Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan Vol. 25 No. 2-2020
ISSN:1813-1786 (Print) 2313-7770 (Online)

A Novel Evaluation of Motif Detection in Protein Sequences of p53 and DNA Sequences of RHAG Gene using Big Data Analytic Techniques

R.Tallat[1], M. Farhan[2], M.M. Iqbal[3], Y. Saleem[4]

[1,2] *Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Pakistan,*
[3]*Department of Computer Science University of Engineering and Technology, Taxila, Pakistan,*
[4]*Department of Computer Science & Computer Engineering, University of Engineering and Technology, Lahore, Pakistan.*

raiha.tallat@live.com

Abstract- Big data has attracted a broad spectrum of attention from researchers and data scientists. Huge batches of data when adequately processed using appropriate algorithms in accordance with the required output prove to be very fruitful in the process of distilling information related to business, health, mechanics and various other domains. Data, when provided with an interface in an interpretable form, is the key to acquiring knowledge from that data. In literature, graph visualization analysis is one of the appropriate techniques for data interpretation, especially in the case of genomics because genomic sequences are comprised of motifs, which can be best, understood and analyzed in graphical form. Research shows that previously graph motif detection has been performed via graph partitioning detection algorithms to retrieve recommendations for binding sites in case of DNA sequences and active site for enzymes in case of protein sequences. Motif detection in protein and DNA sequences, using a partitioned approach is intricate. This paper is based on the protein sequences of p53 known as the guardian of DNA and the RHAG gene sequence responsible for the mutation found in the Rh null system. Detection of the motif in the protein and DNA sequences is discovered by using the MM algorithm implemented in the Multiple EM for Motif Elicitation (MEME) tool for both protein and DNA sequences, and matches are found using the TOMTOM comparison technique. The desired motif is searched across the available thirteen databases like JASPER, Homo Sapiens, and so on. The shortest motif of width was found in all databases except the DAP database. The calculated results have an e-value of 2.05e. The mixture model used for the algorithm showed different processing times for DNA and Protein sequence analysis.

Keywords- Motifs Detection, PPI, Big data analytics, graph analytics

I. INTRODUCTION

Data analysts are considered as the mine diggers of data. This big data can be visualized and interpreted and understood by following the correct sequence as well as the correct algorithms and tools. Data are increasing exponentially by its volume. It means some drastic changes need to be made in the methods of management of big data. As it becomes more worthy and more valuable, it requires more effort and attention to be managed. Making the biological data available for humans to interpret and for computers to analyze is called biocuration [1]. The field of biocuration has opened up new career pathways for data scientists as well as bioscientists. Biological databases have proved to be very resourceful in the detection and analysis of different biological entities for example proteins, chemicals, genomes, and species. The heterogeneous nature of data requires more efficient working algorithms for processing and retrieving information, for example in genomics, 500,000 microarrays are publicly available [2]. This heterogeneity of data is considered as a challenge for data scientists, and it demands some specific approach for manipulation of data and as proposed by a particular community of researchers who demand an efficient platform for Minimum Information for Protein Functionality Evaluation (MIPFE) [3]. Different protein annotations proved to be very resourceful and useful for identification of functional traits of proteins. Annotations are descriptions of the significant part of the genome that describes the region of coding its structures and traits, for this purpose many proteinpedia are available [4]. All of this rich data needs analysis so that some conclusive results can be derived. Data interpretation is an essential part of data analysis. Visualization tools have proved themselves as the sailors in the huge sea of data. Researchers ultimately had to embrace the tools supporting visualization.

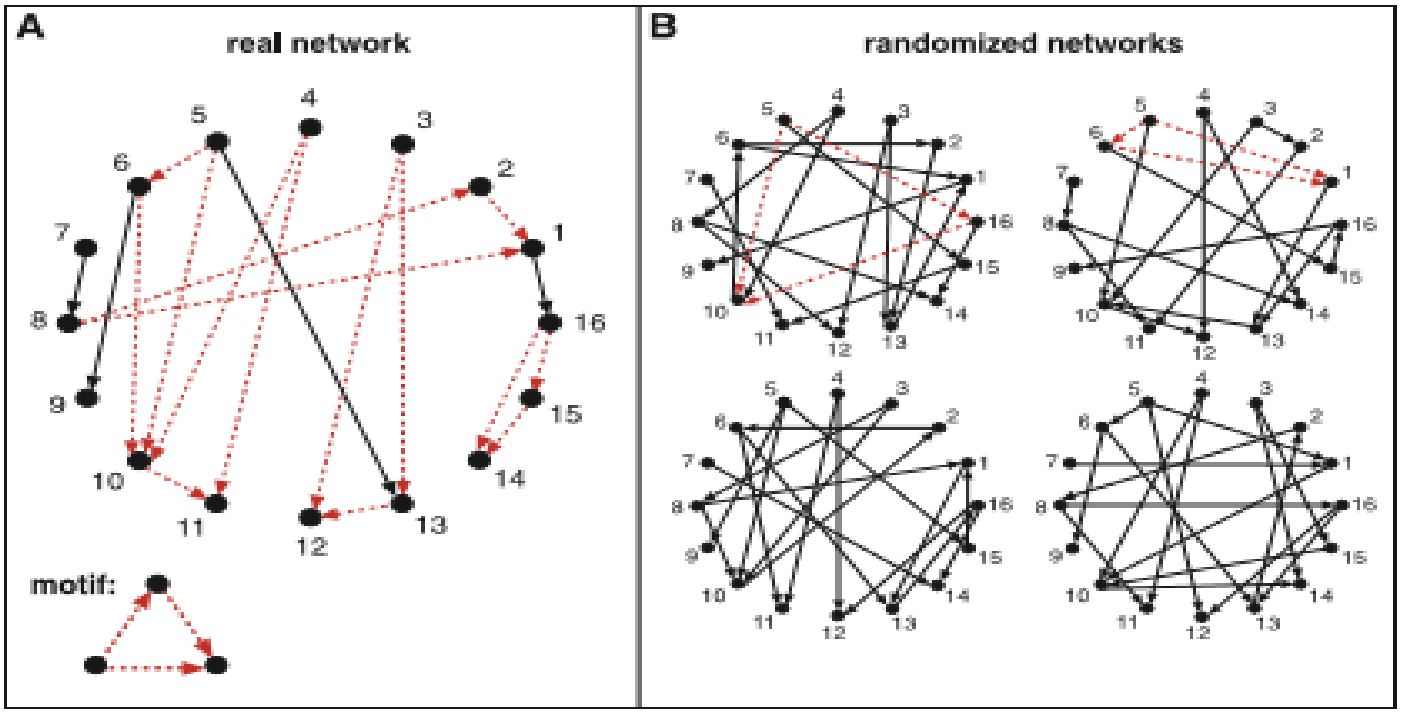

Figure 1: Motif expression in the real network and randomized network

The question that rose was that "Only visualization can prove to be useful?" What if! Visualization could not meet the expectation of the analyst. For this purpose, more visualization techniques were used upon existing graph visualization techniques for detection and interpretation of graph data [5] like visual graph motif detection, where motifs are the repetitive patterns in a graph usually formed in protein-protein interaction networks (PPI) [6]. A view of network motifs can be shown in Fig 1.

