Criterion based Two Dimensional Protein Folding Using Extended GA

T. Kalai Chelvi, P.Rangarajan

Abstract*— In the dynamite field of biological and protein research, the protein fold recognition for long pattern protein sequences is a great confrontation for many years. With that consideration, this paper contributes to the protein folding research field and presents a novel procedure for mapping appropriate protein structure to its correct 2D fold by a concrete model using swarm intelligence. Moreover, the model incorporates Extended Genetic Algorithm (EGA) with concealed Markov model (CMM) for effectively folding the protein sequences that are having long chain lengths. The protein sequences are preprocessed, classified and then, analyzed with some parameters (criterion) such as fitness, similarity and sequence gaps for optimal formation of protein structures. Fitness correlation is evaluated for the determination of bonding strength of molecules, thereby involves in efficient fold recognition task. Experimental results have shown that the proposed method is more adept in 2D protein folding and outperforms the existing algorithms.

***Index Terms*— classification, CMM, criterion analysis, EGA, protein folding, sequence gaps**

I. INTRODUCTION

Extensively, protein folding is the method by which a protein structure deduces its functional conformation. Proteins are folded and held bonded by several forms of molecular interactions. Those interactions include the thermodynamic constancy of the complex structure, hydrophobic interactions and the disulphide binders that are formed in proteins. The folded state of protein is defined as the compact and ordered structure, whereas the unfolded state is substantially less ordered and significantly larger. The mode in which this myriad of unsystematic folds of comparable conformations is a complex issue that still remains.

The primary structure of a protein is given by its linear sequence of amino acids and the position of disulfide bonds. Protein fold recognition is a substantial approach to structure detection that may rely on sequence similarity [4]. In other words, protein structure prediction is defined as the determination of tertiary protein structure by using the information of its primary structures [8]. There described that there are two important issues in protein structure prediction. The first issue is designing a structure model and the second is the design of optimal technology. While considering about structure model, Amino acids are the building blocks of proteins and that is defined as the molecule contains an amine and carboxile groups. Depending upon the structure, size, electric charge and solubility constraints of amino acid side chains, they can be classified under either hydrophobic or hydrophilic. The hydrophobic and hydrophilic can also be termed as the residues of proteins. The energy determination for protein structure model is based on the counting of every two hydrophobic residues that are non-successive in the protein sequence and adjacent neighbors on the lattice [1]. Figure 1 reveals the sample protein residue chain with energy -4. The white square presents hydrophilic residue, while the black represents the hydrophobic. The solid line depicts the protein sequences, whereas the dashed line identifies hydrophobic-hydrophobic (HH) contacts. HH interactions can be stated into two types namely, Connected H (Covalent Bond) and non-connected H (non-Covalent Bond).

On the other hand, computational examination of biological data acquired in genome sequencing is essential for the assimilation of cellular functions and the innovation of new therapies and drugs. Sequence-sequence and sequence-structure discrimination play a critical role in portending a possible function for new sequences. Sequence positioning is accurate in the discovery of relationships between proteins [22].

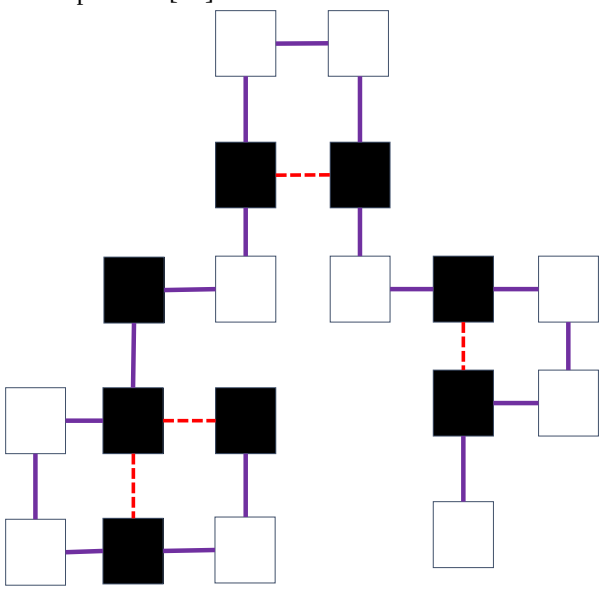

Figure 1: Sample Protein Chain

In general, proteins have various levels of structure, which are described as follows [1]:

1. Primary structure: It is defined as linear structure determined wholly by the number, sequence, and type of amino acid residues.
2. Secondary structure: It is the local structure determined by hydrogen bonding between the amino acids and non-polar interactions between hydrophobic regions.
3. Tertiary structure: This structure is formed with the results of various interactions such as hydrophobic attractions, hydrogen bonding, and disulfide bonding of the amino acids side chains.
4. Quaternary structure: It is determined by the interaction of two or more individual polypeptides (often via disulfide bonds) and results in producing larger functional molecule.

With those basic notes of protein structure analysis, this paper incorporates swarm intelligence based protein structure

Manuscript received on January, 2013.

T. Kalai Chelvi Research Scholar, Sathyabama University, Chennai, India.

Dr.P.Rangarajan Professor& Head/IT R.M.D.Engineering college kavaraipettai,India

formation in order to produce fast and considerably accurate solutions in complex pattern recognition and search problems. Swarm intelligence (SI) is defined as the collective behavior of decentralized and self-organized natural or artificial systems. The bases of evolutionary algorithms are employed with artificial intelligence and there will not follow any centralized control structure. SI algorithms are explored to perform some bioinformatics tasks like micro array data clustering, multiple sequence alignment and protein structure prediction. It is conspicuous that proteins can work together to attain a particular function and often form stable complexes. This feature is correlated with the conceits of evolutionary algorithms and induces in protein folding mechanism. Typically, in protein folding, it involves in predicting the structures of long sequences [16].

