Modeling Networks as Probabilistic Sequences of Frequent Subgraphs

Koenraad Van Leemput[1] and Alain Verschoren[2]

[1] Advanced Database Research and Modelling (ADReM),
[2] Intelligent Systems Laboratory (ISLab),
University of Antwerp, Middelheimlaan 1, B-2020 Antwerpen, Belgium

Abstract. Graphs or networks are used as a representation of data in many different areas, ranging from Biology to the World Wide Web. In this paper, a novel approach to graph characterization based on a probabilistic (de)composition into a linear sequence of frequent subgraphs is presented. The resulting probabilistic models are generative for a family of graphs sharing common structural properties. An evolutionary computing approach is used to learn the model parameters for unknown graph classes. This paper describes the (de)composition procedure and illustrates its use in characterizing and discriminating a number of graph types. To demonstrate its practical usefulness, the method is applied to the problem of modeling transcriptional regulatory networks (TRN).

1 Introduction

Research has demonstrated the existence of recurring small graph structures in many types of networks from domains as diverse as computer science and biology [8]. These recurring subgraph patterns are variously called *network motifs*, *graphlets* or more simply *subgraphs*. It has also been shown that complex networks can be compared and classified into distinct functional families, based on their typical motifs [7].

Moreover, biomolecular networks are hierarchical structures that consist of smaller modules of interacting components [2]. Therefore, global metrics, such as degree distribution and clustering coefficient, can not be used to completely analyze their properties [9] and, as a result, local approaches have become more prominent in the study of networks structure. The **hierarchical** and **modular** nature of biological networks [5, 13] has also been elucidated. Graph motifs aggregate into larger clusters and some of the global topological characteristics of graphs originate from the local combinations of smaller subunits.

Furthermore, close investigation into the structure of transcriptional and metabolic networks of *E. coli* and *S. cerevisiae* has suggested that this combination of motifs is not random. There appears to be a type of **preferential attachment** where homologous motifs cluster together [6]. All this implies that a network's large-scale topological organisation and its local subgraph structure mutually define and predict each other and that networks need also to be evaluated beyond the level of single subgraphs, at the level of subgraphs clusters [13].

The above observations served as inspiration for the creation of a new **network model** that integrates **global** knowledge about the presence of network motifs and their **local** combinations. More specifically, global statistical knowledge about the presence of subgraphs is combined with local knowledge about the specific way in which these motifs are interconnected. Two additional ideas were incorporated to allow better integration of this knowledge into a practical analysis method. The first is linearization of the network into an **ordered sequence of motifs**. Secondly, a **probabilistic** approach was chosen that incorporates the ideas of **growth** and **preferential attachment** together with knowledge about recurring structural elements to allow both decomposition of existing graphs and composition of similar graphs.

2 The model

The translation from graph to sequence is accomplished using a probabilistic model that describes the occurrence of, and connections between *motifs*. Because of the probabilistic nature of the model a sequence of symbols describes instances of a graph family, rather than one single graph. Using the probability distributions as a central source of information, this method can be used to both *decompose* existing graphs in to sequences, and *compose* new instances starting from such a sequence.

Motifs are detected in an existing graph by mapping their edges onto edges in the graph. Additional motifs are connected to already detected motifs by merging some of their vertices. As a result, each edge in the graph belongs to exactly one motif in the sequence, while graph vertices can belong to multiple motifs.

All of the information needed to construct the model is contained in a *set of motifs*. It is important to clarify that the term *motif* takes on additional meaning relative to its use in existing literature [11, 8]. Intuitively a motif can be understood to be a small graph with additional information that specifies which vertices can *attach* to other vertices and associated rules that govern the way it can connect to other motifs. These preference rules are expressed as probability distributions.

