J. Appl. Environ. Biol. Sci., 7(11)214-221, 2017

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Towards Exploring the Combined DNA and RNA Sequences with Motif Pair Detection Technique

Umar Draz[1], Tariq Ali[1], Sana Yasin[1], Low Tan Jung[1], M. Ayaz Arshad [2]

[1]CS. Department, (CIIT) Sahiwal, Pakistan
[2]Communication & sensor centre, UOT, KSA

Received: August 7, 2017
Accepted: November 18, 2017

ABSTRACT

Motif detection is a very challenging task in the field of bioinformatics because a variety of motifs exists in the protein. A lot of algorithms and techniques are available to identify the motifs. These techniques detect the motifs through many supervised algorithms. Existing algorithms are not suitable for the simultaneous searching of motifs that contain various type of length in some gaped and un-gapped fashion. To find out the gapped motifs aretime-consuming task due to the blast of different type of possible combination that occurs by the consideration of the long gaps. In this paper, we introduce a new approach that detects the motifs of different lengths with gaps and without gaps through the unsupervised algorithm. The comparison is also done to analyze the computational time and the quality of the result.
KEYWORDS—Motif recognition; Motif Detection; Protein motifs; Gapped motifs; Protein sequences; Bioinformatics; DNA & RNA sequences

I. INTRODUCTION

Big Data is evolutionary field due to rapid generation of data day-by-day. Data is growing not only with every passing day but with every passing second, having different forms and categories. The social media are the biggest source of data nowadays. With the help of visualization, the key principle of the data is that provides information that can easily be interpreted, that's why the field of graph visualization has gained much more attention in research. A lot of algorithms and techniques are available to identify the motifs. Basically, motifs are the small connected entities that take the whole information about the complex network. Due to motifs, there is no need to read and explore the whole graph[1-3]. Motifs cover all the information inside of the graph. Through many techniques that are used for the detection of the motifs like supervised and unsupervised algorithms. Existing algorithms are not suitable for the simultaneous searching of motifs that contain various type of length due to a large number of sequences[4]. To find out the gapped motifs istime-consuming task due to the blast of different types of possible combination that occurs by the consideration of the long gaps. DNA has a number of bases that contain enough information about the patterns of matching and genes classification. Despite the significant efforts on the motif discovery, motif detection in DNA structure remains a difficult task for computer scientists and biologists. A lot of encouraging tools and algorithm is purposed in this field to make progress. Huge attempts have been done for the enlargement of the computational techniques for the identification of the sequence motifs in proteins and DNA sequences. DNA motifs are short persistent patterns that are recognized to have natural function. Two types of motive discovering algorithm exist in the literature namely supervised and unsupervised algorithms. The major benefits of using supervised algorithm are that they are easily found out motifs in well-disciplined manners. Unsupervised algorithms don't need any type of prior knowledge about the motif sequence. These find the novel sequence motif that does include unexpected high-frequency occurrence and similarities among different types of motifs [5]. Detection of wide variety of biological motifs is a very challenging task. In[2] three classes of biological motifs are introduced. The class First deals with the functional sites of biopolymers that contain short motifs cleavage and binding sites are one of the examples of the first class of biological motifs. Second class deals with the globular structural domains that often occur due to the divergent evolution. This class contains long proteins motifs that are associated with other motifs. The third class deals with the recurring motifs that often appear due to the evolutionary duplications. These classes can't be handled through single motif searching technique because these are too much complex and improbable. Both these two classes play part and parcel role to find out the suitable patterns under the supervised and unsupervised algorithms[6].

DNA and proteins have a large number of motifs that contain enough information about the patterns of whole sequence information. Protein-Protein interaction, DNA sequences and protein co-expression techniques, all have some repetition patterns in the form of common information. Due to motifs these can easily describe the well and defined information inside the networks[7]. The whole network consists of a large number of heterogeneous information about the network. It is not possible to

***Corresponding Author:** Umar Draz, CS. Department, (CIIT) Sahiwal, Pakistan.
Email: Sheikhumar520@gmail.com

understand the whole sequence that is present in the network in this case motifs play role and consist of concise information about the particular networks[8]. Due to motifs study, there is no need enough knowledge to understand the large sequence and protein-protein relationship about the network. In any complex network, the study of motifs plays our role to find out the correct sequence about the inter-connections of the module[1, 2, 9, 10].