Motifs can be found in a single network or different networks, in case of multiple networks, a set of different graph act as the input upon which then similar patterns or motifs are detected. The term "motif" actually has great significance in the world of genomics. Gene structure can be represented in the form of visual graphs due to the repetition of protein sequences in a genome that causes the motifs to form. Mutation in the genome and exome sequence is the cause of many diseases [7]. One such rare genetic mutation in the absence of antigens from the bloodstream. The red blood cells in our blood are covered with 342 antigens. A combination of these is responsible for the decision of different blood types people have. The unique factor occurs when some amount of antigens are missing from the bloodstream, for example, a rare blood type that is the cause of mutation is the absence of 62 antigens from the Rh system thus forming the rarest blood type known as Rh null or hematologists call it "The Golden blood" [8]. This phenotype is further classified into regulator and amorph [9, 10]. This rare genetic mutation is highly significant as it causes an individual to have a rare blood type. Another mutant protein p53 is responsible for the extensive growth of cells thus causing cancer. The p53 protein is found in almost every living organism.

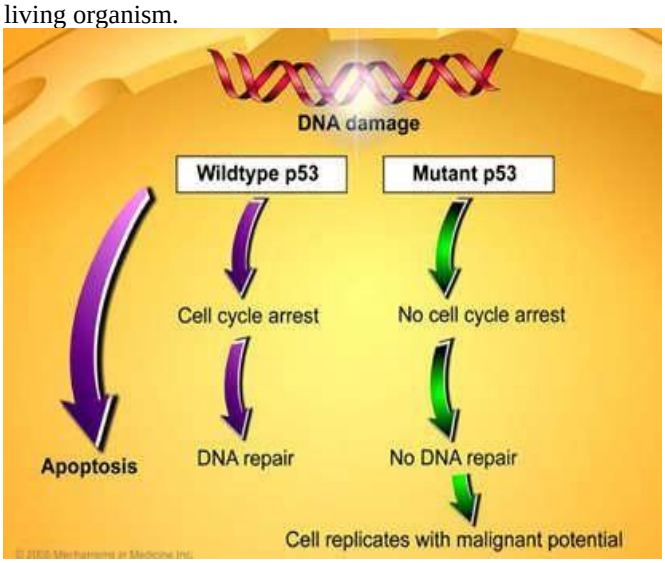

Figure 2: mutant p53 cell cycle

Fig 2 represents a generic representation of the mutation caused in p53 responsible for the origination of different types of cancer in living organisms. This mutation is responsible for various diseases in different living organisms, and its genetic sequence carries significant importance.

A large number of tools like DAVID [11, 12], are used for the functional analysis of such mutated genes. The tools work on the principal and algorithms of Data Science. One such

tool is DNA motif identification and analyses (DMINDA), or now can say DMINDA[2] [13]. De-novo motif identification and associated computational analyses have a great significance in motif identification. The uniqueness of DMINDA is that it contains 5 motif analysis algorithms. Previously tools were focused on the identification of motifs but missing the analysis, for example, MEME suite [14], Motif Finder, PATLOC, AIMIE and so on. DNA motif identification and analyses (DMINDA) provides.

II. Related Work

Gene mapping has been considered as one of the challenging tasks for researchers who belong to the field of bioinformatics and data analysis. Previously such tools are developed, which are used for gene analysis and gene mapping, for example, MAPMAKER [15]. MAPMAKER has been applied to the construction of linkage maps in some organisms; including the human. E. S. Lander et al. [16] proposed this efficient algorithm for constructing linkage maps in Humans and other species. A limitation regarding the linkage map in the past was that it was only possible to determine the linkage maps of the most widely studied organisms due to the prominent mutations found in their genetic markers. For comparing these mutations with mutations found in other organisms, a large number of knowledge bases are available. More than one order of marker is known as a sequence. For example: Sequence (1 2) 3 4 5. Now the MAPMAKER maps it accordingly with (1 2) 3 (4 5) which is also considered as a sequence. A variety of tools is available for data visualizations depending on the size and type of data. A very concentrated area for visual analysis is protein simulation using molecular dynamics. Proteins are considered as the essential building blocks for the human body; they are involved in muscular development. Mutations in its structures can cause various diseases for this purpose simulation in proteins can reveal many new structures.

One of the techniques introduced for protein unfolding is Steering MD [18]. The reduced glycoprotein in the R_h systems can result in the mutation of the RHCE gene mentioned in the introduction section. Several tools for detecting motifs in the graphs are available like FANMOD [19], MAVISTO, PAJEK, MFINDER, and so on. MAVISTO uses the force directed graph layout. The uniqueness of FANMOD is that it helps the detection of large motifs in complex graphs. Another algorithm known as NemoFinder helps to find size-12 motifs in PPI (protein-protein interaction) networks [20]. Graph visual motifs are also helpful for distinguish between applications protocol and to determine the known behavior of unlabeled traffic [21]. The most widely used algorithm for motif discovery is the MEME. MEME is a complete suite and performs a series of operations on the dataset thus discovering, analyzing, finding enrichment and comparison with the existing motif databases [22]. Some other tools like DMINDA[2], Ensemble genome browser also performs a sequence of operations [23]. Rest of the paper is organized as:
Section II describes the Relate Work, section III contains the Methodology used for carrying out the analysis, results are discussed in section IV, and finally, section V sums up the paper with Conclusion.

III. Proposed Methodology

This section has explained the proposed method used for the analysis of sequences performed by using Multiple EM for Motif Elicitation (MEME) and DMINDA[2]. Sequence clusters are downloaded from the UniProt database. The sequences were of p53 and Rh null gene. Both sequences have their significance in genomics. They were selected due to the unique feature of being the "guardian of the DNA," an example of mutation caused in Homo sapiens.

