R3D Prediction Model: Remote Homology Prediction using Protein Contact Map with Dimensionality Reduction

A.Hepsiba, R.Balasubramanian

Abstract*: *Bioinformatics is one of a developing field that utilizations evaluation to extort information from Biological Data. Bioinformatics approaches are regularly utilized for significant activities that create expansive informational collections. Two basic tremendous scale practices that use bioinformatics are genomics and proteomics. Proteins are the far reaching, complex atoms that are essential for normal working of cells. 20% of the human body is involved proteins. Proteins are involved smaller units called amino acids, which are building squares of proteins. Protein remote homology identification and recognition are focal issues in bioinformatics. Sequence homologies are a vital source of data about proteins. In this research, the framework propose different strategy that diminishes the high dimensionality of the vector representation in remote homology detection by utilizing models that are characterized at the 3D level and consequently are very structurally and practically related. Subsequently, the 3D models are mapped from the protein primary sequence. The framework proposes to address the issue of remote homology identification by reducing 3D structure models. The new technique, called remote homology identification by the Reduction of 3D models (remote-R3D), is introduced and tested on various protein families.

Keywords* : *Remote Homology, Protein Contact Map, R3D model, 3D Structure Prediction, Reduced Homology.

I. INTRODUCTION

Bioinformatics includes the combination of computers, programming, and databases to focus the biological difficulties. Bioinformatics approaches are regularly utilized for significant activities that produce substantial informational indexes. Two significant large scale activities that utilize bioinformatics are genomics and proteomics. Genomics alludes to the examination of genomes. A genome can be thought of as the total arrangement of DNA groupings that codes for the inherited material that is passed on from generation to generation. These DNA arrangements incorporate all of the genes (the practical and physical unit of heredity go from parent to posterity) and transcripts (the RNA duplicates that are the underlying advance in unraveling the hereditary data) included within the genome. In this way, genomics alludes to the sequencing and assessment of these genomic substances, including genes and transcripts, in a living being. 20% of the human body is comprised of proteins. Proteins are the substantial, complex atoms that are basic for ordinary working of cells. They are important for the structure, capacity, and control of the body's tissues and organs. Proteins are comprised of littler units called amino acids, which are building blocks of proteins. They are joined to each other by peptide bonds framing a long chain of proteins. An amino acid contains both a carboxylic gathering and an amino gathering. Amino acids that have an amino gathering reinforced specifically to the alpha-carbon are referred to as alpha amino acids. Each alpha amino acid has a carbon molecule, called an alpha carbon, Cα ; clung to a carboxylic acid, – COOH gathering; an amino, – NH2 gathering; a hydrogen molecule; and a R assemble that is distinctive for each amino acid.

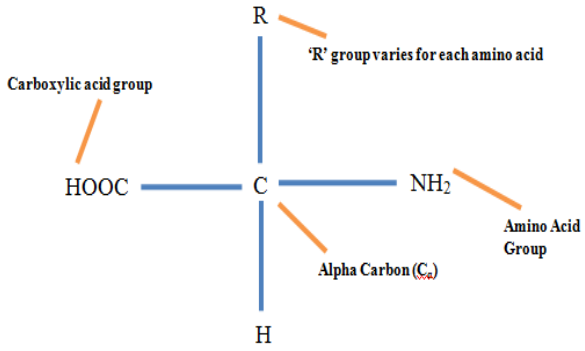

Figure 1: Protein Interactions

Fundamental amino acids are the amino acids which you require through your eating routine on the grounds that your body can't make them. While superfluous amino acids are the amino acids which are not a basic portion of your eating regimen since they can be combined by your body. Homology modeling is utilized to anticipate the 3D-structure of an obscure protein in view of the known structure of a comparative protein. Along with progress, sequence changes considerably quicker than structure. It is conceivable to recognize the 3D-structure by taking a gander at an atom with some sequence similarity. Sequence character is required with a specific number of adjusted residues to achieve the protected homology modeling zone. For a sequence of 100 buildups, for instance, a sequence similarity of 40% is adequate for structure prediction. At the point when the sequence character falls in the protected homology modeling zone, we can expect that the 3D-structure of the two sequences is the same.In this manner, even proteins that have separated considerably in sequence yet at the same time share perceptible similarity will likewise share normal basic properties, especially the general fold.

Revised Manuscript Received on October 15, 2019.

Mrs.A.Hepsiba, Associate Professor, Department of Master of Computer Applications in Karpaga Vinayaga College of Engineering and Technology.

Dr.R. Balasubramanian,& Dean, Department of Computer Applications Karpaga Vinayaga College of Engineering & Technology

Since it is troublesome and tedious to acquire exploratory structures from techniques, for example, X-beam crystallography and protein NMR for each protein of intrigue, homology modeling can give helpful auxiliary models to producing hypotheses about a protein's function and coordinating further experimental work.

This paper contains the following sections as an organization of the paper: section 2 discussed the review literature of this research. In section 3, describes the motivation and problem statement of this paper. Research prototypical of this paper is described in the segment 4. The section 5 illustrates the performance evaluation of the proposed model. Finally the paper ends with the conclusion and future work.

II. RELATED REVIEW

spatial association of a protein its structure in 3 dimensions can be a vital aspect for understanding its capacity. Just when a protein is in its right three-dimensional structure, or adaptation, is it ready to perform quickly. A key arrangement in representing however proteins work is that perform is obtained from 3-dimensional structure, and three-dimensional structure is determined by amino acid groupings [1] to foresee the three-dimensional structure of a protein. Auxiliary structure prediction of a protein from its amino acid sequence is a pivotal advance. A few of the overall algorithms utilize the similarity and similarity [2, 3] to proteins with known auxiliary structures inside the protein information Bank, elective proteins with low comparability estimates need single sequence way to deal with the disclosure of their secondary structure.

Z. Zhen, J. Nan [4] arranged a totally exceptional system augmented spiral premise work neural systems for prediction of protein auxiliary structure. To shape the method identical to various optional structure prediction ways, they utilized the benchmark examination informational index of 126 protein chains during the paper. They furthermore considered an approach to utilize transformative data to broaden the forecast exactness. C. Nandini et al.[5] clarifies numerous systems used by various explores for the order of proteins and moreover gives an outline of different protein sequence characterization procedures. From the huge information to determine the concealed learning with the goal that it's utilized in wide choice of regions to configuration sedate, to distinguish disease, and in order of protein sequence and so forth.

David T. Jones, [6] proposed two-arrange neural system to anticipate protein secondary structure based on Position Specific Scoring Matrices (PSSM) created by PSIBLAST. PSI-BLAST is an incredible arrangement looking through strategy. This creates sequence profiles as a major aspect of the pursuit procedure, and here middle of the road PSIBLAST is investigated, as an immediate contribution to an optional structure forecast strategy instead of extricating the successions, and delivering an express numerous sequence arrangement as a different advance.

Daniel C. Berwick expressed three techniques for Protein expectation, for example, 1D, 2D and 3D. 1D incorporates into auxiliary structure and it is dissolvable availability, which buildups are presented to water, which are covered and transmembrane helices. Expectation in 2D incorporates between buildup/strand contacts. At that point at long last forecast in 3D implies homology displaying, crease acknowledgment, sub-atomic elements, section get together and abdominal muscle initio expectation. Likewise in [7], a SVM with a combined bit capacity made out of three RBF, polynomial and straight portions is applied. These parts are combined utilizing a unique weighting technique relegating a load to each single piece in the last portion condition dependent on its exhibition. In [8] five recently contemplated strategies and two recently proposed techniques have been talked about. The recently contemplated techniques depend on single-organize SVM and the recently proposed ones are two-arrange SVM based strategies created joining the single-arrange ones.

