CS466 – Introduction to Bioinformatics
Department of Computer Science
Spring 2021
Grainger College of Engineering, University of Illinois at Urbana–Champaign

---

Mini Project Report

Motif Finding Conclusions

Dr. Saurabh Sinha

Team Members

Liu, Pan *panl3@illinois.edu*

Zhu, Ruike *ruikez2@illinois.edu*

ANALYSIS AND CONCLUSION

From the plots we have, we can see that:

1. Increasing sequence count (SC) can largely increase the prediction performance. The number of overlap sites and positions shows a great increase when SC is larger, and there also appears much more 0 entropy in the relative entropy table.

2. Increasing motif length (ML) can also increase the prediction performance. The number of overlap sites and positions increases when ML is larger, but the influence of ML is smaller than the influence of SC. For the runtime, it shows a stable increase when ML is larger.

3. Higher information content (ICPC) is associated with better predictions. With ICPC increase, though there is a slight decrease when changing ICPC from 1.5 to 2, the number of overlap sites and positions basically shows an increasing tendency when increasing ICPC. From the relative entropy table, there also appears more 0 entropy when ICPC=2. For the runtime, it shows a stable increase when ICPC is larger.

Analysis:

We use the Greedy search in solving the motif finding problem. Greedy search is an effective and cost-saving method in time complexity, but it cannot always find the optimal solutions. This greedy motif search algorithm finds a similar set of patterns across several DNA sequences and those patterns can be eligible candidates for Motifs in DNA. The greedy method can also be performed on any desired length from a big DNA sequence due to its low cost in time complexity.

The 'greedy' lies in the points that it will select the first seen k-mers to create PWM, the second selected motif ($motif_1$) will depend on this position weight matrix (PWM), and the latter selected motifs ($motif\ _i$) will depend on the PWM formed by ($motif_1, motif_2, \dots, motif_{i-1}$). This means the initialization part (the PWM formed by first seen k-mers and the first few selected motifs) is an important step and takes a significant role in finding an optimal solution. If the initialization is not that great, the greedy algorithm can go to local optimal and cannot find the global optimal solutions.

When SC, ML, and ICPC are low, the probability of finding good first seen k-mers will be smaller, which means we will be less likely to form good weights in the position weight matrix. If we increase the value of these parameters, the performance of the greedy algorithm will have great improvement. And this can also explain why there will appear many INF in the relative entropy table when these parameters are low.

For future work we would like to investigate ways to improve the approximation generated by the greedy algorithm. The result that the greedy algorithm produces is highly dependent on the seed found in the first two DNA strands, and the approximation will not be good if the seed is not a good match to the rest of the strands. This problem can be addressed by running the greedy algorithm on many different pairs of strands to generate approximations from many different seeds. The best motif found from all of these approximations should be much more accurate than an approximation generated by running the greedy algorithm only one time.