2.1 Graph decomposition

To illustrate graph decomposition, the example graph depicted in Fig. 1(d) is decomposed using the motif set in Fig. 1(a). This set becomes the alphabet of the decomposition sequence. In this example, the set consists of a 4-node cycle ($4C$), a 3-node cycle ($3C$) and a reflexive edge structure with 2 nodes and 2 edges (R). In general, the choice of motif set can be driven by domain knowledge or by graph mining techniques that compose a set based on a collection of example graphs (see also Section 3.1)

The initial step involves choosing a vertex that will serve as the starting point for motif detection. The *start vertex* can be any vertex in the source graph but

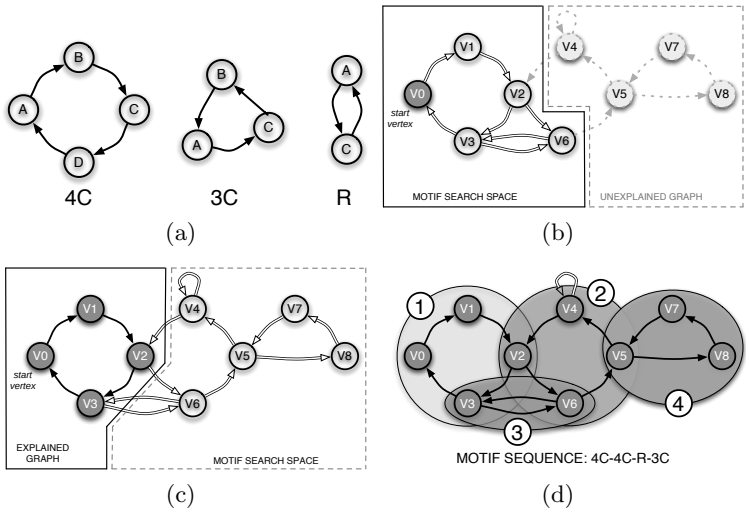


Fig. 1. Example graph decomposition using a sample set of motifs

its choice will influence the decomposition process and the resulting sequence. In the example, vertex $V0$ is chosen as the start vertex. Decomposition is a gradual process, during which motifs are detected in a selected area around the already explained graph. This area of the graph is referred to as the *motif search space*. The depth of the search space depends on the undirected diameter of the biggest motif in the motif set. The motif search space grows by adding edges and vertices, expanding outwards from *fringe vertices*. Initially, the only fringe vertex is the start vertex.

At each step edges and vertices within an undirected distance of E_d from the fringe vertices are added to the search space. In our example the expansion depth of the motif search space at each iteration is two, the largest diameter of any motif in the set. This means that, following the initial expansion from the start vertex $V0$, the *motif search space* contains vertices $V0$ through $V3$, $V6$ and all edges connecting them. Vertices $V2$ and $V6$ are separated from $V0$ by two edges and are therefore not expanded further until they are explained by a motif (situation in Fig. 1(b)).

Edges and vertices that are explained by a motif become part of the *explained graph*. Every consecutive motif in the sequence is required to share its attaching vertices with the part of the graph that has already been explained. The exact number of shared vertices is a function of the motif definition.

At this point, the only matching subgraph is a $4C$ motif, which becomes the first motif in the growing sequence. The only requirement for the initial motif is that it contains the starting vertex. Vertices $V0$, $V1$, $V2$ and $V3$ now become part of the explained graph. It is important to note that only the *edges* of this first motif are removed from the search space. The decomposition is edge-based:

an edge in the source graph can be explained by only one motif in the sequence, while vertices connect motifs and are shared between them.

Starting from this situation, a new expansion is done, and vertices $V4$, $V5$, $V7$ and $V8$ are added to the search space, along with their interconnecting edges (Fig. 1(c)).

Both a second $4C$ motif or an R motif can now be mapped onto the currently unexplained edges. Both would share one vertex ($V2$ and $V3$ respectively) with the first $4C$ motif. In principle, either one could become the next motif in the sequence. To make the choice between *candidate motifs*, it is necessary to introduce the mechanism that can express relative preference for each of the candidates. This is accomplished using the concept of *preferential attachment*.

Preferential attachment rules (Fig. 2) provide a way to evaluate the likelihood of candidate *motifs* and their interconnections, during both graph composition and decomposition. Three essential concepts are combined to determine which candidate should become the next motif in the sequence. Each aspect is described using a probability distribution. These distributions will become the central pieces of information for both graph *decomposition* and *composition*.

Motif-set prior. The first component is a prior preference for specific types of motifs in the motif set. In the example that we are discussing this is a uniform distribution over the motifs in the set because no particular preference has been assigned to any of them.

