Webology (ISSN: 1735-188X)
Volume 19, Number 2, 2022

A Novel Soft Computing Technique For Fusion Protein Structure Prediction

P.Bhargavi [1] , U.Subhashini[2] , S.Jyothi[3]

[1,2]Assistant professor, [2]Research Scholar,
[3]Professor

[1,2,3] Department of Computer Science Sri Padmavati Mahila Visvavidyalayam
Tirupati-517 502

Abstract— Fusion protein structure prediction is one of the most significant and difficult issues in bioinformatics. Nowadays, structural bioinformatics majorly focuses on predicting the 3D structure of proteins by various test methods, for example, nuclear magnetic resonance (NMR), electron microscopy or X-ray diffraction. In this circumstance, fusion proteins are complex proteins because they have many structural domains. Determining the fusion protein structure utilizing experimental methods is costly because of the costs of NMR, electron microscopy or crystallography and time consumption. Machine learning methods were utilized to address the problem, and they've had a lot of success in this field. However, there is still opportunity for improvement in terms of approaching the limit. This chapter proposes a new technique for fusion protein structure forecast based on an Enhanced Fuzzy Logic (EFL) soft computing technique. Given a protein sequence database, this proposed work has to construct candidates of length equal to the query fusion protein sequence, find a overall figure of fuzzy matching techniques for query with provided fault tolerance, also applying this overall figure of matches to the secondary structure prediction. Experimental outcomes demonstrated that the EFL technique effectively predicts fusion protein structure.

Keywords — Peptide, Linker, Linker Flexibility, Fusion protein, Fuzzy logic.

1. INTRODUCTION

Proteins play an important role in nearly entire biological procedures; They are the foundation of life. For instance, they keep the constructional honesty of the cell, the transfer and room of little molecules, signalling, control, catalysis, and the immune system [1]— Proteins in nature contain twenty different amino acids. Amino acids are sequentially linked to the carboxyl group of an amino acid that forms a peptide bond by the amino group of the subsequent amino acid [2]. At first, proteins are derived from the cells of microorganisms, plants, and animals. There is a significant increase in the reproduction of natural proteins by recombinant DNA (rDNA) technology. [3]. It also concentrated on cultivating “de novo” proteins, which are not natural and are called fusion proteins. A fusion protein is a protein constructed by merging minimum of two kinds of protein domains. Fusion proteins enhance biological functions through a wide range of biopharmaceutical and biotechnological devices. The protein structure is necessary for comprehending protein function [4]. To

identify the protein functions in the atomic stage, it is occasionally essential to decide their 3D structure [5]. Predicting structures from protein sequences exactly and dependably is a very difficult task in computational biology [6].

1.1 STAGES OF PROTEIN STRUCTURE:

There are four different stages of protein structure.

1) Primary structure
2) Secondary structure
3) Tertiary structure
4) Quaternary structure

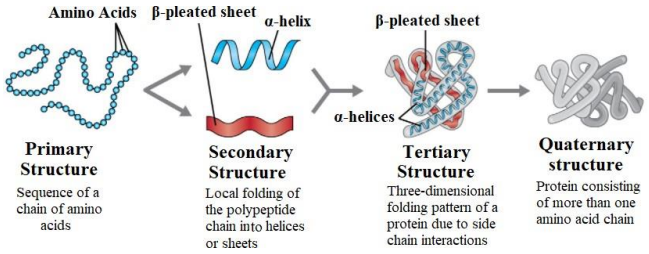

Figure 1: Structure of Protein

Primary structure

The prime structure represents the chain of amino acids in the polypeptide chain [7]. This is fused with peptide bonds, which are produced throughout protein biochemical processes. The two endings of the polypeptide chain is called the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) [8]. The number of residues ever begins at the N-terminal end (NH2-group), which results in the amino group not being involved in peptide bonding. The gene corresponding determines the main formation of the protein.

