A tour of biological network motif mining

Carlos Oliver (carlos@ozeki.io)
Department of Biosystems Science and Engineering
ETH Zürich

McGill University: COMP 564
April. 12, 2022

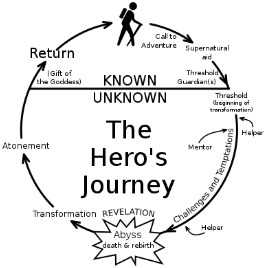
Outline

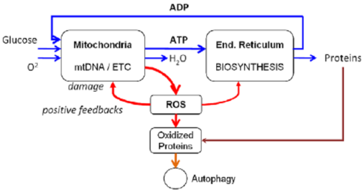
- Part I - Groundwork
  - What is a network motif?
  - Why are network motifs interesting?
- Part II - Methods
  - How does one find motifs?
    - Combinatorial Methods
    - Representation Learning Methods
- Part III - Conclusions

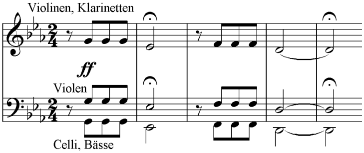
> 'What has been will be again, what has been done will be done again; there is nothing new under the sun.
>
> — Ecclesiastes 1:9

What is a motif?

A motif is a pattern that recurs in multiple observations of a system.

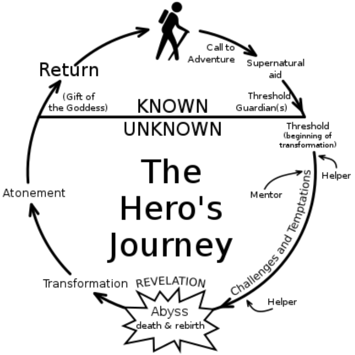


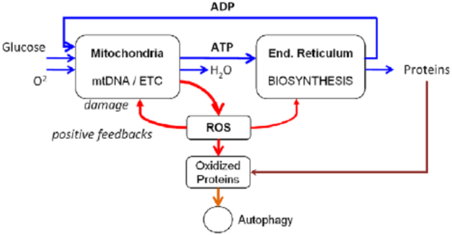




Source: Wikipedia

What is a motif?

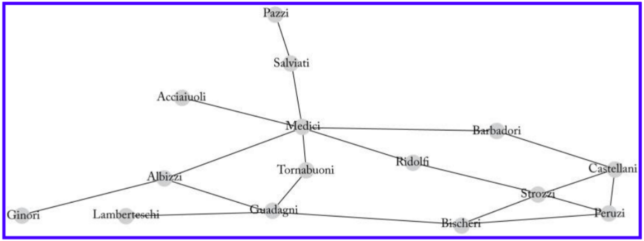




It is often easier to **recognize** an instance of the motif than to define the underlying pattern.

What is a network?

A **network** is a collection of entities and the relations between them. E.g. a group of people and friendship status.

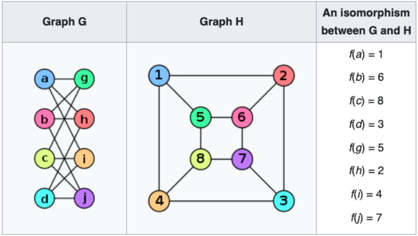


[Ham20]

A **graph** (a.k.a a graph) $G = (V, E)$ is a pair of sets where V is a set of entities called 'nodes' and $E \subset V \times V$

A **subgraph** of G is another graph formed by a subset of the nodes of G. It is **connected** if there is a path between all pairs of nodes.

A pair of graphs G, H is **isomorphic** ($\cong$) if a bijective mapping $f : V_G \to V_H$ such that any two vertices u and v of G are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent in H



Source: Wikipedia

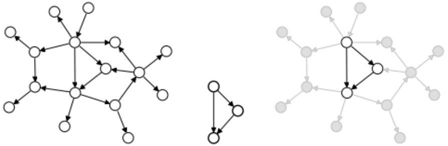
What is a network motif?

High-level:

A pattern which manifests itself as groups of related subgraphs.

