Improved Protein Structural Class Prediction Using Genetic Algorithm and ArtificialImmuneSystem

Improved Protein Structural Class Prediction Using Genetic Algorithm and Artificial Immune System

Sitanshu Sekhar Sahu, Ganapati Panda and Satyasai Jagannath Nanda
Department of Electronics and Communication
National Institute of Technology,Rourkela, Orissa, India
Email: sitanshusekhar@gmail.com, ganapati.panda@gmail.com , nanda.satyasai@gmail.com

Abstract

Predicting the structure of a protein from primary sequence is one of the challenging problems in Molecular biology. In this context, protein structural class information provides a key idea of their structure and also other features related to the biological function. In this paper we present a new optimization approach based on Genetic algorithm (GA) and artificial immune system (AIS) 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 structures. The effectiveness is evaluated by comparing the results with that obtained from other existing methods using a standard database. Especially for all α and $\alpha + \beta$ class protein, the rate of accurate prediction by the proposed methods is much higher than their counterparts.

1. Introduction

In the post genomic era the study of sequence to structure relationship and functional annotation has played an important role in molecular biology. In this context protein fold prediction is one of the major tasks in protein science. The functions of protein are relevant to its 3D structure which can be efficiently determined by the protein sequence and structure analysis. The knowledge of protein structural class can provide useful information towards the determination of protein structure [1][2]. 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. There is a challenge to develop automated methods for fast and accurate determination of the structures of proteins in order to reduce this gap. Therefore the development of computational methods for identification of structural classes of newly found proteins based on their primary sequence is essential. The structural class has become one of the most important features for characterizing the overall folding type of a protein and has played an important role in molecular biology, medicine, rational drug design and many other applications.

The concept of protein structural classes was proposed by Levitt and Chothia [3] 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 $\alpha + \beta$. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands respectively. The α/β and $\alpha + \beta$ classes contain both α-helices and β-strands which are mainly interspersed and segregated. The α class proteins contains more than 45% α-helices and less than 5% β-strands whereas the β class proteins contain less than 5% α-helices and more than 45% β-strands. The $\alpha + \beta$ and α/β classes are characterised by more than 30% α-helices, 20% β-strands with dominantly anti-parallel and dominantly parallel β-strands respectively. These class definitions have been well accepted and are still in common use by many researchers.

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. [5] 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 [2] - [4]. Several classification methods are also proposed such as distance classifier, component coupled methods, principal component analysis [7]and support vector machine [8]. 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 an optimization approach for the prediction of protein structural class using the evolutionary algorithms such as genetic algorithm (GA) and artificial immune system (AIS).

The paper is organized as follows. Section 2 deals with

the maximum component coefficient algorithm for protein structural class prediction. Section 3 proposes the basics of the two evolutionary computing algorithms, GA and AIS used for protein structure prediction. Section 4 contains the performance of the proposed methods and discussion about it. Finally the conclusion of the paper is reported in section 5.

2. Methods

2.1. 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 $\alpha+\beta$ 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 $f_i(x)(i = 1, 2, \cdots, 20)$ 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

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

The protein x in the composition space is then defined as

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

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

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

where δ is $\alpha, \beta, \alpha/\beta$, or $\alpha + \beta$ and x_i is the average composition of the 20 amino acids occurring in the set of each class, defined as

$$x_i = \frac{1}{n} \sum_{k=1}^{n} x_{k,i} (i = 1, 2, \cdots, 20) \qquad (2)$$

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

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

where $\|\theta\|$ 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).

$$\text{D(P,X}_\delta) = Min \left\{ D(P, X_\alpha), D(P, X_\beta), D(P, X_{\alpha/\beta}), D(P, X_{\alpha+\beta}) \right\}$$

In this paper the classification problem is presented as an optimization framework proposed by Zhang and Chou [6]. The query protein is decomposed into four component vectors, each of which corresponds to one of the four standard vectors $X(\alpha), X(\beta), X(\alpha/\beta)$ and $X(\alpha+\beta)$. Hence the query protein is written as

$$\begin{aligned} P(x) &= a_\alpha X(\alpha) + a_\beta X(\beta) + a_{\alpha/\beta} X(\alpha/\beta) + a_{\alpha+\beta} X(\alpha+\beta) \\ &= \sum_j a_j x_i(j) \end{aligned}$$

where $i = 1, 2, \cdots, 20$ and $j = \alpha, \beta, \alpha/\beta,$ or $\alpha + \beta$
The variables a_α, a_β, $a_{\alpha/\beta}$, $a_{\alpha+\beta}$ are the four component co-efficients of the corresponding class with the conatraints

$$\begin{aligned} \sum_j a_j &= 1 \\ 0 \le a_j &\le 1 \end{aligned} \qquad (4)$$

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.

