Protein Structural Class Prediction Using Differential Evolution

Sitanshu Sekhar Sahu[1], Ganapati Panda[1], Satyasai Jagannath Nanda[1] and Sudhansu Kumar Mishra[1]
[1] Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela, India
Email: sitanshusekhar@gmail.com, ganapati.panda@gmail.com, nanda.satyasai@gmail.com, sudhansu.nit@gmail.com

***Abstract*—protein structural class prediction has been a challenging problem in protein science for many years. In this paper we present a new optimization approach using the Differential evolution (DE) for predicting the protein structural class. It uses the maximum component coefficient principle in association with the amino acid composition feature vector to efficiently classify the protein domains. The effectiveness is evaluated by comparing the results with that obtained from other existing methods using a standard database. Especially for all α and α +β class protein, the rate of accurate prediction by the proposed methods is much higher than their counterparts.**

***Index Terms*—Protein, Differential Evolution, Structural class**

I. Introduction

The functional and structural annotation of protein domains is one of the important problems in bioinformatics. In this context, protein structural class information provides a key idea of their structure and also features related to the biological function [1]. The exponential growth of the newly discovered protein sequences by different scientific community has made a large gap between the number of sequence-known proteins and the number of structure-known proteins. Hence there is a challenge to develop automated methods for fast and accurate determination of the structures of proteins in order to reduce this gap.

The concept of protein structural classes was proposed by Levitt and Chothia [2] on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. They proposed ten structural classes, four principal and six small classes of protein structure. But the biological community follows the first four principal classes which are all α, all β, α/β and α+β. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands respectively. The α/β and α+β classes contain both α-helices and β- strands which are mainly interspersed and segregated.

The development of predicting protein structural classes from the primary sequence are mainly focused on the two aspects. Effective representation of the protein sequence and the development of the powerful classification algorithms to efficiently predict the class. Many in-silico structural class prediction algorithms and methods have been developed in recent past. There are number of amino acid indices and features are used for the assignment of the protein sequence. Nishikawa et al. [3] first indicated that the protein structural classes are strongly related to the amino acid composition (AAC). Also auto-correlation functions based on non bonded residue energy, polypeptide composition, pseudo AA composition and complexity measure factor have been used by many researchers [3] [5]. Several classification methods are also proposed such as distance classifier, component coupled methods, principal component analysis [5] and support vector machine [6]. Although satisfactory results have been reported, still there is a need of further improvement in the prediction performance. To achieve such objective the present paper proposes a novel optimization approach for the prediction of protein structural class using differential evolution (DE).

The paper is organized as follows. Section II deals with the maximum component coefficient algorithm for protein structural class prediction. Section III proposes the basics of differential evolution algorithms used for protein structure prediction. Section IV contains the performance of the proposed methods and discussion about it. Finally the conclusion of the paper is reported in section V.

.

II. Methods

A. *Formulation of Maximum Component Coefficient Algorithm as an Optimization framework*

Consider that there are N proteins forming a set S which is the union of four subsets i.e.

$$S = S_\alpha \cup S_\beta \cup S_{\alpha+\beta} \cup S_{\alpha/\beta}$$

where subset S_α contains only all α proteins, S_β contains only β proteins, $S_{\alpha/\beta}$ contains the α/β proteins and $S_{\alpha+\beta}$ contains the α + β proteins. Each protein is represented by a 20-Dimensional feature vector in Euclidean space. The protein corresponds to a point whose co-ordinates are given by the occurrence frequencies of the 20 constituent amino acids.

For a query protein x, let $f_i(x)$ $(1,2,....,20)$ represents the occurrence frequencies of its 20 constituent amino acids. Hence the composition of the amino acids (A_k) in the query protein is given by

$$A_k(x) = \frac{f_k(k)}{\sum_{i=1}^{20} f_i(x)} \quad i,k = 1,2,\ldots\ldots,20 \tag{1}$$

The protein x in the composition space is then defined as

$$P(x) = [A_1(x), A_2(x), \dots \dots, A_{20}(x)]$$

A standard unit vector for each class is defined to represent the norms of the four protein structural classes.

$$X^{\delta} = [x_1 x_2 \dots \dots x_{20}]$$

where δ is α, β, α/β, or α + β and x_i is the average composition of the 20 amino acids occurring in the set of each class, defined as

$$x_i = \frac{1}{n}\sum_{k=1}^{n} x_{k,i} \quad (i = 1,2,\dots\dots,20) \tag{2}$$

where n is the number of proteins in the corresponding structural class. Then the structural class of the query protein can be predicted by computing the Euclidean distance between the protein and each of the standard vector. The Euclidean distance is evaluated as

$$D_{\delta} = \| P(x) - X^{\delta} \| = \left\{ \sum_{i=1}^{20} \{A_i(x) - x_i^{\delta}\}^2 \right\}^{1/2} \tag{3}$$

where ||θ|| represents the norm of the vector. Hence the protein x belongs to the δ-class if the distance D_δ is the smallest among all the distances given by (3).