Rest of this paper organized as related work is discussed in section II and problem statement of the motif detection technique is presents in section III. The proposed methodology against the problem statement is represented in section IV. At the end, the results are derived that is presented in section V. finally conclusion is discussed in section VI.

II. RELATED WORK

Recently the research on proficient motif mining of DNA and RNA sequences has gained attraction in all big data techniques. With respect to motif detection; big data play part and parcel role to ensure the presence of the motif in ant complex graph and structure. Today data is generating at a very high speed, for example, the DNA and its relation with RNA have a long chain structure that is not easy to understand[11]. To read about the whole graph and structure it's not possible to identify the type of information that belongs to the specific domain are present inside it. So, there to need to investigate to identify those patterns that repeat inside the graph in some well-disciplined manner. Motif theory; in this regard play our role because motif has a unique information about the network, structure and any type of graph. The repetition of information from top to bottom that causes the creation of motif. Many algorithm and techniques are present to identify this unique information in the form of motifs. Motif detection is a very difficult task to decide the motif has all information about the network. In case of DNA and RNA sequences and their relationship then he supervised algorithm is well suited [12]. Previously, recurrently emerging patterns called 'motifs' has what type of motif have related information, received much concentration in the field of Big Data and Bioinformatics. These Motifs are useful for various time-series and data mining tasks. The relation between proteins and DNA is a key motivating force. Motif detection is some of the highly focused topics among the researchers in the big data research community. Binding of DNA-RNA sites and the specially targeted proteins are two important moves to understand the concept of biological activities. A Lot of techniques that gives high-through put has recently purposed that try to enumerate the similarity between DNA motifs and RNA proteins. In spite of the strong achievement, these techniques have some limitations and go down towards the strict classification of motifs. As a result, need further critical analysis of DNA and proteins sequences to dig out useful and modifiable information from a stack of strident and raw data. In [13]Motif Mark algorithm is purposed to find out the regular motifs in DNA sequence. This algorithm based on the graph theory and machine learning that finds binding sites of DNA sequences. It also analyzes investigational data that is derived from universal proteins against two of the most precise motif detection methods. In [14] analysis is done on the whole genomes to find out the repetitive DNA sequences called non-B motifs. These motifs are capable to predict the noncanonical structure of the DNA and can autonomously report for deviation in mutation density. In [15] the detection of the motif is performed through FANMODE tool, the possible results are the detection of motifs but in limited numbers. All the related application of motif theory are discussed in [16-18] that shows that the motif is important when the strength of the network and its sequences are in large numbers. in this paper, we present the motif detection through the unsupervised approach in which different lengths of motif have to identify when gaps of proteins, between DNA and RNA sequences, are large. To identify the motifs without gaps between the proteins also a difficult task. At the end perform the comparison between gaped and un-gapped proteins that present inside the DNA and RNA sequences[1, 16].

III. PROBLEM STATEMENT

Due to a large number of DNA and RNA networks, it is difficult to predict the whole information about the network. As DNA and RNA; both have a large number of sequences chaining that consist of different types of bases. The repetition of bases controls the whole strategy about the network. Some sequences gape and some are not; due to this the protein-protein relation has a different meaning. If gapes and un-gapped sequences are present in the same network then to find out the whole information about the network is a big challenge. So there is need to investigate and proposed some type of approach that used to identify the sequences and their patterns. Gaped and un-gapped sequences are also needed to evaluate the information in the form of motifs. With the help of motifs detection technique; there is possible to differentiate the gaped and un-gapped sequences within the defined domain and range of the specific network.

