IJCRR
Vol 05 issue 12
Section: Technology
Category: Research
Received on: 26/01/13
Revised on: 15/02/13
Accepted on: 05/03/13

# FUNCTION PREDICTION USING CLUSTER ANALYSIS OF UNANNOTATED ALIGN SEQUENCES

Anjan Kumar Payra

Dr. Sudhir Ch. Sur Degree Engineering College, Dumdum, Kolkata, India

E-mail of Corresponding Author: anjan.payra@gmail.com

## ABSTRACT

Proteins are responsible for nearly every task of cellular life, including cell shape and inner organization, product manufacture and waste cleanup, and routine maintenance. Proteins also receive signals from outside the cell and mobilize intracellular response. Experimental procedures for protein function prediction are inherently low throughput and are thus unable to annotate a non-trivial fraction of proteins that are becoming available due to rapid advances in genome sequencing technology [15]. This has motivated the development of computational techniques that utilize a variety of high-throughput experimental data for protein function prediction. So, there is need to design algorithm to find similar functional proteomic sequence from large set of sequence database. Here we present a novel unsupervised method, called Function Finder (in short F-Func) for identification function of unannotated proteomic sequence. F-Func uses clustering of sequence information represented by numerical features, performing filtering, assigned score and meet with the criterion produces decision. Using help of producing result estimate success rate of F-Func method. Estimated success rate of F-Func methods, which is almost 70%.

**Keyword:** Sequence, Homology, motif, F-Func, Prediction, Cluster.

## INTRODUCTION

Proteins are macromolecules that serve as building blocks and functional components of a cell, Proteins are responsible for some of the most important functions in an organism.

Such categorization of the types of functions a protein can perform has been suggested by Bork [1] et al. [1998]:

- **Molecular function**: The biochemical functions performed by a protein, such as ligand binding, catalysis of biochemical reactions and conformational changes.
- **Cellular function**: Many proteins come together to perform complex physiological functions, such as operation of metabolic pathways and signal transduction.
- **Phenotypic function**: The integration of the physiological subsystems, consisting of various proteins performing their cellular functions, and the interaction of this integrated system.

In order to predict function we need to study amino acid sequence, where concept of central dogma is crucial. The central dogma [2] of molecular biology is the conversion of a gene to protein via the transcription and translation phases as shown in Fig. 1. The result of this process is a sequence constructed from twenty amino acids, and is known as the protein’s primary structure. This sequence is the most fundamental form of information available about the protein since it determines different characteristics of the protein.

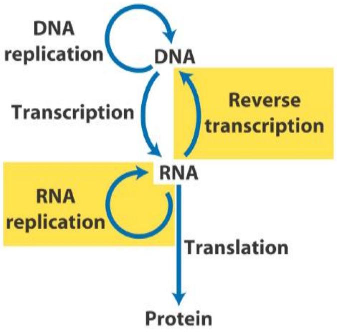


**Fig. 1. Central dogma: Sequence construction**

Here using data sequences are closely related with protein phosphorylation, mediated via a group of enzymes (called kinases) that performs addition of phosphate (PO4) group usually to serine (S), threonine (T), tyrosine (Y) residues, is one of the most frequent forms of post-translational modification mechanism [3].

Function of a gene or protein can predicted using different approaches .Those are given below –

❖ Homology-based Function Prediction :

Homologous genes or proteins are derived from a common ancestral sequence, as given in Fig. 2, where always searching for same constituent. Homologous proteins within or among species are similar in sequence and are likely, but not guaranteed to have a similar function. Comparison of an un-annotated sequence to known homologous sequences is therefore a good starting point for predicting function. Generally, a new sequence could be used to query protein databases using BLAST, to find proteins of known function with high sequence similarity, and their function is transferred to the query sequence.

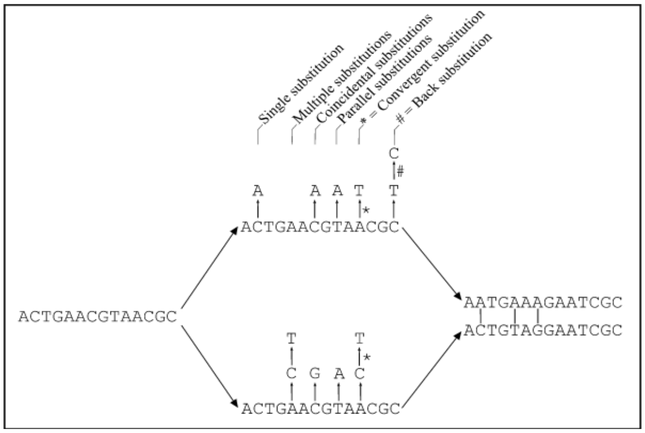


**Fig. 2. Homology based Genomic Sequence matching**

Several approach have proposed a multi-step strategy for functional annotation based on clustering of protein sequences according to their similarities [Xie et al. 2002; Abascal and Valencia 2003; Sasson et al. 2006]. Figure. 3 show the flowchart of the basic strategy adopted in these two studies. The algorithm starts with the construction of the similarity matrix that stores the BLAST similarity values between the protein sequences in the original training set [15]. This matrix is then used to cluster these sequences, and the annotation of a sequence in these approaches depends not on individual homologous sequences but a cluster composed of many such sequences.

