Journal of
Data Mining in Genomics & Proteomics

Nikam, J Data Mining Genomics & Proteomics
2016, 7:3
DOI: 10.4172/2153-0602.1000200

Research Article Open Access

Suffix Graph - An Efficient Approach for Network Motif Mining

Rahul Nikam* **and Usha Chauhan**

Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal-462003, India

***Corresponding author:** Rahul Nikam, Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal-462003, India, E-mail: rahul.nikam1907@gmail.com

Received date: May 19, 2016; **Accepted date:** Jun 02, 2016; **Published date:** June 09, 2016

Abstract

Network motif is a pattern of inter-connections occurring in complex network in numbers that are significantly higher than those in similar randomized network. The basic premise of finding network motifs lie in the ability to compute the frequency of the subgraphs. In order to discover network motif, one has to compute a subgraph census on the original network that calculates the frequency of all the subgraphs of certain type. Then there is a need to compute the frequency of a set of subgraphs on the randomized similar network. The bottleneck of the entire motif discovery process is therefore to compute the subgraph frequencies and this is the core computational problem. The proposed work is to present the Suffix-Graph, a data structure that store graphs efficiently and to design an algorithm to retrieve subgraph efficiently that detects network motifs and apply them to transcriptional interactions in *Escherichia coli*.

Keywords: Network motif; Graph theory; Suffix array; Data structure; Transcriptional interactions

Introduction

Many natural structures are intuitively represented by complex networks, which have received increased attention by the researchers in recent years [1-10]. In 2002, Milo et al. [5] noted that some sub networks appeared with a much higher frequency in the studied networks than it would be expected in similar randomized networks, i.e., with the same degree sequence. These overrepresented topological patterns were named as network motif.

Network motifs are important in the analysis of networks from several domains, particularly in the biological domain [3]. For example, it has been demonstrated that they can have functional significance in transcriptional regulatory networks [6] or protein-protein interaction networks [1]. They have been applied in other biological areas, like brain networks [7] or food webs [8] and they are also significant in networks from other domains, like electronics circuits [9] or software architecture [10]. In this paper we do not try to position ourselves in that conceptual discussion Instead we focus our attention on the algorithmic aspect of finding network motifs, in order to obtain more efficient methods that can bring new insight into this topic.

Finding network motifs is a computationally hard task, since at its core, we are basically dealing with the graph isomorphism problem. Current methods count the frequency of all sub graphs of both original network and random networks. The problem is that the execution time increases exponentially when we increase the size of motif. Sampling has been introduced by Kashtan et al. [10] to improve accuracy in terms of time, but still it is very time consuming.

In this paper we are presenting a new approach to discover network motif efficiently in terms of time with the help of suffix tree and suffix graph.

Preliminaries

Network

A network can be modelled as a graph G composed of set of V (G) of vertices or we can say nodes and the set of E (G) of edges or connections. The size of a graph is the number of vertices in the graph.

A sub graph G' is a graph in which V' (G') is a subset of V (G) and E' (G') is a subset of E (G).

Network motifs

We start by formally defining the network motif for simplicity we will use motif in place of network motif. The motifs concept has several variations. In this paper we are concentrating on the standard definition established by Milo et al.

Basically motifs are patterns of interconnections occurring in a complex network are significantly higher frequency then in similar random networks.

Related work

We are aware of seven different main algorithms strategies for finding motifs mfinder, ESU, FPF, Grochow, G-Tries, Kavosh and MODA.

Conceptually, these strategies are divided into three main categories:

- Network-centric methods, that produce a sub graph census by enumerating all subgraphs of size k of original graph (mfinder, FPF, ESU and Kavosh)
- Single subgraph methods, those excel in computing the frequency of pre-defined single individual k-subgraph (grochow and MODA)
- Subgraph-set method, that specialize in counting the frequency of set of pre-defined k-subgraphs,that do not necessarily have to be all possible k-subgraphs (G-Tries) (Table 1)

Comparison of old methods and proposed method

	Old Method	Proposed Method
Storage	Graph	Suffix array
Enumerating subgraph	Backtracking	Based on frequency count
Random network generation	Switching operation	Switching operation
Subgraph isomorphism	Standardise by tough method	Subgraph isomorphism not required

Table 1: Comparison of old methods and proposed method.