IV. METHODOLOGY

In this paper, we introduce a new approach that detects the motifs of different lengths with gaps and without gaps through the unsupervised algorithm. The comparison is also done to analyze the computational time and the quality of the result. A large number of protein sequences are used as a data set namely the alternate ID, for example, NFYA, KIF7 and TBP. The datasetconsists of a collection of gaped and un-gapped sequences. We used similarity indexing to decrease the query time. Initially, the unique ID is assigned to every motif inside the graph. Another alternate ID is assigned to motifs have the same sequence number. If the repetition of motifs is randomly replaced then alternate ID is helpful to identify the unique ID name. The

alternative ID usually assigns the same tag while all remaining ID is assigned the unique numbers. This strategy helps us to identify the entire motif that is notrepeated in some graph or network. Sequence name is associated the total number of characters that present inside the network. To deal the 'strands' inside the DNA and RNA sequences two types of indicator are used like '+' and '-'. The '+' is used when the strand is present in DNA and RNA and '–' are used when there is no strand is present. In this paper; both types of strands are used to identify which type of sequence has a strand sequence or not. This is the uniqueness of our proposed approach to identify those sequences which strands or easily identify in which types of motifs. The starting and ending point of the associated sequences is also presented that where the sequences start and where it ends.

To make the proposed approach effective that are used to identify the motifs inside the network is work with the sequence starting and ending numbers. The main purpose of this approach is that those motifs that are detected under the range of defined sequence i-e; domain and range. Those motifs that are detected under the specified range are not being considered for next time and for next iterations. Two values are assigned namely 'p' and 'q' value. Both values are denoted the motif sequences under the RNA and DNA strands. If these two values are under the defined range at the same time thesequence is matched otherwise not.

To analyze the proposed approach and its attribute the Vienna Development Method-Specification Language (VDM-SL) toolboxis used. Formal methods are implemented in VDM-SL through various attributes. The use of VDM-SL window is to check the syntax and semantic scene of the proposed approach. Only the attributes checking is performed in this toolbox. The proper implementation of proposed approach through formal methods is our future work.

V. THE ALGORITHM OF MOTIF PAIR DETECTION

Here is the algorithm that describes the motif pair detection model and its alignment. Different variables are used in this algorithm that describes the movement of motifs inside the network.

Algorithm1. The pseudo code of the Motif Pair Detection Algorithm

INPUT: *Set of related Proteins of DNA and RNA sequences*

OUTPUT: A motif pair detection of DNA and RNA sequences in gapped and un-gapped sequences and their corresponding locations

1. *Take the set of sequence A,B and C as an input. (collection of gapped & un-gapped sequences)*
2. $\forall$ *sequence set* $S_N for N = 1,2, \ldots\ldots n$
3. $\boldsymbol{for\ All\ Inputs\ the\ Alt.ID\ is\ assigned\ as: NFYA, KFI7\ and\ TBP}$
4. *Strands: '+'= present; '-' = absent.*
5. *Initialize: Starting point & ending point where:*
   $\forall: \boldsymbol{S} < E$ $(\boldsymbol{place\ the\ marker\ where\ it\ start\ and\ end\ for\ next\ iteration})$
6. *P= "DNA"* ***&*** *q= RNA*
7. ***if***
8. ***{***
9. ***(S<< E&& p=q)*** **|| ('+' && '-')** *// sequence is matched*
   Else if {
   (S<< E&&p!=q) **|| ('+' || '-')** *// sequence is matched*
   Else if {
   (S >>E&&p!=q) **|| ('+' || '-')** *// sequence is matched*
   {
   else
   $\forall$ *Sequence does not matched*
   }
10. *Motif Discovery (sequences) // Link each sequence with motif*
11. $\forall\ \boldsymbol{Motif\ M_k\ for\ K = 1,2, \ldots\ldots n}$
12. ***If*** *Motif* **<***enrichment score* ***than***
13. ***gapped-motifs*** $(\boldsymbol{R_K})$ **||** ***Best Motifs***
14. *Motif Integration(motifs)*
15. *If (repetition of motif >> enrichment score)****then***
    Gapped-between motifs pair
    Else
    Un-gapped motifs pair
16. *This process continues until non-redundant motifs are discovered*
17. ***End While***
18. ***End if***
19. ***End For***