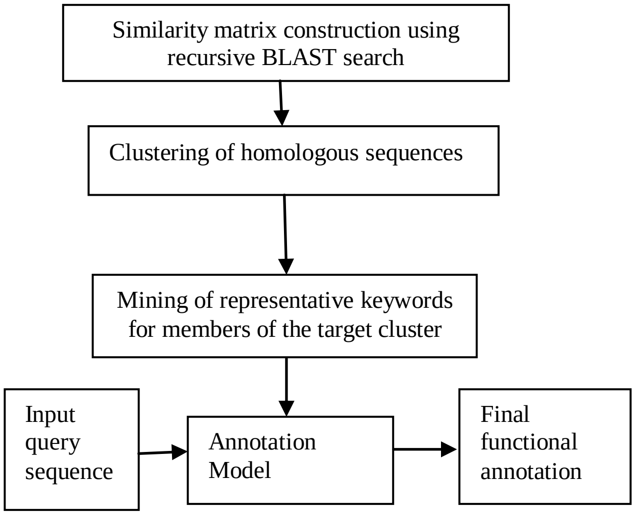

Similarity matrix construction using recursive BLAST search
Clustering of homologous sequences
Mining of representative keywords for members of the target cluster
Input query sequence
Annotation Model
Final functional annotation

**Fig. 3. Flowchart on multi step & protein sequence**

❖ Function Prediction Using Sequence Motifs : Motifs are defined as sub-sequences which are conserved across a set of protein sequences belonging to a family [Bork and Koonin 1996]. Owing to their conservation property, they are candidates for functional sites in proteins, such as sites for ligand binding, DNA binding and interactions with other proteins, and thus are useful as clues for predicting the function of a protein [Bork and Koonin 1996; Huang and Brutlag[2001].

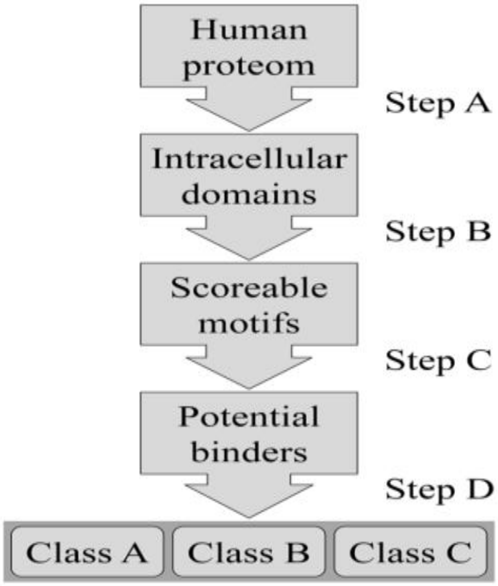


**Fig 4. Function prediction using motif score**

Motif databases such as PROSITE ('database of protein domains, families and functional sites') can be searched using a query sequence as given in Fig. 4. Motifs can, for example, be used to predict sub cellular localization of a protein (where in the cell the protein is sent after synthesis).

❖ Structure-based Function Prediction:

3D protein structure is generally better conserved than protein sequence, structural similarity is a good indicator of similar function in two or more proteins. Many programs have been developed to screen an unknown protein structure against the Protein Data Bank. The prediction of function from structure, and the ways these perspectives have been formulated in various approaches

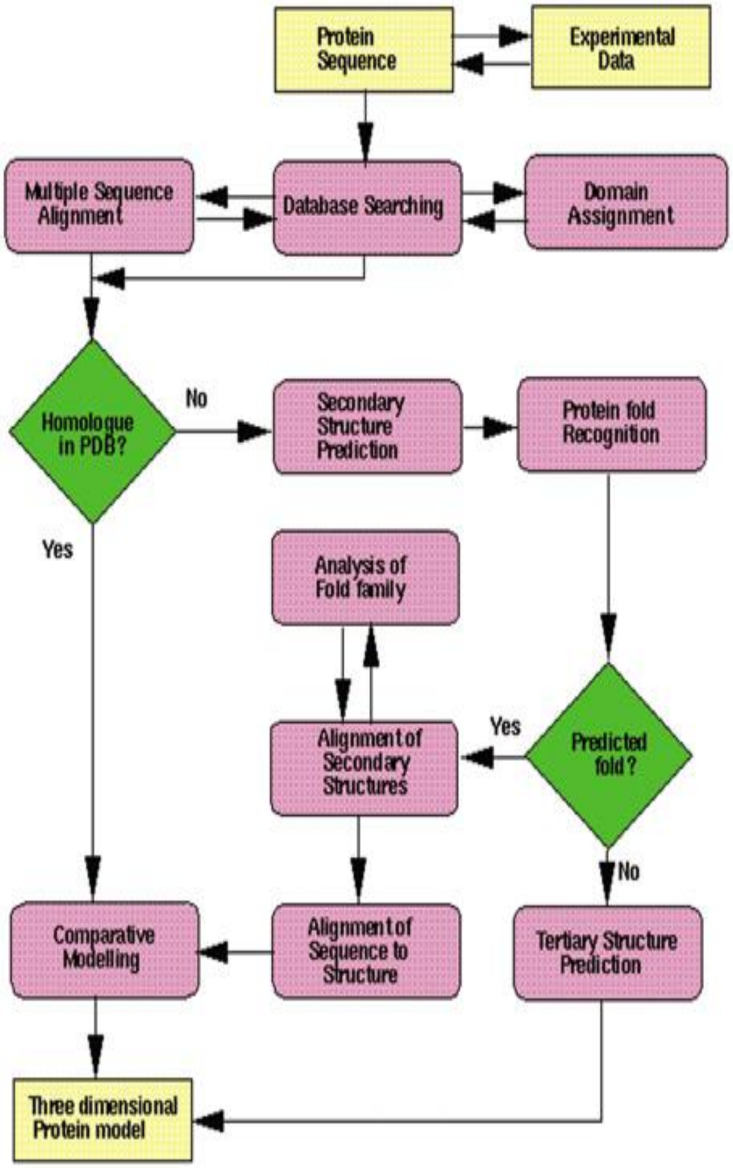

Protein Sequence
Experimental Data
Multiple Sequence Alignment
Database Searching
Domain Assignment
Homologue in PDB?
No
Secondary Structure Prediction
Protein fold Recognition
Yes
Analysis of Fold family
Alignment of Secondary Structures
Yes
Predicted fold ?
No
Comparative Modelling
Alignment of Sequence to Structure
Tertiary Structure Prediction
Three dimensional Protein model

**Fig 5. Function prediction using 3D structure study**

❖ Genomic Context-based Function Prediction**:** The protein function prediction are not based on comparison of sequence or structure as above, but on some type of correlation between novel genes/proteins and those that already have annotations. Also known as phylogenetic profile, these genomic context based methods are based on the observation that two or more proteins with the same pattern of presence or absence in many different genomes most likely have a functional link. Whereas homology-based methods can often be used to identify molecular functions of a protein, context-based approaches can be used to predict cellular function, or the biological process in which a protein acts [15]. For example, proteins involved in the same signal transduction pathway are likely to share a genomic context across all species, as given in Fig. 6.