In the proposed work, the protein structures are classified using Bayesian approach and analyzed with the parameters such as fitness correlation, sequence gaps, identity and similarity. The fitness value should be considerably high and the sequence gaps should be significantly reduced to form efficient folding. The main intent of this paper is to produce two dimensional protein folding rather than 3D. MAX-MIN parameters of SI are highly considered for 2D protein folding. They signify the exponent of the pheromone levels and the heuristic measures in the random proportional rule, respectively and involves in emphasizing the differences between the arcs. Extended genetic algorithm is framed for 2D folding of long protein sequences with target fitness and sequence length. The adduced work provides a framework for 2D protein folding using the adept conceits in precise manner.

The remainder of this paper is organized as follows. Section 2 provides a deliberation on the related work. Section 3 presents the system architecture and design with an implementation of the affirmed system. Section 4 presents the experimental results and Section 5 concludes the paper with pointers to future work.

II. RELATED WORKS

Myriad researches have been made protein structure and protein folding prediction. Here, some of the works that induced this research is summarized. There was a study about the folding problems with the incorporation of motion planning technique [6]. The paper composed PRM (Probabilistic Roadmap Method) since it produces better results in exploring high-dimensional configuration spaces in protein folding. The paper [14] was discussed about the high-level simulation approach that manages the chemical interactions between all atoms present in the aminoacids. The work was focused on predicting tree dimensional native conformation of protein. Concurrent constraint programming was induced with the simulation based technique. Speed limit of protein folding was analyzed in [17]. Approximately, the speed limit of protein folding for a generic N-residue single domain protein was given as N/100 µs, both in theoretical and experimental approaches. The paper also stated that α proteins folding is faster than β or αβ protein folding. There was also a discussion about the theoretical approaches for the determination of protein folding speed limit such as Polymer Collapse Theory and Kramer's Theory of unimolecular reaction rates.

STAPL (the Standard Template Adaptive Parallel Library) was adopted for parallel protein folding [20]. The paper comprised roadmap analysis, potential energy calculations to achieve effective parallel folding. Sequential codes were utilized to obtain scalable speedups. Guided Genetic Algorithm was presented in [3] for protein folding prediction in two dimensional Hydrophobic-Hydrophilic (HP). The shape of H-core was given for effective boundary determination. New operators such as diagonal move and tilt move were included to form the core boundary. With that, possible sub conformation layers are determined to form the HP mixed layer. Generally, the proteins folding of consecutive chain of aminoacids provide a 3D structure. 2D HP model is applied to achieve this structure. The mechanism could be extended with the analysis of some additional parameters. In addition to that, the paper [5] explained about the inverse protein folding problem on 2D and 3D lattices using the Canonical model. Shifted slice and dice approach was also incorporated to design a polynomial time approximation scheme that solves the inverse protein folding problems and paves a way to analyze the protein landscapes. Moreover, in protein structure prediction problem, lattice model had been utilized for effective folding mechanism. The FCC (Face-Centred-Cube) HP lattice model provided the most compact core and that could map closest to the folded protein [10]. Hybrid Genetic Algorithm that supports square and cube lattice model was adopted for framing the 3D FCC model. The authors have produced optimum conformation, crossover and mutation in 3D, whereas our focus is on 2D protein folding. A different method for appropriate protein folding has been given in [7]. According to that, in low resolution model, twin removal from the genetic algorithm population provided a great impact on effective conformational searching. Since twins cause a population to lose diversity and resulting in both the ineffectual crossover and mutation operations. The HP model given in the paper stated that for an aminoacid sequence S of length N, the protein sequence prediction involves in finding the conformation g, where,

$g^* \in G(S)$ and $E^* = E(G)$

As far as the protein folding is concerned, HP model plays a substantial role and many researchers focused on that. The paper [11] measured the protein structural similarity based on the HP forces. The paper described that aminoacids contained within a k-sized sub-conformation and developed two algorithms namely HP sub-conformation similarity prediction algorithm and HP shape analysis algorithm.

On the other hand, a review paper [13] described about the evolutionary algorithms involved in protein folding problem and their current trends. The authors defined that the proteins are complex macro molecules that accomplish vital function in living organisms. While constructing the protein structure, it must have the following features, according to the computational model.

- A model of protein must be defined by a set of entities representing atoms and their interactions.
- A set of rules that are defining the possible conformations of protein should be included.
- A computationally feasible function should be contained for evaluating the free-residue of each possible conformation.

The paper also examined about the computational approaches such as molecular dynamics, approximation algorithms, genetic algorithms, encoding, fitness functions, Ant Colony Optimization (ACO), etc., for protein folding. Following that, the paper [21] an algorithm called hybrid population based ACO algorithm for protein folding problems. With the base of ACO, the approach considered the

pheromone information that stores information about better solutions and transferred from an iteration of the process to the next. Another approach for protein folding based on BCO (Bacteria Chemotaxis Optimization) was developed for 2D protein folding using lattice model. Foraging behavior of bacteria has taken into the account for framing the model. The algorithm has been applied effectively for proteins with small chain and become ineffectual on long chain protein sequences. An EDA (Estimation of Distribution Algorithm) based method for protein folding was given in [15]. The algorithm replaced the traditional fitness function of HP model with the composite fitness function to enhance the prediction performance. Furthermore, a set of guided operators have been used to increase the diversity of GA population. Backtracking mechanism also invoked to repair for operating with long sequence protein instances. Nonetheless, the proposal was not appropriate for 2D folding.