VI. RESULTS

The MEME Suite allows to discover motifs in collections of unaligned nucleotide or protein sequences and to perform a wide variety of other motif-based analyses. The MEME Suite supports motif-based analysis of DNA, RNA and protein sequences. It provides motif discovery algorithms using both probabilistic (MEME) and discrete models (MEME), which have complementary strengths. It also allows discovery of motifs with arbitrary insertions and deletions. The whole sequence is dividing into small chunks and motif is shown the sequence pattern. Here is the first case in which the motif is shown against a different number of bits through MEME toolbox.

The motifs are identified only where the matched sequence is present. Strands are present in those places where motif detection is easily identifying the p and q value within the defined range. In table 1, the alignment of all possible result is described.

Table 1: Motifs and its alignment

Motif ID	Alt ID	Sequence Name	Strand	Start	End	p-value	q-value	Matched Sequence
MA0060.1	NFYA	chr2	-	60221163	60221178	3.36e-09	0.00194	CTCGGCCAATCAGAGC
MA0060.1	NFYA	chr3	-	54858838	54858853	4.77e-09	0.00194	ATCAGCCAATCAGCGG
UP00093_1	Klf7_primary	chr2	+	172266018	172266033	1.06e-08	0.00496	GCGACCCCGCCCCTTT
UP00093_1	Klf7_primary	chr11	+	65981465	65981480	1.31e-08	0.00496	TTGACCCCGCCCCTCA
UP00020_1	Atf1_primary	chr3	+	65470135	65470150	2.63e-08	0.0222	GCTGTGACGTCACCGC
MA0060.1	NFYA	chrX	+	51925276	51925291	3.55e-08	0.00583	TTCAGCCAATCAGCGC
MA0060.1	NFYA	chrX	+	52016291	52016306	3.55e-08	0.00583	TTCAGCCAATCAGCGC
UP00093_1	Klf7_primary	chr1	+	93302476	93302491	3.87e-08	0.00898	TCGGCCCCGCCCCTCC
MA0060.1	NFYA	chr15	-	94234692	94234707	4.18e-08	0.00583	GTCGACCAATCAGCGG
MA0060.1	NFYA	chr4	-	106584009	106584024	4.83e-08	0.00583	GGCGGCCAATCGGCGC
MA0060.1	NFYA	chr19	-	56896602	56896617	5.46e-08	0.00583	ACGAGCCAATCAGCGC
MA0060.1	NFYA	chr15	-	103164696	103164711	5.72e-08	0.00583	ATCAGCCAATCAGAGT
UP00093_1	Klf7_primary	chr1	-	44158990	44159005	6.12e-08	0.00898	TCGGCCACGCCCCTCG
MA0060.1	NFYA	chr14	-	79987508	79987523	6.79e-08	0.00616	CGTAGCCAATCAGCGG
UP00093_1	Klf7_primary	chr9	-	54434205	54434220	6.99e-08	0.00898	GTGGCCCCGCCCCTAG
UP00093_1	Klf7_primary	chr14	-	79987408	79987423	7.12e-08	0.00898	CCGGCCCCGCCCCTAC
UP00020_1	Atf1_primary	chr2	-	49474775	49474790	8.05e-08	0.0256	ACGGTGACGTCACTGC
UP00093_1	Klf7_primary	chr7	-	25267019	25267034	9.06e-08	0.00979	TCGACCCCGCCCCCGA

A. Case 1: Basic motifs in DNA sequence

The DNA sequence is under study in which total motifs are shown as an individual category. The position 1 and 4 is unique where the information of the network is repeated against a different number of bits which means in every place where the distance of one information is repeatedpresent in 4 places. Figure 1 shows that the motifs detection at 1, 4, 8 and so on the present.