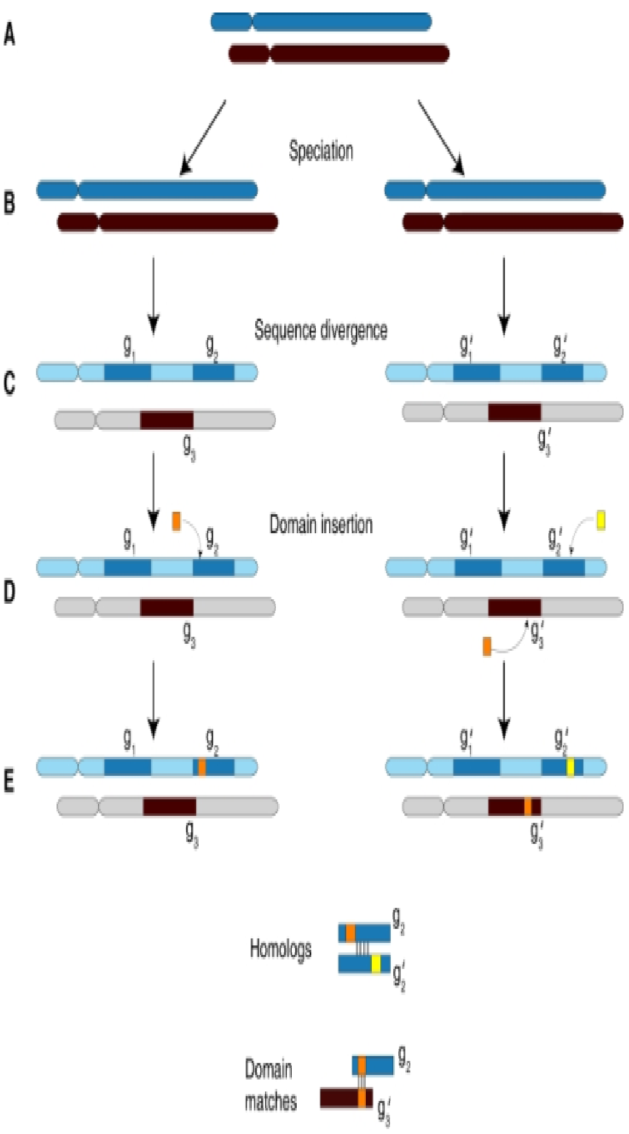


**Fig 6. Function prediction homologous and domain based**

❖ **Protein interaction networks and protein complexes:** A protein almost never performs its function in isolation. Rather, it usually interacts with other proteins in order to accomplish a certain function. At the highest level, they can be categorized into genetic and physical interactions. Genetic interactions occur when the mutations in one gene cause modifications in the behavior of another gene, which implies that these interactions are only conceptual and do not occur physically in a genome. Of particular interest in our project are the physical interactions between proteins, since they are more directly related to the process through which a protein accomplishes its functions [15]. The interactions can be structured to form a network, and hence the name protein interaction networks as shown in Fig. 7 and Fig. 8.

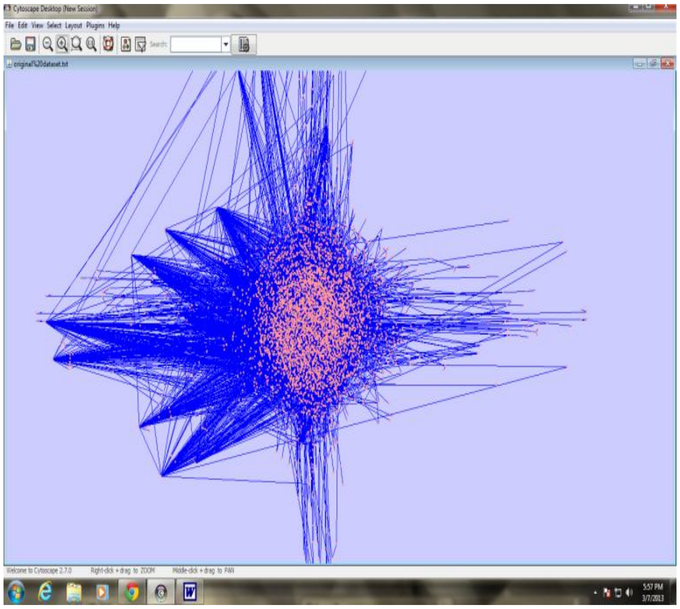

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search:
original%20dataset.txt
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN
5:57 PM
3/7/2013

**Fig. 7 Organic View (Cytoscape) of a PPI data set**

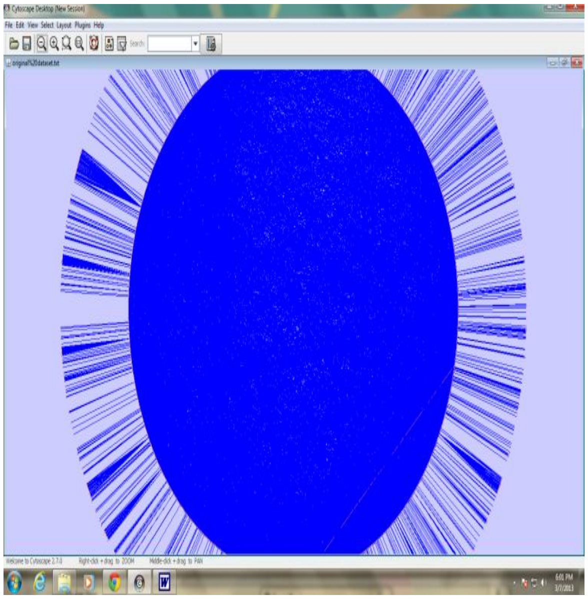

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search
original%20dataset.txt
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN
6:01 PM
3/7/2013