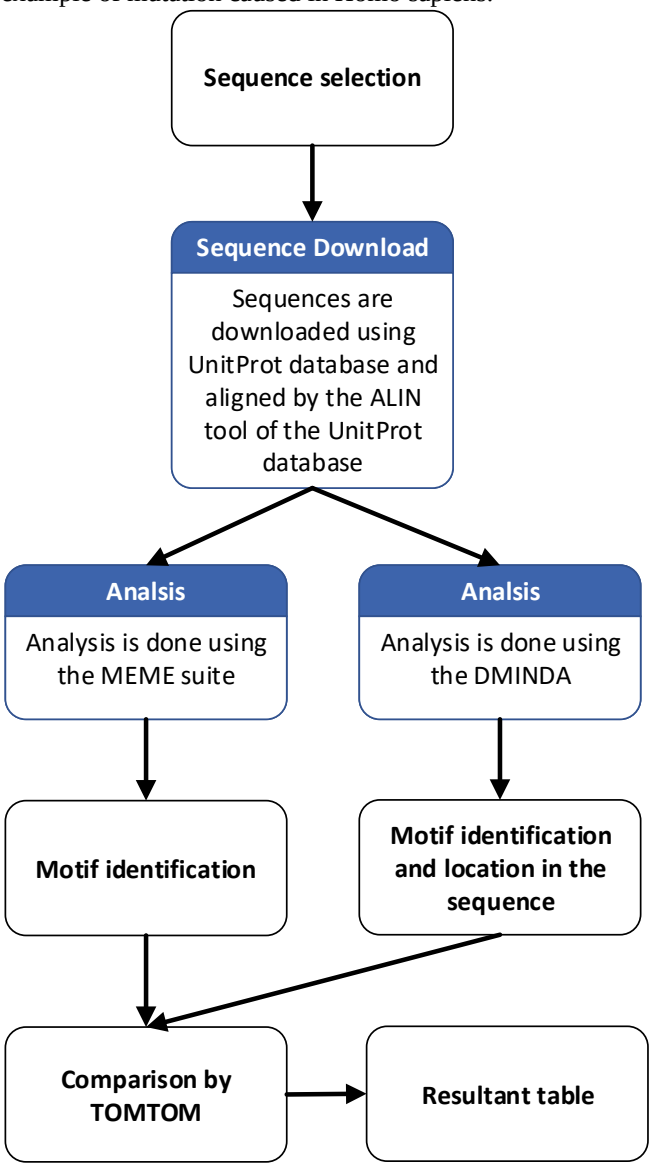

Figure 3: Methodology for Motif elicitation in the desired sequence

MEME discovery algorithm is used for finding the motif within the sequences. 10 such motifs are discovered, and the shortest motif found is then compared using TOMTOM, and a resultant table is derived according to the number of matches along with the DNA motifs that are found using DMINDA. The motif which is found in a maximum number of sequences is compared with the existing database and results are derived. The methodology is given in fig 3.

The match threshold should definitely have an E-value of 10 or smaller to be considered as a match. Following were the results when the given motif was searched in the mentioned knowledgebase. The sequence "CATACCCCG" got a few hits in the motif database. The TOMTOM algorithm provides the p-value and q-value which are the measures of false discovery rate of the match.

IV. Result and Discussion

In this section, the findings by DMINDA[2] are discussed, as well as the results are evaluated. The UniProt database was used for downloading sequences of p53 and the Rh null syndrome of the species Homo sapiens and submitted for a job in two different tools MEME suite and DMINDA [24].

A. Motif Detection by, MAST output with similarity matrix:

MOTIFS

	Logo	Name	Alt. Name	Width	Motif Similarity Matrix 1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1.		HFTQTPLDPGSQVLVRVDWERRFDHMQQHSGQHLITAVADHLFKLKTTSW	MEME-1	50	--	0.14	0.17	0.12	0.13	0.14	0.14	0.19	0.19	0.19
2.		NMCCGTHVSNLSDLQVIKILGTEKGKKNRTNLIFLSGNRVLKWMERSHGT	MEME-2	50	0.14	--	0.19	0.15	0.15	0.22	0.2	0.23	0.15	0.33
3.		QNSTKILQKNNLNLLRDLAVHIAHSLRNSPDWGGVVILHRKEGDSEFMNI	MEME-3	50	0.17	0.19	--	0.14	0.18	0.14	0.2	0.22	0.17	0.3
4.		FTTTVVSCCPAELQTEGSNGKKEVLSGFQVVLEDTVLFPEGGGQPDDRGT	MEME-4	50	0.12	0.15	0.14	--	0.16	0.23	0.13	0.22	0.12	0.22
5.		FRSAIELDTPSMTAEQVAAIEQSVNEKIRDRLPVNVRELSLDDPEVEQVS	MEME-5	50	0.13	0.15	0.18	0.16	--	0.17	0.15	0.22	0.13	0.26
6.		PPASVETLGPRVAEVLEGKGAGKKGRFQGKATKMSRRMEAQ	MEME-6	41	0.14	0.22	0.14	0.23	0.17	--	0.17	0.21	0.19	0.26
7.		IEDHRIVFSCKNADGVELYNEIEFVAKVNSKDSQDKRSSRSITCFVRKWK	MEME-7	50	0.14	0.2	0.2	0.13	0.15	0.17	--	0.22	0.2	0.31
8.		IANEIGSEETLLFLTVGDEKGGGLFLLAG	MEME-8	29	0.19	0.23	0.22	0.22	0.22	0.21	0.22	--	0.16	0.22
9.		TKEDIKPVWLSVDFDNWRDWEGDEEMELAHVEHYAELLKKVSTKRPPPAM	MEME-9	50	0.19	0.15	0.17	0.12	0.13	0.19	0.2	0.16	--	0.23
10.		RGLPDDHAGPIRVVNIEGVDS	MEME-10	21	0.19	0.33	0.3	0.22	0.26	0.26	0.31	0.22	0.23	--

Figure 4: Motif detection by MAST, 10 best Motifs, and similarity matrix.

B. Motif detection in p53 protein sequence, MEME output:

MEME showed the shortest motif found in the sequence with the minimum width 21, no motif in the sequence is found with a width less than 21 [25]. The sequence logo for the shortest is given in fig 4. The description of the logo is as shown in figure 5. The details are summarized in table 1.

i. The height of the Letter ≈ fraction of the time that letter is observed at a specific position.
ii. The height of all the letters in a column ≈ to which extent the amino acid is conserved.