A novel tensor based method was introduced for performing a Spatio-temporal analysis of protein folding pathways [18]. The resultant of the approach revealed three regions of protein depicted similar and collective attributes across multiple simulations, and also represented valuable dynamic invariants in the protein folding process. Besides, in order to avoid the conformational deformities with hydrophilic residues, an extended HP model has been developed [19]. The conventional HP lattice model provided high degeneracy and that would be extended with FCC lattice configuration, which provides closest resemblance of the real folded 3D protein instance. The results could be amended with further enhancements of the algorithm.

Perhaps, the crossover operators of GA resulted in invalid conformations [12]. While combining that with DFS (depth first search), the potential pathways were revealed and invalid crossover were turned into valid. Random conformations were often applied for maintaining the diversity level. The paper directed its enhancement with the exploration of biological significances. Another work in [9], introduced ABC (Artificial Bee Colony) optimization for 2D protein folding by applying it to HP lattice model. The reliability of the process could be further improved by banding some efficient conceits.

III. PROPOSED WORK

Folding of protein is an intricate and abstruse mechanism. While solving protein folding prediction, the proposed work incorporates Extended Genetic Algorithm with concealed markov model. The main goal is to ascertain the protein fold by conceding both the amino acid sequences and the secondary structure of protein. The amalgamation of the structural and sequential information of proteins is accomplished by the concealed markov model. Moreover, sequence is made with modified Bayesian classification method based on domains. Optimization has been done with the examination of protein sequence parameters such as sequence gaps, identity, similarity and fitness correlation.

Functionally effective proteins are the sequences of amino acids that fold promptly under appropriate conditions, described in chemical point of view. Here, the core of CMM is utterly based on the notion of local structure, which may have different representations. Such representations can be captured by the equivalent classes, production rules or by any clustering scheme.

Apparently, the elementary process of the proposed methodology is to obtain the protein sequences from large repositories. Those sequences may often differ in its amino acids chain length. Perhaps, there is a possibility of the occurrence of erroneous identified interactions and potential false positives. Hence, the protein interactions obtained from the database needs to be re-affirmed by preprocessing. In order to further proceed from the acquired protein interactions, preprocessing of amino acid sequences for classification is much substantial. It normalizes the sequences having variant lengths, eliminate suspect edges and false interactions. In such a way that the enriched protein sequences are obtained and transformed for further process called classification.

3.1 Domain based Classification

The process involves in grouping of protein sequence into its related domains such as Myoglobin, T4-Lysozyme and H-RAS etc. Modified Bayesian classification method is used here, which greatly supports better understanding and predict the structure of proteins. In that, estimation is based on both the number and content of classes. Moreover, the method involves in maximizing the net information gained. Each class in the potential classification is described by a set of attribute mean and variances. Regarding the adduced Bayesian classification, data are defined by the probability distribution. Probability is calculated that the data element 'A' is a member of classes C, where C= $\{C_1, C_2, \dots, C_N\}$

$$P(A \in C) = \frac{P_C(A)}{\sum_{C_1}^{C_N} P_{C'}(A)} \quad (1)$$

Where $P_C(A)$ is the given as the density of the class C evaluated at each data element. Prior probability is also determined that a class described by a specific set of domains (d) exists. It is responsible to measure the relevancy rate of the required sequences categorized under a particular class.

While applying Bayesian classification process, the probability of the domains (d) and scores (s) of a sequence is maximized. When the probability of the applied data is constant, the Modified Bayesian classification yields,

$$P(s, d|A) = \frac{P(s,d)P(A|s,d)}{P(A)} \quad (2)$$

This is achieved based on the theory of finite mixtures. It means that, the observed distribution has been drawn from a population that consists of a number of distinct classes. Correlation between the structural classification and amino acid sequence is optimally used to verify the classification accuracy.

3.2 Concealed Markov Model (CMM)

After classification based on domains, conceits of concealed markov model is applied for training and testing the sequences. The concept emphasizes the correlation between the parts of an entity and the whole. The main focus is that the complex protein pattern $B = \{B_1, B_2, \dots, B_T\}$ can be considered as the sequence of constituents of B_i that is made of strings of symbols $B \in \Sigma$ interrelated in some way. Here, it is assumed that each B_i is assigned to the local structure called C_k. The figure presented below represents the graphical representation f concealed markov model.