**Fig. 8 Circle View (Cytoscape) of a PPI data set**

Present growth of the protein sequence and structure databases, there remains only a small fraction of proteins whose functions have been experimentally characterized. It is sometimes possible to infer the function of uncharacterized proteins by comparison to the sequences or structures of functionally annotated homolog. Common descent does not necessarily imply functional similarity, however (Hegyi and Gerstein 1999[4] ; Devos and Valencia 2000 [5]; Todd et al. 2001[6]) and functional annotation transferred from one homologous protein to another. Several investigators have considered the problem of functional site prediction using multiple sequence alignments (Casari et al. 1995[7]; Andrade et al. 1997[8]; Hannenhalli and Russell 2000; Li et al. 2003).Casari et al. (1995), applied principal component analysis to a vector representation of protein sequences in a multidimensional “sequence space,” to derive subfamily-specific residues involved in protein function. Andrade et al. (1997) proposed a rigorous clustering algorithm based on a self-organizing map as a means to identify protein subfamilies and retrieve characteristic sequence patterns. Some methods of functional site prediction use phylogenetic analysis to identify residues associated with functional divergence (Lichtarge et al. 1996[9]; Sjolander 1998[10]; Aloy et al. 001[11]; Madabushi et al. 2002[12]).

## PRESENT WORK

**Motivation:** Many approaches have been discussed in the previous section over sequence. After studying and going through various papers it can be analyzed that very few assessments had been pursued on clustering this analyzation has encouraged us to work it. I am using homology based similarity to create different clusters to predict functional group of unannotated sequences in F-Func methods.

**Database and Tools:** Data has been collected only serine –phophorylated peptides of length 13 (i.e., 13-mers centered at serine) from the Phospho.ELM database, which are experimentally determined to be substrates of different kinases. Motif –X [13] and F-motif [14] are used for analyzing my algorithms, which are given in next section.

### Propose Method

To predict function of an unannotated sequence using F-Func, we divide three major portions. First phase is to form clusters based on matching criteria. Next two are different approaches to predict.

**Algorithm: *Cluster formation***

***Step1:*** Consider sequence ($Seq_i$) $=\{S_1,S_2,\ldots..S_n\}$, $n\leq 25000$.

***Step 2:*** Eliminate all repetitions from $Seq_i$ and obtained non repeated sequence ( $Seq_i$).

***Step 3:*** Consider central (j=6) residue element (……S……) of a sequence of length 13. Extract left pattern ($L(Seq_{ij}),i\leq n,\ 0\leq j\leq 5$ ) and right pattern($R(Seq_{ij}),i\leq n,\ 7\leq j\leq 13$ ) for each $Seq_{i.}$

***Step 4:*** Comparing the $Seq_i$ with left pattern and generate left cluster ($LC_P$ ,p<n ) with minimum occurrence of the pattern in the $Seq_i$ is at least N, where N is positive integer (eg. N=50).Similarly, generate right cluster ($RC_q$ ,q<n ). Allocate functions ($f_r$, $r\leq p+q$) for each clusters.

***Step 5:*** Select random sequence from $LC_P \cap RC_q$ and consider it as unannotated sequence ($Seq_{un}$).

After cluster formation, our objective is to predict functions of unannotated sequences logically or mathematically, which should be new and well justified. Here, I will discuss two approaches of predicting sequences. 1st approach is given below-

**Algorithm: *Assigned functional group to unannotated sequence ($Seq_{un}$) using mean value calculation.***

***Step 1:*** Consider one of the cluster from collection of clusters ($LC_P$ or $RC_q$ ) and select sequences from it. Count the number ($N_0$) of sequences for selected cluster.

***Step 2:*** Count maximum occurrence ($S_{count}$) of a constitutes (eg. S, G, R, P, A etc) present in the sequences of respectively clusters.

***Step 3:*** Estimate mean (M) value -

$$M = {}^{S_{count}}/_{N_0}.$$

The value of M is assigned as score of the corresponding cluster.

***Step 4:*** Repeat the step 1 to 3 till there no cluster is left. $LC_P$ € Ø, $RC_q$€ Ø.

***Step 5:*** Merge two clusters if selected constitutes (eg. S) and score values are same for both.

***Step 6:*** Select a sequence ($S_i$) from $Seq_{un}$. Estimate occurrence of constitutes in it. Assigned the maximum value to $E_{max}$ . If $S_i$ can't find suitable clusters to estimate distance for $E_{max}$ and constitutes, than select next maximum occurrence of constitutes for $S_i$ and continue.

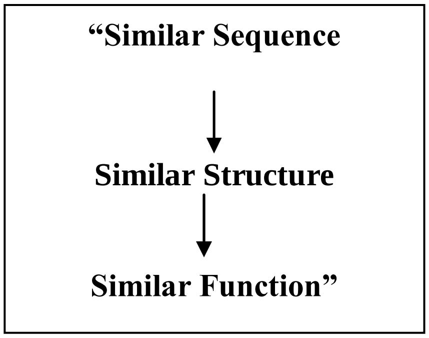
***Step 7:*** Calculate distance of $S_i$ from clusters-

$$d_j = E_{max} - M,\ \ 0 < j \le\ \ p + q, d_j \ge 0$$

***Step 8:*** Sequentially $S_i$ is more nearer to the cluster with minimum $d_j$ value. $S_i$ is more functionally related with the cluster. The sequence $S_i$ is placed and function is assigned of the respective cluster.

***Step 9:*** Finally, estimate success rate of prediction.

Through multiple sequence alignments, patterns of characteristic residues may emerge. Above algorithm is based on basic paradigms "Homology". Here basic principal is-



**"Similar Sequence**

**Similar Structure**

**Similar Function"**

Success rate is always crucial while predicting functional group of a sequence. Here, key in my approach is principal.