Secondary structure

The secondary structure represents the most common local substructure in the true polypeptide vertebral chain [9]. The two major kinds of resultant construction are α-helix and β-strand or-sheets. These resultant structures are explained through the pattern of hydrogen bond among the major chain peptide groups. Both the α-helix and the β-sheet refer to the completion of all hydrogen binding donors and receptors in the peptide vertebrae [10]. A few sections of the protein are arranged, however, do not create regular structures. They supposed to not be puzzled with a random coil, an unfolded polypeptide sequence with no permanent three-dimensional construction.

Tertiary structure

The tertiary structure represents the three- dimensional structure created by a one protein molecule (single polypeptide chain) [11]. It would contain one or more domains. The α-helix and β-pleated-sheets are folded into a small universal structure. In addition, the fold is driven by not detailed hydrophobic contacts,

burying the hydrophobic residue from the water. However, the structure remains constant only when partition of a protein domain are locked by particular tertiary contacts, for example, hydrogen bonds, salt bridges, and the tight packing of side chains and disulfide bonds [12]. Disulfide bonds are very uncommon in cytosolic proteins because the cytosol (intracellular fluid) is normally decreasing surroundings.

Quaternary structure

The quaternary structure is a 3D structure that acts as a multimeter of two or more separate polypeptide chains (subunits) [13]. The consequence multimer is ensured by the similar non-covalent communications and disulfide bonds like in the tertiary structure. There exist a number of achievable quaternary shape organizations. Composite of two or extra polypeptides (such as several subunits) are recognized as multimers. Particularly, it is referred to as a dimer if it has two subunits, a trimer if it has three subunits, a tetramer if it has four subunits and a pentamer if it has five subunits. Subunits are frequently related with symmetrical functions, for instance 2-fold axis on a dimer. Multimers with identical subunits are represented by the prefix "homo- " and those with a number of subunits are represented by the prefix "hetero-", for instance, a heterotetramer, for example, two alpha and two beta hemoglobin chains [14].

2. A NOVEL SOFT COMPUTING TECHNIQUE FOR FUSION PROTEIN STRUCTURE PREDICTION

After the fusion protein is fused to the linker, the prediction of its function is significant since separate proteins contain an explained shape and structure. The result is that a closer look at the fusion protein system reveals what it derives from its parent domains. Nowadays, structural bioinformatics is majorly concentrated on predicting the 3D structure of proteins by various test methods, for example, NMR, electron microscopy or X-ray diffraction. In this circumstance, fusion proteins are complex proteins because they have many structural domains. Determining the fusion protein structure utilizing experimental methods is costly because of the costs of NMR, electron microscopy or crystallography and time consumption. To overcome this problem, this work proposes a new technique for fusion protein formation forecast based on the Enhanced Fuzzy Logic (EFL) soft computing technique. Secondary structures participate a major responsibility in the protein structure and protein folds. In particular, these structures are more foldable than others. In addition, these structures are very significant for the forecast of tertiary and quaternary structure. Therefore, this work focuses on forecasting the secondary structure of the fusion protein utilizing the EFL technique.

Fusion protein secondary structure prediction presents data on fusion protein function and relationships. The secondary structure of fusion protein represents the polypeptide vertebrae of local, compatible fusion proteins. There are two conventional secondary structure levels, the α-helix (H) and the β-strand / sheet (E), and an irregular secondary structure level, the coil region (C). Sander [15] implemented a dictionary of Secondary Structure of Proteins

(DSSP) that spontaneously allocates the secondary structure to eight states (I, G, L, S, T, B, E, and H) along with hydrogen-bonding methods. These eight states are commonly made simpler into three states of the helix, sheet/strand and coil. The most broadly utilized convention is that helix is appointed as G, H and I; sheet/strand as B and E; and all other states are appointed as coils. In general, the secondary structure prediction issue is designed as follows: Given a fusion protein sequence with amino acids, forecast whether

or not every amino acid is at the α-helix (H), β- sheet/strand (E), or coil region (C). Fusion protein secondary system prediction is carried out via Q3 precision, which calculates the % of residues for three-tier secondary structures is accurately estimated.

