Formal Analysis of Coherent Non-Redundant Partition-based Motif Detection Algorithm for Data Visual Analytics

Sana Yasin[1], Tariq Ali[1], Umar Draz[1], Low Tan Jung[2], M. Ayaz Arshad[3]

[1]CS. Department, (CIIT)Sahiwal, Pakistan
[2]CIS. Department, UTP, Malaysia
[3]Communication & sensor center, UOT, KSA

ABSTRACT

Recently the interest in motif detection increased rapidly in the field of data visual analytics and bioinformatics. The significant improvement of motif detection in data visual analytics, but still many challenges are needed to overcome the issues for visual analytics applications. In most off the publishedwork, the focus is on motif detection as a solution of data visual analytics the role of motif detections and its functional non-functional requirements in the field of the large and complex network are still ignored. This paper introduces a proactive approach with the help of formal specification and analysis of partitioning motif detections algorithm (PMDA). In data visual analytics (DVA), we present a new model in data visual analytics with the use of VDM-SL to describe the formal specification of this algorithm to separate the non-redundant motifs from sparsely connected protein sequences. Various invariants are usedto validate the algorithm in all pre-post conditions confirm the correctness working of the proposed algorithm. VDM-SL is the formal specification language that is used for the implementation of a software system on detail level examination. The PMD algorithm specification is implemented, verified and validated with respect to syntactically and semantically through the VDM-SL toolbox.

KEYWORDS: Motif detection; DVA; PMDA; Verification & Validation; formal specification

I. INTRODUCTION

Data is growing rapidly in visual analytics the analysis of large grasps is playing a well-known role in various fields of visual analysis. Success full visual analysis of graph required appropriate visual presentation. It has captured the attention of public and media and researcher of the particular related field. Data visual analytics[1, 2] has successfullyreplaced the data analytics. A lot of application is present like visual semantic, a visual representation of information, visual decision trees, visual time critical model, concretive and perceptual sciences, and all interdisciplinary field of research due to some limitation in data analytic recently data visual analytics are introduced with some unique technique called motif detections. Basically, themotifis small connected sub-networks that are present inside the network and have to significantly display the higher frequency that would be expected for a random network. Motif has recently gathered attention as a concept to uncover the structure design and principles of complex biological networks [3, 4]. All the biological networks consist of proteins-proteins interaction PPI, DNA sequences, co-expression of genes, the basis of genes, human attraction pattern, DNA graph theory and neural networks. Many tools are used for detection the motifs in different networks. FANMODE is the tool for motif detection that uses for fast motif detections algorithm. In bigger and complex network to abstract the hidden information,FANMODE[5] is used many biological network use FANMODE due to its Rapid response for uncovering the structure design principle and biological networks[6]. The analytics of network motifs has to lead to very interesting result about the area of protein-protein interaction, hierarchal network decomposition, and analysis of temporal gene patterns. To find the motif in the network consists of three computationally sub-tasklike find out all the occurrence of subgraph as an input and in what number, secondly with of these graphs are isomorphic, and at the last, all subgraphexists in some random session with a higher number of repetition patterns[7].

The whole time that isused to spend to find the motif inside the network has thelarge contribution of the second subtask in which find out the topological equivalent isomorphic graph. In order to find the subgraph, the nonuniform network is considered. However, DVA also present in nonuniform fashion there for it is easy to detect the motif of these types of analytics. A topological layout is considered to make the proposed algorithm. The nature of the network is analogous and relative to the graph theory. All the vertices and edges are shows the relationship of data among each other and one another. For example, the vertices represent the data nodes which are assumed as data visual nodes and the edges perform as the relation between the data visual nodes. To find out the motif inside the network is assumed that the repetition of nodes and edges is present in the whole network.

