' 2014. Springer, 2015, 753-63. https://doi.org/10.1007/978-3-319-11313-5_66.

[57] Mezura-Montes E, Velázquez-Reyes J, Coello CC. Modified differential evolution for constrained optimization. 2006 IEEE International Conference on Evolutionary Computation 2006:25-32.
[58] Mohamed AW, Sabry HZ. Constrained optimization based on modified differential evolution algorithm. Information Sciences 2012;194:171-208. https://doi.org/10.1016/j.ins.2012.01.008.
[59] Apolloni J, Leguizamón G, Alba E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl. Soft Comput. 2016;38:922-32. https://doi.org/10.1016/j.asoc.2015.10.037.
[60] Khushaba RN, Al-Ani A, Al-Jumaily A. Differential evolution based feature subset selection. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on 2008:1-4. https://doi.org/10.1109/ICPR.2008.4761255.
[61] Khushaba RN, Al-Ani A, Al-Jumaily A. Feature subset selection using differential evolution and a statistical repair mechanism. Expert Systems with Applications 2011;38:11515-26. https://doi.org/10.1016/j.eswa.2011.03.028.
[62] Wenwen L, Xiaoxue X, Fu L, Yu Z. Application of Improved Grid Search Algorithm on SVM for Classification of Tumor Gene. International Journal of Multimedia & Ubiquitous Engineering 2014;9:181-8. https://doi.org/10.14257/ijmue.2014.9.11.18.
[63] Ding S, Zhang S, Li Y, Wang T. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie 2012;94:1166-71. https://doi.org/10.1016/j.biochi.2012.01.022.
[64] Huang CL, Wang CJ. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Systems with Applications 2006;31:231-40. https://doi.org/10.1016/j.eswa.2005.09.024.
[65] Pal M, Foody GM. Feature selection for classification of hyperspectral data by SVM. IEEE Transactions on Geoscience and Remote Sensing 2010;48:2297-307. https://doi.org/10.1109/TGRS.2009.2039484.
[66] Garcia-Nieto J, Alba E, Apolloni J. Hybrid DE-SVM Approach for Feature Selection: Application to Gene Expression Datasets. Logistics and Industrial Informatics, 2009. LINDI 2009. 2nd International 2009:1-6. https://doi.org/10.1109/LINDI.2009.5258761.
[67] Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization 1997;11:341-59. https://doi.org/10.1023/A:1008202821328.
[68] Zou D, Liu H, Gao L, Li S. A novel modified differential evolution algorithm for constrained optimization problems. Computers & Mathematics with Applications 2011;61:1608-23. https://doi.org/10.1016/j.camwa.2011.01.029.
[69] Yildiz AR. Hybrid Taguchi-differential evolution algorithm for optimization of multi-pass turning operations. Applied Soft Computing 2013;13:1433-9. https://doi.org/10.1016/j.asoc.2012.01.012.
[70] Wong K-C, Wu C-H, Mok RK, Peng C, Zhang Z. Evolutionary multimodal optimization using the principle of locality. Information Sciences 2012;194:138-70. https://doi.org/10.1016/j.ins.2011.12.016.
[71] Kurgan LA, Zhang T, Zhang H, Shen S, Ruan J. Secondary structure-based assignment of the protein structural classes. Amino Acids 2008;35:551-64. https://doi.org/10.1007/s00726-008-0080-3.
[72] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 1999;292:195-202. https://doi.org/10.1006/jmbi.1999.3091.
[73] Liu N, Wang T. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. FEBS letters 2006;580:5321-7. https://doi.org/10.1016/j.febslet.2006.08.086.
[74] Cheng J, Tegge AN, Baldi P. Machine learning methods for protein structure prediction. IEEE reviews in biomedical engineering 2008;1:41-9. https://doi.org/10.1109/RBME.2008.2008239.
[75] Sun X, Zhang L, Tan H, Bao J, Strouthos C, Zhou X. Multi-scale agent-based brain cancer modeling and prediction of TKI treatment response: Incorporating EGFR signaling pathway and angiogenesis. BMC bioinformatics 2012;13:218. https://doi.org/10.1186/1471-2105-13-218.
[76] Huang C-L, Dun J-F. A distributed PSO–SVM hybrid system with feature selection and parameter optimization. Applied Soft Computing 2008;8:1381-91. https://doi.org/10.1016/j.asoc.2007.10.007.