The Key of Network Motif Mining

There are five steps involved in our approach 1) Form of storage a way to store biological network; 2) Converting given input graph into sequence using graph traversal algorithms; 3) Generation of random graph and random sequences for each random graph; 4) Finding repeating sequences in original graphs sequence as well as in random generated sequences; 5) Motif Identification in this step we designate frequently repeating sequence as network motif whose frequency is higher than given threshold (Figure 1).

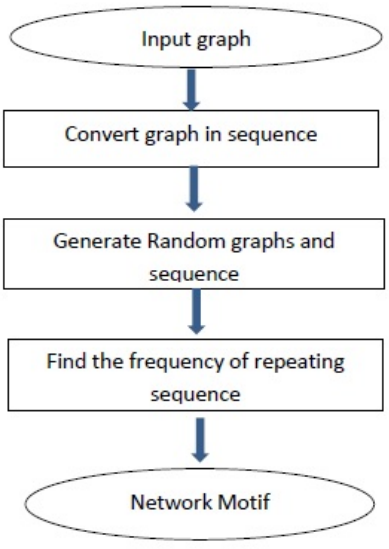

Figure 1: Steps involved in network motif generation.

Algorithm 01

Input: Graph/biological network

Output: Sequence

1. Function: DFS (G,v)

2. Stack S := {}//start with empty stack

3. Push S, v;

While (S is not empty) do

{ u := pop S;

if (not visited[u]) then

visited[u] := true;

for each unvisited neighbour w of u

push S, w;

end if

end while

}

Algorithm 02

For identification of motif, we need to generate random graphs with the same numbers of vertex and edges. That have the same features of input graph generation of random graph is very important step in our method.to generate random graph we use switching operation. In switching operation we choose any two edges and exchange their end points.

After generation of random graph we pass this random network as input to our sequence generator programmer.

Then this sequence is given as input to substring finder which gives as the most repeating subsequence and position of that subsequence.

The main motto of this proposed approach is that we convert graph into sequence by doing so we can able to find repeating subsequence or we can say conservative sequence that is motif in linear time.

Result and Discussion

As *Escherichia coli* is the most well-known bacterial model about the function of bacterial genes. The proposed work will eventually aid to the study being done by the researchers in the field of genetics and other related ones. *Escherichia coli* contains approximately 4,000-5,500 genes and several of genes correspond to transcription factor (Figure 2).

By applying proposed computational approach to the 10 of regulatory groups present in *E. coli*, the seed could be found out. Here in this approach, the seeds are variable lengths and thus the user is not restricted to fixed length, which may give a hypothetical solution. In this way, the proposed approach leads to the more appropriate results in finding network motifs (Figure 3).

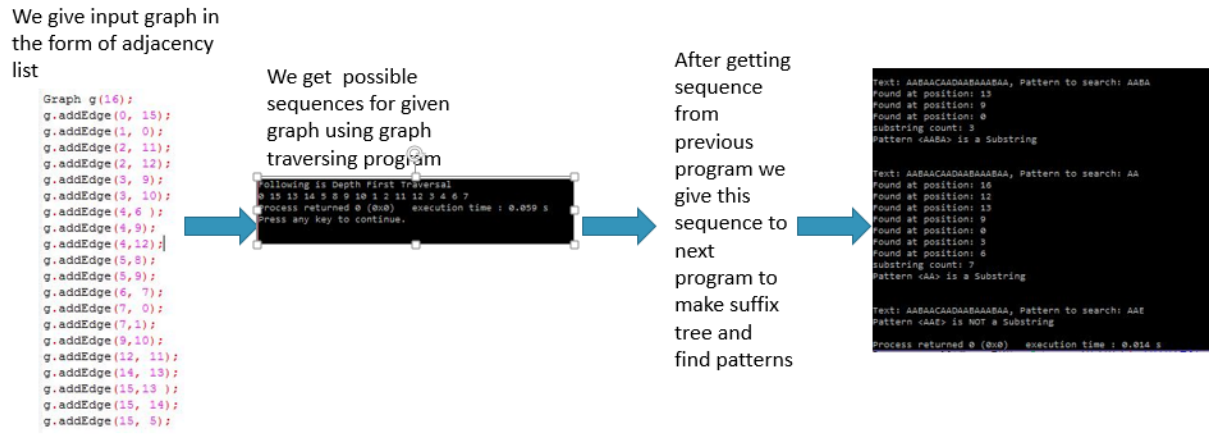

Figure 2: Sequence.

