A Guided Sampling Algorithm for Identifying Network Motifs in a Transcription Regulatory Network

Raymond Wan[†] rwan@kuicr.kyoto-u.ac.jp
Nelson Hayes[†] nelson@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Hiroshi Mamitsuka mami@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

1. Introduction

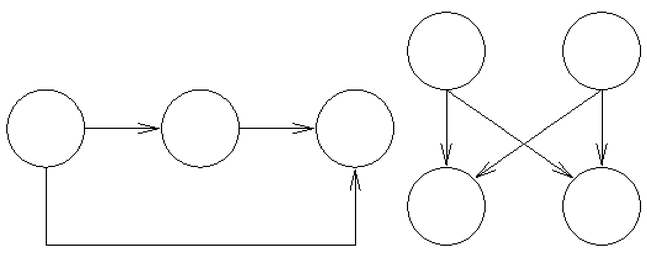
Transcription regulatory networks use conserved network motifs to dynamically control the rate and progress of gene expression in a cell. Genes are activated and transcribed into RNA when one or more gene-specific transcription factors associate with upstream DNA recognition sites. Thus, the identification of important functional network motifs in transcription regulatory networks is an important problem in bioinformatics. To avoid a computationally intensive brute-force approach, alternatives such as graph sampling have been proposed. In this poster, we extend the graph sampling method by taking into account variation in node degree in a graph to guide sampling. We examine the method's potential with some preliminary experiments on the *E. coli* network.

2. Hypothesis

In a network with variation in node degree, statistically significant motifs can be detected by partitioning the network into n buckets of nodes sorted by degree and then sampling an equal proportion of nodes from each bucket.

3. Related Work

Milo et al. [2002] defined network motifs as being statistically significant subgraphs relative to a random network. They applied a brute-force technique to discover motifs in transcription regulatory networks (represented as directed graphs) and showed that the feed-forward loop and the bi-fan motifs (below) are significantly over-represented, when compared to random graphs.



(a) Feed-forward loop (b) Bi-fan

Figure 1: The feed-forward loop and the bi-fan.

Improvements over the brute-force technique:

- Kashtan et al. [2004] obtained motifs by randomly selecting neighboring edges.
- This technique was further improved by enumerating all subgraphs and then filtering to attain unbiased sampling [Wernicke, 2005].
- Wang et al. [2005] developed a parallel algorithm.
- Network motifs have also been considered by Chen et al. [2006] for undirected protein-protein interaction networks.

4. Method

In this poster, we extend the random sampling approach by focusing on the initial selection of nodes. Graphs such as transcription regulatory networks contain localized clusters of highly connected nodes [Artzy-Randrup et al., 2004]. As a consequence, we explored the possibility of guiding the sampling process by explicitly factoring in the unequal distribution of node degree. We score each vertex based on its degree as well as the degrees of its neighbors and calculate the cumulative score across all vertices and divide by n.

Then we sort the vertices by score and divide them into n buckets such that the cumulative score of each bucket is roughly the same. By sampling an equal proportion from each bucket, we are able to normalize the probability of sampling from regions of high and low connectivity.

5. Algorithm

1: **Input:** Graph $G(V, E)$, motif size m, bucket count n, sampling percentage p.
2: **Output:** Set of motifs M.
3: Read in G.
4: Score each $v \in V$.
5: Sort V by score.
6: Separate V into buckets $B_1 ... B_n$.
7: **for** $i = 1$ to n **do**
8: $\quad S_i \leftarrow p$ percent of B_i at random.
9: $\quad$ **for each** $v \in S_i$ **do**
10: $\quad\quad M \leftarrow M \cup \text{DFS}(v)$
11: $\quad$ **end for**
12: **end for**
13: $M \leftarrow M-$ isomorphic motifs in M.
14: **print** M.

6. Results

Preliminary experiments with the *E. coli* transcription regulatory network (Figure 1) have shown that our method can identify the feed-forward loop as significant using 25% of the vertices. In the experiments shown below, 16 iterations were employed to obtain 95% confidence intervals.

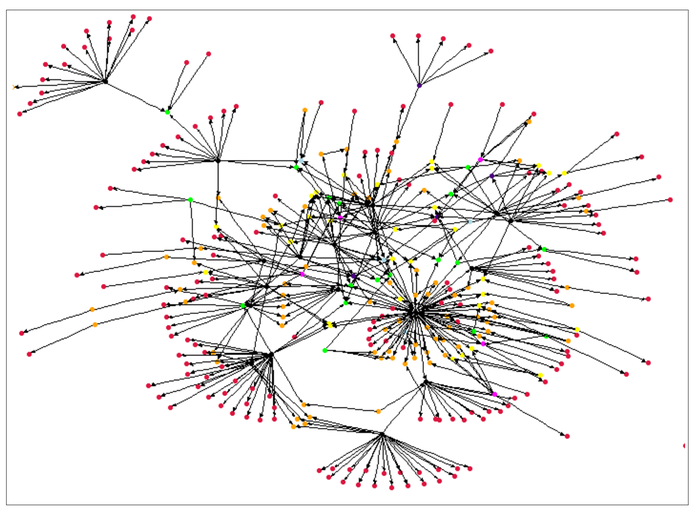


Figure 2: The *E. coli* network has 423 nodes and 519 edges [Mangan et al., 2003].