Semi-Formal:

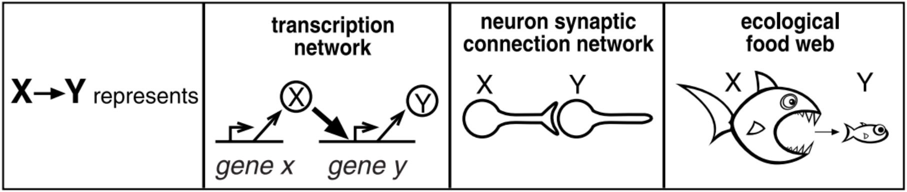
We say a pattern (small connected graph) is a motif given a set of graphs $\mathbb{G}$ if it is isomorphic to more subgraphs of $g \in \mathbb{G}$ than expected.



[SS10]

Why are motifs interesting?

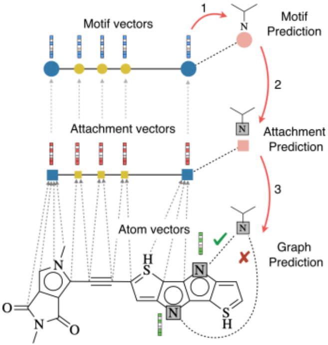
Network motifs raise the hope that network function can be understood in terms of basic computational building blocks. - Uri Alon 2003



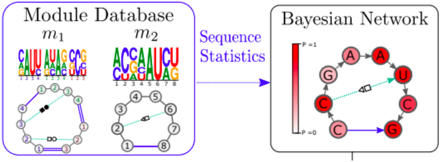
[MSOI+02]

Why are motifs interesting?

- Motifs can be applied to solve many computational problems.
  - Graph generation
  - Function prediction
  - Structure prediction



[JBJ20]



[SGYR+20]

Questions?

Motif mining is not easy

To identify the set of recurrent patterns in a graph (or set of graphs) you need to solve 3 sub-problems:

1. **Pattern search:** Enumerate all possible subgraph patterns.
2. **Pattern detection:** For each pattern, count # of appearances.
3. **Significance testing:** For each occurrence count, measure significance. (Repeat 1 and 2 on null model)

[SS10]

The arrival of computational motif mining [MSOI+02]

- Uri Alon & co sparked the network motif mining field in 2002.

Network Motifs: Simple Building Blocks of Complex Networks

R. Milo,[1] S. Shen-Orr,[1] S. Itzkovitz,[1] N. Kashtan,[1] D. Chklovskii,[2] U. Alon[1*]

> *"None of the network motifs shared by the food webs matched the motifs found in the gene regulation networks or the World Wide Web.... Different motif sets were found in electronic circuits with different functions. This suggests that motifs can define broad classes of networks, each with specific types of elementary structures."*

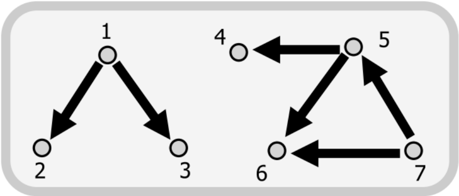
M-Finder Sampling Algorithm [KIMA04]

1. Pick a graph pattern p size k
2. Sample random k-subgraphs $\mathcal{S}$, compute adjusted concentration estimate

$$C_p = \frac{\sum_{s \in \mathcal{S}} \mathbb{1}_{[s \cong p]}}{|S|}$$

3. Repeat for a randomized graph.

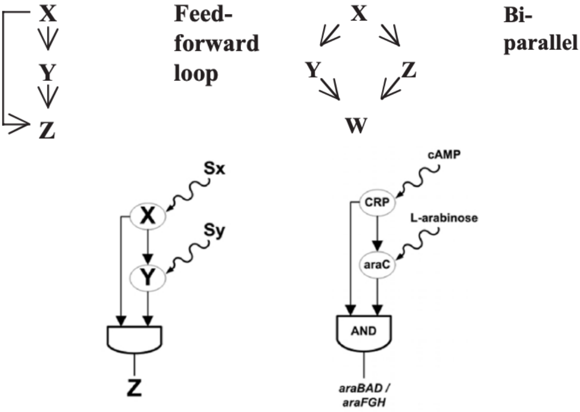
Toy Network:



Probability to sample {1,2,3}:
There are 2 possibilities to sample {1,2,3}:
1. Pick first (1,2): Pr=1/E=1/6.
then pick (1,3): Pr=1.
Pr[(1,2) then (1,3)]=1/6*1=1/6.
2. Pick first (1,3):Pr=1/E=1/6.
then pick (1,2): Pr=1.
Pr[(1,3) then (1,2)]=1/6*1=1/6.
In Total: Pr[{1,2,3}] = 1/6 + 1/6 = 1/3=12/36

Probability to sample {4,5,6}:
There are 2 possibilities to sample {4,5,6}:
1. Pick first (5,4): Pr=1/E=1/6.
then pick (5,6): Pr=1/2.
Pr[(5,4) then (5,6)]=1/6*1/2=1/12
2. Pick first (5,6): Pr=1/E=1/6.
then pick (5,4): Pr=1/3.
Pr[(5,6) then (5,4)]=1/6*1/3=1/18.
In Total: Pr[{4,5,6}] = 1/12 + 1/18 = 5/36

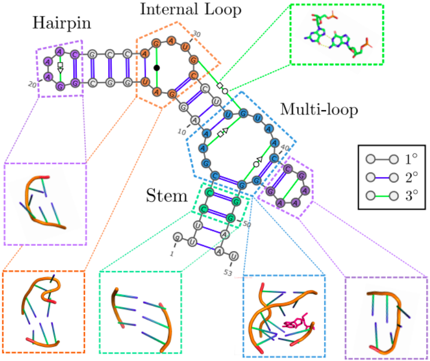
Mined motifs uncover design principles. [MZA03]



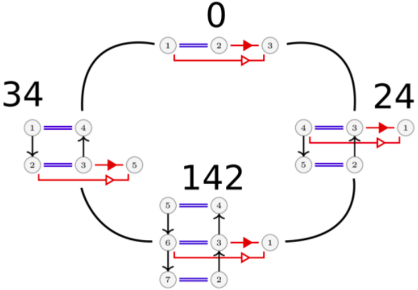
- Need to be careful when attributing function to motifs [ISS06].
  - Randomization
  - External context

CaRNAval exploits domain knowledge to identify application-specific motifs. [RSW+18]

- Look for new motifs connecting known motifs.



RIN A-minor Type I



[RSW+18]

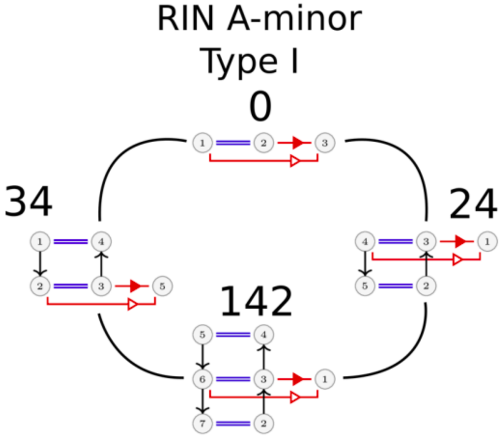
Questions?

Part 2.1 Representation Learning

Bottlenecks in combinatorial approaches: motif size and variability.

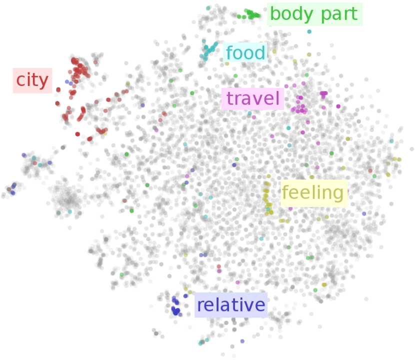
Motifs are often related or come from noisy data

Bottlenecks in combinatorial approaches: motif size and variability.