Given a sequence database (training dataset) with N sequence with secondary structure, query mode of length p and % of fault tolerance allowed k. For every sequence in the database, we would like to find out the total number of candidate patterns that almost match the query pattern within the tolerance (maximum k). Initially, candidates of length similar to the query are created. For a given sequence, candidates with n lengths can be created by sliding the window of n characters on the entire sequence. This progress is carried out character by character. By each movement, the subsequence of n characters is taken out.

Once the candidates are created, we will match each candidate to the query to discover the cosine distance (let the candidate sequence and query sequence, the cosine distance will be computed among s and t). Cosine similarity is a measurement utilized to compute how similar the two sequences are regardless of their size. Mathematically, the cosine of the angle between two vectors is obtained from the dot product of two vectors, divided through the product of the magnitude of the two vectors shown in Eq. (2.1).

$$\textbf{Similarity}(p, q) = \cos\Theta = \frac{p \cdot q}{|p|.|q|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \quad \textbf{(2.1)}$$

Since we find the cosine of the two vectors, the output is always -1 to 1, where -1 shows that the two items are different and one shows that the two items are identical. If the distance is close to 1 or suitable for a particular tolerance, the candidate sequence will be selected as an approximate/fuzzy match. This procedure is repeated for the entire database. Once all the approximately matching patterns have been extracted, we use the cosine distance between these approximate matching sequences with the query sequence. After that, we find the sequence with the cosine distance closest to 1 from the query sequence. Finally, we extract the secondary structure of the found sequence from the sequence database.
Example: Consider the following sequence Sequence1: CCGGGSTGGA

Let the uncertainty be CGSTG of span five, and lenience permitted is 30%. To find the structure of a given sequence, we have first to create candidates of length equal to query, and then we encompass to discover cosine space among each candidate for the uncertainty as exposed in Table 1. Then, gather whole fuzzy matches within the given tolerance (Given tolerance is 30%. Therefore, distance should be greater than 0.7). Fuzzy matches within the given tolerance are marked as bold in Table 1. This procedure is repetitive for the whole database and collects all fuzzy matches.

Table 1: Candidates discovered and their distance with query CGSTG for the example Sequence 1

Candidates Generated with Length 5	**Cosine Distance**
CCGGG	0.6
CCGGS	**0.8**
CCGGT	**0.8**
CCGGA	0.6
CCGST	**0.8**
CCGSG	**0.8**
CCGSA	0.6
CCGTG	**0.8**
CCGTA	0.6
CCSTG	**0.8**
CCSTA	0.6
CCSGG	**0.8**
CCSGA	0.6
CCTGG	**0.8**
CCTGA	0.6
CGGGS	**0.8**
CGGGT	**0.8**
CGGGG	0.6
CGGGA	0.6
CGGST	**0.9**
CGGSG	**0.8**
CGGSA	**0.8**
CGGTG	**0.8**
CGGTA	**0.8**
CGSTA	**0.8**
CGSGG	**0.8**
CGSGA	**0.8**
CGTGG	**0.8**
CGTGA	**0.8**
CSTGA	**0.8**
CSGGA	**0.8**
CTGGA	**0.8**
GGGST	**0.8**
GGGSG	0.6
GGGSA	0.6
GGGTG	0.6

GGGTA	0.6
GGGGG	0.4
GGGGA	0.4
GGSTG	**0.8**
GGSTA	**0.8**
GGSGG	0.6
GGSGA	0.6
GGTGG	0.6
GGTGA	0.6
GSTGG	**0.8**
GSTGA	**0.8**
GSGGA	0.6
GTGGA	0.6
STGGA	**0.8**

Followed by, we calculate the Cosine distance between all collected Fuzzy matches to query sequence. Finally, we find a fuzzy matching sequence that has a cosine distance close to 1. Followed by, we extract the secondary structure of this discovered sequence from the sequence database. Table 2 shows the cosine distance between the query sequences to all collected fuzzy matches.