Figure 1: Basic motif detection at DNA un-gapped sequence

B. Case2: Basic motifs in RNA sequence

As RNA is the extension of DNA; all the information of DNA also has some repeating pattern in RNA so the chain of RNA is consist of DNA sequence and its patterns. Figure 2 shows that RNA un-gapped sequence where the information is suited tolinear chain patterns. The very first motif detect in at the 9th position which means a lot of information is skipped at DNA sequence and only into account the RNA sequence. The 6thposition contains the DNA sequence in the form of motifs but in RNA the 9th, 18th, 27th and so on position is associated.

Figure 2: Basic motif detection at DNA un-gapped sequence

C. ***Case 3: combined DNA and RNA sequenece (gapped)***

Figure 3 shows that the combined gapped sequence network in which both types of the sequence is present in DNA and RNA sequence. In some places where the motif detects is bold and large. By and large, the network consists of two motifs that are associated with DNA and RNA sequence.

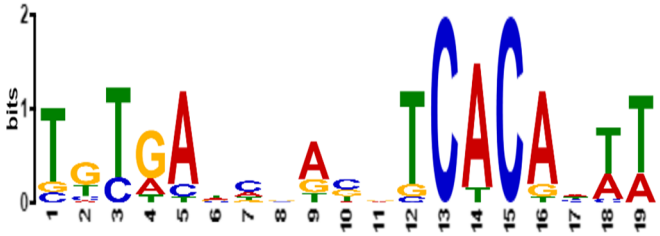

Figure 3: Basic motif detection at DNA & RNA gapped sequence

D. ***Case 4: combined DNA and RNA sequenece (un-gapped)***

After analysis the overall separate sequence of DNA and RNA there is the graph in which both DNA and RNA sequence is present in un-gapped fashion. Initially, the first 50 results of motifs are shown. Due to a large number of sequence, it is not possible to show all the results. Remaining results also follow' the same mechanism that is shown in figure 4.

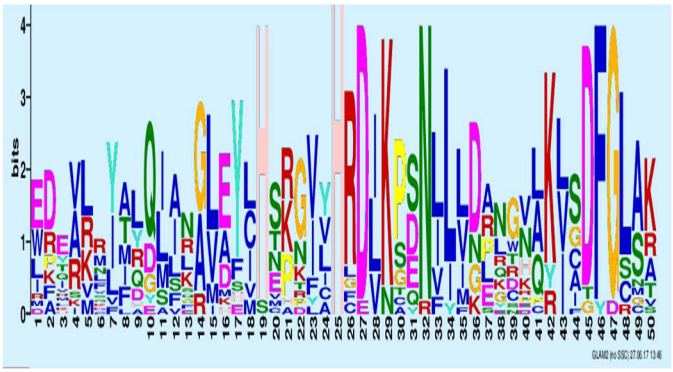

Figure 4: Basic motif detection at DNA & RNA un-gapped sequence

E. ***Case 5: location of motifs***

In this paper, some of few graphs are shown but overall in the data set the possible location of motifs are shown in figure 5. Red bars have associated those motifs that re-identify under the un-gapped proteins. The green bars are shown those motifs that are associated the gapped proteins. Other motifs location is those in which not decide either it is under the gapped position or un-gapped positions. So it can be acceptable on the borderline of gapped and un-gapped proteins. Figure 5 shows that the location of motifs in the whole dataset.