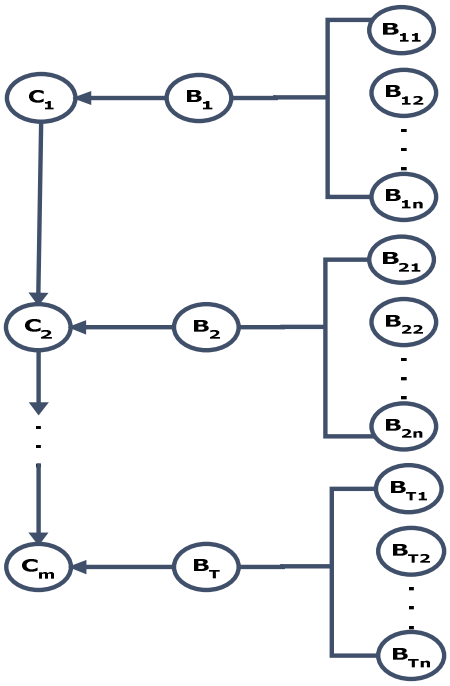

Figure 2: Graphical representation of Concealed Markov Model

The main advantage of CMM in protein folding problem is that the method is capable of predicting the future state with respect to the analysis of the past state. Training of protein sequence is made by the CMM training model with the consideration of local structures obtained from a long sequence. In such a way that all the sequences obtained from the database are trained with local structures, given as C_k. Following that, testing is also made based on specific domains, wherein the trained sequences are tested for accurate folding process. Figure 3 reveals the overall flow of proposed mechanism.

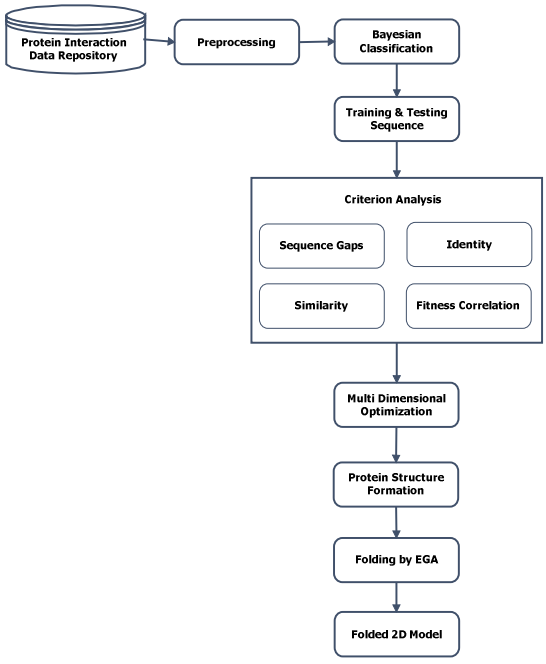

Figure 3: Overall flow of Proposed Mechanism

3.3 Criterion Analysis and Optimization

In the proposed criterion based folding approach, the main parameters that are taken for analysis is sequence gap, identity, similarity and fitness correlation. Depending upon the sequence alignment, the consequences of structural, functional and evolutionary relation between the sequences are obtained. The description for the criterions used here are as follows:

- *Sequence Gap:*

Sequence gap in protein interactions is defined as any maximal or consecutive run of spaces in a single sequence of a given alignment. For example, consider the following sequence alignment,

S = a t t c - - g a - t g g a c c

T = a - - c g t g a t t - - - c c

This alignment has four gaps containing a total of eight spaces. The alignment would be described as having seven matches, no mismatch, four gaps and eight spaces.

- *Identity:*

Identity of a sequence is defined as the identical position of the protein interaction or length of the obtained protein sequence.

- *Similarity:*

Structural similarity of obtained sequences is evaluated for effective folding. Moreover, protein sequences are said to be similar when they have the same arrangement of major secondary structures and having the same topological connections.

- *Fitness Correlation:*

Fitness correlation is evaluated to determine the bonding strength of the protein sequence. Threshold value will be assigned here for examining whether the particular protein sequence is fit to fold. If the fitness correlation value is less than the threshold, the corresponding protein sequence cannot be folded. The criterion is described with an example as,

$$F(x,y) = \begin{cases} +2 & ,x = y \\ -1 & ,else \end{cases}$$

Where x and y are the protein sequences.

Additionally, MAX-MIN parameters, which are given as α (lower bound) and β (upper bound) are also determined based on the protein sequence arcs. Mainly α and β are the Ant system parameters, where it includes in swarm intelligence. The parameters involve in increasing the convergence of folding mechanism and lead to find the iteration best solution for framing the appropriate protein structure. Following the criterion analysis, the protein sequences are optimized for adept two dimensional folding. Multi dimensional optimization has been done to increase the specificity to the novel cofactors and substrates. Implicitly, optimization reduces the sequence gap between the protein interactions and increases its fitness, which is responsible for folding.

3.4 Protein Structure Formation

Swarm intelligence is applied anew for typical protein structure formation. The principles of some of the evolutionary algorithms such as ACO (Ant Colony Optimization) and BCO (Bee Colony Optimization) and ABC (Artificial Bee Colony) optimization [8, 9] are incorporated for structure formation. The procedure is based on the behavior of swarms that moves towards the food source in an optimal way. According to that, the structure has been formed to construct a relative good solution based on the criterions calculated above.

3.5 EGA based protein folding

Consecutively, protein structure has been molded to 2D protein folding using the novel algorithm called Extended Genetic Algorithm (EGA). The EGA algorithm is presented in Table 1. The algorithm begins with long pattern protein sequence (S) and length of the sequence (L) to obtain the 2D

folded sequence. The process of classification, training and optimization is accomplished by the procedures stated above. In the algorithm, F(S) and G(S) define the fitness correlation value and sequence gap respectively, and it is stated that if the fitness value is greater than the target fitness (threshold value), then the sequence is proceeded with the folding process. Else the protein sequence is not fit for folding and it will be discarded.