III. MOTIVATION AND PROBLEM STATEMENT

Assistant upon whether relative structures are found in the PDB library, the protein structure estimate can be sorted into design based exhibiting and free modeling. The vital issues/attempts in the field of protein structure assumption include: first, for the groupings of equivalent structures in PDB (especially those of evacuated homologous association with the target), how to recognize the correct structure and how to refine the organization structure closer to the neighborhood; second, for the arrangements without realistic outline, how to gather models of right structure without promptness.

Protein remote homology identification and recognition are focal issues in bioinformatics. Sequence homologies are a vital wellspring of data about proteins. Different Sequences alignments of protein sequence contain much data with respect to developmental procedures. An extremely successful method for deriving the structure or capacity of a formerly un-commented on protein is by means of sequence homology with at least one protein whose structure or function is as of now known. In this research, a novel strategy for protein remote homology detection has been introduced.

Motivation

- The concept of homology modeling is involved in the observation the protein 3D structure is better preserved than amino acid sequence.
- Protein remote homology detection is a key technique to explore the structures and functions of the proteins based on their primary sequence information, which is important for both basic research and practical applications (such as modeling the structures of target proteins for computer aided drug design).

IV. RESEARCH MODEL

Protein remote homology identification and recognition are focal issues in bioinformatics. Sequence homologies are a critical source of data about proteins. In this paper, a novel technique for protein remote homology identification has been proposed is shown in Figure 2. In this paper, the framework propose different strategy that diminishes the high dimensionality of the vector representation in remote homology detection by utilizing models that are characterized at the 3D level and consequently are very structurally and practically related.

Subsequently, the 3D models are mapped from the protein primary sequence. The framework proposes to address the issue of remote homology identification by reducing 3D structure models. The new technique, called remote homology identification by the *Reduction of 3D models (remote-R3D)*, is introduced and tested on various protein datasets.

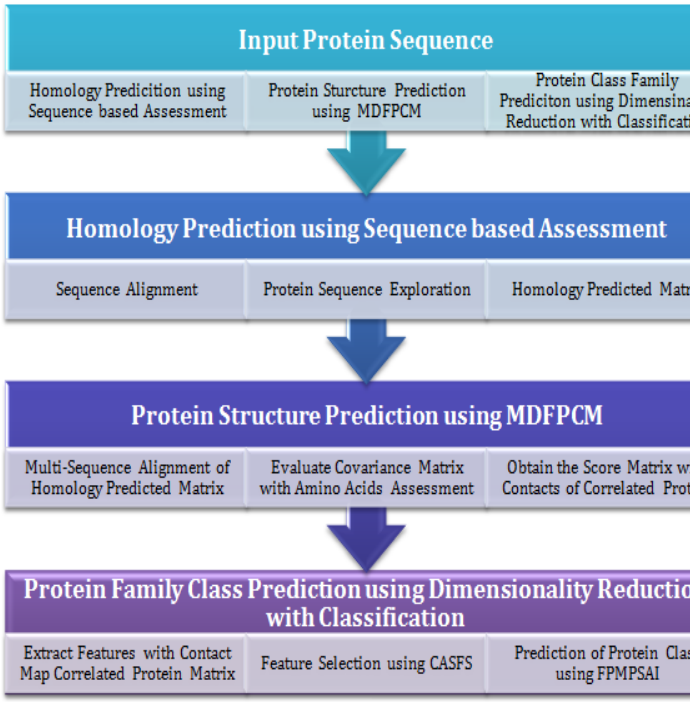

Figure 2: Architectural Design of the Proposed Research Model

The overall process flow for the proposed system is illustrated in figure 3. Considering the execution accomplished by the diverse methodologies on distinctive datasets the paper demonstrate that the vast majority of the proposed multiclass based classification approaches are very successful in unraveling the remote homology identification and fold recognition issues and that the schemes that utilize predictions from binary models built for hereditary classifications inside the protein progressive system have a tendency to prompt reduce error rates as well as diminish the quantity of blunders in which a superfamily is allocated to a totally unique fold and a fold is anticipated as being from an alternate protein class. The outcomes likewise demonstrate that the restricted size of the training data makes it difficult to learn complex second-level models, and that model of direct unpredictability prompt reliably better outcomes. Incorporating protein sequence and structure data has along these lines turn into an objective, particularly in the field of protein structure identification from sequence, by methods for homology modeling (HM) techniques. In this framework, the homology evaluation is performed utilizing three layers:

- Sequence Based Homology Modeling
- Contact map based Structure Prediction
- Feature based Protein Family Prediction

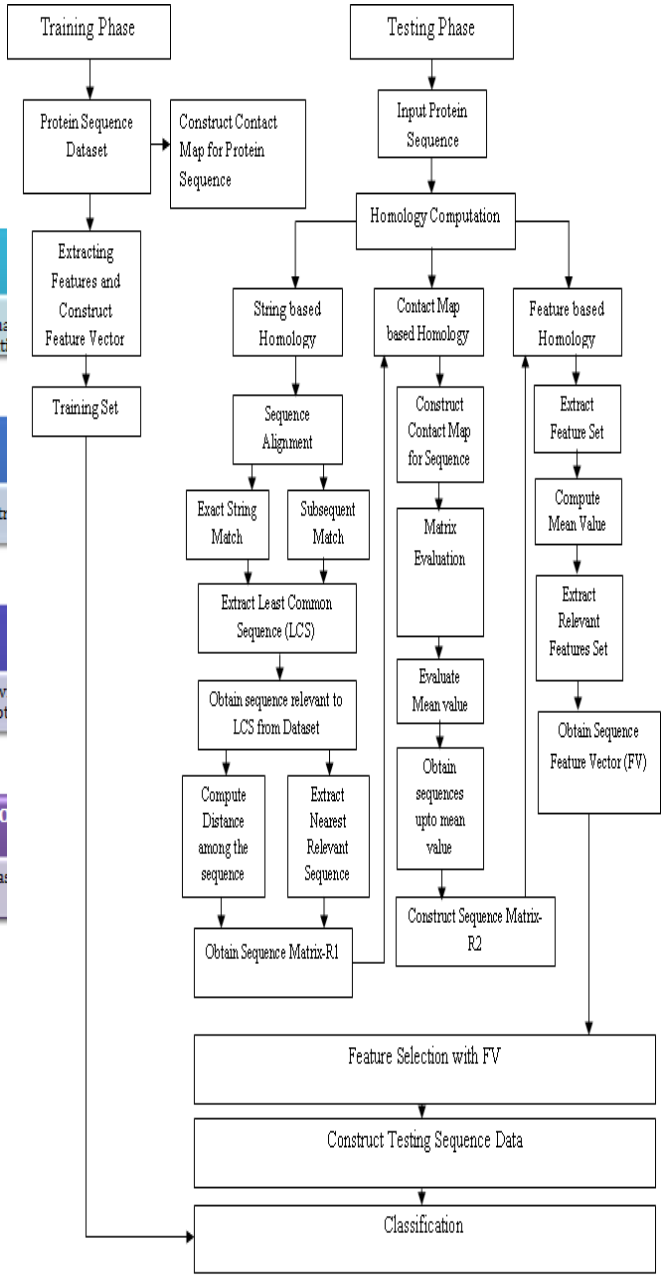

Figure 3: The Overall Process Flow of the Proposed Model

A. Predicting Protein Homology of the Protein Sequence using Sequence Exploration Assessment

The expression "Homology Modeling", likewise called near demonstrating or here and there Template based Modeling (TBM), alludes to demonstrating a protein 3D structure utilizing a known trial structure of a homologous protein (the layout). Auxiliary data is dependably of extraordinary help with the investigation of protein work, elements, communications with ligands and different proteins. The "low-determination" structure given by homology presenting contains adequate data about the spatial game plan of essential deposits in the protein and may control the outline of new investigations, for instance site-coordinated mutagenesis. The layers are structured to compute and regularize the homology of the protein sequence.