Table 2: Cosine distance between query sequence CGSTG to all collected Fuzzy matches

Sequence Id	Collected Fuzzy Matches	Cosine Distance
1	**CGGST**	**0.9**
2	CGSHL	0.6
3	GECGS	0.8
4	SCGSV	0.6
5	SCGSV	0.6
6	GCGSS	0.8
7	GCGSC	0.8
8	QCGSC	0.6
9	SCGSP	0.6
10	YCGSF	0.6

From Table 2, the Cosine distance between query sequence CGSTG to CGGST is 0.9. It is close to 1 compared with all. Therefore, we extract the secondary structure of CGGST from the sequence database as the final result. Algorithm 1 shows the fusion protein structure prediction. This algorithm first counts the length of the Query fusion protein sequence. It applies Algorithm 2 for Secondary Structure Prediction Based on Enhanced Fuzzy Logic (EFL). Furthermore, the algorithm shows secondary structure.

Algorithm 1: Fusion protein structure prediction

Input : Protein sequence database (PSD), Query fusion protein sequence (QS), Threshold length (TL)

Output : Predicted secondary structure

```
QL = count length of QS
SST = Enhanced Fuzzy Logic (EFL) based
Secondary Structure Prediction (QS, PSD)
// Algorithm 2
If (QL > TL) Then
PAR[] = split(QS, TL)
SS[] = split(SST, TL)
a = 0, i = 0
For each partition P from PAR and each
structure S from SS
QS = P
SST = S
If (i == 0) Then
a = length of QS
Else
a = a + length of QS
End If
End If
```

Algorithm 2 predicts secondary structure using Enhanced Fuzzy Logic (EFL). It first computes all combinations for each protein sequence then calculates cosine distance between these protein sequences with query fusion protein sequence. Then we fetch the secondary structure of a protein sequence, which has the highest distance of all.

Algorithm 2: Enhanced Fuzzy Logic (EFL) based Secondary Structure Prediction

Input : Query fusion protein sequence (QS), Protein sequence database (PSD)

Output : Predicted secondary structure (SST)

```
QL = count length of QS
SP[] = Extract each sequence with
secondary structure from PSD
For each sequence S and secondary
structure,
SST from SP
LS = count length of S
If (LS >= QL) Then
For i = 0 to LS-1
For j = i+1 to LS
combination1 = S.substring (i, j)
combination2 = SST.substring (i, j)
```

Print(" AA: "+QS+" - "+a)

// AA – Target Sequence

Print("Pred: "+ SST) // Pred – Predicted Secondary Structure

Print ("VS Pred: "+ SST.replace All("C", "- ").replace All ("H", " ⚡ ").replace All ("E", "➽"))

// VS Pred – Visually Showed Predicted Secondary Structure

End For

Else

Print(" AA: "+QS+" - "+a)

// AA – Target Sequence

Print("Pred: "+ SST) // Pred – Predicted Secondary Structure

Print ("VSPred: "+ SST.replace All("C", "- ").replace All("H", " ⚡ ").replace All("E", "➽"))

// VS Pred – Visually Showed Predicted Secondary Structure

End For

If (length of combination1 == QL) Then distance = Cosine Distance(QS, combination1) // Eq. (2.1) DT[k] = distance

DWC[k] = distance + "#" + combination1 + "#" + combination2

k++

End If

End If

End If

End For

Sort DT based on descending order

For i = 0 to k

D[] = Split DWC[i] based on “#” If D[0] == DT[0]

SST = D[2]

Break End If

End For

return SST

---

3. Results & Discussion:

The derived composition of the fusion protein could be computed using its atoms' 3D coordinates once the 3D structure of the fusion protein is resolved to utilize NMR or X-ray crystallography. In general, DSSP is a tool utilized to compute the derived structure and allocates one of the subsequent secondary construction kinds to each amino acid in a protein:
S: Bend T: Turn I: π-helix
G: 3-helix B: β-bridge H: α-helix E: β-strand
C: Loops and irregular elements
But, NMR or X-ray is luxurious. Thus, we would desire to forecast the secondary structure of a fusion protein using its primary sequence straightly. For secondary structure forecasting, it is general to make simpler the eight states mentioned above (Q8) into three (Q3) through combining (C, S, T) into C, (H, G, I) into H, and (E, B) into E.