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

2) The objective function defined in (5) is minimized using the GA and AIS, outlined in section 3.
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 $\alpha, \beta, \alpha/\beta,$ or $\alpha + \beta$. If $j = \alpha$, 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.

3. The proposed GA and AIS based Approach for structural class prediction

Biologically inspired computing algorithms, theories and techniques have been playing an important role in many fields like optimization,pattern recognition, classification,clustering etc. These are the heuristic search methods that does not fall to local minima and ensures global convergence. The optimization problem pertaining to protein

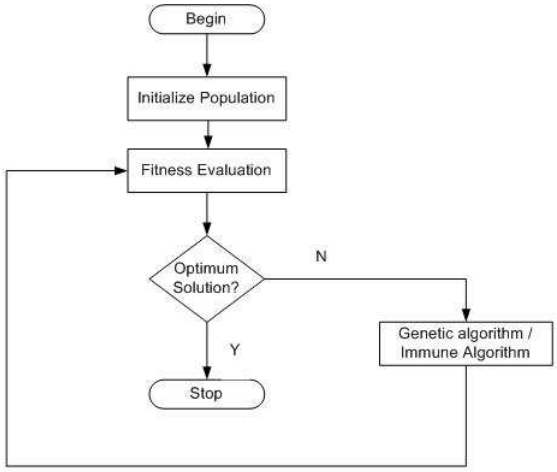

Figure 1. GA/AIS iteration cycle

structural class prediction explained in section 3, is effectively simulated using GA and AIS outlined in this section. The flow graph of these algorithms is shown in Fig. 1.

3.1. The Genetic Algorithm

The Genetic algorithm is an evolutionary computing alorithm based on the concept of survival of the fittest. It mimics the evolution process of the nature and provides a computing technique to get optimal solution [9] [10].The operations like initialization of population, mate selection, crossover, mutation and population replacement constitute a canonical GA. The steps involved in GA are outlined in brief.

3.1.1. Initial population generation. The parameters of the prediction model to be optimized are taken as chromosomes. A set of N chromosomes (individuals) are initialized each of which consists of m bit bianry string. The chromosomes are converted to decimal values and are constrained to the condition defined in (4).

3.1.2. Evaluation of fitness. After the initial population generation, the fitness of each individual is determined using (5). Fitness is a numeric index to measure the effectiveness of each individuals of the population.

3.1.3. Selection Operation. A pair of best fit individuals is selected from the current population for mating.

3.1.4. Crossover Operation. A singlepoint crossover with probability P_c is applied to the selected (parents) individuals to generate a pair offsprings.

3.1.5. Mutation Operation. Random mutation with probability P_m is applied to the newly generated offspring to prevent from premature convergence. It randomly alters the gene from zero to one or from one to zero.

Following the above steps the final set of best population is chosen which replaces the initial population. The cycle is continued till the best fit population is obtained.

3.2. The Artificial Immune System

Artificial immune system is a newly emerging bio-inspired technique that mimics the principle and concepts of modern immunology. The current AISs observe and adopt immune functions, models and mechanisms, and apply them to solve various problems like optimization [11], data classification and system identification [12]. The four forms of AIS algorithm reported in the literature are immune network model, negative selection, clonal selection and danger theory. In this paper the optimization aspect of clonal selection principle is used for protein structural class prediction. The steps involved in the clonal selection algorithm are much similar to GA with slight exception.

3.2.1. Initial population. A binary string which corresponds to a immune cell is initialized to represent a parameter vector and N number of such vectors is taken as initial population each of which represent probable solution.

3.2.2. Fitness Evaluation. The fitness of the set of population is evaluated using (5) to measure the potential of individual solution.