The three layers of this modeling integrate the relevant homology computation and reduction in features for the class detection. The sequence based homology modeling is evaluating the similarity among the protein sequences, if similarity between two proteins is detectable at the sequence level, then structural similarity can usually be assumed. Thus the proposed model of homology prediction *Correlated Extensive-Subsequent Searching based Homology Prediction (CESHP) Model* is involved in the evaluating homology prediction.

At first this modeling as a rule begins via scanning the PDB for known protein structures utilizing the objective grouping as the inquiry. This searching is for the most part done by contrasting the objective sequence and the sequence of every one of the structures in the database. The strategy acknowledges single or various record ways and utilizes the right document handlers, iterator items, and parsers for the user. The input sequences are parsed. In a representative sequence each base or buildup monomer in each sequence is denoted by a letter. This depends on the presumption that the joined monomers equally separated along the single measurement of the particle's essential structure. Starting now and into the foreseeable future the system will allude to an alignment of two protein sequences.

VKLSEDQEHYIKGVWKDVDHKQITAKALERVFV
VYPWTTRLFSKLQGLFSANDIGVQQHADKVQRA
LGEAIDDLKKVEINFQNLSGKHQEIGVDTQNFKLL
GQTFMVELALHYKKTFRPKEHAAAYKFFRLVAE
ALSSNYH

: : ‖ | || : ‖ |: | : : || | | : : | :| | |::| | : |

---A-----F----------TA-C-E----------K-Q---T---IG----
--KI--A--Q-V--LAK--------S---PE----A--Y---G----AE-C
L----------A-RLF--V-----T--H

Figure 4: Sequence Alignment Output

Each component in a sequence trace is either a match or a gap. The folding among one of the two aligned sequences are identical to its associate with the other related amino-acid letter codes in the two sequences are vertically aligned in the sequence trace match. At the point when a residue in one sequence alignment appears have been erased since the expected difference of the sequence from its associate, its "nonappearance" is labeled by a dash in the inferred sequence. When a residue seems to have been embedded to create a longer sequence a dash shows up inverse in the unaugmented sequence. Since these dashes denotes to "gaps" in one or other grouping, the activity of embeddings such spacers is known as gapping.

The sequence parsing performs the parsing of the sequence with complete format. The sequence of the protein contains header information and sequence data information. The parser that extracts the sequence header information and sequence data information from the format. The extracted information is store in separate format for the further identification of the protein sequence. The string matching algorithms are used to find out the sequence in dataset.

A string denotes the characters of the sequence. The proposed model will represents a string with 0-indexed array S. Thus the string S =”ASDFRGGR” is indeed an array [‘A’,’S’,’D’,’F’,’R’,’G’,’G’,’R’]. The string (Sequence of character) length is denoted by S[i].

Table I: Algorithm for Exact Sequence Matching

Algorithm ExactMatching(S - Protein Sequence)

{
ReadString(S)
Find the length of the character N=StrLen(S)
For each i =1 to N
{
For each j=1 to i
{
A[]={s_1, s_2,….s_i} ϵ S
ExractSubStr(A[i,j])
}
If(A[i,j] isequals s_i)
{
SP=Return the exact matched sequence Pattern
}
Else
{
ExtractSubStr(A[i,j])
}
}
Return SP
}

A substring is a sequence of continuous adjacent components of a string, it will indicate the substring beginning at i and consummation at j of string S by S[i...j]. A prefix of a string S is a substring that begins at position 0 and postfix a substring that closure at |S|-1. An appropriate prefix of S is a prefix that is distinctive to S. So also, an appropriate postfix of S is an addition that is distinctive to S. The + operator will speak to string concatenation.

To measure the similarity of two or more sequences, its longest common subsequence should be detected. Longest Common Sequence (LCS) of two sequences is a subsequence, of maximum possible length, which is common to both the sequences and obtains the relevant sequence matrix. Then the computation progress to the subsequent string matching for the least sequence with common aspect is performed. Let's consider an example. Suppose we have the following two protein sequences: HJSUGFLSGJSGK and SDJGSDGSDJGDSF. The LCS of the two sequences is SDFGHE, which can be obtained from the following alignment.

SDFSFGSRTJKIYUIUDFBSFAFQETRYRTKGNDFBS
GWRYXCVBFGJYUYWRQWQDACAADSDGSGDHFJF
IOPUWTFDQAVVSBDGFNBBXBTURRYWGADVSDS
BBSFHYKTMFSAFAFAGSGRRYUYITYJETSSFBDGN
KUTISRHWBSFBSFRYURTWSFGFDH

The other probable common sequence with shorter length is available. Although the multiple LCS is possible while evaluating the sequences.

Table II: Algorithm for Sub-Sequent Matching

```
Algorithm Subsequent(S – Protein Sequence)
{
ReadString(S)
L=StrLen(S) // compute the string length
For each i = 1 to L // searching for the possible subsequent string
{
For each j = 1 to i
{
ExtractSubStr(S[i,j])
If(S [i,j] ! = 0 )
{
Search(S[i,j], s1 ϵ S)
}
Else
{
ExtractSubStr(S[i,j])
}
}
Return S[i,j]
}
}
```

Then this LCS is assessed with the dataset used to obtain the relevant sequence. The extracted relevant sequence is then evaluated with the input sequence to obtain a relevant sequence matrix (R1). Compute distance among the sequence and evaluate relevant matrix for the input sequence, to regularize the integrated sequencing matrix.

Table III: Algorithm for CESHP (Proposed Homology Model)

```
Algorithm CESHP: Correlated Extensive-Subsequent Searching based Homology Prediction
{
Input: S – Protein Sequence
   P – PDB File
Output: R – Resultant Matrix
ReadPDBFile(P') from P
Convert PDB file P' to FASTA data
ParseSeq(S) from P' // Parsing the protein sequence from the FASTA data
ExtarctSeq(S)
Compute Sequence Alignment for the set of protein sequence extracted
SeqAlign(S' ϵ S)
Sequence Alignment gap is filled by sequence searching
ExactMatching(S') // check for exact matching by using algorithm ExactMatching()
A[]=Subsequent(S') // Check for sub-sequent matching sequence using Subsequent()
L=Strlen(A[]) // compute substring length to find LCS (Longest Common Sequence)
For each i=1 to L
{
If(Strlen(A[i,j] >=L) // to find the LCS
{
FecthSeq(A[i,j]) from S' //fetching that sub-string form the sequence sub-string set
StrComp(A[i,j], S'[i,j]) // finding the common sub-sequence
R[i,j]=[A[i,j],Strlen(A[i,j])]
C[i,j]=[S'[i,j],Strlen(S[I,j])]
If (Mismatch(R[i,j],C[i,j])
{
SetScore(0)
}
Else
{
SetScore(A[i,j]) + 1
SetScore(C[i,j]) + 1
In similar way compare the substring of the sequence
Evaluate the scoring matrix S_Score
ComputeDist(A[i,j],C[i,j])
}
}
}
}
```

Database search to identify homologous sequences based on similarity scores. Ignore position of symbols when scoring. Similarity scores are additive over positions on each sequence to enable DP. Scores for each possible pairing, e.g. proteins composed of 20 amino acids, 20 x 20 scoring matrix. The biological correlation among the amino-acids is represented in the scoring matrix. The common ancestor shares or the one sequence is the ancestor of another that represents the possible positioning of the two AA's appearance in homologous evaluation.