$$D(P, X_{\xi}) = Min\{D(P, X_{\alpha}), D(P, X_{\beta}), D(P, X_{\alpha+\beta}), D(P, X_{\alpha/\beta})\}$$

In this paper the classification problem is presented as an optimization framework proposed by Zhang and Chou [4]. The query protein is decomposed into four component vectors, each of which corresponds to one of the four standard vectors X(α), X(β), X(α/β) and X(α+β). Hence the query protein is written as

$$P(x) = a_{\alpha}X(\alpha) + a_{\beta}X(\beta) + a_{\alpha+\beta}X(\alpha+\beta) + a_{\alpha/\beta}X(\alpha/\beta) = \sum_{j} a_j x_i(j)$$

where i = 1, 2, · · · , 20 and j = α, β, α/β, or α + β

The variables a_α, a_β, $a_{\alpha/\beta}$, $a_{\alpha+\beta}$ are the four component coefficients of the corresponding class with the constraints

$$\sum_{j} a_j = 1$$
$$0 \le a_j \le 1 \tag{4}$$

Hence the structural class prediction is treated as an optimization problem with the following steps

1. The distance between the query protein *P*(x) and the composite component vector of that protein which is defined as the cost or objective or fitness function is calculated.

$$S(x) = \sum_{i=1}^{20}\left[A_i(x) - \sum_{j} a_j x_i(j)\right]^2 \tag{5}$$

2. The objective function defined in (5) is minimized using the DE, outlined in section III.
3. At the minimal cost function *S*, a protein belongs to the class whose component coefficient is maximum. In other words

$$a_j = \max(\alpha, \beta, \alpha+\beta, \alpha/\beta)$$

Where j is α, β, α/β, or α+β. If j = α, then it concluded that the weight of component coefficient a_α is largest and hence the composition of alpha attribute is more in the query protein and it belongs to the alpha class.

III.THE PROPOSED DE BASED APPROACH FOR STRUCTURAL CLASS PREDICTION

The differential evolution (DE) algorithm proposed by storn and price [7] is a population based stochastic search technique, which is an efficient global optimizer in continuous search domain. The DE is often a reliable candidate for providing performance for a wide range of optimization problems and thus employed for various applications such as clustering, nonlinear identification etc. The steps involved in DE algorithm are outlined below in brief.

A. *Initialization of population*

The individuals of the population are represented as real valued vectors given by

$$Q_{i,G} = [q_{i,G}^{1} q_{i,G}^{2} \dots \dots q_{i,G}^{\psi}]$$

where i = 1,2,……M and M represents the entire set of population. The generation number is G and ψ is the dimensionality of the problem. In this problem we have taken M=20 and ψ=4. The initial population vector is chosen randomly constrained by (4).

B. *Mutation*

In this process a new parameter vector is generated by adding the weighted difference between two individual vectors to a third vector. Fig.1 illustrates the process in detail. For example the weighted difference between two individuals Q_{r2} and Q_{r3} is added to a third individual Q_{r1} to yield the mutant vector ($V_{i,G}$).

$$V_{i,G} = [v_{i,G}^{1} v_{i,G}^{2} \dots \dots v_{i,G}^{\psi}]$$

$$V_{i,G} - Q_{r_1,G} + F * (Q_{r_2,G} - Q_{r_2,G}) \tag{6}$$

where the random indexes r_1, r_2, r_3 ∈1, 2,…..M are mutually different integers and the scaling factor *F* is a positive control parameter.

C. *Crossover*

The crossover operation is applied on the target vector $Q_{i,G}$ and its corresponding mutant vector $V_{i,G}$ to produce a new individual

$$T_{i,G} = [t_{i,G}^{1} t_{i,G}^{1} \dots \dots t_{i,G}^{\psi}]$$

where

$$t_{i,G}^{j} = \begin{cases} v_{i,G}^{j}, & if\ (rand \le CR)\ or\ (j = j_{rand}) \\ q_{i,G}^{j}, & otherwise \end{cases} \tag{7}$$

where j = 1, 2,….ψ. The crossover rate CR is a constant taken within the range [0,1] and j_{rand} is a randomly chosen integer in the range [1,ψ].