3.2.3. Selection. The parameter vector (corresponding cells) for which the objective function is minimum is selected. Here those cells with low affinity or self-reactive receptors are eliminated.

3.2.4. Clone. The parameter vector (corresponding cells) which yields best fitness value is duplicated.

3.2.5. Mutation. Mutation operation introduces variations into the immune cells. Probability of mutation P_m indicates that the operation occurs occasionally. Here the fitness as well as the affinity of the antibodies gets changed towards the optimum one.

The best fit population (known as memory cells) obtained by the above process replaces the initial population and the cycle continues till the objective is achieved.

4. Results and Discussion

4.1. 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 $\alpha+\beta$. The average sequence similarity scores in the protein classes are 21% for all α, 30% for all β, 15% for α/β and 14% for $\alpha+\beta$. Hence most of the proteins in the dataset are not similar to each other.

4.2. Results

In statistical prediction and classification problems, cross validation tests are very offen used to examine the power of the predictor or classifier. There are three commonly used cross validation tests as independent dataset, sub-sampling test and Jackknife test. Among these 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 proposed methods are implemented using MATLAB in a 2.8GHz Pentium IV computer. In GA and AIS the population of chromosome/immune cells is taken as 20 and the component coefficients are taken randomly. In computational point of view these two algorithms are simpler than the other existing algorithms. Although the two algorithms exhibits equivalent performance, the AIS performs faster than GA. The consumption of CPUtime incase of GA is 1.17 sec whereas that of AIS is 0.80 sec.The comparison of the success rate of the methods are listed in Table 1.

Method	α	β	$\alpha+\beta$	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%
Genetic Algorithm	89%	94%	82%	80%	86.25%
Artificial Immune Syatem	90%	94%	82%	80%	86.50%

Table 1. Comparison of success rates obtained by the Jackknife test for the 204 protein

From Table 1, it is shown that the proposed GA and AIS 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 $\alpha+\beta$ classes which is atleast 7% and 4% higher respectively. Eventhough 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.

5. Conclusion

In this paper the problem of protein structural class prediction is formulated as a constrained optimization problem. The GA and AIS are used as optimization tools to minimise the cost 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 GA and AIS. The proposed techniques achieve the optimum minimal objective function of the geometric distance providing maximum composition of the structural class in the protein to be predicted. 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. 44,57-59,2001

[2] J. P. Klein and C. Delisi, "*Prediction of protein structural class from the amino acid sequence*", Biopolymers, vol. 25, no. 9, pp.1659-1672, 1986.

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

[4] Chou P.Y, "*Prediction of protein structural classes from amino acid compositions*", In Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D., Ed.), pp. 549–586. Plenum Press, New York,1989

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

[6] Zhang T. and Chou K.C.,"*An optimization approach to Predicting protein structural class from amino acid composition. protein secondary structure prediction*", J. Mol. Bi01. 225, 1049–1063,1992

[7] 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 Biomolecular Structure and Dynamics,Vol. 23,pp. 635-640, 2006

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

[9] M. Srinivas, and L. M. Patnaik, "*Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms*", IEEE Transaction on Systems, Man and Cybernetics, Vol. 24, No. 4,pp.657–667, April,1994

[10] Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L, *"Prediction of mitochondrial proteins based on genetic algorithm partial least squares and support vector machine*", Amino Acids 33, pp. 669-675,2007

[11] L N de Charsto and J. V. Zuben, *"Learning and Optimization using Clonal Selection Principle "*, IEEE Trans on Evolutionary Computation ,Special issue on Artificial Immune Systems, vol. 6,issue 3 , pp.239–251,2002.

[12] Satyasai Jagannath Nanda, G. Panda and Babita Majhi,*"Improved Identification of Nonlinear Dynamic systems using Artificial Immune system*",IEEE international conference on Control, Communication and Automation (INDICON-08), pp. 268-273, IIT Kanpur,India, 2008.

[13] L N de Charsto and J. Timmis , *"An Artificial Immune Network for Multimodal Function Optimization "*,IEEE Congress on Evolutionary Computation (CEC'02), vol. 1,pp.699-674,Hawaii,May,2002.