Table IV: Scoring Matrices

Sub Sequence Match	Distance of Sequence	Score of String Match
-16.5606	-45.4773	-62.0379
-19.3168	-49.2334	-68.5502
-18.6519	-50.1519	-68.8037
-19.6104	-51.9437	-71.5541
-18.5398	-48.1231	-66.6629
-27.2145	-59.6312	-86.8457
-36.0727	-69.3227	-105.395
-22.6664	-54.0831	-76.7495
-22.4917	-54.6584	-77.1501
-23.262	-55.6787	-78.9407
-28.2255	-60.0588	-88.2843
-24.8644	-58.6977	-83.5621
-26.6415	-57.8082	-84.4497
-36.7462	-68.4128	-105.159
-25.3618	-55.9451	-81.3069
-27.1492	-59.1492	-86.2985
-24.3101	-56.8934	-81.2035
-4.1383	-60.0549	-64.1932
-25.1182	-57.3682	-82.4864
-24.4451	-56.2784	-80.7236

In this section, the sequence based computation of homology is evaluated to generalize the homology vector for the structure prediction. This layer deals with the resultant matrix R1 to generalize the vector for contact map prediction. The matrix R1 is processed by homology computation to extract the nearby sequence. The nearby sequencing is obtained by analyzing and applying chaining among the sequences in matrix R1. To obtain the reduction vector the mean value for the correlated matrix is evaluated. Then the numerical vector is classified that satisfies the mean value. It would return the relevant sequence to the input sequence; it will diminish the searching process complexity. The sequence matrix R2 is constructed with the relevant sequences with essential contacts.

B. Predicting Protein Structure using Contact Map with Resultant Sequence with Homology

Protein Contact Map is the representation of the amino acids sequence in the protein is a trending area of research. The representation of the amino acids folding which reduces the formation of probable amino-acids contacts from the structure. Then the research of contact map prediction process is begins with detecting the folding from the sequence. The contact map prediction methodologies are introduced to predict the structure of the large amount of protein sequence. The protein 3D structure prediction is based on the resultant sequence matrix. The Homology sequence matrix is processed with the proposed contact map method to predict the protein 3D structure.

This framework likewise presents five parameters, based on physical contemplations that can be utilized to judge the basic closeness between proteins through contact maps. To anticipate the overlap of a given amino acid sequence, it predicts a contact map will adequately deduced the structure of the relating protein. Then access the similarity of this contact map the delegate contact map of each overlay; the fold that compares to the nearest match is our predicted overlap for the input sequence. The system has discovered that our practicality measure can separate amongst doable and infeasible contact maps. Further, this novel approach can expect the folds from sequences altogether superior to an irregular indicator.

In the research of development, proteins demonstrate a wonderful preservation of their three-dimensional structure and their natural capacity, prompting solid transformative requirements on the sequence inconstancy between homologous proteins. The strategy goes for extricating such requirements from quickly aggregating sequence data, and in this way at deriving protein structure and function from sequence information alone. This research proposed a very efficient *Multi-Variant based Direct Folding Prediction using Contact Map (MDFPCM) scheme* for the folding analysis.

This research considers on two main aspects of protein structure prediction: (1) the presumption of protein folding contacts for the large multi-sequences of the protein homology and (2) concerning the atomic determination of the protein interactions and identification. Thus, this research provides an alternative approach to consider these aspects of protein structure prediction. The detailed description for the approach is discussed in this section and figure 5 illustrates the process flow contact map prediction for protein interactions.

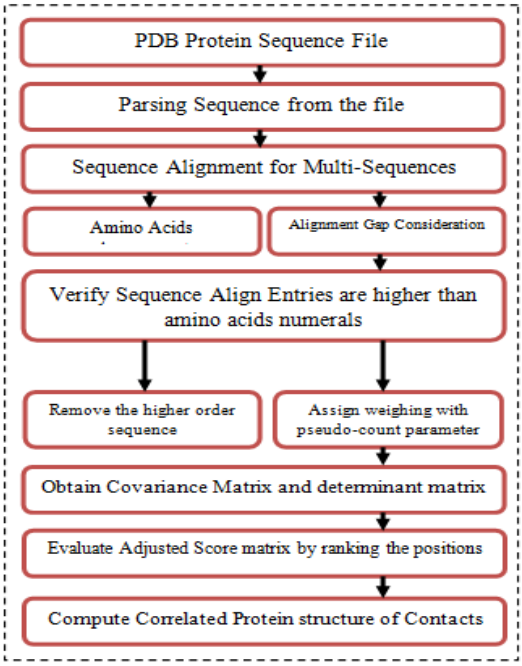

Figure 5: Protein Structure Prediction using Contact Map Flow

The identification of protein structure by contacts of the proteins is obtained with the protein sequence. The protein is sequence data is parsed with PDB format. The PDB file contains the sequence of the protein with header and molecular information of the protein. The input file of sequence is parsed to obtain the sequence and the molecular information for prediction process. Then the input sequence data is aligned if the multi-sequence is determined for the large protein families. The multi-sequence alignment S of the protein sequence length L is considered. The various amino acids N for the protein sequence are assigned with numeric values respectively. The alignment gaps are considered with the amino acids values for multi-sequence alignment.

The MDFPCM modeling is highly structured, due to the fact that at most one amino-acid is present in each position of each sequence. In the input data, for the consequence of the protein sequence residues and has two states of amino acids x and y with constrain of $x \neq y$. For instance, the two different amino acids of the same sequence are uncorrelated, and then the parameters folding of the protein among the different amino acids values are generated for obtained uncorrelated values. To end with the resultant matrix this depicts the positions and ranking pairs of the protein sequences.

Contact prediction utilizing direct collapsing of the proteins depends on positioning sets of acids positions as indicated by its immediate collaboration quality. As referenced previously, two positions cooperate by means of an N x N matrix. To look at two position sets ‘m and m’, the system have to map the matrix to a single scalar amount. At that point the change of the framework to get the immediate data of the protein folding among the groupings is estimated. After the evaluation the sub-matrix is acquired with: (1) Changing the determinant of the cooperation to such an extent that the entirety of each column and row is 0. (2) Removing the column and row comparing to the symbol of gap.

Table V: Algorithm for Sub MDFPCM (Proposed Model for Contact Map Prediction)

Algorithm MDFPCM (Multi-Variant based Direct Folding Prediction using Contact Map)

```
{
Input: S – Protein Sequence
Output: C – Direct Folding Prediction
Read(S') from S //Read the PDB sequence file
ParseOfSequence(S')
SA=SeqAlign(S') in S // multi-sequence alignment of the protein sequence
Assigning numeral values to the different amino acids N
FillGap(SA) with N // alignment gaps are considered with amino acids numerals
If (AlignGap(SA) >= 90) // verify that sequence align entries are high or not
{
Remove (SA) from the Sequence Set
}
Else
{
Assign (SA, N' ϵ N) // assigning amino acids numerals to align the gaps in multi-sequence
W (SA) =AssignWeigh(SA)
Computing the pseudo-count correction for weighed attribute W (SA)
Estimate the covariance matrix ∑ by means of pseudo-count correction
Compute the determinant |∑| of ∑ matrix and obtain N x N blocks
Select Amino Acids from N =Select (a,b) from SA
For each 1 <= m and n <= L // choosing the positions m, n with length of sequence L
{
Evaluate Score matrix SM
e_mn(x,y) = |∑| (m-1)N+x(n-1)N+v
Obtain adjusted score matrix SM_mn with average product of (m,n)
Ranking all pairs of (m,n) based on SM_mn
Evaluate correlated families of the contacts SM_mn
}
}
}
```

C. Predicting Protein Class using Dimensionality Reduction with Classification

In this level of the approach, the dataset is preprocessed before it is utilized in the preparation or testing stage. Selected datasets are then changed into a feature vector space described in Table 6, using a component encoding method that concentrates relative features from protein arrangements. Sequence encoding techniques amazingly impact the quality and relevance of the AI systems.