2nd approach is quite similar with 1st approach, it will provide more accurate result as compare previous one. The approach is given below-

**Algorithm: *Assigned functional group to unannotated sequence ($Seq_{un}$) using max value calculation of sequence constituents.***

***Step 1:*** Consider one of the cluster from collection of clusters ($LC_P$ or $RC_q$ ) and select sequences from it. Count the number ($N_0$) of sequences for selected cluster.

***Step 2:*** Count maximum occurrence ($S_{count}$) of a constitutes (eg. S, G, R, P, A etc) present in the sequences of respectively clusters.

***Step 3:*** Estimate max ($C_M$) constitutes value to assign score of a cluster. [eg. Cluster (N) contains S=8, G=3,R=2.Then $C_M$ =8 for S of Cluster(N)]

***Step 4:*** Repeat the step 1 to 3 till there no cluster is left. $LC_P$ € Ø, $RC_q$€ Ø.

***Step 5:*** Merge two clusters if selected constitute (eg. S) and score values are same for both.

***Step 6:*** Select a sequence ($S_i$) from $Seq_{un}$ .Estimate occurrence of constitutes in it. Assigned the maximum value to $E_{max}$ . If $S_i$ can't find suitable clusters to estimate distance for $E_{max}$ and constitutes, than select next maximum occurrence of constitutes for $S_i$ and continue.

***Step 7:*** Calculate distance of $S_i$ from clusters-

$$d_j = C_M - E_{max},\ \ 0 < j \le\ \ p + q,\qquad d_j \ge 0$$

***Step 8:*** Sequentially $S_i$ is more nearer to the cluster with minimum $d_j$ value. $S_i$ is more functionally related with the cluster. The sequence $S_i$ is placed and function is assigned of the respective cluster.

***Step 9:*** Finally, estimate success rate of prediction.

**Illustration using example:**

Algorithms of F-Func are applied over pre-aligned data set of length 13 mers. So, here obtain motif using F-motif [13] and motif-X [14] tools, generate web logo, which is given in Fig. 9. Cluster formation, assign scores to the clusters using above

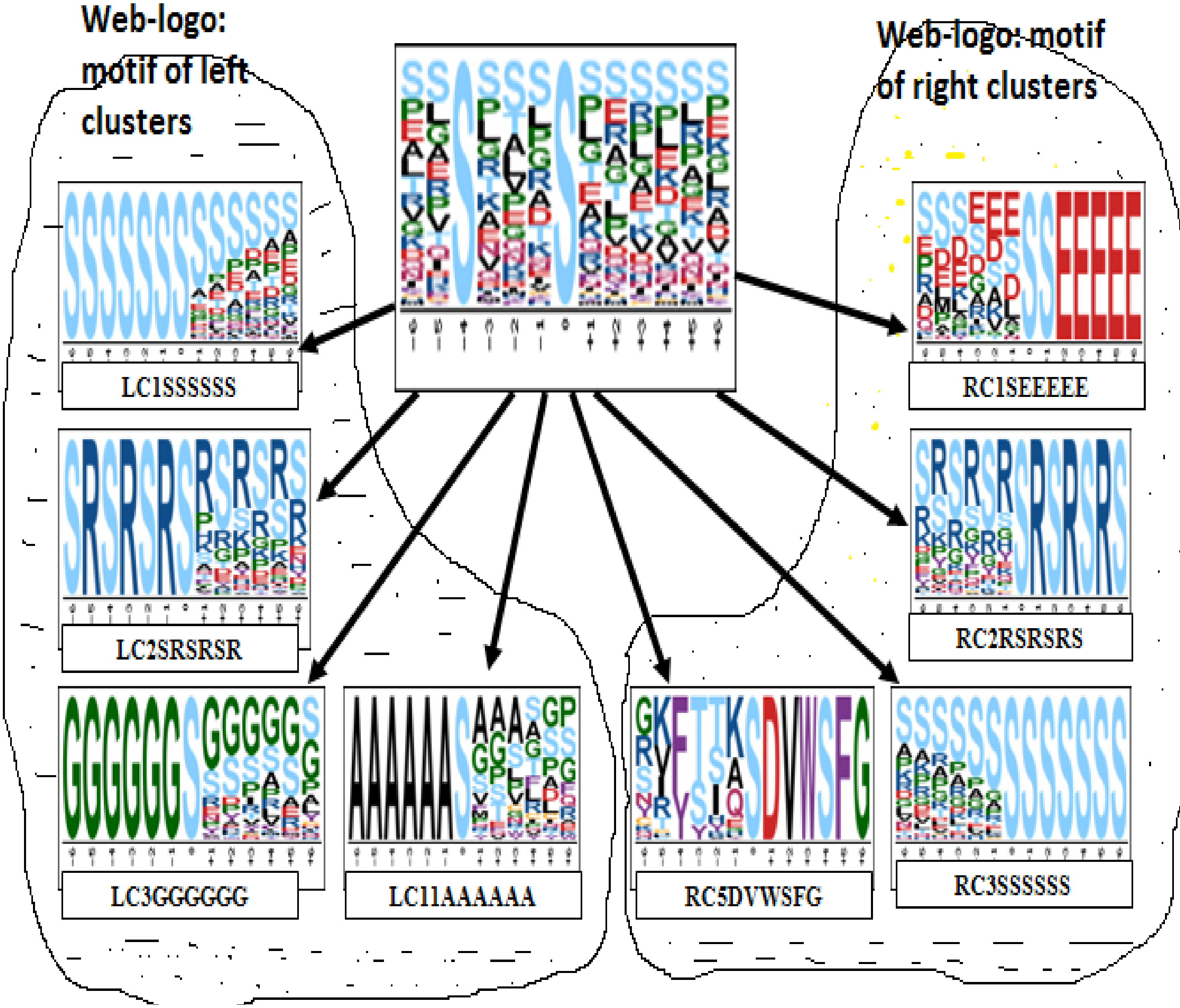


**Fig. 9 Web logo of left and right clusters motif**

given algorithms are given below. Depending on left and right match sequence form left and right clusters assigned scores using mean and max value calculation. In tabular format, it is given in table 1