Motif-Vertex preference. As described in the example, motifs are regarded in the context of their connection to adjacent motifs in the graph. Such connections give rise to a partitioning of a motif's vertices into a set of *attaching vertices* $\mathcal{A}$ and a set of *non-attaching vertices*. *Attaching vertices* serve as connection points to already explained motifs, while *non-attaching vertices* are mapped onto previously unexplained vertices in the current *motif search space*. Every *attaching vertex* has an associated probability distribution over all possible vertices in the motif set, indicating its affinity for specific motifs and vertices belonging to them. The distribution effectively defines which motifs and vertices are preferred candidates for attachment.

Sequence Distance Rule. The final concept is the *sequence distance rule*. Each attaching vertex of a motif contains an additional probability distribution governing its affinity for a target based on a concept of distance in the (de)composition sequence. Differential preference can be given to attachment between motifs depending on the number of motifs in the sequence that separate them.

During decomposition, the likelihood that any newly discovered motif becomes the next one in the sequence is evaluated in the context of the already explained graph and the growing sequence. A newly detected candidate motif m is connected by its attaching vertices to a set of already detected motifs. Each attaching vertex $v \in \mathcal{A}$ is merged with a vertex v' belonging to a motif m' earlier in the sequence. The distance $d(m, m')$ between motif m and motif m' is defined as the number of motifs separating them in the sequence.

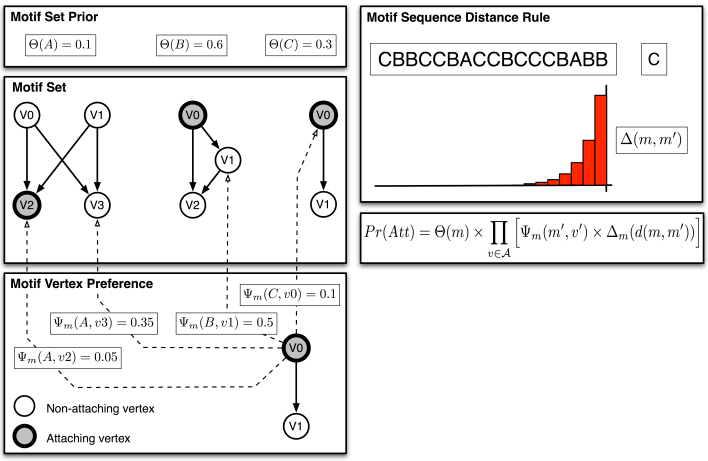


Fig. 2. Example motif set and preferential attachment rules (motif prior Θ, motif-vertex preference Ψ_m and sequence distance rule Δ_m) for one of the motifs.

Definition 1. ***Probability of attachment.*** *The probability of the attachment is the product over all attaching vertices of the three preferential attachment components: the motif prior Θ, the motif-vertex preference Ψ_m and the sequence distance rule Δ_m:*

$$Pr(Att) = \Theta(m) \times \prod_{v \in \mathcal{A}} \left[\Psi_m(m', v') \times \Delta_m(d(m, m'))\right]$$

Back in our example, the choice between a $4C$ and an R motif is resolved by calculating the likelihood for each attachment. First, it is necessary to check whether every *attaching vertex* is mapped onto an already explained vertex in the discovered subgraph. If we accept that the example $4C$ motif has only one attaching vertex and the R motif has two, the only valid decomposition sequence is $4C - 4C - R$. Should the *motif set* contain multiple variants of the $4C$ and R motifs with a different number of attaching vertices, both candidates could be valid. Their *preferential attachment rules* would then determine their order in the sequence. If candidates are equally likely, the choice is made randomly.

Continuing this procedure, and given that the correct sequence is $4C - 4C - R - 3C$, all vertices in the source graph have now been explained. However, this still leaves one edge unexplained, the self-loop at $V4$. This demonstrates that even with a well-chosen motif set that quite accurately captures the structural characteristics of the source graph it may not be possible to completely decompose a graph. To deal with this, *glue motifs* can be introduced into the motif set to collect edges or vertices that can not otherwise be mapped with the conventional motifs. An adequate choice of the motif set would limit the necessity for *glue motifs*.