The EGA based folding algorithm is processed with the best fitted sequence. It is also substantial to verify whether the fitness correlation is present between the MAX-MIN values that are determined earlier. Mainly, the two basic operations performed here are,

1. Crossover
2. Mutation

By crossover, the new offsprings are created from the parental sequences. The sequence is converted into its corresponding binary value by the binary encoding mechanism and then, the crossover is made with initial two binary digits, and iterated with consecutive digits till attains the adept results.

After performing crossover, mutation takes place. This prevents falling all solutions into a local optimum population of solved problem. Further, mutation alters the new offspring in randomized manner. The mutation depends on both the encoding and crossover resultants.

Input: Long Pattern protein Sequence (S), Length of the sequence (L)
Output: Best fitted 2D folded sequence

1. ***Begin Initialization***
   a. Id ->0, F->0, G->0;
2. **end**
3. ***Begin Classification***
4. ***for each*** sequence S
5. Determine score (s)
6. Find domain D
7. Classify by Modified Bayesian Classification Theorem
8. ***end***
9. ***Begin Training***
10. ***for each*** Domain D
11. ***do***
    a. ***while*** No of Domain n(D) = No of Trained Domains $n(D_T)$
    b. Perform- Training domain
    c. Perform- Testing domain
    d. Train sequence until end of the sequence reached
    e. ***end***
12. ***end***
13. ***Begin Criterion based analysis***
14. ***for each*** Sequence S
    a. Evaluate G(S)
    b. Determine Id(S)
    c. Calculate F(S)
    d. Evaluate Similarity
15. ***end***
16. ***Begin Optimization***
17. ***do***
18. ***for each*** sequence (S)
    a. ***if*** F(S) > target fitness;
    b. ***return*** True;
    c. ***else*** discard sequence;
19. ***end***
20. ***Begin Folding***
21. ***for each*** best fitted sequence (S)
22. ***do***
23. ***evaluate*** folding parameters α & β
24. ***while*** α < F(S) < β
    i. crossover;
    ii. mutation;
    iii. ***evaluate*** F(c);
25. ***end***
26. ***Obtain*** 2D fold
27. ***end***

Table 1: Algorithm for Criterion based 2D folding Mechanism

Then, the fitness correlation of each character representation of the protein interactions is evaluated for perfect 2D fold. With those accomplishments, the two dimensional protein fold is obtained. The resultant protein folding is achieved with less sequence gap and highly bonded manner.

IV. EXPERIMENTAL RESULTS

For affording the efficiency of the proposed work, experimentation has been made with the dataset that is obtained from SCOP (Structural Classification of Proteins) database. It is the PDB-49D dataset that is developed by the authors of [23]. They used the protein features based on the statistical information on aminoacids such as transition, composition and distribution. After acquiring the large patterns of protein sequences, it is preprocessed for accurate classification and structure analysis. The sample preprocessed large pattern protein sequences obtained for the process are given below.

Figure 4: Sample Large pattern protein sequences

These sequences are based on some specific domains of protein structures like Myoglobin, Protein (Ribonuclease), T4-Lysozyme, Asparagin-Synthetase, H-RAS-P21, etc. The protein interactions are classified under its domain using Modified Bayesian Classification model, explained in section 3.1. Moreover, to outperform the related approaches, the adduced work provides an extended genetic algorithm with concealed markov model.

4.1 Training and Testing

For the sake of explanation, 27 protein classes are considered in the database; hence, 27 CMM training models have been built. Training of protein sequences is made with the consideration of four types of secondary structures namely, 'helix', 'sheet', 'turn' and 'extended'. Thus, the local structure is fixed here as 4. The input protein sequences are trained based on the four determined local structures along with its classified domains. Approximately, the dataset holds 990 aminoacid sequences. For measuring the power of generalization of CMM based classifier, m-fold cross-validation estimation technique is used. Further, the 990 aminoacid sequences are divided into 5 sets, wherein each contains 198 sequences.

Then, among the 5 sets, one set is selected for testing and the other 4 sets are fixed for training. Iteration takes place by selecting different sets for testing in the same manner until accomplish the testing for final test. Each amino acid sequence has been tested with all 27 CMM models and also

regarding its domains. The one that attains the highest score is the class assigned to that protein sequence. While testing, the optimal model amongst the 27 is the one that adapts the time series sequence of amino acid in a better way. Besides, global accuracy of CMM is the mean of the results obtained from the 5 test sets.

4.2 Performance Evaluation

Following training and testing, criterion based analysis has been performed. According to our consideration, the sequence gap, identity, similarity and fitness correlation is evaluated to capture the best fitted sequence for 2D folding. Figure 5 presented below represents the determined values for those criterions. With those values, the sequences are optimized and examination has been made with the results obtained before and after optimization.

