Biological Network Clustering and Network Motif Finding
Wooyoung Kim, Yi Pan
Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA

Abstract

Motivation: Network motifs are defined as over-represented connected subgraphs in a network. Biological data can be structured with biological networks, and it has been shown that network motifs carry significant functions in the network. The identification of distinct network motifs provide information about the typical patterns in different types of biological networks. To find network motifs, three steps involved: First, listing or sampling k-node subgraphs from the target network, and secondly identifying all isomorphic graphs for each sample. Then, each isomorphic k-node subgraph found in the target network is evaluated with its statistical significance compared with the generated networks. Therefore, computational time of finding k-node subgraphs increases exponentially even as the size of target network or k grows linearly. Various algorithms have been developed to find network motifs, including exhaustive search and approximation search algorithms. Those algorithms, however, result in discovering relatively small size of network motifs, due to its high computational complexity.

Results: In this work, we propose to combine network parallelization and query parallelization for feasible computing. First we cluster a large network into smaller sub-networks for network parallelization, then query parallelization strategy is used to recover missing subgraphs from the removed edges after clustering. As we exploit existing network clustering algorithms currently, a key challenge is developing an algorithm to search missing subgraphs from the removed edges without overlapped search. We show how the clustering can speed up searching k-node subgraphs and present an algorithm to search missing subgraphs in complete parallel way. We also plan to design clustering networks with biological significance to see if we can speed up the process of finding biologically meaningful network motifs.

Process of Network Motif Discovery
1) Find which subgraphs occur and how many times.
   - Exact algorithms: enumerate all the possible k-node subgraphs: exponential search
   - Approximation algorithms: approximately search k-node subgraphs: not exact results.
2) Determine isomorphic subgraph groups.
   - McKay’s nauty algorithm: each subgraph is labeled uniquely.
3) Determine which group are frequent and unique compared in random graphs.
   - Use p-value or z-value to determine if the subgraph is unique.

Parallel Approach

Query parallelization
- Assigns a subset of query subgraphs to each worker. Then each worker search the isomorphic subgraphs to the query in the target network.

Network parallelization
- Divide the target network into a number of subnetworks.
- Assigns the subnetworks to workers and search for the same query set.

Fig.2 Network clustering. Four edges are removed.

Biological Network Clustering

- Network Parallelization

Fig.4 Recover missing subgraphs from removed edges in parallel.

Conclusion and Future work
We use clustering algorithms to make the process feasible with parallel approach. Since the sub-networks are not overlapped and since we can recover missing subgraphs from each removed edge, parallelization is trivial. Future work, however, includes clustering with biological weights and labeling network motifs with GO terms as well.