2.2 Graph composition

Graph composition is governed by the same *preferential attachment rules* as graph decomposition. Starting from a sequence of motifs a graph instance is generated by probabilistically adding edges and vertices as determined by the motifs in the sequence. Attaching motif vertices are merged with graph vertices created by motifs earlier in the sequence. Non-attaching motif vertices create new graph vertices. The number of edges in the resulting graph equals the sum of the number of edges of all motifs in the sequence.

When adding the next motif in the sequence, every attaching vertex has to be mapped onto a graph vertex. To do so, all of the graph vertices are evaluated as potential candidates, using the *preferential attachment rules*. Two additional constraints guide this process: an attachment may not introduce parallel edges and two vertices belonging to the same motif may not be merged, since this would fundamentally alter the structure of the motif.

The likelihood for each valid attachment point is calculated and an ultimately one is chosen using roulette-wheel selection.

3 Experimental evaluation

3.1 Learning

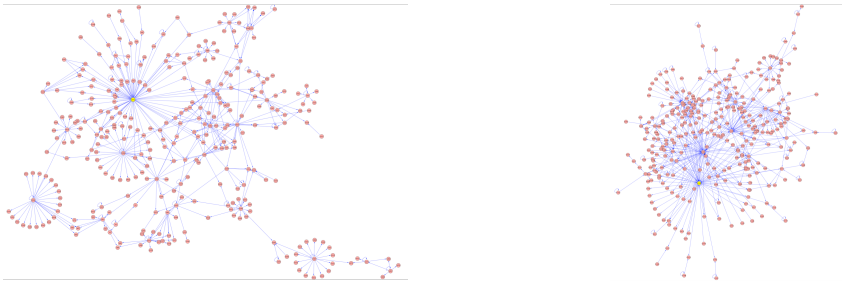
In order to use the system in a new setting — for example the characterization of gene regulatory networks — it is necessary to construct a motif set that adequately characterizes the desired graph family. Because for complex networks it is not feasible to construct the model manually, a *machine learning* was chosen in this work.

Given a training set of positive examples of a certain graph class, an evolutionary algorithm [4] was used to learn a motif set that can generate similar graphs and classify graphs as belonging to this class. In one experiment the largest connected component of the *E. coli* transcriptional regulatory network as described by [11] was used as a training set (Fig. 3(a)). Fig. 3(b) shows an example graph composed with the learned model.

3.2 Classification

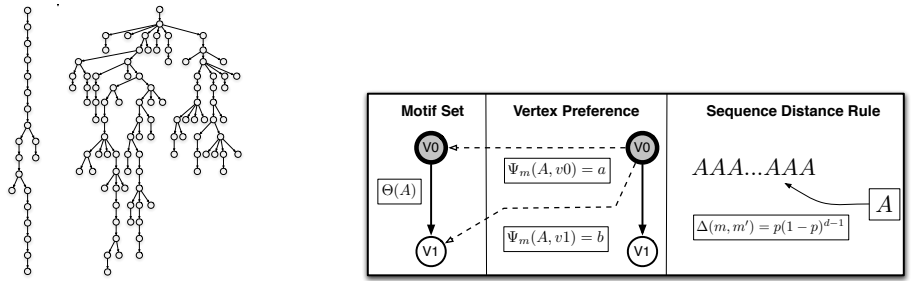
Starting with the same motif set (Fig. 4(c), $a = 0$, $b = 1$) a variety of *trees* was composed from a motif sequence by changing the sequence distance rule. When using a geometric distribution with $p = 0.9$, the trees that are generated are very chain-like, with very few and short branches (Fig. 4a). In this case it is extremely likely that new motifs in the sequence attach to adjoining motifs while with $p = 0.1$ attachments further up the chain are much more likely, resulting in the creation of many branches(Fig. 4b).

One thousand chain-like and one thousand highly-branched trees were generated from a one hundred motif sequence with appropriate parameters. Every tree was decomposed, starting from the root vertex, using the motif set that



(a) (b)

Fig. 3. (a) Largest connected component of the *E. coli* TRN used as training set for learning the TRN model. (b) Example of a graph composed with the learned model.



(a) (b) (c)

Fig. 4. A chain-like tree (a), characterized by a low number of very short branches and a highly-branched tree (b). Motif set used to generate them (c).