Sequence 1	Sequence 2	Identity	Similarity	Gaps	Fitness Correla...
MVLSEGEWQLVLHVWAKVE...	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIG...	14/43 (32.56%)	20/43 (46.51%)	12/43 (27.91%)	34.00
MVLSEGEWQLVLHVWAKVE...	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIG...	22/91 (24.18%)	32/91 (35.16%)	27/91 (29.67%)	34.50
MVLSEGEWQLVLHVWAKVE...	RGIFEMLRIDEGLRLKIYKDTEGYYTIGIG...	15/67 (22.39%)	25/67 (37.31%)	22/67 (32.84%)	29.50
MVLSEGEWQLVLHVWAKVE...	GQHSMLRIDEGLRLKIYKDTEGYYTIGIG...	13/41 (31.71%)	19/41 (46.34%)	12/41 (29.27%)	30.00
MVLSEGEWQLVLHVWAKVE...	AHVALRIDEGLRLKIYKDTEGYYTIGIGHL...	22/91 (24.18%)	32/91 (35.16%)	27/91 (29.67%)	34.50
MVLSEGEWQLVLHVWAKVE...	MKTAYIAKQRQISFVKSHFSRQLEERLG...	24/107 (22.43%)	38/107 (35.51%)	40/107 (37.38%)	46.00
MVLSEGEWQLVLHVWAKVE...	AYIAKQRQISFVKSHFSRQLEERLGLIEV...	24/107 (22.43%)	38/107 (35.51%)	40/107 (37.38%)	46.00
MVLSEGEWQLVLHVWAKVE...	RQISFVKSHFSRQLEERLGLIEVQAPILSR...	24/107 (22.43%)	38/107 (35.51%)	40/107 (37.38%)	46.00
MVLSEGEWQLVLHVWAKVE...	APILIAKQRQISFVKSHFSRQLEERLGLIE...	24/107 (22.43%)	38/107 (35.51%)	40/107 (37.38%)	46.00
MVLSEGEWQLVLHVWAKVE...	GAVFLVTAYIAKQRQISFVKSHFSRQLEE...	13/55 (23.64%)	20/55 (36.36%)	24/55 (43.64%)	38.00
MVLSEGEWQLVLHVWAKVE...	KESAAAKFERQHMDSGNSPSSSSNYCN...	4/14 (28.57%)	8/14 (57.14%)	0/14 (0.00%)	21.00
MVLSEGEWQLVLHVWAKVE...	FERQHMDSGNSPSSSSNYCNLMMCCR...	4/17 (23.53%)	10/17 (58.82%)	0/17 (0.00%)	22.00
MVLSEGEWQLVLHVWAKVE...	HESLADVKEKFERQHMDSGNSPSSSSN...	6/16 (37.50%)	11/16 (68.75%)	1/16 (6.25%)	25.00
MVLSEGEWQLVLHVWAKVE...	SESAAKFERHESLADVSGNSPSSSSNYC...	7/15 (46.67%)	9/15 (60.00%)	5/15 (33.33%)	23.00
MVLSEGEWQLVLHVWAKVE...	SGNSPSSSSNYCNLMMCCRKMTQGKC...	13/55 (23.64%)	21/55 (38.18%)	14/55 (25.45%)	27.50
MVLSEGEWQLVLHVWAKVE...	MTEYKLVVVGAGGVGKSALTIQLIQNHF...	19/81 (23.46%)	32/81 (39.51%)	23/81 (28.40%)	26.00
MVLSEGEWQLVLHVWAKVE...	KLVVVGAGGVGKSALTIQLIQNHFVDEY...	17/70 (24.29%)	27/70 (38.57%)	16/70 (22.86%)	24.00

Figure 5: Results of Criterion based Analysis

Figure 6 depicts the graphical representation for similarity gap between sequences. It is apparent from the graph that, the similarity between sequences is much reduced after optimization.

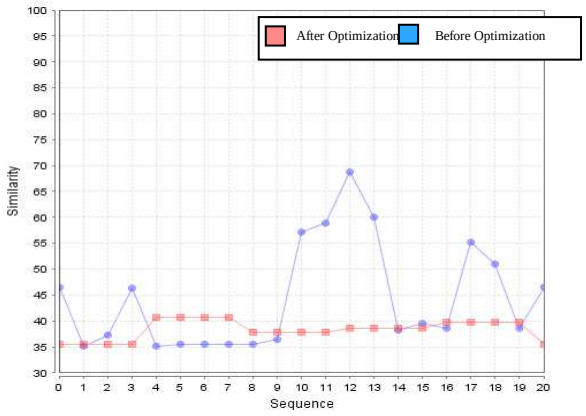

Figure 6: Similarity between sequences

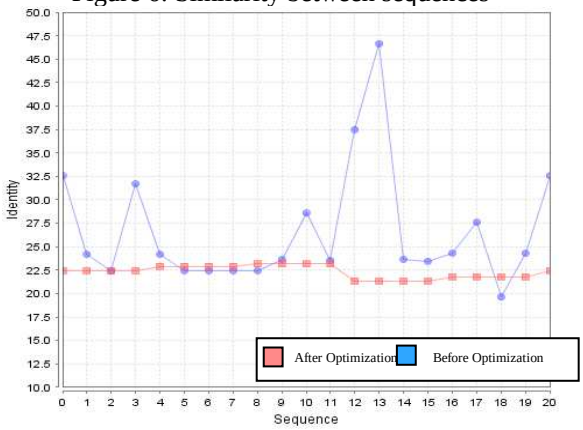

Figure 7: Identity analysis between sequences

The above figure demonstrates identity analysis among sequences, wherein the length of the protein sequence optimized. The following figure represents the results of sequence gap evaluation acquired before and after optimization. Obviously, the sequence gap is considerably reduced by multi dimensional optimization.