**Table 1 : Clusters with mean and max values**

| **LEFT CLUSTERS** | | | | |
|---|---|---|---|---|
| **NAME OF CLUSTER** | **ELEMENT FOR MEAN CALCULATION** | **MEAN VALUE** | **ELEMENT FOR MAX CALCULATION** | **MAX VALUE** |
| ***LC1SSSSSS*** | **S: 2664** | **9.0** | **S: 13** | **13** |
| ***LC2SRSRSR*** | **S: 697** | **5.0** | **S:5** | **5** |
| ***LC3GGGGGG*** | **G: 367** | **8** | **G:12** | **12** |
| ***LC4PPPPPP*** | **P: 190** | **7** | **P:11** | **11** |
| ***LC5ASSSSS*** | **S: 210** | **8** | **S: 12** | **12** |
| ***LC6RRSRSR*** | **R: 194** | **6** | **R:8** | **8** |
| …………. | ………… | …….. | ………. | ……… |

……………………………………………………………….
……………………TOTAL 11 LEFT CLUSTERS………

| **RIGHT CLUSTERS** | | | | |
|---|---|---|---|---|
| **NAME OF CLUSTER** | **ELEMENT FOR MEAN CALCULATION** | **MEAN VALUE** | **ELEMENT FOR MAX CALCULATION** | **MAX VALUE** |
| ***RC1SEEEEE*** | **E: 149** | **6** | **E: 8** | **8** |
| ***RC2RSRSRS*** | **S: 729** | **5** | **S: 10** | **10** |
| ***RC3SSSSSS*** | **S: 2870** | **9** | **S: 13** | **13** |
| ***RC4GSSSSS*** | **S: 197** | **8** | **S: 11** | **11** |
| ***RC5DVWSFG*** | **S: 106** | **2** | **S: 5** | **5** |
| ***RC6SGSSSS*** | **S: 194** | **8** | **S: 12** | **12** |
| …………. | ………… | …….. | ………. | ……… |

…………………………………………………………………
……………TOTAL 9 RIGHT CLUSTERS………..

After assigned score values to clusters, next select sequence from intersection of left and right cluster randomly and treated it as unannotated sequence. Considering an unannotated sequence to predict functional group, count max occurrence of constituent and assign as a score of the sequence. Estimate distance values ($d_j$ ) from all clusters ($LC_P$ and $RC_q$ ) using above given algorithms. It is given below in Fig. 10.

Unannotated sequence will assign by the functional group of the cluster whose distance value is minimum.

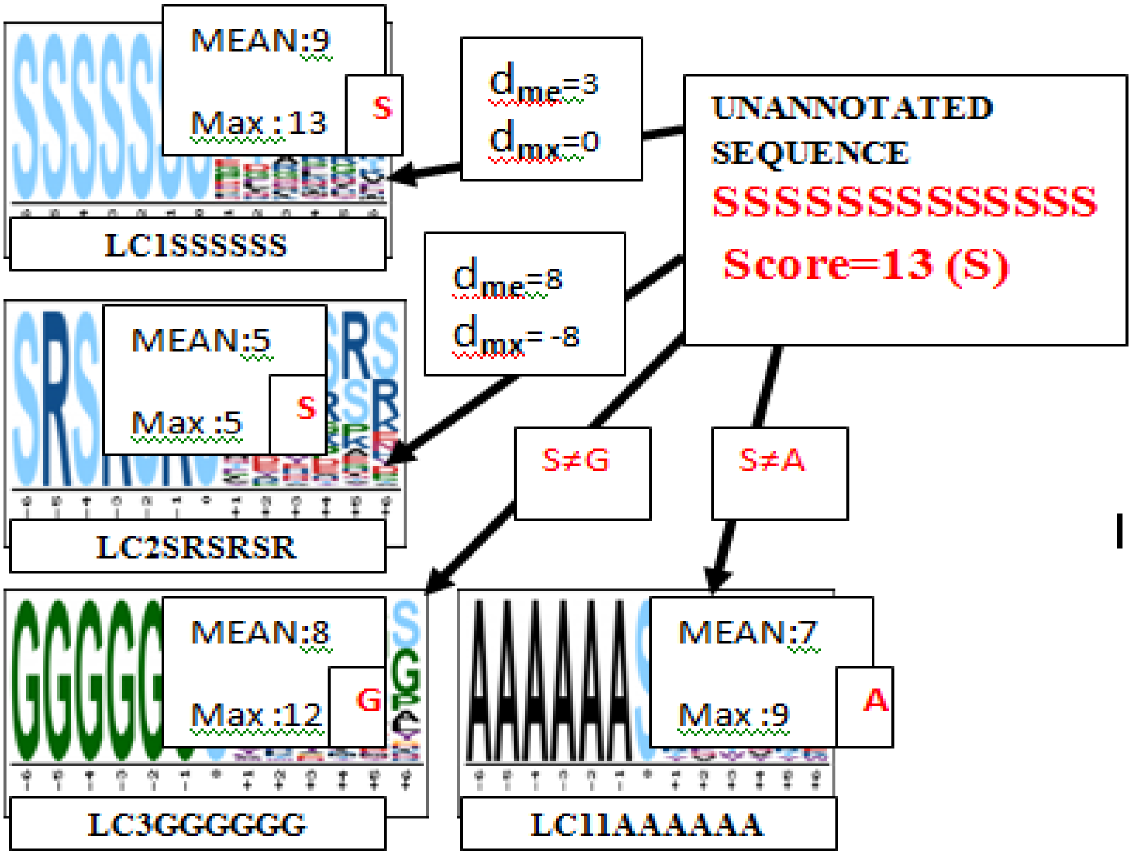


**Fig. 10 distance calculation of unannotated sequence from left clusters motif**

Here, distance values: *min {$d_{me}$: due to mean values of clusters}, min {$d_{mx}$: due to max values of clusters}* are represented in Fig. 10. I can conclude that, functional group of unannoated sequence is assigned by functional groups of LC1SSSSSS.