Figure 5: the Shortest motif found in the sequence, MEME output

Table 1: Parameters for shortest motif detection in sequences

Name	Start	p-value	Sites
sp|Q9BTE6-3|AASD1_HUMAN	360	2.97e-26	DDPEVEQVSG**RGLPDDHAGPI RVVNIEGVDS**NMCCGTHVSN
sp|Q9BTE6-2|AASD1_HUMAN	299	2.97e-26	DDPEVEQVSG**RGLPDDHAGPI RVVNIEGVDS**NMCCGTHVSN
sp|Q9BTE6|AASD1_HUMAN	186	2.97e-26	DDPEVEQVSG**RGLPDDHAGPI RVVNIEGVDS**NMCCGTHVSN

C. Motif Detection in RHAG by, DMINDA[2]:

The motif detection in DMINDA is performed upon the same sequence as in MEME; the unique functionality of DMINDA shows the occurrence of the motif in a sequence.

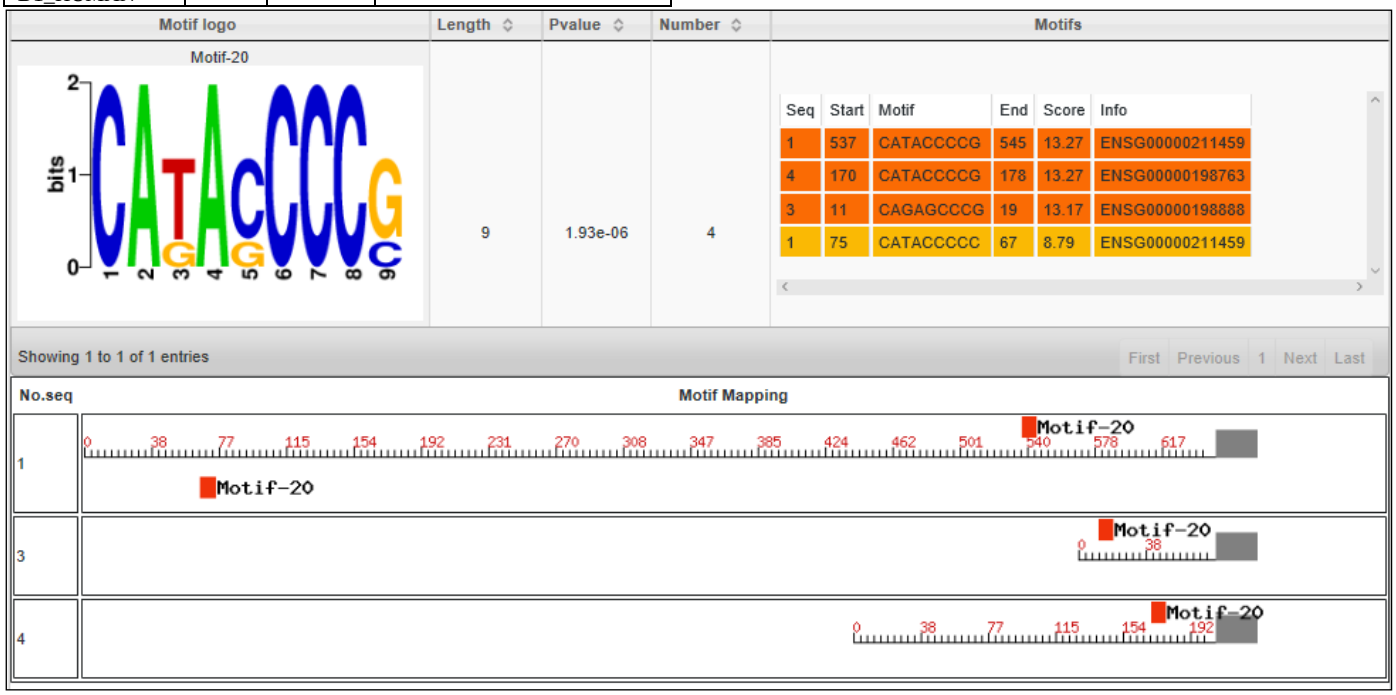

Figure 6: Motif detection in RHAG by DMINDA[2]

Motif 20 shown in fig 4 is the only overlapping motif that is repeated 3 times in 3 of the given sequences.

D. Motif Detection in RHAG protein sequence by, MEME:

Protein sequences of RHAG obtained by the UniProt database were analyzed by the MEME tool to determine the motifs present in the sequence [26]. The fig 6 shows the motif containing the minimum number of residues, the detected motifs can be used for determining the binding pockets [27].

E. Advantages of the MM algorithm, as per calculated Results

The MM algorithm has two fundamental advantages:

i. Advanced classification of input sequences is not required. It searches for the motif in a data set as well as the occurrence frequency.
ii. The MM algorithm also allows the user to select various parameters for the motif discovery via the Opal interface.

F. Comparison of the motif with JASPER, Eukaryotic, and Malaria database using TOMTOM algorithm :

The match threshold should have an E-value of 10 or smaller to be considered as a match. The following were the results when the given motif was searched in the mentioned knowledgebase. The sequence CATACCCCG got a few hits in the motif database. The TOMTOM algorithm provides the p-value and q-value which are the measures of the false discovery rate of the match.

G. TOMTOM Algorithmic Matches

Motif discovery is indeed a vital biological activity, but after the discovery of the motifs, the next research question that pops up is whether the discovered motif is already known or has some genomic significance [28]. For this purpose, TOMTOM algorithm has been used to find the significance of the query motif. The TOMTOM algorithm is capable of finding relative offsets in case of protein motifs and reverse compliments in case of DNA motifs [29]. Moreover, the program also determines those motifs which are highly redundant in a motif database and also has the merging capability of those motifs as shown in fig 7.

P-values are calculated on the basis of Motif Score. A comparison between p values of both the motif scores is computed. Where a motif L of width w, a score function $\hat{s}$ ($L_{i,}b$) that produces a score which is a positive integer for the similarity of the i^{th} column of L and the letter $b \in A$.For

computing the p-value of a motif TOMTOM identifies an offset as well and relative orientation for which the offset *P* value is minimal. The probability of observing a minimum *P* value of *P** among a collection of *M* independent *P* values is:

$$1 - (1 - P^*)^M .$$

The expression the calculated p-value for a motif. The calculated p-values are shown in table 2.