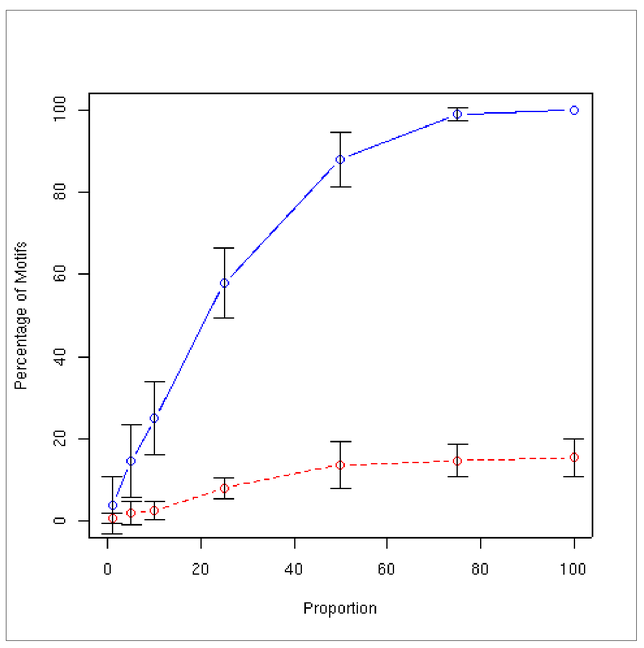
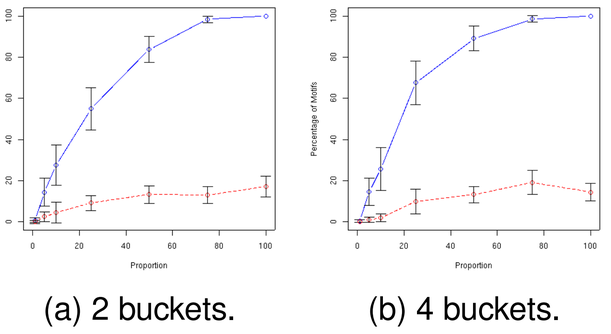


Figure 3: Feed-forward loop discovery without using buckets. Blue represents the *E. coli* network and red represents random networks.



(a) 2 buckets. (b) 4 buckets.

Figure 4: Feed-forward loop discovery using 2 and 4 buckets.

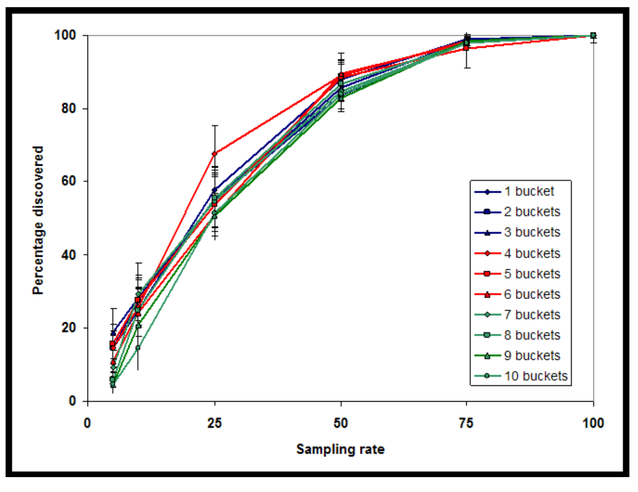


Figure 5: Relationship between sampling rate and motif detection rate using varying bucket sizes.

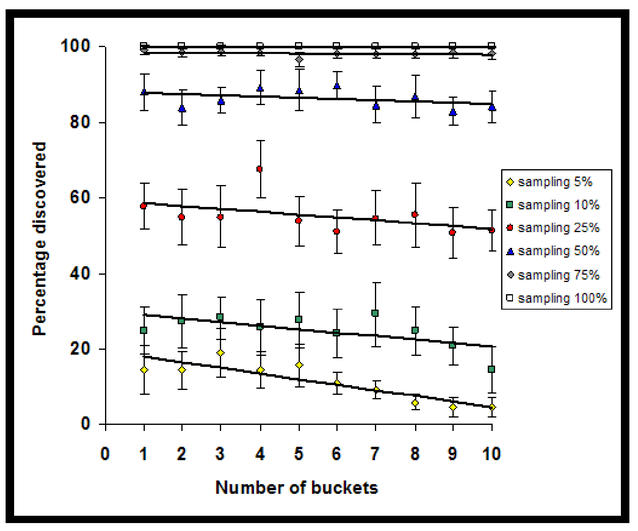


Figure 6: Relationship between the number of buckets and motif discovery rate using different sampling rates.

7. Discussion

Our results show that statistically significant motifs in a network can be detected by sampling a fraction of the nodes (e.g. 25%). Sampling efficiency can be improved by using buckets to ensure that roughly equivalent proportions of high- and low-complexity nodes are sampled. Increasing the number of buckets past 4 appears to offer only marginal improvement.

8. Future Work

We plan to explore the following:

- Improve the vertex scoring mechanism.
- Investigate alternate random graph algorithms.
- Expand our experiments to larger motifs and other transcription regulatory networks.
- Determine the optimal bin size dynamically based on degree distribution.
- Determine the sampling proportion as a function of the number of vertices in each bucket.
- Incorporate more efficient tree-traversal methods, as developed by others.

Acknowledgments: Both R. W. and N. H. were supported by fellowships from the Japan Society for the Promotion of Science.

The 5th Asia Pacific Bioinformatics Conference (**APBC 2007**), January, 15-17, 2007, Hong Kong, China. ([†]These authors contributed equally to this work.)