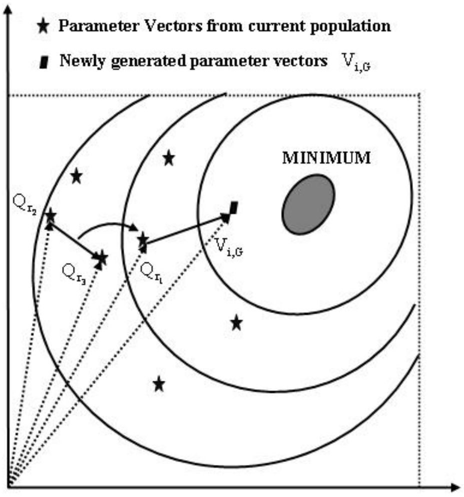
Parameter Vectors from current population
Newly generated parameter vectors $V_{i,G}$
MINIMUM
Q_{r_2}
Q_{r_3}
Q_{r_1}
$V_{i,G}$

Figure 1 Schematic view of the movement of the individuals towards the solution

D. Fitness evaluation

Fitness is a numeric index to measure the effectiveness of the individuals of the population. After the above process the fitness of each individual is determined as defined in (5).

E. Selection operation

The population which will continue as parent vector for next generation is selected.

$$Q_{i,G+1} = \begin{cases} T_{i,G}, & if\ S(T_{i,G}) \le S(Q_{i,G}) \\ Q_{i,G}, & otherwise \end{cases} \quad (8)$$

The steps III-B to III-E are repeated until a specific termination criterion is satisfied. The termination criteria can be either a specific number of generations or a best fit population.

IV. RESULTS AND DISCUSSION

A. Dataset

In order to compare the efficiency of the proposed method with the other existing methods we used the standard data set constructed by Chou for the analysis. The dataset contains 204 proteins, of which 52 are all α, 61 are all β, 45 are α/β and 46 are α+β.

B. Results

In statistical prediction and classification problems, cross validation tests are very often used to examine the power of the predictor or classifier. The jackknife test is the most desirable and useful test used by the researchers to test the efficiency of the method. We have tested the proposed method using the Chou's dataset by the Jackknife test and compared with the existing distance based classifier methods (Hamming distance algorithm and Euclidean distance algorithm), amino acid principal component analysis and support vector machine. The comparisons of the success rate of the methods are listed in Table I.

From Table I, it is shown that the proposed DE based protein structural class prediction is superior to the other existing methods in classifying the protein structural domains. Especially it provides better results for all α and α+β classes which is atleast 6% and 4% higher respectively. Even though the evolutionary methods provide better result, still it is far away from the accuracy of prediction. This accuracy can be further improved by introducing the amino acid sequence order, length and autocorrelation information.

TABLE I
COMPARISON OF SUCCESS RATES OBTAINED BY THE JACKKNIFE TEST FOR THE 204 PROTEIN

Method	α	β	α+β	α/β	Overall
Euclidean Distance	73%	82%	57%	49%	67%
Hamming Distance	71%	89%	57%	49%	68%
AAPCA	82%	97%	78%	82%	85%
SVM	75%	90%	64%	64%	74.5%
Differential Evolution	88.46%	91.8%	82.6%	82.2%	86.27%

CONCLUSIONS

This paper discussed the protein structural class prediction as a constrained optimization problem. The Differential evolution (DE) is used as a potential optimization tool to minimize the objective function. The present study demonstrated that the structural class of a protein is strongly correlated with its amino acids composition. It explores the idea of maximum component coefficient methods by the use of DE. The potential of the proposed method is observed by comparing the predicted results with that of the existing methods and it shows superior performance in the structural class prediction.

REFERENCES

[1] Zhou G.P., Assa-Munt N.," Some insights into protein structural class prediction", PROTEINS: Struct., Funct., Genet. 44,57-57, 2001

[2] M.Levitt and C.Chothia," Structural patterns in globular proteins", Nature, vol.261, No. 5561, pp. 552-558,1976

[3] Nakashima H., Nishikawa K., and Ooi T.," The folding type of a protein is relevant to its amino acid composition", J. Biochem, pp.153-159,1986

[4] Zhang T. and Chou K.C," An optimization approach to predicting protein structural class from amino acid composition, protein secondary structure prediction", J. Mol. Bio. 225,pp. 1049-1063,1992

[5] Qi-Shi Du, Zhi-Qin Jiang,Wen-Zhang He, Da-Peng Li, Kou-Chen chou," Amino Acid Principal Component Analysis(AAPCA) and its Applications in Protein Structural Class Prediction", Journal of Bimolecular Structure and Dynamics, Vol.23, pp. 635-640, 2006

[6] Cai YD, Liu XJ, Xu XB, Zhou GP ,"Support vector machines for predicting protein structural class", BMC Bioinformatics pp. 15, 2001

[7] R. Stron and K. V. Prince," Difference Evolution- A simple and efficient heuristic for global optimization over continuous space", J. global Optim. Vol.11,pp.341-359,1997