The data visual analytics are yet unable to overcome many functional and operational challenges, particularly for critical and large-scale applications. A major reason is that a focus of today research is basically the performance evaluation of data visual

---

*Corresponding Author: Sana Yasin, CS. Department (CIIT)Sahiwal, Pakistan. Email: sanayaseen42@yahoo.com

analytics with motif detections are simulation based. The simulationdoes not confirm the correctness approach which is very important In motif detections and data visual analytical system[4]. However, simulation ensures the qualitative performance of the system. Recently mathematical techniques like formal methods have been used for verification the function attributes of any system and in short the correctness of the approach. The basic function of the formal method transforms the informal to the formal requirement in terms of property and specifications. That's why PDMA is employed in this work to model the data visual analytics as a dynamic approach that transforms the associated formal specification using formal methods. In this work, formal method base technique means VDM-SL is employed as a formal specification language for the implementation of PMDA. The PMD algorithm is based on partitioning approach that first partition the whole data into two submodules with an equal number of repetition patterns. With the help of proposed algorithm motif is find out in each module and extract the hidden information in the form of motifs with a defined wide range of analytics[8, 9].

Up to our best knowledge this kind of work in the field of DVA in firstly described in the formal specification by using VDM-SL. The composite object, sets, operations and some static variables are used to describe the dynamic behaviors of the proposed algorithm. It is to be noted that every aspect of any kind of system can be formalized define with the help of VDM-SL notations.

The rest of this paper is organized as: section II discusses the critical related work and formal specification of the proposed PMDA represented in section III. High-level pseudo code and algorithm are described in section IV. The correctness of the proposed model and its analysis is discussed in section V. At the end conclusion are discussed in section IV.

II. RELATED WORK

 The major challenge is to identify the repatriation patterns inside the network, graph or any type of interaction. Many characteristics patterns have enough information of the graph and network that resemble the whole hidden patterns. For example, the protein-protein interaction and DNA sequences have a large graph due to its large sequences. These large sequences do not easy to understand and difficult to translate the hidden information[10]. Many algorithms and techniques are in working stage [11-13]that elaborates the sequences in the form of Motif. Motif has enough information that is easy to understand the whole information and status of the graph. Motif detection is a complete phenomenon in which small cluster networks have large and concise information. To find the motif detection inside the network and any data visual analytics technique is much difficult. In [14] using the technique of motifs the system of active visual analysis for large graph aggregation is performed. To draw the useful insight relationship between graph motif detection technique plays our role. To enhance the graph layout by a graph analysis stage is a difficult task. It might be difficult to understand the hidden structures of the complex network, with motif graph analysis visualization the motif is easily detected the whole scenario that converts the theme of the complex network. The simulation result is shown the motif filtering is the best example to withdraw the complex information from inside the network. In [10] all the major principle and practices about the motif detection is discussed. To find the motif in the biological network there are three fundamental motif detection principles; for example, to find out the motif hierarchy, sub0graph frequency, graph isomerism, significance motif frequency, random graph, association graph and visualization of the graphs. In [15]motif algorithms are discussed for large expression of graphs. A large expression of graph usually presents the protein-protein interaction, co-expression of the genes, DNA sequences and neural networks. To deal these types of field-motif detection is so much beneficial for finding the correct graph patterns about the several or particular disease. All the application of the medical field is discussed in [14]. The working motif centric digital algorithm is presented in which all the function of motifs can be described in the digits form. To calculate about the how many patterns exist inside the network, to find out the how many motifs may be present, gene patterns, DNA bases structures, and the neural information is calculated through this algorithm. The possible graphical results shown are [16]. All the simulation is done with the help of NeMo finder and Mavisto algorithm[7, 17]. Through this technique easily built up the pattern growth tree in which all the particulate and relative information are present as leaf nodes. These sizes of motif also have a unique role to find out the best relationship between the motifs detection.