H. Calculation of p-values in MEME:

Table 2: Calculated p-value for motif discovery in input sequences

Sr. #	***p-value***	***2.04e-04***
1.	*e-value*	*2.04e-04*
2.	*q-value*	*2.04e-04*
3.	*Overlap*	*7*
4.	*Offset*	*0*
5.	*orientation*	*Normal*

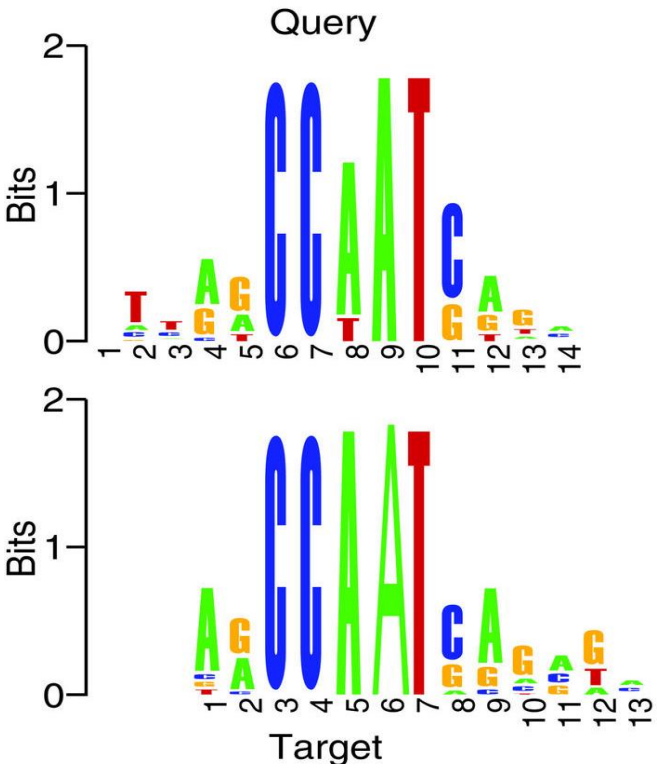

Figure 7: TOMTOM Motif comparison between Query and Target motif

Table 3: No of matches found by the motif determined by DMINDA in different DNA databases

Sr. #	***Knowledgebase***	***No of Matches***
1.	Malaria (Plasmodium falciparum)	9
2.	JASPER	4
3.	Homo Sapiens	8
4.	FLY (combined Drosophila database)	2
5.	CIS-BP Single species	1
6.	Prokaryote DNA (CoLlecTF (bacterial TF motifs))	5
7.	Ray 2013 all species (DNA Encoded)	3
8.	YEASTRACT	3
9.	SwissRegulon e coli	9
10.	DAP Motifs	0
11.	Vertebrates (in Vivo and in Silico)	1
12.	MOUSE	3

I. P-value calculation using DMINDA:

DMINDA[2] has the capability of finding motif co-occurrences in DNA sequences in order to identify joint regulation connections by multiple transcription factors. Calculation of P-value is done by using k as a variable where k is the representation of the regulatory regions containing common motifs. Further calculations are done using the BoBro algorithm of computational genomics as shown in fig 9. The details are given in table 3.

The term methodology defines system of methods or basic rules to perform something in a systematic way. Furthermore, methodology described detailed systematic and theoretical analysis process by implementing techniques and methods of some field of study. In this study, the proposed methodology to empirically evaluate graph databases id discusses in detail. This proposed methodology consists of eight steps, starts from data collection and end on elevation processes. In first step a detail description of all collected data from biological data repositories are described. In second step, data transformation process performed due to data heterogeneity as data collected from different biological data sources.

In the field of science, there are two rapidly changing phenomena. The first thing under consideration is increasing data volume (size) form terabytes to petabytes and beyond. The second one is the advancement of biological interaction networks which produced piles of highly throughput data regarding proteomes, interactomes, transcriptomes, metabolomes, genomes and much more. As field of biology has become increasingly important due to its data intensiveness, therefore biology and computer science have become complementary to each other bridged by other branches of science such as statistics, mathematics, physics, and chemistry. Hence, the emergence of knowledge versatility of these domains caused an advent of Network Biology, Big Data Biology or many other biology branches.

