Computational Inference of Cis-Regulatory Modules Using Evolutionary Computation

Will Gantt

September 15, 2016

1 Goal

The goal of my honors project is to develop a tool that uses evolutionary computation to infer candidate cis-regulatory modules (CRMs) more effectively than existing methods. This is a direct continuation of my work in the Congdon lab this past summer, and I consequently have the benefit of eight weeks' experience with the subject going into my independent study this fall.

My hope for this project is that it will yield a program that improves on typical approaches (viz. multiple alignment) to CRM inference by taking into account the possible variation in the order of regulatory elements within a module across species. Moreover, if the results prove promising, I hope to publish them, as a peer-reviewed publication would improve my competitiveness as an applicant to top graduate programs in computer science.

I will know that I have succeeded in my goal if my program can outperform common methods of CRM inference in a series of tests. Short of this, I would hope, at the very least, to have created by the end of the year a tool that is of some practical use to other members of the Congdon lab.

2 Context

Transcription factors are proteins that regulate gene expression by binding to noncoding regions of DNA near the gene to be regulated. They help either to activate or to silence the transcription of that gene by enabling or blocking (respectively) the recruitment of RNA polymerase to that site.

Typically, transcription factors work in groups to regulate a gene. A *cis-regulatory module* (CRM) is the set of sites on the DNA where the transcription factors in such a group will bind.

Developing computational methods to infer individual candidate transcription factor binding sites (*motifs*) is a well-studied problem in bioinformatics. The value of addressing this problem consists in its potential to vastly reduce the time and money required of biologists working in the lab by enabling them to restrict their search to just a small set of candidate sites. However, the list of candidates generated is often still too large to be searched exhaustively. By grouping motifs into modules, we can further hone the list of candidates, and give a better indication of which transcription factors are involved in the regulation of which genes.

3 Experience and Methodology

The inspiration for my project this summer came from Jeff Thompson, a former member of the lab who wrote GAMI-CRM — a collection of bash scripts that infers candidate modules using output data from Genetic Algorithms for Motif Inference (GAMI), a tool designed by Clare and her lab. However, my program differs from Jeff's most notably in that it will, as mentioned, allow for differences in the order of regulatory elements within modules in different species.

During my fellowship, I implemented an extremely rudimentary version of my program in C, using John Grefenstette's GENEtic Search Implmentation System (GENESIS) as the basis for the genetic algorithm component. The program showed some promise on randomly generated test sequences, but I did not get the chance to run it on real GAMI post-processor data. Rather than begin from scratch this semester, I intend to improve and build on what I have already implemented.

4 To-Do

I have grouped project objectives into short-, medium-, and long-term categories. The level of specificity varies accordingly.

Short-term (September - Mid-October)

- Review code from the summer.

- Give a presentation to the CS faculty and honors students on my project.
- Make a poster for the President's Summer Research Symposium and present it.
- Read relevant chapters in Krane and Raymer's *Fundamental Concepts of Bioinformatics*.
- Begin literature review. Specifically, look for:
  - Papers that have proposed alternatives to multiple alignment.
  - Papers that propose methods for computational inference of CRMs. These will serve as the basis for a review paper (begun by Jeff Thompson) that summarizes the existing approaches to CRM inference.
- Write summaries for each method of CRM inference that I read about.

Medium-term (Mid-October - December)

- Finish writing the review paper on existing approaches to CRM inference, and submit for publication
- Resume development and testing of the basic algorithm.
  - Run program on actual GAMI post-processor data
  - Compare performance to other methods
  - Use results to improve the algorithm
- Continue literature review and prepare presentations as they arise.

Long-term (January - May)

- Continue testing (*aggressively*).
- If preliminary results are encouraging, work with Clare to try to get them published.
- Write thesis.
- Prepare defense.

5 Challenges

I anticipate that this project will present more than a few challenges. Apart from the inherent difficulty of the problem itself — of developing creative and effective algorithms for CRM inference — I am most concerned about the availability of test data. For measuring the effectiveness of my program, I will be limited to sequences known to contain CRMs, and I am unsure how much of this kind of data is readily accessible. Ideally, the program would be run on unexamined sequences and the results validated (or invalidated) in the lab, but I do not expect to get to that point.