The structure and functions of large size of upcoming proteins is classified into existing super families of protein would assist in prediction. Due to the large amount of features obtained from the protein sequence results in lack of protein class prediction. In this work, a statistical metric-based feature selection technique has been proposed in order to reduce the size of the extracted feature vector.

Table VI: Feature Vector

Feature Vector
Sum of Amino Acid Pairs Descriptor
Median of Amino Acid Composition Descriptor
Mean of Alpha Value
Sum of Amphiphilic Pseudo Amino Acid Composition Descriptor
Mean of Binary descriptor
Mean of Beta Value
Sum of Atomic Components
Mean of Composition descriptor
Mean of Distribution descriptor
Mean of Transition descriptor
Mean of Amino Acids Count
Median of Dipeptide Composition Descriptor
Mean of hydrophobicity

In this research, the limitations of high-dimensional sequence information by involving a statistical metric for dimensional reduction for more number of features are the main objective. The proposed method chooses the most esstia highlights which lead to improved grouping results on different sorts of superfamilies arrangements acquired from the publically known benchmark datasets. In overabundance of existing sequence procedures, the proposed system is like arrangement free protein succession techniques for characterization and furthermore has basic advantages while condemnation. In addition, the proposed strategy is basic, quick, solid, and robust and requires an exceptionally short preparing time.

A feature is an individual experimental property of the procedure being detected. Utilizing a collection of features, any machine learning techniques can perform classification. In the past years in the utilizations of machine learning or pattern recognition, the area of features this system planned a unique technique for choosing a salient set of features by exploitation Conditional Attribute Set (CASs) concepts. The feature selection problem tries to seek out a set of features that has 2 main characteristics, the chosen features should have most relevance with the class labels and have minimum redundancy among the selected features, at the same time. Suppose the dataset which has n instances and d features (Dnxd), for generating CASs we want a universe of discourse. The notation i S is employed for representing the ith feature. Also, the set is employed as the universe of discourse. The Conditional Attribute Set primarily based Feature selection algorithm (CASFS) is conferred.

The CASFS contains 3 essential elements. The primary part provides a set of features that they need minimum redundancy among themselves and the second a part of the algorithm finds a subset of features which have most relevance to class labels. The third a part of the algorithm makes an attempt to calculate the intersection of two above sets. In that case the remaining features once intersection step have each fascinating aspects (maximum relevancy and minimum redundancy) simultaneously. the primary part of the algorithm includes steps, for generating CASs on outlined universe of discourse some experts' opinion is considered, for addressing this, 3 completely different proximity measures is employed.

After generating CASs the formula makes an attempt to calculate the composition matrix and within the next step the CASFS tries to calculate the binary matrix by selecting appropriate C. Choosing C is vital and has direct impact on the experiments, in all of our experiments, during this paper we did try and error to choose the best C. within the last section of the primary a part of the algorithm features are clustered into some clusters and the biggest cluster with respect to the scale of members has some non-redundant features.

The second part of the CASFS just like the first part has half-dozen main steps, within the first and second ones the CASs are generated, for having the various experts' opinions we use 3 different ranking algorithms and putting their weights on the CASs. This section use Fisher, Relief and information Gain [16] as 3 totally different experts' opinions, that the CASs is outlined in the 10th equation. Once generating the C the conditional coefficients among CASs is computed. within the fourth step of the formula the composition matrix of correlation is calculated and in the fifth step of the algorithm the acceptable C is decided by try and error and in the last step clusters are appeared, and the largest cluster is chosen here as the subset of the features that have most relevance. in the last step of the CASFS the calculation of intersection between 2 mentioned sets will be determined and optimum features set will be calculated .has been expanded from tens to various aspects in various applications.

The challenging tasks of reducing irrelevant and redundant attributes various techniques are introduced to address this problem. The concept of feature selection (attribute elimination) assists in observing information of the protein, reduction of dimensionality and enhances the performance of prediction.

Table VII: Extracted Features

-4.8112	0.0496	141.2766	2140	141	0.2143	4.0032	0.1526	141	146.73	0.9718	1.2556
-5.264	0.0483	238.5093	3743	238	0.2076	6.4639	0.1429	238	230.26	1.0094	1.2698
-6.054	0.0442	145.2636	2341	147	0.2123	4.0627	0.1399	147	155.05	1.0224	1.2949
-9.5296	0.043	154.357	2393	151	0.2146	3.9322	0.1433	151	159.29	0.9997	1.2575
-8.7089	0.0524	191.6296	3009	191	0.214	5.1234	0.1436	191	188.37	0.9992	1.2938
-0.0827	0.0379	65.3798	1162	66	0.2	1.6336	0.1268	66	81.31	0.9791	1.2603
-13.8953	0.0395	184.9789	3015	190	0.198	5.1332	0.1502	190	204.85	0.9977	1.3256
-15.7105	0.0514	147.3956	2259	146	0.197	4.0655	0.1546	146	161.88	1.0269	1.373
-23.5922	0.0374	148.4298	2100	147	0.1993	3.9869	0.1502	147	154.12	0.9528	1.1333
11.0937	0.0394	124.8243	1994	127	0.1988	3.5777	0.1563	127	116.95	1.0617	1.212
0.8319	0.0497	149.7129	2461	151	0.2105	3.8299	0.135	151	160.77	1.024	1.3556
1.7881	0.0282	68.1046	1076	71	0.2265	1.928	0.1299	71	67.08	0.892	1.2342
-37.2327	0.0517	487.074	7572	484	0.197	12.8799	0.1459	484	513.96	0.9948	1.2111
-14.3769	0.0558	214.1663	3294	215	0.1909	5.8631	0.1515	215	212.74	1.0374	1.2275
-10.7993	0.0422	157.2727	2436	154	0.2132	3.9876	0.1419	154	160.79	0.9933	1.2515
-16.2799	0.0401	165.4919	2420	162	0.2085	4.4409	0.1574	162	168.19	0.9659	1.1896

a. Dimensionality Reduction based Class Prediction

This process of constructing new features can be followed by or combined with a feature subset selection process the original feature set is first extended by the newly constructed features and then a subset of features is selected. The best features selected for the class prediction of the protein is shown in Table 8. The table depicts that various features are selected for prediction for various feature selection methodologies. The proposed methodology CASFS selects the optimal set of features for the set of protein sequence data.