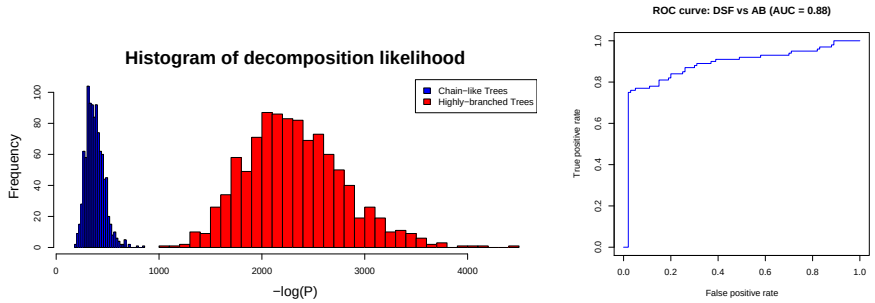
generates chain-like trees. As explained in Section 2.1, during the decomposition, the likelihood that any newly discovered motif becomes the next one in the sequence is evaluated in the context of the already explained graph and the growing sequence. The sum of these likelihoods, expressed as negative log-probabilities (Def. 2), can be interpreted as an overall score for the plausibility of the decomposition.

Definition 2. ***Likelihood of decomposition.*** *The likelihood L of decomposition of a graph G, given a motifset $\mathcal{M}$ into a motif sequence $S(\mathcal{M})$ is defined as*

$$L = \sum_{a \in S(\mathcal{M})} -log(Pr(Att)_a)$$

This likelihood can also be seen as a measure for the probability of generating the decomposed graph from the sequence, given the specific motif set. Fig. 5(a) shows the histogram of the resulting log-probability scores for all decompositions. As expected, the likelihood of the decomposition is much higher for the chain-like trees than for the highly-branched trees. The introduction of new branches

results in lower scores because new motifs do not attach to the most recently discovered motifs, but to motifs further away in the sequence, which is unlikely given the geometric distribution with $p = 0.9$.



(a) (b)

Fig. 5. (a) Histogram of decomposition likelihood for two families of trees: *chain-like (blue)* and *highly-branched (red)*. (b) Receiver operating characteristic (ROC) curve for two-way classification between AB and DSF graphs.

A similar experiment using a motif set learned from the *E. coli* TRN [11] was used to decompose a series of random graphs generated with the Albert-Barabási (AB)[1] and the directed scale free (DSF) [3] models. The likelihood of the decomposed sequence was then used as a score for the overall plausibility of the decomposition. Fig. 5(b) shows that it is possible to distinguish these different graph classes using the learned model.

4 Conclusions

This paper presented a model that allows characterization of graph families through a (de)composition method based on probabilistic sequences of motifs. Given a motif set, a sequence can probabilistically produce many graphs by sequentially combining the motifs. Both decomposition and composition are governed by the same probability distributions, that dictate the order of motif detection or combination.

The feasibility of using a machine learning approach to construct a suitable motif set for a new family of graphs was demonstrated. These learned motif sets can then be used to distinguish between different classes of graphs.

Acknowledgments This work was supported by an SBO grant (IWT-60045) of the Flemish Institute supporting Scientific-Technological Research in industry (IWT) and has been partially funded by the EU contract IQ FP6-516169. The authors also wish to thank Bart Goethals, Kris Laukens, Bart Naudts and Piet van Remortel for their valuable input and feedback concerning this work.

References

1. R. Albert and A. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:4797, 2002.
2. A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101113, 2004.
3. B. Bollobás, C. Borgs, C. Chayes, and O. Riordan. Directed scale-free graphs. In *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms*, page 132139, 2003.
4. D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, Reading, MA, USA, 1989.
5. H.-W. Ma, J. Buer, and A.-P. Zeng. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, 5:199, 2004.
6. H.-W. Ma and A.-P. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):14231430, 2003.
7. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):15381542, 2004.
8. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824827, 2002.
9. N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):35083515, 2004.
10. N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
11. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31:6468, 2002.
12. T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7:43, 2006.
13. A. Vazquez, R. Dobrin, D. Sergi, J. Eckmann, Z. Oltvai, and A. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, 101(52):1794017945, 2004.