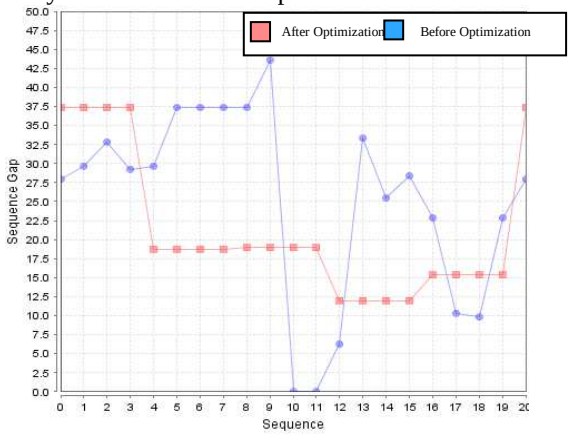

Figure 8: Sequence gap evaluation

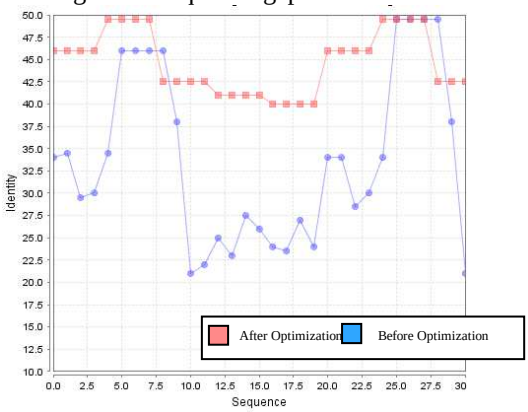

Figure 9: Fitness Correlation between sequences

Figure 9 shown above demonstrates the fitness correlation among sequences. Since the most fitted sequence can be applied for folding, Fitness value has been enhanced effectively. Then, the value gained for each sequence is verified with the target or threshold fitness for 2D folding. Following that, the folding operation is carried out with best fitted sequences according to the proposed Extended Genetic Algorithm. Figure 10 shows the sequence gap of protein interactions before folding and Figure 11 depicts the level of sequence gap after folding.

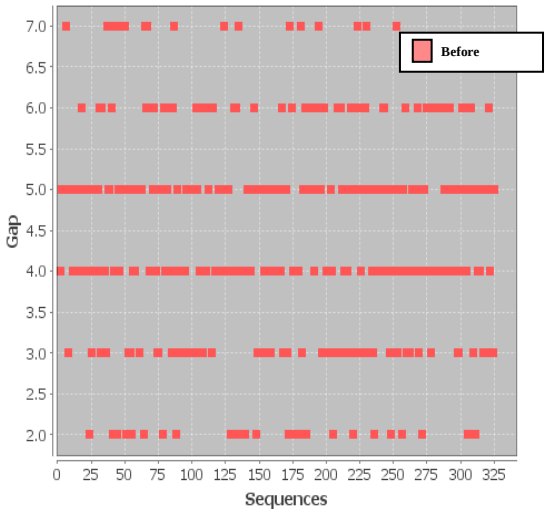

Figure 10: Sequence gap obtained before folding

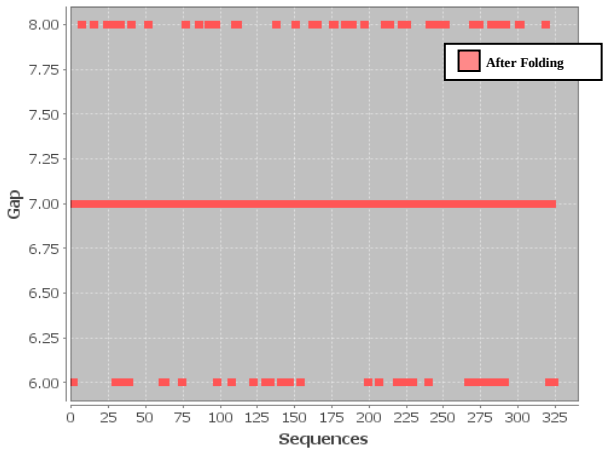

Figure 11: Sequence gap obtained after folding

Amalgamation of the adept conceits in the proposed work leads to the effective folding mechanism. By comparing the two graphical representation presented above, it is substantiated that the adduced folding algorithm reduces the sequence gap between the aminoacid sequences considerably. Thus, the appropriate 2D folding has been obtained for long pattern protein sequences and also it is affirmed that the methodology outperforms the efficiency of the related works.

V. CONCLUSION AND FUTURE WORK

The predominant intention of this paper is to provide a congruous method for protein fold recognition application. With that concern, the proposed work incorporates novel conceits for two dimensional protein folding of large pattern amino acid sequences with varying lengths. Initially, the protein interactions are classified using modified Bayesian classification method that affords appropriate domain based classification results. For training and testing the sequences, CMM based training model has been developed and enforced. Criterion analysis significantly evaluates the fitness correlation of obtained sequences that are to be folded. Accordingly, the sequences are optimized and protein structure is formed with best fitted sequences based on the core of swarm intelligence. Hence, it exploits the relationship among secondary structure of a protein, which is much vital for the recognition of protein 2D fold. Then, the 2D folding process is carried out with the framed EGA. The experimental results show that the proposed work affords precise 2D protein fold with extremely reduced sequence gaps of protein interactions.

Though the results obtained here are very encouraging, future investigation is still necessary. With respect to future enhancements, this work is open for distinctive research areas where beneficial contributions can be done.

REFERENCES

[1] Yudong Zhang, Lenan Wu, Yuankai Huo and Shuihua Wang, "Chaotic Clonal Genetic Algorithm for Protein folding model," In the Proceedings of International Conference on Computer Application and System Modeling, 2010, Vol. 3, pp. 120-124.

[2] Yudong Zhang and Lenan Wu, "Bacterial Chemotaxis Optimization for Protein Folding Model," In the Proceedings of Fifth International Conference on Natural Computation, 2009, Vol. 4, pp. 159-162.