Table VIII: Feature Selection

Methods	K-Value	Selected Features
MRMR	5	PFRID-1, PFRID-4, PFRID-7, PFRID-9, PFRID-11
BAT	5	PFRID-1, PFRID-2, PFRID-5, PFRID-7, PFRID-12
ACO	5	PFRID-3, PFRID-5, PFRID-6, PFRID-8, PFRID-10
ACO+BAT	5	PFRID-1, PFRID-3, PFRID-5, PFRID-11, PFRID-12
CASFS	5	PFRID-2, P FRID-6, PFRID-10, PFRID-11, PFRID-12

Table IX: Algorithm for CASFS (Proposed Feature Selection Technique)

Algorithm CASFS

{
Input: D-Dataset={x_1,x_2,…x_n}
d- Feature Count
n-Count of instances
F-set of features ={F_1,F_2,…F_D}
C-Target Class
Output: F' –an optimal feature sub-set
Calculate Similarity Measure based on the target class *C*
Create CASs dataset d_1 of similarity measures from the attribute set F
Calculate the conditional coefficients among CASs
Generate conditional coefficient matrix by

$$(\overline{\rho_{ij}}) = \max_k \left\{ \min \left\{ \rho_{ik}, \rho_{kj} \right\} \right\}, \quad i, j = 1, 2, \ldots, m$$

Clustering similar columns of the conditional coefficient matrix
Measuring relevancy and redundancy among dataset and return the weights
Construct d_2 CASs dataset for weighed columns
Compute conditional coefficient among CASs and construct matrix
Find the optimal set and create a matrix F'
Clustering similar columns and return the weights
F'= The features are combined which is obtained from d_1 and d_2
Return F'
}

A predictive model with a numerical target uses regression rules, now not a class set of rules. The best form of classification problem is binary classification. In binary class, the target feature has most effective viable values: as an instance, excessive credit score rating or low credit rating. Multiclass objectives have more than one value: for instance, low, medium, high, or unknown credit score. In the model construct (training) method, a classification set of rules probes relationships among the values of the predictors and the values of the target.

In this segment, the framework portrays the proposed *Feasible Prediction of Multi-Class Protein utilizing Artificial Intelligence (FPMPSAI)* framework for prediction. This investigation at first targets enhancing the exactness of multi-class classifier by recognizing the subset of best informative features and assessing the best qualities for regularization of part parameters for multi-class model. So as to accomplish this man-made consciousness based upgraded system is utilized. FPMPSAI algorithm obtains AI thoughts by enhancing the parameters of multi-class utilizing artificial intelligence. Artificial Intelligence beginnings with n-arbitrarily chose components and looks for the optimal elements iteratively. Every component is an m-dimensional vector and speaks to a competitor arrangement. FPMPSAI classifier is worked for every applicant answer for assess its presentation through the cross approval strategy.

Table X: Algorithm for FPMPSAI (Proposed Classification Model for Protein Class Prediction)

Algorithm FPMPSAI

{
Input: F- Protein Features Set
T-Testing Feature Set
v- Element Momentum Rate
l_{best}- Local best element
g_{best} – Global best element
f_{best} – Element best fitness value
Output: P_c – Protein Class
Initialize the target class C and the class coefficient γ
Randomize initial elements from the feature set F and T
Compute the class coefficient value γ
Calculate element momentum rate v for each class features
If(γ <= v) // condition to be satisfied when the coefficient value is limited to element momentum rate
{
For each F_i from F
{
Evaluate FitnessVal(f_{best}) and FitnessVal(f_{lbest})// computing fitness value for features set
If (f_{best} < g_{best})
{
Determine l_{best} and g_{best} from the elements fitness computation
g_{best} = f_{best} and l_{best} = f_{lbest}
}
// Training the multi-class classifier with the feasible training set
TrainClassifier(FPMPSAI, g_{best})
}
}
TestClassifier(FPMPSAI, T_i ϵ T)
If (Class (T_i) in Class (FPMPSAI,C_i)
{
P_c = PredictProtein(T_i)
}
Return P_c
}

Artificial Intelligence technique manages the choice of potential subsets that lead to best prediction accuracy. The algorithm utilizes the fit components to add to the up and coming generation of n-competitor components. Accordingly, on the normal, each progressive populace of applicant components fits superior to its ancestor. This procedure proceeds until the presentation of multi-class merges. Man-made reasoning is utilized to discover ideal component subsets by finding the best element blends as they fly inside the issue space from the handled datasets.

V. PERFORMANCE EVALUATION

In this section, the performance analysis for this research concept is evaluated. The evaluation of the performance metrics is performed for particular objective of this thesis. The Protein R3D Homology prediction is implemented in MATLAB2014a to analysis the results and its performance. The chapter includes a detailed description of the dataset used for the prediction process.

Then the performance metrics that are involved in measuring the performance analysis of the main objectives of the R3D is illustrated. In the following sections, the dataset description, Performance analysis for Homology Modeling, Contact Map Structure Prediction and Protein Class Prediction are depicted.

A. Dataset

The protein sequence formatted input dataset is described in this section. Table 11 depicts the dataset for the protein family. The Table includes the protein families name with protein family ID. There are various amount of protein families are described in the table.

Table XI: Sample Protein Family Dataset

Family	Protein Family ID	Description
Kunitz_BPTI	PF00014	This proteins for inhibit the function of protein degrading enzymes or, more specifically, domains of Kunitz-type are protease inhibitors.
KH_1	PF00013	It is first identified in the human heterogeneous nuclear ribonucleoprotein (hnRNP) K. This family protein conserved sequence of around 70 amino acids.
HSP70	PF00012	Heat shock Proteins This protein exists in living organism.
SH3_1	PF00018	This family has small size of protein about 60 amino acides.

Table 12 depicts the proteins included in protein families and its functionalities. The table describes the proteins name with its family included. This dataset is involved in prediction of the class of the particular protein family.

Table XII: Sample Proteins and Its Functionalities

Protein(s)	Description	Family
AMBP	Protein AMBP is a protein that in humans is encoded by the AMBP gene.	Kunitz_BPTI
APLP2	Amyloid-like protein 2, also known as APLP2, is a protein that in humans is encoded by the APLP2 gene. APLP2 along with APLP1 are important modulators of glucose and insulin homeostasis.	Kunitz_BPTI
TFPI2	Tissue factor pathway inhibitor 2 is a protein that in humans is encoded by the TFPI2 gene.	Kunitz_BPTI
ANKRD17	Ankyrin repeat domain-containing protein 17 is a protein that in humans is encoded by the ANKRD17 gene.	*KH_1*

B. Performance Analysis of Homology Modeling

This section describes the performance evaluation of the homology prediction modeling using subsequent searching. The performance analysis of the various homology models (GGSearch and RAPSearch) with the proposed model (SCSHM) is evaluated with some performance metrics such as Accuracy Ratio and Time complexity while predicting the homology of the particular protein sequence from the large amount of protein families' dataset. In the following section the performance metric analysis is illustrated.

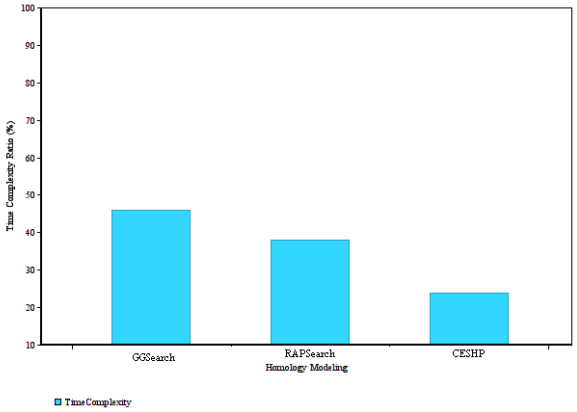

Figure 6: Homology Modeling Time Complexity Analysis

The estimated time complexity of the proposed homology prediction model and other methodologies are shown in Figure 6. The figure depicts that the proposed prediction model SCSHM consumes less time complexity while predicting the homology for a set of protein sequence than the other models such as GGSearch and RAPSearch.

