Motif Evaluation and Location Methods for Genetic Algorithms for Motif Inference (GAMI)

R a c h e l T e o

University of British Columbia

Abstract

Functional sections of DNA (called "motifs") in non-coding DNA can have a regulatory impact on the expression of genes, which can enhance or suppress the production of proteins. Due to the large amount of non-coding DNA as compared to coding regions of DNA, it is difficult to identify these functional sections through lab work alone. In addition to a conservation of patterns across evolutionary divergent species, positional conservation suggests the presence of motifs in a sequence. Positional bias is a tendency for motifs to remain approximately the same distance from a gene across multiple species.

Genetic Algorithm for Motif Inference (GAMI) is a genetic algorithm that evolves a population of motifs that appear to have been conserved across species that have evolved divergently. This paper explores the impact of adding positional bias as a means of further narrowing down potential motifs for lab work.

Introduction

Until fairly recently, non-coding DNA has been considered to be "junk DNA", or nonsensical DNA accrued from generations of evolutionary mutations. However, fairly recent research has suggested that non-coding DNA could impact downstream genes in the genome[1].

Past studies find these significant portions of non-coding DNA (motifs) by manually comparing sections of non-coding DNA across multiple species[2], but this is tedious and requires sifting through thousands of basepairs. Genetic Algorithm for Motif Inference (GAMI)[3,4] uses a genetic algorithm to find motifs that have a high likelihood of being functional in a shorter amount of time. These motifs can then be passed on to biologists to for further analysis as to the actual biological functionality of the motifs.

An issue that arises as a result of using genetic algorithms in this manner is that the final population is frequently so large that it is still difficult for all the motifs found to be analyzed in lab work. Position bias has been suggested to be an indicator of functional elements[5] and as such, would be a valid means of further narrowing down viable motifs for lab work.

Another interesting factor is that in many cases, motifs remain functional even when portions of the motif do not match the "ideal" motif. In many cases, functional motifs allow for certain levels of damage tolerance[6], which allows for a number of errors within a motif. That is to say, a motif could be functional even if several basepairs within the motif does not

match the “ideal”. A different means of calculating overall scores for motifs that take into account this factor would allow for further analysis of the data.

My goal in doing this project was to develop a better means of locating and scoring motifs in non-coding DNA by further building on GAMI. With that in mind, I wrote 9 methods, 8 of which focused on scoring motifs more effectively by taking positional bias into consideration, and the last of which directly impacted the motifs found by changing the scoring value in the genetic algorithm. This last method focused on the idea that incomplete motifs[7] are often still functional.

Genetic Algorithms

Genetic algorithms (GAs) are loosely inspired by evolution in nature. GAs take a starting population and mutates them using various operators to create the next generation. Each member of each population is scored against a fitness function that measures how well they measure up against a set of criteria. In this case, the fitness function measures how well and how frequently each motif in the population finds a match in the data given. At the end of each cycle, the parents for the next generation are chosen randomly, with a preference for members that scored better.

The typical operators used in GAs are mutation and crossover. In GAMI, mutation works similarly to point mutation in nature, with a replacement of a single nucleotide in the motif. Likewise for crossover, with the exception that while crossover can occur at multiple points in nature, crossover occurs only once in GAMI.

Special operators used in GAMI are slide mutation and directed mutation. Slide mutation removes a nucleotide at one end of a motif and adds a random nucleotide at the other end, and directed mutation mutates a motif by replacing a nucleotide with another nucleotide that occurs more frequently in that location. In figure 1, for example, the fourth base mutated from a T to an A.

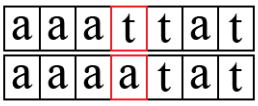

Figure 1: The base that mutated is highlighted in red.
The thymine nucleotide was replaced with an adenine nucleotide.

With each new generation, the members of the population would score better on average against the fitness function, and at the end of a set number of generations, the population would consist of better motifs than could otherwise generally be found within the same amount of time.

System Design

The existing post-processor for GAMI returns an HTML file for each motif, with a summary of matching motifs found in each sequence **(fig. 2a)**, points in the motif at which the motifs found in the sequence differs from the motif in question **(fig. 2b)**, as well as a graphical and numerical summary of the locations in the sequence at which motifs matching the motif in question were found **(fig. 2c)**.

Motif: **aaaaaaaaaaga / tctttttttttt**
Scores: GAMI Score: 225.00 PP Score: 195.00
MC: 195.00/228.00 is a **85.5%** match
PP scoring function: mc

Figure 2a: The motif and its reverse complement are given as well as various scores and the percentage match of the motifs found.