[3] Md. Tamjidul Hoque, Madhu Chetty and Laurence S Dooley, "A New Guided Genetic Algorithm for 2D Hydrophobic-Hydrophilic Model to Predict Protein Folding," In the Proceedings of Congress on Evolutionary Computation, 2005, Vol. 1, pp. 259-266.

[4] D. Bouchaffra and J. Tan, "Protein Fold Recognition using a Structural Hidden Markov Model," In the Proceedings of 18th International Conference on Pattern Recognition, 2006, Vol. 3, pp. 186-189.

[5] Piotr Berman and Bhaskar DasGupta, "The Inverse Protein Folding Problem on 2D and 3D Lattices," Journal on Discrete Applied Mathematics, 2007, Vol. 155, Issue. 6-7, pp. 719-732.

[6] Guang Song and Nancy M. Amato, "A Motion Planning Approach to Folding: From Paper Craft to Protein Folding," IEEE Transactions On Robotics And Automation, 2004, Vol. 20, Issue. 1, pp. 60-71.

[7] Md Tamjidul Hoque, Madhu Chetty, Andrew Lewis, and Abdul Sattar, "Twin-Removal in Genetic Algorithms for Protein Structure Prediction using Low Resolution Model," IEEE/ACM Transactions On Computational Biology And Bioinformatics, 2011, Vol. 8, Issue. 1, pp. 234-245.

[8] R. F. Mansour, "Applying an Evolutionary Algorithm for Protein Structure Prediction," American Journal of Bioinformatics Research, 2011, Vol. 1, Issue. 1, pp. 18-23.

[9] Yudong Zhang, LenanWu, "Artificial Bee Colony for Two Dimensional Protein Folding," Advances in Electrical Engineering Systems, Vol. 1, Issue. 1, pp. 19-23.

[10] Md Tamjidul Hoque, Madhu Chetty and Abdul Sattur, "Protein folding prediction in 3D FCC HP lattice model using genetic algorithm," In the Proceedings of IEEE Conference on Evolutionary Computation, 2007, pp. 4138 – 4145.

[11] Trent Higgs, Bela Stantic, Md Tamjidul Hoque, and Abdul Sattar, "Hydrophobic-Hydrophilic Forces and their Effects on Protein Structural Similarity," Supplementary Proceedings [of the] Third IAPR International Conference on Pattern Recognition in Bioinformatics, 2008.

[12] Md Tamjidul Hoque, Madhu Chetty, Andrew Lewis, Abdul Sattar and Vicky M Avery, "DFS generated pathways in GA crossover for protein structure prediction," In ScienceDirect Journal of Neurocomputing, 2010, Vol. 73, Issue. 13-15, pp. 2308-2316.

[13] Heitor Silv´erio Lopes, "Evolutionary Algorithms for the Protein Folding Problem: A Review and Current Trends," Journal on Computational Intelligence in Biomedicine and Bioinformatics, 2008, Vol. 151, pp. 297-315.

[14] Luca Bortolussi, Alessandro Dal Palu, Agostino Dovier and Federico Fogolari, "Protein Folding Simulation in CCP," In the Proceedings of 20th International Conference on Logical Programming, 2004, Vol. 20, pp. 1-19.

[15] Benhui CHEN, Long LI and Jinglu HU, "A Novel EDAs Based Method for HP Model Protein Folding," IEEE Congress on Evolutionary Computation, 2009, pp. 309-315.

[16] Swagatam Das, Ajith Abraham and Amit Konar, "Swarm Intelligence Algorithms in Bioinformatics," Journal of Computational Intelligence in Bioinformatics, 2008, Vol. 94, pp. 113-147.

[17] Jan Kubelka, James Hofrichter and William A Eaton, "The protein folding 'speed limit'," Journal of Current Opinion in Structural Biology, 2004, Vol. 14, Issue. 1, February 2004, pp. 76-88.

[18] Arvind Ramanathan and Christopher J. Langmead, "Dynamic Invariants in Protein Folding Pathways Revealed by Tensor Analysis," In the Proceedings of 8th Annual International Conference on Computational Systems Bioinformatics, 2009.

[19] Tamjidul Hoque, Madhu Chetty and Abdul Sattar, "Extended HP Model for Protein Structure Prediction," Journal of Computational Biology, 2009, Vol. 16, Issue. 1, Pp. 85–103.

[20] Shawna Thomas Nancy M. Amato, "Parallel Protein Folding with STAPL," Journal of Concurrency and Computation: Practice and Experience, 2005, Vol. 17, Issue. 14.

[21] Torsten Thalheim, Daniel Merkle, Martin Middendorf, "A Hybrid Population based ACO Algorithm for Protein Folding," Proceedings of the International MultiConference of Engineers and Computer Scientists, 2008, Vol. 1, pp. 19-21.

[22] Julia Hockenmaier, Aravind K. Joshi and Ken A. Dill, "Routes Are Trees: The Parsing Perspective on Protein Folding," Journal of Proteins: Structure, Function, and Bioinformatics, 2006, Vol. 66, Issue. 1, pp. 1-15.

[23] L. Lo Conte, B. Ailey, T. hubbard, S. Brenner, A. G. Murzin, and C. Chothia, "Scop: a structural classification of proteins database," Journal of Nucleic Acids Research, 2000, Vol. 28, pp. 257-259.