The Size of motifs is in the form of different patterns where the amount of data is easily available[18]. The neural networks, bioinformatics network where genes repeat our sequences and have repetition patterns. Theses repetition patterns further divide the isomerism scenario where one motif have enough information for one cluster. That is the reason for the large width range there in only need to find the only 2 to 3 motifs. In any repetition case study there is enough to detect the motif with only large range but have a small number of motifs information. The tabular form of the result is discussed[19]. All the challenges of visual data analytics can be removed with motif finding. In the dark and semantic web, there is a large scope of motif finding because all the information that consists of the dark web needs to extract with motifs.

A lot of application is present like information and geospatial analytics, scientific and statistical analytics, knowledge discovery and interaction of informatics, science and bioinformatics, data management, distributed management, presentation of a graph, production of graph and dissemination, cognitive and perceptual science are the major field of visual analytics. The details survey about this field is presented in [7]. The solution of financial data with the visual analytics technique is discussed in [20] in which the related current information is stored in the form of visual analytics. Today the age of big data, the handling of data is becoming the hot issue in which need to separate all the related and unrelated data. All the data-intensive application, challenges, application and related techniques are discussed in [16]. Due to the era of big data, there is enhancing the scope of motif

technique because the entire scene where the amount of data is present, the need of motifs is also increased. Toward the concept of uncovering structural design, the principle of network detection with motif technique also encourage. Among PPI networks, the motif detection technique is so much impressed due to the large and complex scenario of PPI networks. In this paper, the PPI networks are under observation in which find out the motif with the respective wide inclusive range. For experiments, the range of PPI networks can be considered with the zero background knowledge. At the end, the motif tells about the occurrence frequency and the respective range of data that present in PPI networks.[19, 21-23].

   Most of the work in this domain is in simulation-based that have no ensured about either the simulation criteria is well suited to the problem. In other words, the proof of correctness is missing in all type of simulation-based work. So; therefore the formal method; due to its abstract and details level explanation of the particular work is used for this proposed approach. The model analysis gives the correctness of the work and all type of syntax and semantic is easily checked in VDM-SL Toolbox. Today the trend of a formal method is increased due to its vast application in every field of life, particularly in acritical system. Many of the work are already formalized in VDM-SL toolbox like formal verification of hybrid connectivity model are discussed in [24, 25]. Towards the formalism of earthquake detection and flood detection with the help of VDM-SL is presented in [25, 26]. To find the motif in its correct operation and alignment inside the network the formal methods play part and parcel role. In this paper; the formal specification of PMDA is used to convert the informal hierarchy of motif detection to the formal layout. This is the first work is done in this domain with the help of formalspecification as up to our best knowledge. At the end, the model analysis is also shown that give the true spirit of this proposed work.

III. MOTIF DETECTION ALGORITHM

This section represents the High-levelpseudo-codefor the motif detection in the structure of the protein that is frequently available in the sequences. It takes proteins sequence as an input and detects the redundant and non-redundant motifs in these sequences respectively. The nature of the algorithm is proactive so that the whole sequence is partition into two submodules.
Algorithm1. The pseudo code of **Partitioning Motif Detection Algorithm (PMDA)**

INPUT: *Set of related Proteins sequences*
OUTPUT: *A coherent non-redundant set of motifs per initial input group*
1. *Take the set of sequence A,B and C as an input.*
2. $\forall$ *sequence set$S_N$forN = 1,2, … … n*
3. $\forall$ $\left(SET\ OF\ SEQUENCES\ (X,Y) \right) \ni M \cong Sn \ \leftrightarrow M(x|y)$
4. *Partitioning sequence: X,Y/2=x || y*
5. **If** $(x \in (X,Y)\&\&\ (Y \in (X,Y)$ **Motif** M_k **for** $N = 1,2, … … n$
6. *Motif Discovery (sequences)   // Link each sequence with motif*
7. **Else**
8. **Motif doesn't discover**
9. $\forall$ *Motif* M_k *for* $K = 1,2, … … n$
10. **Redundant Motifs** *(motifs);*
11. $\forall$*Redundant motif* R_k*forK = 1,2, … … n*
12. *While* R_k*<n Do   where* R_k*=Redundant Motif, Do =Distribution motif*
13. *Split tree (* $R_{k \rightarrow M\ (x|y)}$ *)*
14. **If** *Motif <enrichment score* **than**
15. **Refined-motifs (R$_{fK}$)**
16. *Motif Integration(motifs)*
17. *This process continues until non-redundant motifs are discovered*
18. **End While**
19. **End if**
20. **End For**