The similarity among the sequence searching for the prediction of homology using various sequence searching is shown in figure 7. The existing searching for the homology prediction provides less similarity ratio while searching than the proposed system CESHP.

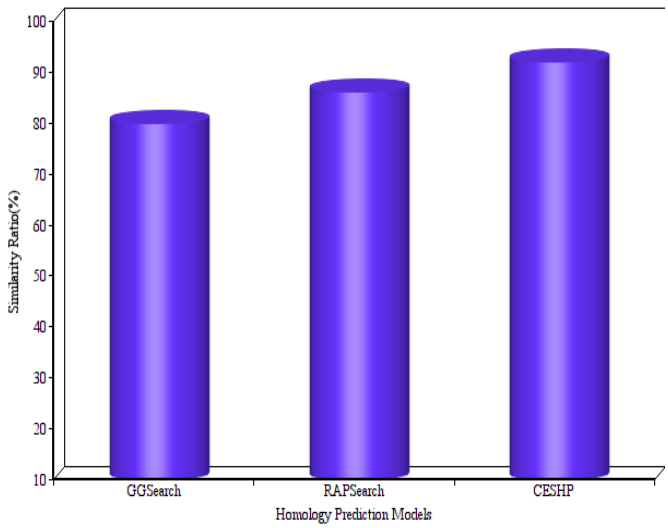

Figure 7: Similarity Ratio for Homology Prediction

C. Performance Analysis of Contact Map Structure Prediction

The protein 3D structure prediction is performed using contact maps. The contact map represents the folding of the amino-acids in the proteins and it would predict the protein structure among the protein family. While predicting the structure of the protein, the performance of the proposed model would be estimated. In this section, the performance evaluation for the contact map prediction model is illustrated.

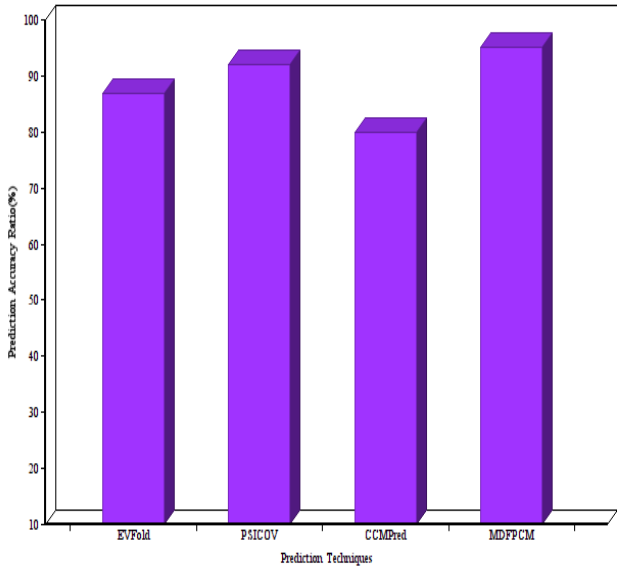

Figure 8: Prediction Accuracy Ratio Measured with respect to Prediction Techniques

The accuracy ratio in predicting the structure of the protein by means of protein sequence is evaluated for prediction methodologies is shown in figure 8. The figure depicts that the prediction accuracy ratio for the proposed technique MDFPCM provides high accuracy ratio of 96% when compared with other methodologies.

Figure 9 illustrates that the average coverage ratio for the various prediction methodologies. The proposed MDFPCM gives more over increased coverage ratio for the large number of protein sequences of the families than the other methods.

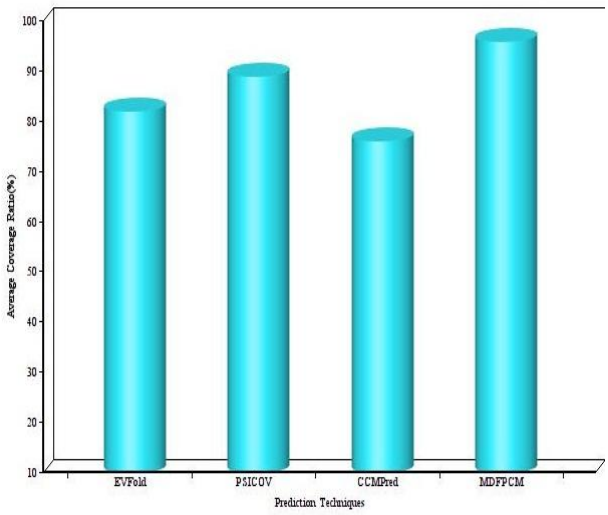

Figure 9: Average Coverage Ratio for the Prediction Methodologies

D. Performance Analysis of Protein Class Prediction using Dimensionality Reduction

In this section, the performance analysis of the classification methodologies is discussed with the illustration. The illustration depicts that our proposed methodology for the classification methods provides better outcomes than existing methodologies. The performance metrics such as reliability ratio, prediction rate, error rate, etc., are evaluated to measure the performance ratio.

The performance ratio for the various feature selection techniques are illustrated in Figure 10. The figure clearly depicts that the proposed methodologies CASFS provides high reliable ratio than the existing methodologies. This reliability ratio depicts that the predicted outcomes for class prediction using dimensionality reduction has more true positive predictions.

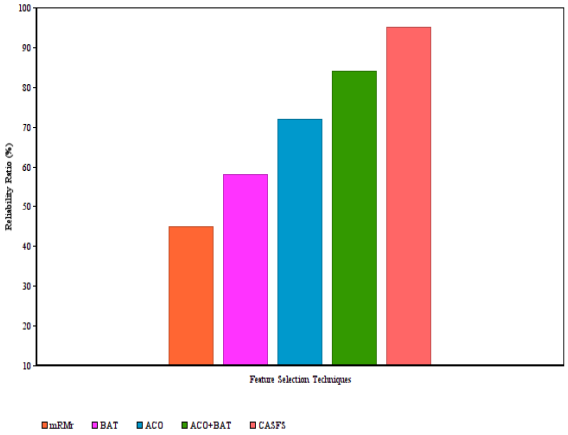

Figure 10: Reliability Ratio Analysis for Dimensionality Reduction Techniques

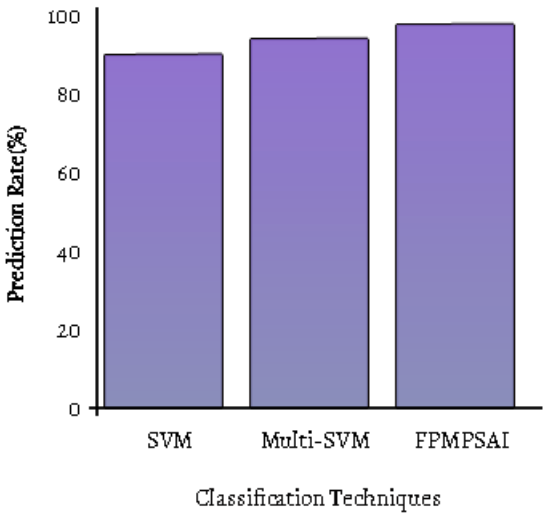

Figure 11: Prediction Rate

The prediction rate for the proposed methodology is evaluated with the existing methodologies is shown in Figure 11. This illustrates that proposed method FPMPSAI gives high prediction rate i.e. provides more true prediction values.