Representation Learning

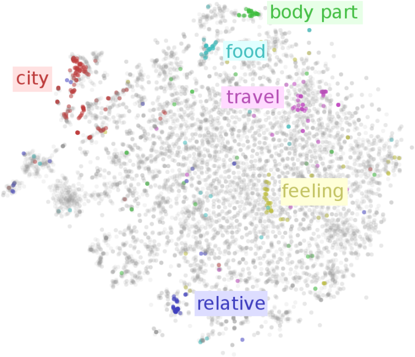
Idea: work in a vector space instead of discrete space.



[Ola]

Representation Learning

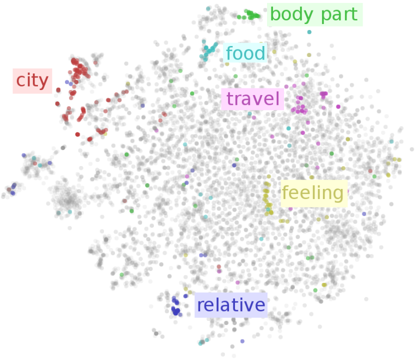
Learn the parameters θ of a function $f_\theta(X)$ which maps input data X to a vector space with useful properties.



[Ola]

Representation Learning

We choose θ by backpropagation to minimze a **differentiable** penalty function known as the **loss function**.

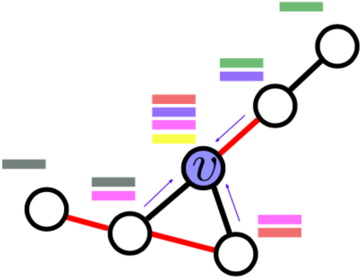


[Ola]

Graph representation learning maps discrete structures to continuous spaces

Message passing lets us represent graphs in an order invariant manner.

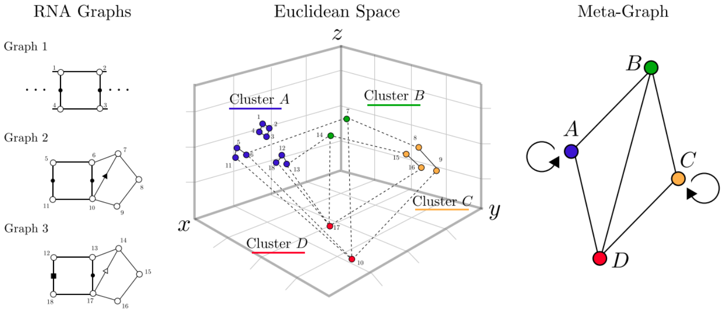
$$z_v^{(k)} = \sigma\left[\mathbf{W}_{\text{self}} z_v^{(k-1)} + \mathbf{W}_{\text{neigh}} \sum_{v' \in \mathcal{N}(v)} z_{v'}^{(k-1)} + \mathbf{b}\right]$$



VeRNAl uses a continous embedding to mine fuzzy motifs [OMP+22]

1. Use the continuous properties of representation techniques to capture continuous relationships.
2. Use motifs as building blocks for larger motifs.

$$\mathcal{L}(\theta, u, v) = \sum_{u,v \in V(G)} \left[K(u,v) - f_\theta(u)^T f_\theta(v)\right]^2 \tag{1}$$

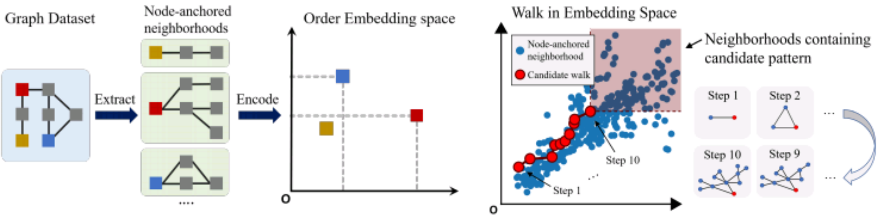


SP-Miner encodes subgraph-supergraph relationships [RAJJ20]