IV. FORMAL SPECIFICATION BY USING VIM-SL

This section presents the formal specification of detection of redundant and non-redundant motifs in proteins structure. Analysis of different motifs in the sequences is done to uncover the structure design and principles of complex biological networks. Various invariants are used to validate the algorithm in all pre-post conditions that confirm the correctness working of proposed

approach that is used to detect the motifs. This specification includes different types of composite objectsand some efficient, proactive operations that detect the best motifs successfully.

```
types
Motif=token;
sequence=token;
non-Redundant motifs=token;
Redundant motif=token;
Link_Type = <Link> | <Not_Link>
Type::motif:Motif
motif:Motif
inv mk_link(motif1,motif2) == motif1 <> motif2;
values
LIMIT:nat=600000;

state PROTEINS of
sequence:set of sequence
sequence _link:set of links
motif_link:set of links
inv mk_PROTEINS(sequence)== card sequence<=LIMIT
inv mk_PROTEINS(sequence,motif_links)
== forall links in set motif_link& link.motif1 in set motif and
link.motif2 in set motifs and forall moif in set sequence
init mk_PROTEINS(sequence)==sequence={}
end
```

Total sequence function returns the total number of sequences that are available for the detection of motifs. This function is very helpful for the motif detection because accurate sequence strength makes the detection process easy.

```
TotalSequences()total:nat
ext rd sequence: set of Motifs
pre true
post  total= card sequence;
```

Remove motifs function facilitate the user to remove some motif from the sequence of the protein. This function is very helpful to find out the best motifs from the sequence because it removes the unnecessary motifs from the sequences. Search motif function provides facility to the user to search out any motif that they want. To search the motifs from the sequence *motif id* is required to give input. This motif search function returns the motifs that are id is matched with the given id.

```
Remove motif (MotifIn:Motif)
ext wr sequence: set of Motifs
post sequence = sequence~ \ {MotifIn}
post dying motifblock= dying motifblock~ union {MotifIn};

Search_Motifs(motifIn:Motifs)query:bool
ext rd sequence: set of sequence
ext rd Motifs: set of Motifs
pre true
post query <=>motifIn in set motif;
```

Redundant motifs detection functions readout all the motifs and checks out the redundant motifs in the sequence. This function gives the Boolean results. It only tells either redundant motif is available in the sequence or not.
Non-redundant motifs work like the redundant motif detection function with a minor difference. Non-redundant motifs detection function returns the Boolean results against the non-redundant motifs and checks its availability in the sequence.

```
Redundant motifs_Detection(motifIn:Motif)query:bool
ext rd Redundant_motifblock: set of Motifs
pre true
post query <=>motifIn in set Redundant_motifblock;

UnRedundant motifs_Detection(motifIn:Motif)query:bool
ext rd unRedundant_motifblock: set of Motifs
pre true
post query <=>motifIn in set unRedundant_motifblock;
```

These two functions create the source and destination motifs in the structure of the protein. Create source motif function take *motif id* as an input and declare it as a source node. Create *destination motif* function take *motif id* as input and declare it as a destination motif. Both these functions are very helpful to establish communication within the motifs. Through these two functions, theseparate biological network can be established that will be able to do communication between the motifs only. The message or control that will be sent by the source motif will be received at the destination motifs.

```
create_source_motif(MotifId:Motif)
ext wr sequence: set of Motifs
pre NetIn not in set sequence and card sequence <1
post sequence= sequence~ union {MotifIn};
post source= source~ union {MotifIn};

create_destinationmotif(MotifId:Motif)
ext wr sequence: set of Motifs
pre NetIn not in set sequence and card sequence <1
post sequence= sequence~ union {MotifIn};
post destination= destination~ union {MotifIn};
```