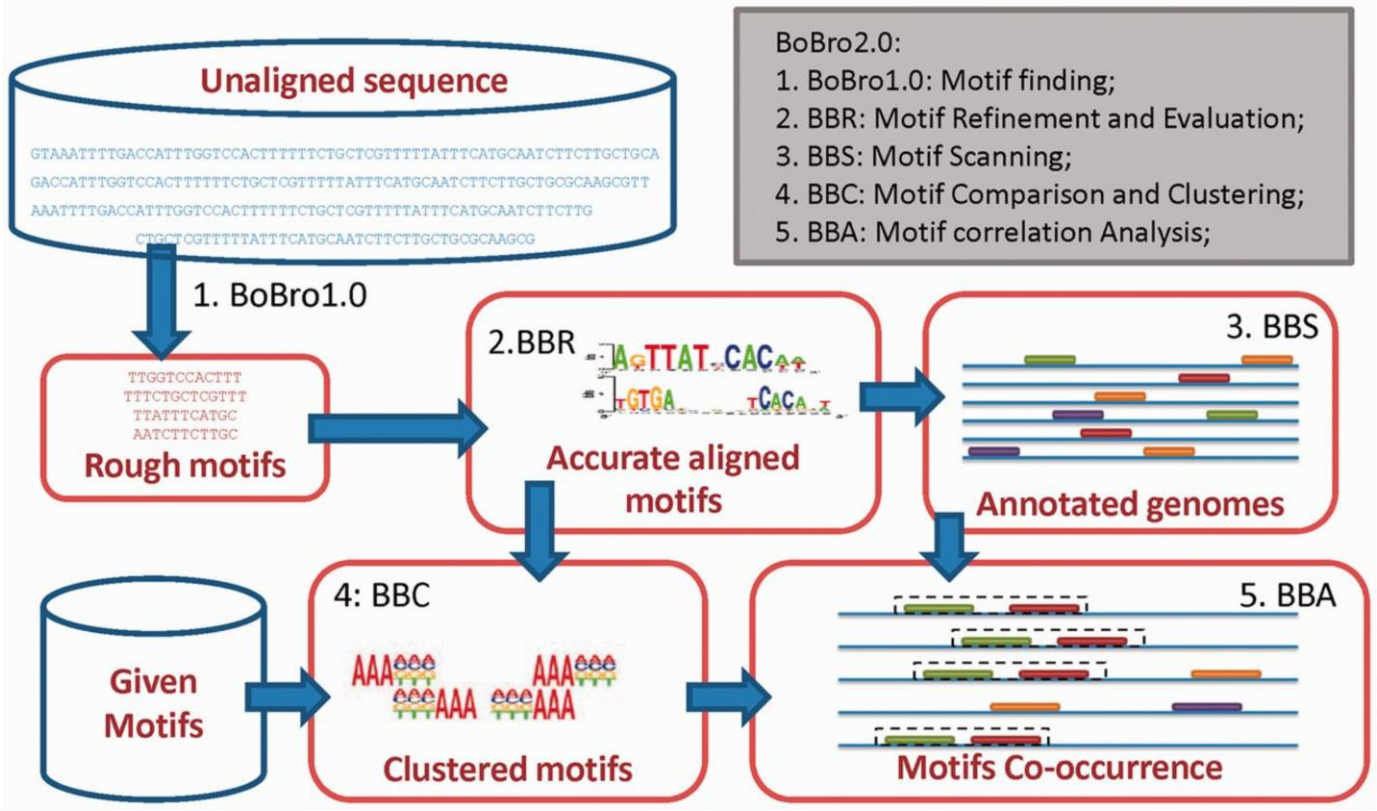

Figure 8: BobRo 2.0 Algorithmic Sequence

Supposedly y be the random variable that represents the number of instances of a particular motif in a given set of sequences of RHAG gene. The distribution probability is expressed by p(x) which can be estimated by using a Poisson distribution. The p-value calculation is done by summing up all the probabilities. By denoting that a motif may have k occurrences, an enrichment score is calculated which is used to evaluate the statistical significance of a motif in input sequences.

$$Z= (S_R-(|R|*S_C)/|C|) / (\sqrt{(R \mid *S_C)} /|C|)$$

Where S_R and S_C are the numbers of instances in R and C, respectively.

To evaluate the graph analytic techniques over biological complex connected data of Protein-Protein Interaction Networks, Gene-Protein Interaction Network and Gene-Gene Interaction Networks. The data sets collected from heterogeneous data repositories are first evaluated on Data Volume, Total No. of Record in each dataset, Distinct No. of Interaction, there Interaction Detection Method and Experimental setup Used, Interaction Type, and many more.

The process of evaluation involves critical analysis of the proposed method or a program. Thus, after defining and describing in the methodology detailed for graph databases (Neo4j, OrientDB, and Titan) in data storage, data querying, and analytics perspective by using biological complex connected interaction networks such Gene-Protein Interaction, Protein-Protein Interaction and Networks Gene-Gene Interaction.

Queries for Centrality Analysis are applied on the created graphs. The queries were targeted for viewing graphs by eliminating specific nodes and relating genes, proteins, and protein and gene sequences with their respective diseases with variations in their properties.

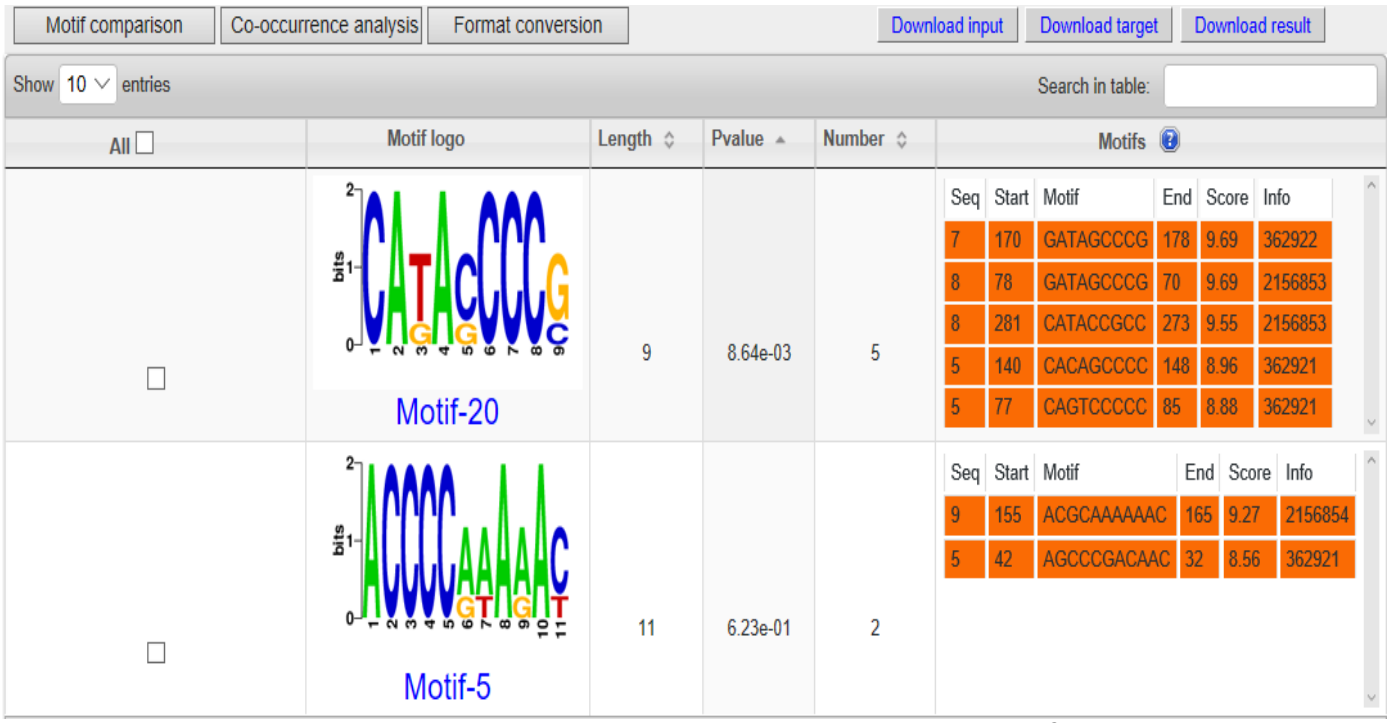

Figure 9: Calculated P-values of candidate sequences in DMINDA[2]

J. Similarity Matrix, p53 DNA sequence using DMINDA:

The similarity matrix for the motif detection in the given input DNA sequences of the RHAG gene is given in fig 8.