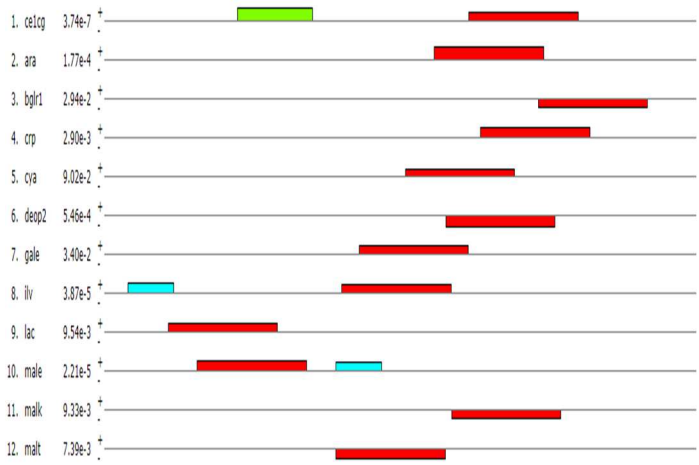

Figure 5: Location of motifs inside the data set

F. Case 6: position of best motif in dataset sequences

After observing the location of motifs, there is need to extract the best motifs inside the network with respect to enrichment score. In the algorithm, if the identified motifs are greater than enrichment score then it is declared as best motifs. Figure 6 shows that all the motifs of TBP have a high score of probability as compared to KFI7 and NFYA. The associated 'p' values are also shown in the figure on the behalf of strand value. The other two like KFI7 and NFYA are gradually decreased due to low 'p' values.

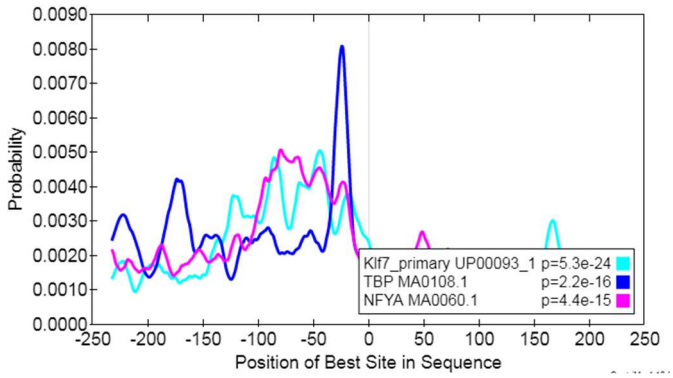

Figure 6: Detection of best motif pair inside the network

VII. ALGORITHM ANALYSIS THROUGH VDM-SL TOOL BOX

To check the correctness, consistency and its integration of the proposed algorithm the VDM-SL toolboxwindow is used. VDM-SL provides the platform in this regard that ensures all the correctness of proposed algorithm [19]. The VDM-SLtoolboxprovides support to check all the related invariants in a different mode. To check all the composite objects, state, function, and operations the VDM-SL checking window provide the syntax check, type check, pretty and integrity check. All the simulation work does not provide the correctness of the model, technique, and algorithm but formal methods are enough flexible and give the proof of the proposed technique or algorithm that is design inside the tool, box. Table 2 describes the verification of the model against all related possible function. It ensures that the proposed algorithm specification is correctly verified and validated.

Table 2: Model analysis of proposed algorithm through VDM-SL

Composite object, State, Function, and Operations	*Syntax Check*	*Type Check*	*Pretty Check*	*Integrity Check*
Object	Yes	Yes	Yes	Yes
Abstract Motif	Yes	Yes	Yes	Yes
DNA sequence	Yes	Yes	Yes	Yes
RNA sequence	Yes	Yes	Yes	Yes
Gapped sequence	Yes	Yes	Yes	Yes
Un-gapped sequence	Yes	Yes	Yes	-
Functions	Yes	Yes	Yes	-
Diagrams & analysis	Yes	Yes	Yes	-
Alternative ID	Yes	Yes	Yes	-
Motif ID	Yes	Yes	Yes	-
Protein sequence	Yes	Yes	Yes	-
Protein-protein sequence	Yes	Yes	Yes	Yes
Values/attributes	Yes	Yes	Yes	Yes
Strands	Yes	Yes	Yes	Yes
Motif pair detection	Yes	Yes	Yes	Yes
Location of motifs	Yes	Yes	Yes	Yes
Execution	Yes	Yes	Yes	Yes
Pre/post conditions	Yes	Yes	Yes	-
Validation and verification	Yes	Yes	Yes	-