This function returns the total number of motifs in the sequence that is very helpful to establish communication within the motifs. It helps the user to define the source and destination motifs in the sequence. The user can declare the first detected motif as a source motif and last detected motif as destination motifs.

```
Total_Motif ()output:nat
ext rd Motifs: set of Motifs
pre true
post output=card(Motifs);

Find_Best_Motifs()output:set of Motifs
ext rd Motifs: set of Motifs
pre true
post total= Best_Motif;
```

Motifs *links function* creates the link between the motifs. This function firstly takes the source motif and its neighboring motifs as an input and creates the link between them. It continues the process of linking till the last destination motif is connected.

```
Motifs_links(source_motif:motif,destination_motifs:motif)


create_link::source_motif:motif
destination_motif:motif;

inv mk_link(source_motif,destination_motif) == source_motif<>destination_motifs
```
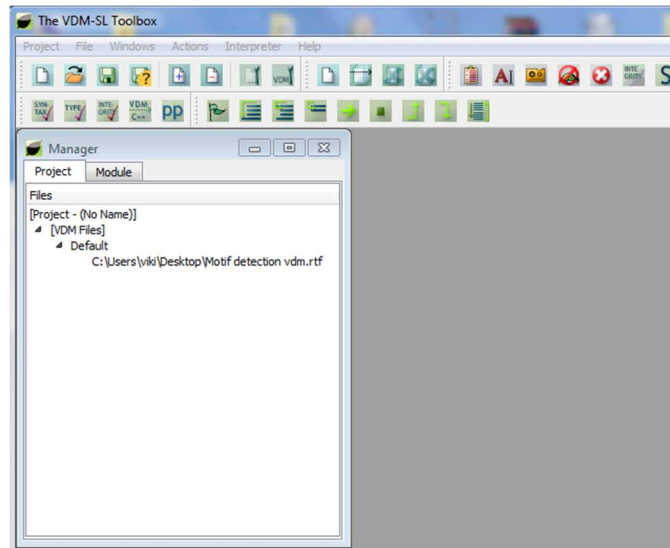
V.   MODEL ANALYSIS

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 [27]. 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 PMDA specification is correctly verified and validated. Table 1 describes the proposed algorithm and its state, function, structures and its operation is well organized in the VDM-SL toolbox. It has been observed that from figure 1 and 2 is the proof of correctness that the proposed algorithm is successfully implemented without any syntax and semantic error.

TABLE 1: ANALYSIS OF STATE, FUNCTIONS, AND OPERATIONS

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
Protein sequence	Yes	Yes	Yes	Yes
Partitioning 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	-
Partition and its attributes	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	-