K. Comparison of MEME and BobRo used in DMINDA in Motif Identification:

A comparative analysis is shown in table 4, representing the

DMINDA 2.0: An integrated DNA motif analysis web server

name	No	Motif-1	Motif-3	Motif-18	Motif-5	Motif-4	Motif-6	Motif-10	Motif-11
Motif-1	1	1.00	0.12	0.21	0.16	0.18	0.12	0.01	0.10
Motif-3	2	0.12	1.00	0.10	0.11	0.06	0.25	0.20	0.06
Motif-18	3	0.21	0.10	1.00	0.17	0.46	0.14	0.03	0.08
Motif-5	4	0.16	0.11	0.17	1.00	0.13	0.27	0.04	0.12
Motif-4	5	0.18	0.06	0.46	0.13	1.00	0.13	0.06	0.11
Motif-6	6	0.12	0.25	0.14	0.27	0.13	1.00	0.08	0.17
Motif-10	7	0.01	0.20	0.03	0.04	0.06	0.08	1.00	0.38
Motif-11	8	0.10	0.06	0.08	0.12	0.11	0.17	0.38	1.00

Figure 10: Similarity matrix calculation in DMINDA[2]

tools and algorithms used in both for determining and evaluating motifs in candidate protein and DNA sequences.

Table 4: Comparison of Motif Detection in MEME and DMINDA[2]

Sr. #	Functions	MEME	BoBro 2.0
1.	Motif Scanning	Nill	*P*-value assessment for all the scanned and input motifs
2.	Motif Refining	FIFO	Strong ability in filtering out noises at a genome-scale
3.	Motif Comparison	TOMTOM	A motif clustering algorithm

L. Graph representation and clustering of obtained Motifs:

The detected motifs are further organized and clustered in the form of graphs. Minimal Spanning Tree is used to generate a graph of all candidate motifs. Initially, a full graph is generated, representing all the candidate motifs represented as a node and connected by edges.

The MST is generated by using Kruskal's algorithm. A methodology for the generation of spanning tree is shown in fig 10. All the motifs are then compared to the Regulon. A resultant regulon tree is obtained describing the motif similarity.

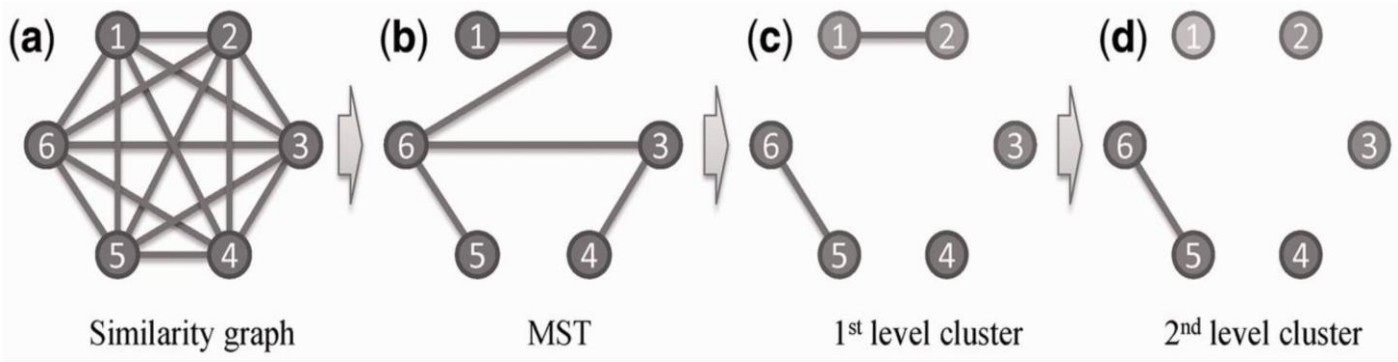

Figure 11: Minimal Spanning levels of clustering

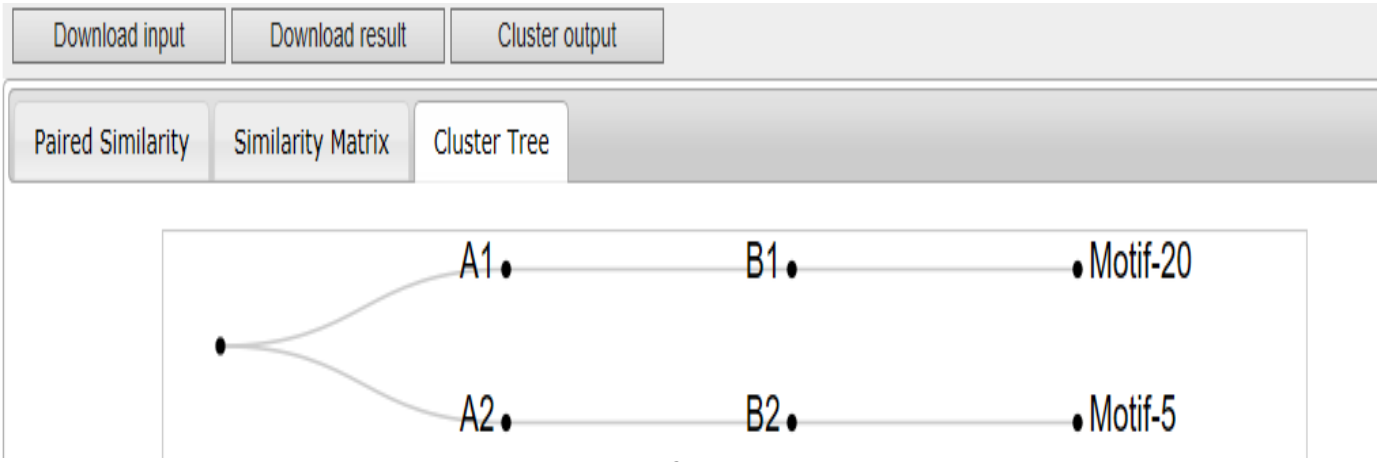

Figure 12: Cluster tree obtained by DMINDA[2] of the candidate DNA sequences of RHAG gene

Fig 11 and fig 12 illustrates the tree obtained after the identification and clustering of motifs in DNA sequences of the RHAG gene responsible for the rare mutation of Rh Null.