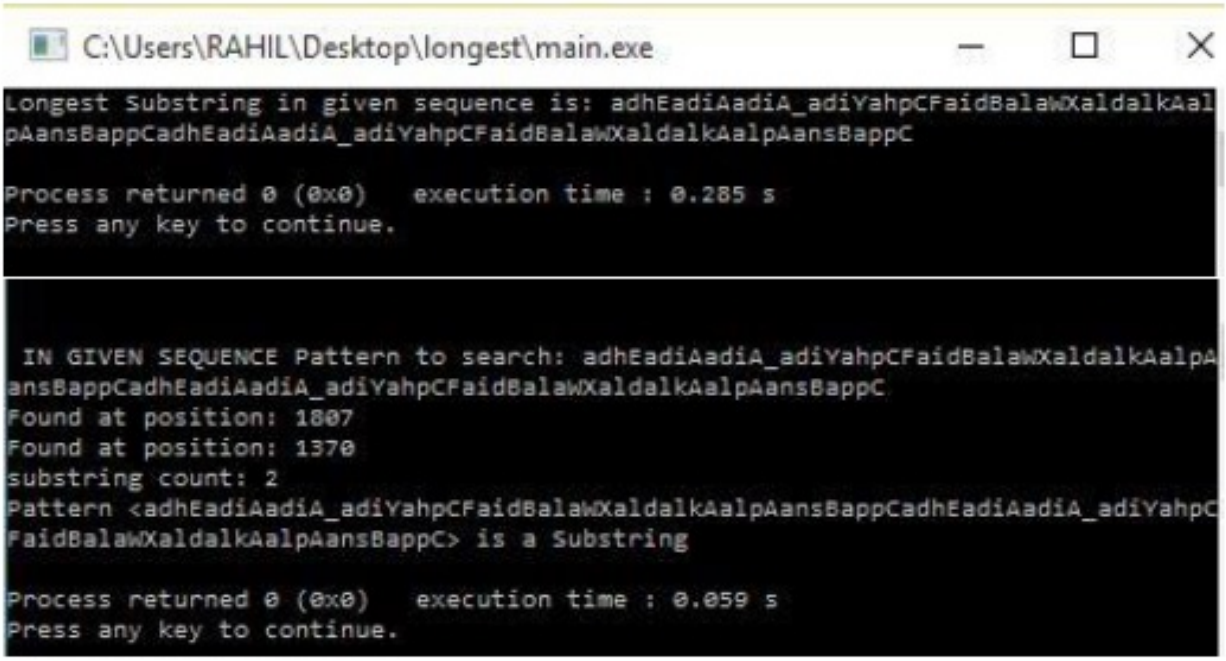

Figure 3: The result after execution of substring frequency counter.

Conclusion

Based on detailed analysis of previous approaches, we have presented new algorithm which allow for a faster detection of network motif and this enable motif detection for larger networks and more complex motifs than was previously possible. Hopefully facilitating future research on network motifs and it's adjoin fields in system biology.

References

1. Albert I, Albert R (2004) Conserverd network motifs allow protein-protein interaction prediction. Bioinformatics 20: 3346-3352.
2. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks, Reviews of Modern Physics 74: 1.
3. Alm E, Arkin AP (2003) Biological natworks, Current Opinion in structural biology 13: 193-202.
4. Ingram PJ, Stumpf MP, Stark J (2006) Network motif:structure does not determine function. BMC Genomics 7: 108.
5. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D (2002) Network motifs: simple building blocksof complex networks. Science 298: 824-827.
6. Shen-Orr SS, Milo S, Mangan S, Alon M (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31: 64-68.
7. Itzkovitz S, Levitt R, Kashtan N, Milo R, Itzkovitz M, et al. (2005) Coarse-graining and self-dissimilarity of complex networks. Physiccal Review E 71: 2-1.
8. Valverde S, Sole RV (2005) Network motifs in computational graphs: a case study in software architecture. Physical Review E 72.
9. Foggia P, Sansone C, Vento M (2001) A performance comparison of five algorithms for graph isomorphism. In: Graph Based Representation in Pattern Recognition.
10. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20: 1746-1758.