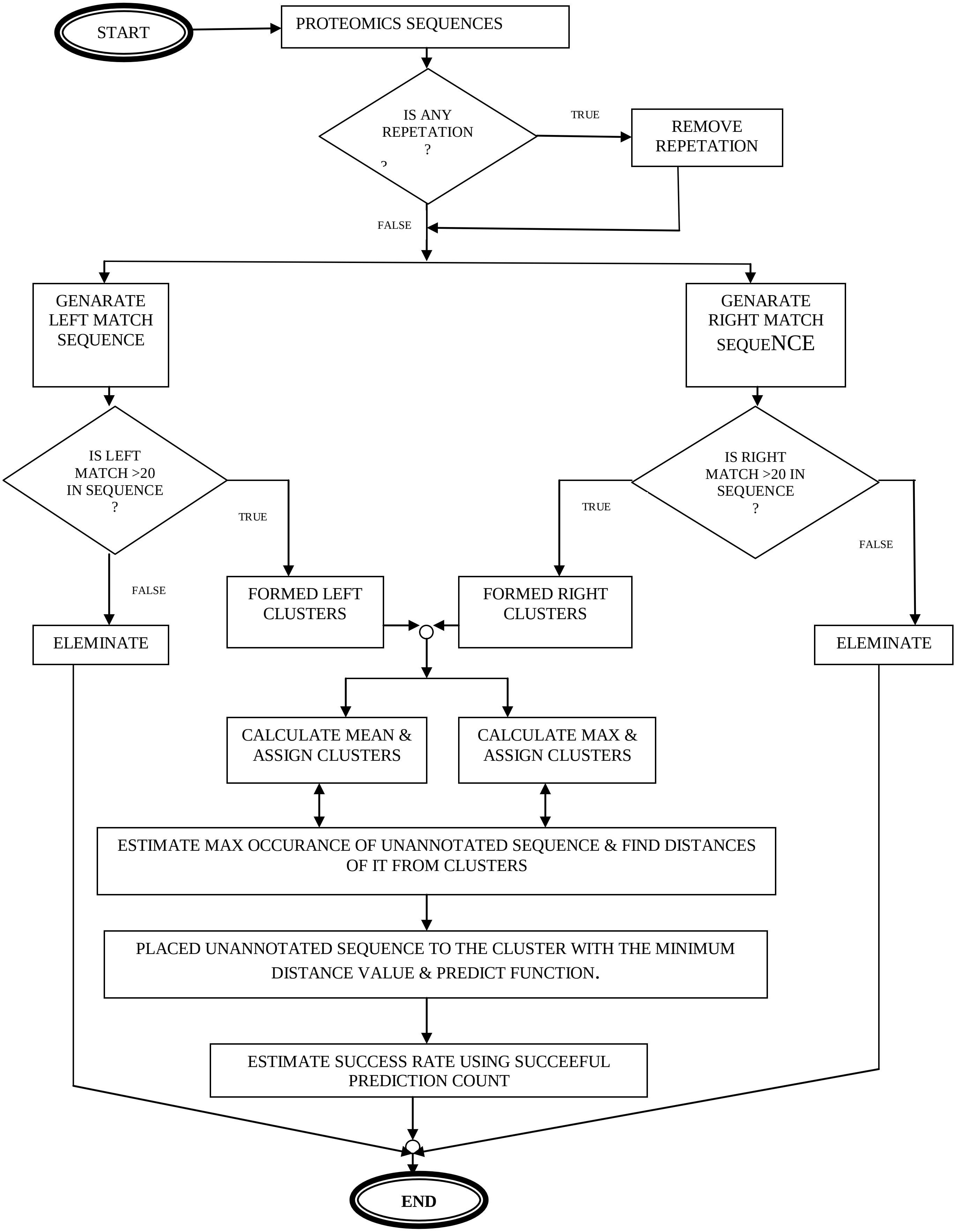
If total no of sequence in dataset is N. Complexity of cluster formation =NXN=$N^2$ . Consider, number of left cluster and right cluster are P and Q respectively, where P<N and Q < N. Unannotated sequence is represented using Un.

Total complexity $\leq$ {(Un X P)+(Un X Q)+$N^2$}

$\leq$ {(P+Q)Un +$N^2$}

= $\Theta$ **($N^2$)**

Flow chart is provided below of the above given algorithms of F-Func in Fig. 11.


START
PROTEOMICS SEQUENCES
IS ANY REPETATION ?
TRUE
REMOVE REPETATION
FALSE
GENARATE LEFT MATCH SEQUENCE
GENARATE RIGHT MATCH SEQUENCE
IS LEFT MATCH >20 IN SEQUENCE ?
TRUE
FALSE
IS RIGHT MATCH >20 IN SEQUENCE ?
TRUE
FALSE
ELEMINATE
FORMED LEFT CLUSTERS
FORMED RIGHT CLUSTERS
ELEMINATE
CALCULATE MEAN & ASSIGN CLUSTERS
CALCULATE MAX & ASSIGN CLUSTERS
ESTIMATE MAX OCCURANCE OF UNANNOTATED SEQUENCE & FIND DISTANCES OF IT FROM CLUSTERS
PLACED UNANNOTATED SEQUENCE TO THE CLUSTER WITH THE MINIMUM DISTANCE VALUE & PREDICT FUNCTION.
ESTIMATE SUCCESS RATE USING SUCCEEFUL PREDICTION COUNT
END