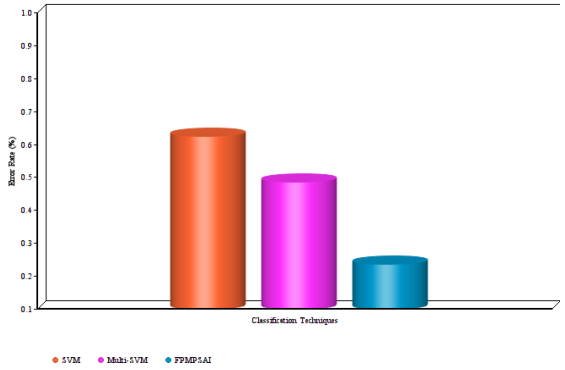

Figure 12: Error Rate Analysis for Protein Class Prediction

The error rate analysis for the protein class prediction is illustrated in Figure 12. The figure depicts that the proposed method provides more accurate results with less time complexity than the other methodologies.

Table 13 describes the prediction models for Homology Modeling, Structure Prediction and Protein Class Prediction are evaluated. The table depicts that the proposed system of R3D works as a framework of protein homology prediction with the use of contact map and protein class prediction. Also the table shows that the other existing systems predict any one of the aspect of the protein data analysis.

Table XIII: Prediction Models Evaluation Table

Model Name	Sequence Based Homology Modeling	Contact Map Based Homology Modeling	Class / Family Prediction
GGSEARCH	√	-	-
RAPSEARCH	√	-	-
CCMPred	-	√	-
EV-Fold	-	√	-
DeepFam	-	-	√
SVM-Prot	-	-	√
R3D	√	√	√

VI. CONCLUSION AND FUTURE WORK

This work has proposed supervised learning strategies which would improve the prediction of protein structure and protein class prediction. The proposed technique models arrangement homology prediction, sequence contact relationship, one for displaying of successive features and the other for demonstrating of protein class detection. The proposed technique is one of a kind in that it predicts all contacts of a protein all the while, which enables us to effectively display high-request association. Examining the exhibition of the proposed technique improves homology detection, contact map prediction, more than presently the best strategies (e.g., Evfold, CCMpred and PSICOV) by an awesome limit. The grouping based investigation of the protein homology expectation gives high related protein arrangement to the homology forecast, which would results the more connected protein with the separated structure that diminish the dimensionality.

The proposed technique performs because of several reasons. To start with, the framework predicts homology of protein, contacts utilizing data just in a classified protein family, while in this strategy takes in understanding structure relationship from a huge number of protein families. Second, the framework considers just pairwise buildup relationship, while our intense system framework can hold high-order correlation exceptionally well. The strategy utilizes a subset of protein highlights utilized by CASFS and FPMPSAI, the CASFS and FPMPSAI outperforms with the prediction of class of the protein.

As the future work, the system can be improved by included more input information and combining more protein sequence families for the training data set. The system will also be work with predicted contact maps to reduce the error rate and increase the recognition rate of the protein structure.

REFERENCES

1. R. N. Chandrayani, K. Manali, “Bioinformatics: Protein Structure Prediction”, 4th IEEECNT, 2013.
2. Zhang Y, Li T, Yang C, Li D, Cui Y, et al. (2011) Prelocabc: A Novel Predictor of Protein Sub-cellular Localization Using a Bayesian Classifier. J Proteomics Bioinform 4: 44-52.
3. Vaseeharan B, Valli SJ (2011) In silico Homology Modeling of Prophenoloxidase activating factor Serine Proteinase Gene from the Haemocytes of Fenneropenaeus indicus. J Proteomics Bioinform 4: 53-57.
4. Z. Zhen, J. Nan , “Radial Basis Function Method For Prediction Of Protein Secondary Structure”, Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, Vol. 3, pp. 1379-1383, 2008
5. C. Nandini, “A Survey on Protein Sequence Classification with Data Mining Techniques”, International Journal Of Scientific & Engineering Research, Vol. 7, No. 7, pp. 1442-1449, July 2016.
6. David T. Jones, “Protein Secondary Structure Prediction Based on Positionspecific Scoring Matrices, ” University of Warwick, Coventry CV4 7AL, United Kingdom, 1999.
7. Battey, J.N., Kopp, J., Bordoli, L., Read, R.J., Clarke, N.D. and Schwede, T., "Automated server predictions in CASP7, ” Proteins 69 (S): 68–82, 2007.
8. M. Hossein Zangooi, S. Jalili, “PSSP with dynamic weighted kernel fusion based on SVM-PHGS”, Elsevier, Knowledge-Based Systems, 2011, 27:424-442.
9. N. Nguyen Minh and J. C. Rajapakse. "Multi-class support vector machines for protein secondary structure prediction." Genome Informatics 14 (2003): 218-227.

10. B. Hafida, B. Messabih, A. Chouarfia. "Effect of simple ensemble methods on protein secondary structure prediction." Soft Computing19.6 (2015): 1663-1678.
11. A. Hepsiba and Dr. R. Balasubramanian, “Remote-R3d: A Novel Technique to Reducing Dimensionality In Remote Homology Finding Using Predicted Protein Contact Maps”, Aust. J. Basic & Appl. Sci., 9(20): 518-526, 2015
12. A. Hepsiba and Dr. R. Balasubramanian,”MPPCM: Combing Multiple Classifiers to Improve Protein-Protein Interaction Prediction“,Asian Journal of Information Technology,15(1), Pp:26-30,m Year:2016 Medwell Journals, ISSN:1682-3915.
13. A. Hepsiba and Dr. R. Balasubramanian, ” Developing A Hybrid Algorithm For Precise Protein Contact Map prediction”,International Journal of Advanced Research in Computer Science and ApplicationsVol. 2, Issue 5, May 2014.Pp:11-17.
14. Wu,S. et al. (2011) Improving protein structure prediction using multiple sequencebased contact predictions. Structure, 19, 1182–1191.
15. Wang,S. et al. (2013) Protein structure alignment beyond spatial proximity. Sci. Rep., 3, 1448.
16. Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins. Proteins Struct Funct Bioinformatics 2008;71:1899–907.
17. P. Maji, “Mutual information-based supervised attribute clustering for microarray sample classification,” IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 127–140, 2012
18. J. Huang, X.G. Hu, X. Geng, An intelligent fault diagnosis method of high voltage circuit breaker based on improved EMD energy entropy and multi-class support vector machine, Electrical Power System Research 81 (2) (2011) 400–407.
19. J. Derrac, S. Garc´ıa, D. Molina, and F. Herrera, “A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms,” Swarm and Evolutionary Computation, vol. 1, no. 1, pp. 3–18, 2011.

AUTHORS PROFILE

Mrs.A.Hepsiba, first author has completed her degree MCA.,M.Phil.,Ph.D.,.She is working as Associate Professor, Department of Master of Computer Applications in Karpaga Vinayaga College of Engineering and Technology. She has teaching experience of 11 years 3 months in this institute.

.R. Balasubramanian, has completed his B.Sc., M.Sc., M.Phil. (Maths)., Ph.D., M.Phil. (Computer Science)., M.Phil. (Management). M.B.A. (Operations Management)., M.S. (Education Management)., C.C.P., D.D.E., P.G.D.C.A., M.A.D.E., D.I.M., P.G.D.I.M., P.G.D.O.M.,. He is working as Professor & Dean, Department of Computer Applications Karpaga Vinayaga College of Engineering & Technology and has experience of 44 years. He had supervised more than 45 students under M.Phil., and guided more than 20 scholars. He has published more than 20 papers in National and International Journals and published more than 25 papers in Seminars and Conferences.