VIII. CONCLUSION

In this paper, we introduce a motif pair detection algorithm is proposed for identify the motifs in gapped and un-gapped protein sequence. To find out the gapped motifs aretime-consuming task due to the blast of different type of possible combination that occurs by the consideration of the long gaps. The MEME Suite allows to discover motifs in collections of unaligned nucleotide or protein sequences and to perform a wide variety of other motif-based analyses. To analyze the proposed approach and its attribute the Vienna Development Method-Specification Language (VDM-SL) toolbox is used. DNA and proteins have a large number of motifs that contain enough information about the patterns of whole sequence information. Protein-Protein interaction, DNA sequences and protein co-expression techniques, all have some repetition patterns in the form of common information. Due to motifs pair detection algorithms these can easily describe well and defined information inside the networks. Result has been shows that the location of motifs inside the protein network is well suited for the long, gapped and un-gapped sequence. The future horizon is about to formalize these entire motif with the help of formal methods implementation.

IX. REFERENCES

1. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-827.
2. Chen, Y. and Y. Chen, *An Efficient Sampling Algorithm for Network Motif Detection.* Journal of Computational and Graphical Statistics, 2017(just-accepted).
3. Atay, Y. and H. Kodaz, *Network motif detection in PPI networks and effect of R parameter on system performance.* International Journal of Applied Mathematics, Electronics and Computers, 2016. **4**(3): p. 78-82.
4. Moses, A.M., D.Y. Chiang, and M.B. Eisen, *Phylogenetic motif detection by expectation-maximization on evolutionary mixtures*, in *Biocomputing 2004*. 2003, World Scientific. p. 324-335.
5. Saxton, W. and J. Frank, *Motif detection in quantum noise-limited electron micrographs by cross-correlation.* Ultramicroscopy, 1976. **2**: p. 219-227.
6. Andrienko, G., N. Andrienko, and S. Wrobel, *Visual analytics tools for analysis of movement data.* ACM SIGKDD Explorations Newsletter, 2007. **9**(2): p. 38-46.
7. Keim, D.A., et al., *Visual analytics: Scope and challenges*, in *Visual data mining*. 2008, Springer. p. 76-90.

8. Marchal, K., et al., *Genome-specific higher-order background models to improve motif detection.* Trends in microbiology, 2003. **11**(2): p. 61-66.

9. MacEachren, A.M., *Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier*, in *Spatial Data Handling in Big Data Era*. 2017, Springer. p. 139-155.

10. Munshi, A.A. and Y.A. Mohamed. *Cloud-based visual analytics for smart grids big data*. in *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*. 2016. IEEE.

11. Steed, C.A., et al., *Big data visual analytics for exploratory earth system simulation analysis.* Computers & Geosciences, 2013. **61**: p. 71-82.

12. Thomas, J.J. and K.A. Cook, *A visual analytics agenda.* IEEE computer graphics and applications, 2006. **26**(1): p. 10-13.

13. Vahdatpour, A., N. Amini, and M. Sarrafzadeh. *Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series*. in *IJCAI*. 2009.

14. Wernicke, S. and F. Rasche, *FANMOD: a tool for fast network motif detection.* Bioinformatics, 2006. **22**(9): p. 1152-1153.

15. Wong, E., et al., *Biological network motif detection: principles and practice.* Briefings in bioinformatics, 2011. **13**(2): p. 202-215.

16. Schreiber, F. and H. Schwöbbermeyer, *MAVisto: a tool for the exploration of network motifs.* Bioinformatics, 2005. **21**(17): p. 3572-3574.

17. Prytuliak, R., et al., *HH-MOTiF: de novo detection of short linear motifs in proteins by Hidden Markov Model comparisons.* Nucleic Acids Research, 2017.

18. Prokop, J.W., *Functional Motif Discovery in Signaling Biology Using a Deep Sequence-to-Structure-to-Function Analysis.* The FASEB Journal, 2016. **30**(1 Supplement): p. 969.31-969.31.

19. Afzaal, H., N.A. Zafar, and F. Alhumaidan, *Hybrid subnet-based node failure recovery formal procedure in wireless sensor and actor networks.* International Journal of Distributed Sensor Networks, 2017. **13**(4): p. 1550147717704417.