## RESULT AND DISCUSSION

**Table 2. Results of unannotated sequences using propose algorithms**

| UNANNOATED SEQUENCE | MAX OCCURANCE OF ELEMENT | PREDICTED LEFT CLUSTER USING MEAN ALGORITHM | PREDICTED RIGHT CLUSTER USING MEAN ALGORITHM | SUCCESSFUL PREDICTION USING MEAN | | PREDICTED LEFT CLUSTER USING MAX ALGORITHM | PREDICTED RIGHT CLUSTER USING MAX ALGORITHM | SUCCESSFUL PREDICTION USING MAX | | SUCCESS RATE USING MEAN CALCULATION | SUCCESS RATE USING MAX CALCULATION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | LEFT CLUSTER | RIGHT CLUSTER | | | LEFT CLUSTER | RIGHT CLUSTER | | |
| SSSSSSSRSRSRS | S=10 | LC1SSSSSS | RC3SSSSSS | ✓ | X | LC10SGSSSS | RC2RSRSRS | X | ✓ | 16<br>OUT<br>OF<br>26<br>=16/26<br>= .615<br>**61%** | 18<br>OUT<br>OF<br>26<br>=18/26<br>=69.7<br>**70%** |
| SSSSSSSSSSSSS | S=13 | LC1SSSSSS | RC3SSSSSS | ✓ | ✓ | LC1SSSSSS | RC3SSSSSS | ✓ | ✓ | | |
| SSSSSSSSGSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | ✓ | X | LC10SGSSSS | RC6SGSSSS | X | ✓ | | |
| SRSRSRSRSRSRS | S=7 | LC2SRSRSR | RC2RSRSRS | ✓ | ✓ | LC2SRSRSR | RC2RSRSRS | ✓ | ✓ | | |
| GGGGGGSGGGGGG | G=12 | LC3GGGGGG | RC9GGGGGG | ✓ | ✓ | LC3GGGGGG | RC9GGGGGG | ✓ | ✓ | | |
| GGGGGGSGSSSSS | G=7 | LC3GGGGGG | RC9GGGGGG | ✓ | X | LC3GGGGGG | RC9GGGGGG | ✓ | X | | |
| ASSSSSSSSSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | X | ✓ | LC5ASSSSS | RC4GSSSSS | ✓ | X | | |
| RRSRSRSRSRSRS | R=7 | LC6RRSRSR | RC2RSRSRS | ✓ | ✓ | LC6RRSRSR | RC2RSRSRS | ✓ | ✓ | | |
| SASSSSSSSSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | X | ✓ | LC7SASSSS | RC4GSSSSS | ✓ | X | | |
| GSSSSSSSSSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | X | ✓ | LC8GSSSSS | RC4GSSSSS | ✓ | X | | |
| SSGSSSSSSSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | X | ✓ | LC9SSGSSS | RC4GSSSSS | ✓ | X | | |
| SSGSSSSSGSSSS | S=11 | LC1SSSSSS | RC3SSSSSS | X | X | LC9SSGSSS | RC6SGSSSS | ✓ | ✓ | | |
| SGSSSSSSSSSSS | S=12 | LC1SSSSSS | RC3SSSSSS | X | ✓ | LC10SGSSSS | RC4GSSSSS | ✓ | X | | |

Proposed algorithms of F-Func are applied to predict functional groups of unannoated sequences. The results are given in tabular format as table 2. It is concluded to results that, Max approach produces better outcomes as compare Mean approach. It is represented graphically in Fig.12.

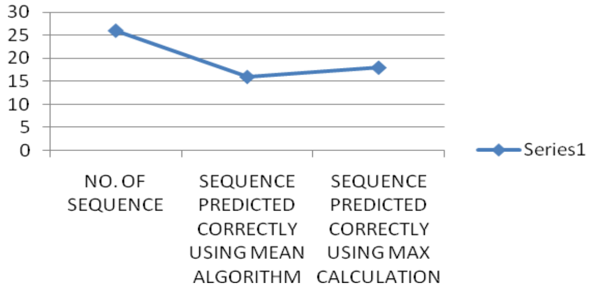


**Fig. 12. Success rate of propose algorithms**

## CONCLUSION

It is quite clear that, using homology we can predict unannoated sequences successfully. Even success rate can be more than 70% as per my algorithms. Hopefully F-Func algorithms will provide better results for larger dataset. My proposed work adds a dimension to existing methods. Hopefully, performance of the F-Func algorithms will increase if length of left-match and right-match is increased.

## ACKNOWLEDGEMENT

It is great help received from the scholars whose articles cited and included in references of this manuscript. I am also grateful to authors / editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed. Hence, I am grateful to IJCRR editorial board members and IJCRR team of reviewers who have helped me to bring quality to this manuscript.

## REFERENCES

1. Bork, P. and Koonin, E.V. 1998. Predicting functions from protein sequences—Where are the bottlenecks? Nat. Genet. 18 313–318.
2. A.K.Payra and S.Saha, IJCET.2013.Generic approach for predicting unannotated protein pair function using protein.page.142-159
3. Cheng, Chung, Aguan, Yang, Wang, N.Paul, PLoS ONE,2011, Dicovery of protein Phosphorylation Motif through Exploratory Data analysis.
4. Anna R. Panchenko, Fyodor Kondrashov, and Stephen Bryant - Prediction of functional sites by analysis of sequence and structure conservation, 2004Devos, D. and Valencia, A. 2000. Practical limits of function prediction. Proteins 41 98–107.
5. Todd, A.E., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307 1113–1143.
6. Casari, G., Sander, C., and Valencia, A. 1995. A method to predict functional residues in proteins. Nat. Struct. Biol. 2 171–178.
7. Andrade, M.A., Casari, G., Sander, C., and Valencia, A. 1997. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol. Cybern. 76 441–450.
8. Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257 342–358.
9. Sjolander, K. 1998. Phylogenetic inference in protein superfamilies: Analysis of SH2 domains. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6 165–174.
10. Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J. Mol. Biol. 311 395–408.
11. Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J. Mol. Biol. 316 139–154.
12. http://motif-x.med.harvard.edu/motif-x.htm
13. http://f-motif.classcloud.org
14. "Computational Approaches for Protein Function Prediction: A Survey"- Gaurav Pandey, Vipin Kumar and Michael Steinbach