3.1 Dataset:

The protein sequence dataset was fetched from the Kaggle machine learning repository [16]. In addition, this dataset lists peptide sequences and their equivalent secondary structures. Furthermore, the description of columns showed in Table 3.

Table 3: Column description

Column	Description
pdb_id	the id utilized to situate its entry on https://www.rcsb.org/
chain_code	when a protein has many peptides (chains), a chain code is needed to find a specific one
seq	peptide sequence
sst8	the eight-state (Q8) secondary structure
sst3	the three-state (Q3) secondary structure
len	peptide length

hasnonstdaa	Does peptide contain substandard amino acids (B, O, U, X, or Z) or not

3.2 Case Study 1:

To predict the structure of fusion protein, YEHDFHHIREWGNHWKNFLAVMGFFTALSTVMSLL TEVETPIRNEWGCRCNDSS

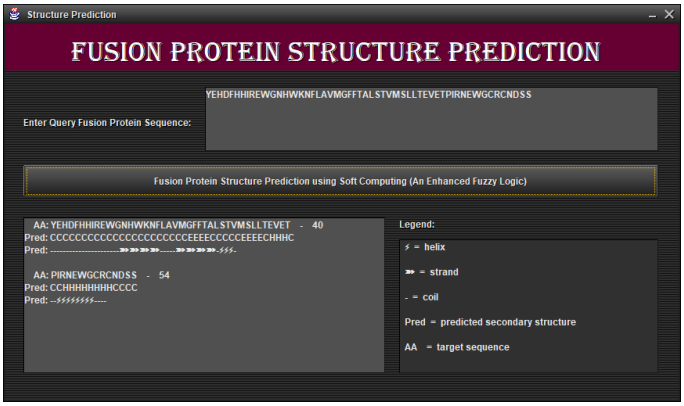

Figure 1: Case Study 1

3.3 Case Study 2:

Fusion Protein Sequence:

To predict the structure of fusion protein, GSHMIKPDETRVKQFLEGFNIETFEMVGTLSNAQGTF ALVKGAGGVHRVRVGDYLGRNDGKVVGISEGKIDVI
EIVPDGEGNWLERPRSLTLKENPGVTQLNRLAAHPPF
ASMGSSHHHHHHSSGLVPRGSHMIEVVVNDRLGKK
VRVKCLGEDSVGDFKKVLSLQIGTQPNKIVLQKGGS VLKDHISLEDYEVHDQTNLELYYL