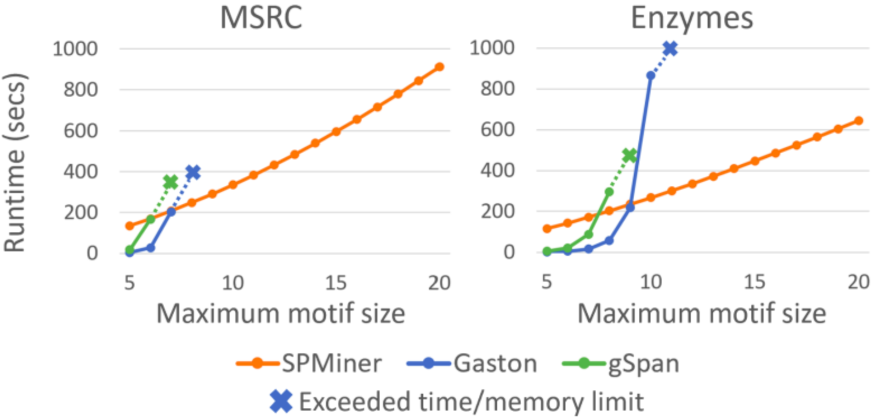
- **Observation:** A motif is a subgraph of many supergraphs.
- **Idea:** Encode subgraph-supergraph structures in representation step.
- When $A \subset B$, $f(A) < f(B)$:

$$E(A, B) = ||\max(0, f_\theta(A) - f_\theta(B))||^2 \tag{2}$$

$$\mathcal{L}(\theta, A, B) = \sum_{(A,B)\in P} E(A, B) + \sum_{(A',B')\in N} \max(0, \alpha - E(A', B')) \tag{3}$$



SP-Miner Identifies large motifs



Summary

1. Network motifs are important features of real-world networks.

Summary

1. Network motifs are important features of real-world networks.
2. Mining network motifs is a rich and interesting problem.

Summary

1. Network motifs are important features of real-world networks.
2. Mining network motifs is a rich and interesting problem.
3. Recent advances are opening new doors for motif mining methods.

Acknowledgements

- Jérôme Waldispühl
- Vincent Mallet
- Pericles Philippopoulos
- Roman Sarrazin Gendron
- William L. Hamilton

References I

William L Hamilton.
Graph representation learning.
Synthesis Lectures on Artifical Intelligence and Machine Learning, 14(3):1–159, 2020.

Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark.
Network motifs: structure does not determine function.
BMC genomics, 7(1):1–12, 2006.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola.
Hierarchical generation of molecular graphs using structural motifs.
In *International Conference on Machine Learning*, pages 4839–4848. PMLR, 2020.

References II

Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon.
Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs.
Bioinformatics, 20(11):1746–1758, 2004.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon.
Network motifs: simple building blocks of complex networks.
Science, 298(5594):824–827, 2002.

Shmoolik Mangan, Alon Zaslaver, and Uri Alon.
The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.
Journal of molecular biology, 334(2):197–204, 2003.

References III

Christopher Olah.
Blog, howpublished = `https://colah.github.io/posts/2015-01-visualizing-representations/`.
Accessed: 2022-04-12.

Carlos Oliver, Vincent Mallet, Pericles Philippopoulos, William L Hamilton, and Jérôme Waldispühl.
Vernal: a tool for mining fuzzy network motifs in rna.
Bioinformatics, 38(4):970–976, 2022.

Ying Rex, Wang A, You J, and Leskovec J.
Frequent subgraph mining by walking in order embedding space.
2020.

References IV

Vladimir Reinharz, Antoine Soulé, Eric Westhof, Jérôme Waldispühl, and Alain Denise.
Mining for recurrent long-range interactions in rna structures reveals embedded hierarchies in network families.
Nucleic acids research, 46(8):3841–3851, 2018.

Roman Sarrazin-Gendron, Hua-Ting Yao, Vladimir Reinharz, Carlos Oliver, Yann Ponty, and Jérôme Waldispühl.
Stochastic sampling of structural contexts improves the scalability and accuracy of rna 3d modules identification.
In *RECOMB 2020-24th Annual International Conference on Research in Computational Molecular Biology*, 2020.

Falk Schreiber and Henning Schwöbbermeyer.
Motifs in biological networks.
In *Statistical and Evolutionary Analysis of Biological Networks*, pages 45–64. World Scientific, 2010.