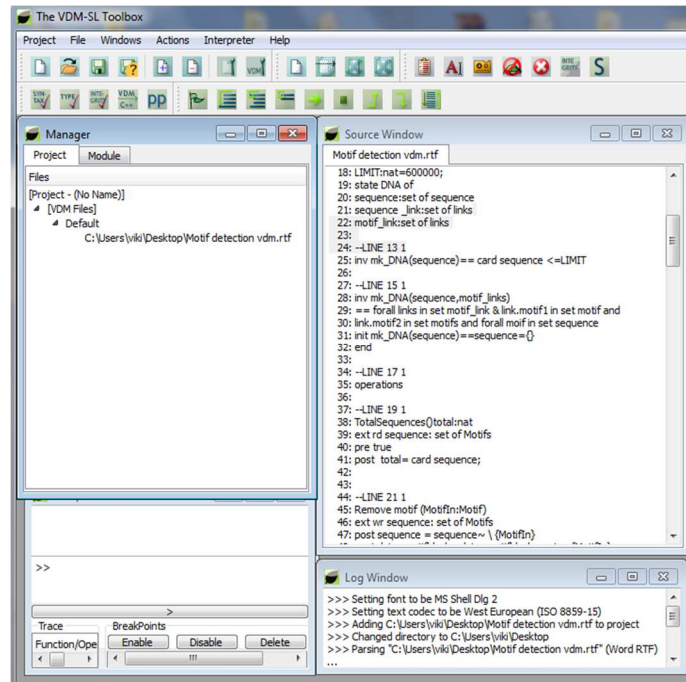***Figure 1: Loading of the project***

Figure 2: Snapshot of model analysis

VI. CONCLUSION

Data is growing rapidly in visual analytics the analysis of large grasps is playing a well-known role in various fields of visual analysis. In most off the publishedwork, the focus is on motif detection as a solution of data visual analytics the role of motif detections and its functional non-functional requirements in the field of the large and complex network are still ignored. This paper introduces a proactive approach with the help of formal specification and analysis of partitioning motif detections algorithm (PMDA). In data visual analytics (DVA), we present a new model in data visual analytics with the use of VDM-SL to describe the formal specification of this algorithm to separate the non-redundant motifs from sparsely connected protein sequences. Various invariants are used to validate the algorithm in al pre-post conditions confirm the correctness working of the proposed algorithm. VDM-SL is the formal specification language that is used for the implementation of a software system on detail level examination.

Up to our best knowledge this kind of work in the field of DVA in firstly described in the formal specification by using VDM-SL. The proposed PMD algorithm is successfully separate the non-redundant motifs from redundant one in a large sparsely connected proteins sequences. With little bit modification, this algorithm is implemented in any type of scenario where the separations of two modules are required. Through formal methods, it is possible to make the type of algorithm that helps us in the field of big data and data visual analytics technique.

VII.   REFERENCES

1. Keim, D.A., et al., *Visual analytics: Scope and challenges*, in *Visual data mining*. 2008, Springer. p. 76-90.

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

3. Saxton, W. and J. Frank, *Motif detection in quantum noise-limited electron micrographs by cross-correlation.* Ultramicroscopy, 1976. **2**: p. 219-227.

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

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

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

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

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

10. Kayastha, S., et al., *Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps.* Journal of Chemical Information and Modeling, 2017. **57**(5): p. 1218-1232.

11. 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.

12. 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.

13. Nguyen, Q.V., et al., *Special issue on Big Data Visual Analytics*. 2017, Elsevier.

14. Zhou, Z., et al. *The Motif Detection of Turning Points in Stock Time Series*. in *International Conference on Mechatronics and Intelligent Robotics*. 2017. Springer.

15. Viégas, F.B. and M. Wattenberg. *Artistic data visualization: Beyond visual analytics*. in *International Conference on Online Communities and Social Computing*. 2007. Springer.

16. Chen, Y. and Y. Chen, *An Efficient Sampling Algorithm for Network Motif Detection.* Journal of Computational and Graphical Statistics, 2017(just-accepted).

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

18. Thomas-Chollier, M., et al., *RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.* Nucleic acids research, 2011. **40**(4): p. e31-e31.

19. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-827.

20. 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.

21. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proceedings of the National Academy of Sciences, 2003. **100**(21): p. 11980-11985.

22. 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.

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

24. Imran, M. and N.A. Zafar, *Formal specification and validation of a hybrid connectivity restoration algorithm for wireless sensor and actor networks.* Sensors, 2012. **12**(9): p. 11754-11781.

25. Zafar, N.A., *Formal specification and validation of railway network components using Z notation.* IET software, 2009. **3**(4): p. 312-320.

26. Afzaal, H., et al. *Battlefield surveillance formalism using WSANs*. in *Electrical Engineering (ICEE), 2017 International Conference on*. 2017. IEEE.

27. 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.