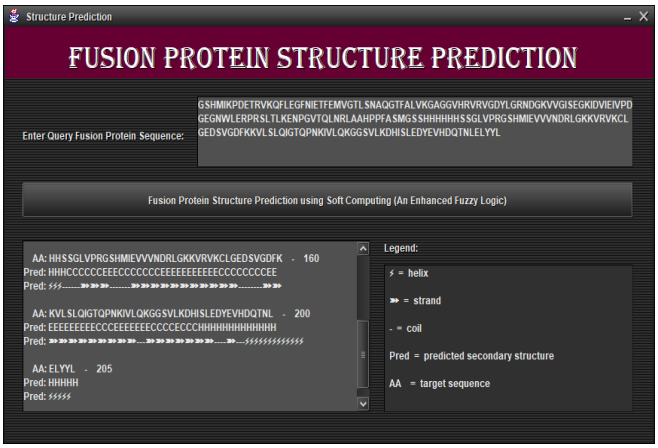

Figure 2: Case Study 2-1

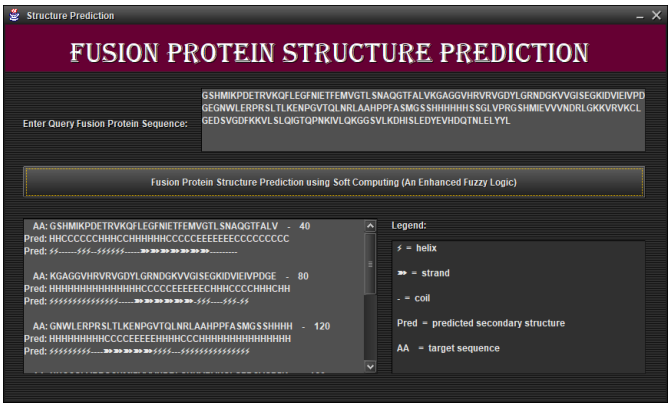

Figure 3: Case Study 2-2

4. Summary:

Once the fusion protein is fused to the linker, the prediction of its function is significant since separate proteins contain an explained shape and structure. The result is that a closer look at the fusion protein system reveals what it derives from its parent domains. Nowadays, structural bioinformatics majorly focuses on predicting the 3D structure of proteins by various test methods, for example, nuclear magnetic resonance (NMR), electron microscopy or X-ray diffraction. In this circumstance, fusion proteins are complex proteins because they have many structural domains. Determining the fusion protein structure utilizing experimental methods is costly because of the costs of NMR, electron microscopy or crystallography and time consumption. To deal with this problem, this paper proposed a new technique for fusion protein structure prediction based on an Enhanced Fuzzy Logic (EFL) soft computing technique. This technique predicts the secondary structure of fusion protein efficiently. In future, the EFL technique may be enhanced to predict the Tertiary and Quaternary structure of the fusion protein.

References:

[1] Ma, Y., Liu, Y., & Cheng, J. (2018). Protein secondary structure prediction based on data partition and semi- random subspace method. Scientific reports, 8(1), 1-10.

[2] Coin, I. (2018). Application of non-canonical crosslinking amino acids to study protein-protein interactions in live cells. Current opinion in chemical biology, 46, 156-163.

[3] Shah, N. J. (2019). Recombinant DNA technology. An Introduction to Basics of Pharmacology and Toxicology (pp. 413-417). Springer, Singapore.

[4] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronne berger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with Alpha Fold. Nature, 1-11.

[5] Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., & Marks, D. S. (2019). Inferring protein 3D structure from deep mutation scans. Nature genetics, 51(7), 1170-1176.

[6] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710.

[7] Lella, M., & Mahalakshmi, R. (2017). Metamorphic proteins: the emergence of dual protein folds from one primary sequence. Biochemistry, 56(24), 2971-2984.

[8] Hein, A., Cole, C., & Valafar, H. (2021). An investigation in the optimal encoding of protein primary sequence for structure prediction by artificial neural networks. In Advances in Computer Vision and Computational Biology (pp. 685-699). Springer, Cham.

[9] Torrisi, M., Kaleel, M., & Pollastri, G. (2019). Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Scientific reports, 9(1), 1-12.

[10] Wardah, W., Khan, M. G., Sharma, A., & Rashid, M.
A. (2019). Protein secondary structure prediction using neural networks and deep learning: A review. Computational biology and chemistry, 81, 1-8.

[11] Farhadi, T. (2018). Advances in protein tertiary structure prediction. Biomedical and Biotechnology Research Journal (BBRJ), 2(1), 20.

[12] Álvarez, Ó., Fernández-Martínez, J. L., Fernández- Brillet, C., Cernea, A., Fernández-Muñiz, Z., & Kloczkowski, A. (2018). Principal component analysis in protein tertiary structure prediction. Journal of bioinformatics and computational biology, 16(02), 1850005.

[13] Nakamura, T., Oda, T., Fukasawa, Y., & Tomii, K. (2018). Template-based quaternary structure prediction of proteins using enhanced profile-profile alignments. Proteins: Structure, Function, and Bioinformatics 86, 274-282.

[14] Bliven, S., Lafita, A., Parker, A., Capitani, G., & Duarte, J. M. (2018). Automated evaluation of quaternary structures from protein crystals. PLoS computational biology, 14(4), e1006104.

[15] Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules, 22(12), 2577-2637.

[16] Kaggle.com. 2021. Protein Secondary Structure. [online] Available at: <https://www.kaggle.com/alfrandom/protein-secondary- structure? select=2018-06-06-pdb-intersect-pisces.csv> [Accessed 28 August 2021].