V. CONCLUSION

Visual graph motif analysis techniques have proved to be very useful for the determination of repetitive functional patterns in protein structure. We showed that different clusters of sequences could have a repetitive pattern in them, thus forming motifs and these motifs can be further grouped in the form of clusters. Clusters are generated based on the similarity matrix in both MEME and DMINDA[2]. We discovered the motifs between a specific range, i.e., $6 \leq x \leq 5$, where x is the length of the motif. The shortest motif discovered in the candidate sequences in MEME was of length 21, which means that 21 is the threshold value for motif length in the sequence under consideration. MEME represents the motifs in the graphical form known as Motif Logo. All the motifs discovered were from the un-gapped sequences. Similarly, RHAG DNA sequences were analyzed

on DMINDA[2] and the most repetitive motif found was then searched for the match in the existing knowledge bases using the TOMTOM algorithm. A different number of matches are derived across different databases. Only one database gave 0 matches across the given motif, which concludes that the motif "CATACCCCG" is unique for the DAP motif database.

Furthermore, the DMINDA discovered motifs were clustered in the tool using the Minimal Spanning tree. Based on this, a similarity matrix and tree were obtained. Our findings represented that among all the discovered motifs Motif-5 and Motif-20 showed the highest similarity and were clustered together, thus generating the required graph. Our analysis concluded that the algorithms used for identification and analysis of motifs have strengths in particular areas, but DMINDA[2] outperforms MEME due to the integration of five robust algorithms and clustering algorithms which converts all the discovered motifs in the form of graph providing visualized results. Another critical aspect of DMINDA is the DOOR2 eukaryotic database. In both the analytic tools similarity is based upon the p-values of the discovered motifs. The method for p-value calculation is explained in Section III of the paper. The genomic importance of the sequences lies in the clustering of the input protein and DNA sequences. Our research shows that these integrated Graph analytic techniques and algorithms being used in MEME and DMINDA[2] plays a vital role in the identification of repeated patterns known as "motifs." These binding pockets play a vital role in DNA and protein development as they are responsible for the cell's regulatory network. For future work, the discovered motifs can be analyzed by genomic experts in order to obtain their biological and genomic importance. The performed empirical evaluation is one of the novel approaches of visual graph analytic techniques.

REFERENCES

[1]. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., & Twigger, S. (2008). Big data: The future of biocuration. *Nature*, *455*(7209), 47-50.

[2]. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

[3]. De Marco, A. (2008). Minimal information: an urgent need to assess the functional reliability of recombinant proteins used in biological experiments. *Microbial cell factories*, *7*(1), 20.

[4]. Mathivanan, S., Ahmed, M., Ahn, N. G., Alexandre, H., Amanchy, R., Andrews, P. C., & Björling, E. (2008). Human Proteinpedia enables sharing of human protein data. *Nature biotechnology*, *26*(2), 164-167.

[5]. Frankel, F., & Reid, R. (2008). Big data: Distilling meaning from data. *Nature*, *455*(7209), 30-30.

[6]. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594), 824-827.

[7]. Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., ... & Basehore, M. J. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, *13*(3), 255-262.

[8]. Hyland, C. A., Cherif-Zahar, B., Cowley, N., Raynal, V., Parkes, J., Saul, A., & Cartron, J. P. (1998). A novel single missense mutation identified along the RH50 gene in a composite heterozygous Rh null blood donor of the regulator type. *Blood*, *91*(4), 1458-1463.

[9]. Qureshi, A., Salman, M., & Moiz, B. (2010). Rhnull: a rare blood group phenotype. *Journal of the Pakistan Medical Association*, *60*(11), 960.

[10]. Yu, S., Xiao, J., Xu, H., Lin, Y., Chang, J., Li, Y., & Miao, T. (2017). The best strategy of blood transfusion in patients with Rh null syndrome.

[11]. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), 1-13.

[12]. Lane, W. J., Westhoff, C. M., Uy, J. M., Aguad, M., Smeland-Wagman, R., Kaufman, R. M., & Silberstein, L. E. (2016). Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion*, *56*(3), 743-754.

[13]. Yang, J., Chen, X., McDermaid, A., & Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics*.

[14]. Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME suite. *Nucleic acids research*, *43*(W1), W39-W49.

[15]. Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., & Newburg, L. (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, *1*(2), 174-181.

[16]. Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., & Lim, J. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, *42*(D1), D142-D147

[17]. Rysavy, S. J., Bromley, D., & Daggett, V. (2014). DIVE: A graph-based visual-analytics framework for big data. *IEEE computer graphics and applications*, *34*(2), 26-37.

[18]. Toofanny, R. D., & Daggett, V. (2012). Understanding protein unfolding from molecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *2*(3), 405-423.

[19]. Wernicke, S., & Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics*, *22*(9), 1152-1153.

[20]. Chen, J., Hsu, W., Lee, M. L., & Ng, S. K. (2006, August). NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 106-115). ACM.

[21]. Wright, C. V., Monrose, F., & Masson, G. M. (2006, November). Using visual motifs to classify encrypted traffic. In *Proceedings of the 3rd international workshop on Visualization for computer security* (pp. 41-50). ACM.

[22]. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, *37*(suppl_2), W202-W208.

[23]. Fernández, X. M., & Birney, E. (2010). Ensembl genome browser. In *Vogel and Motulsky's Human Genetics* (pp. 923-939). Springer Berlin Heidelberg.

[24]. Yang, J., Chen, X., McDermaid, A., & Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. Bioinformatics, 33(16), 2586-2588.

[25]. Danciu, C., Muntean, D., Alexa, E., Farcas, C., Oprean, C., Zupko, I., ... & Hancianu, M. (2019). Phytochemical characterization and evaluation of the antimicrobial, antiproliferative and pro-apoptotic potential of Ephedra alata Decne. hydroalcoholic extract against the MCF-7 breast cancer cell line. Molecules, 24(1), 13.

[26]. Wen, J., Verhagen, O. J., Jia, S., Liang, Q., Wang, Z., Wei, L., ... & Ji, Y. (2019). A variant RhAG protein encoded by the RHAG* 572A allele causes serological weak D expression while maintaining normal RhCE phenotypes. Transfusion, 59(1), 405-411.

[27]. Gao, X., Gui, H., Lan, M., Zhu, J., Xie, Y., Zhan, Y., ... & Wu, G. (2020). Identification and preliminary characterization of chemosensory-related proteins in the gall fly, Procecidochares utilis by transcriptomic analysis.

[28]. Li, Y., Ni, P., Zhang, S., Li, G., & Su, Z. (2019). ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. Bioinformatics, 35(22), 4632-4639.

[29]. Wylie, D., Hofmann, H. A., & Zemelman, B. V. (2018). SArKS: de novo discovery of gene expression regulatory motifs and domains by suffix array kernel smoothing. BioRxiv, 133934.