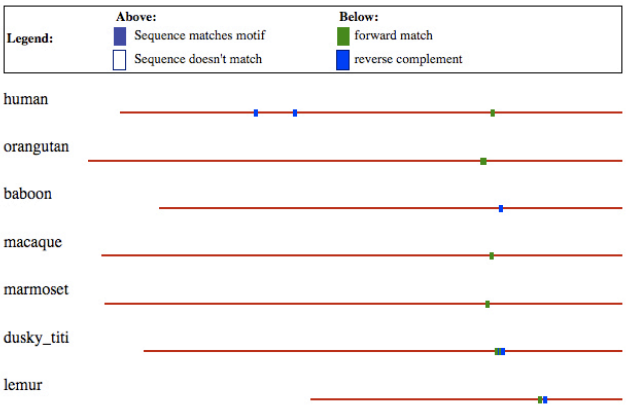

Figure 2c: The red lines indicate the total length of the DNA being analyzed, and the blue and green bars are sequences which match the motif.

Motif	a	a	a	a	a	a	a	a	a	a	g	a
orangutan							t				c	
orangutan							c				c	
orangutan				c							t	
baboon							g					
baboon			t									
marmoset											a	
marmoset											a	
marmoset											a	
marmoset											a	
marmoset											t	
dusky_titi							t				a	
dusky_titi						t					a	
dusky_titi				t							c	
dusky_titi						c						c
dusky_titi				t		t						
dusky_titi			g				g					
lemur	t	t			c							
lemur	t					t					c	

Figure 2b: The matching sequences are laid out in a table with the mutations highlighted.

I wrote several methods during this process. The first method (Method 1) looked solely at the rightmost motifs in the upstream DNA of nineteen different evolutionary divergent species. This method used the rightmost motif of the first species as a basis for comparison and compared the rightmost motif of all subsequent species to this first motif. If the motifs were within 500 basepairs of each other, the score was incremented by one. If a motif was found that was more than 500 basepairs away from the first motif, then the method ends. The final score assigned to the motif indicates the species that last fell within 500 basepairs of the first motif.

One of the problems that arose with Method 1 was that some species, such as the fugu, were so short that the motifs seldom came within 500 basepairs of the first sequence, which in the data was the human sequence. Method 2 deals with that problem by using a percentage comparison. Two motifs were considered positionally conserved if they fell within one percent of each other, with the percentage calculated using the second sequence. Another consideration was that the motif might have drifted due to evolution, therefore Methods 3 and 4 progressed down the phylogenetically ordered list in a pairwise fashion; that is to say, rather than using the first species as a basis for comparison for all other species, each

species was compared to the species directly above it in the list. Methods 5 and 6 calculated the total number of species that had a positional bias using a comparison margin of 500 basepairs and one percent respectively. Method 7 improves on the previous methods by looking at all available motifs in the subsequent species rather than focusing solely on the rightmost motif, and Method 8 capitalizes on Method 7 by repeating Method 7 on all motifs in the first sequence and returning the location and score of the motif with the best conservation. These methods could potentially be used in the fitness function to factor positional conservation into the GA.

Another method was also written that dealt directly with the calculation of the overall score for each motif. In this method, the first sequence was analyzed in the same manner as before, with the entire sequence being searched for motifs that match the given motif. The differences occur in the following sequences, with only a subsequence being searched for motifs. This subsequence is calculated using the percentage location of the rightmost motif in the first sequence, with a range of one percent to either side of the percentage calculated. This method finds motifs that might not match as the given motif as closely, but are definitely candidates for positional conservation.

Results

Overall, doing a pairwise comparison down the list of species seems to make more sense than scoring all other species against one specific species, but further testing with real data should be performed to verify this. Likewise, when comparing the distance of similar motifs using a percentage value rather than a specific number, using a percentage value seems to make more sense given that some species have far shorter genomes than others, but again, further testing on real data should be performed.

Although the first 8 methods written were successful and could potentially be useful in various situations, the last method is most likely to be the most useful. This method renders the previous 8 methods moot since it is able to look at a given region and find multiple partial matches.

Notes

---

[1] Frederick P. Roth, Jason D. Hughes, Preston W. Estep & George M. Church. "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." *Nature Biotechnology* **16** (1998): 939 – 945.
[2] Michael Z. "Ludwig, Functional evolution of noncoding DNA." *Current Opinion in Genetics & Development* **12.6** (2002): 634 – 639.
[3] C. B. Congdon, C. W. Fizer, N. W. Smith, H. R. Gaskins, J. Aman, G. Nava, C. Mattingly. "Prelimenary Results for GAMI: A Genetic Algorithms Approach to Motif Inference" *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (2005).
[4] Clare B. Congdon, Joseph C. Aman, Gerardo H. Nava, H. Rex Gaskins, & Carolyn J. Mattingly. "An Evaluation of Information Content as a Metric for the Inference of Putative Conserved Noncoding Regions in DNA Sequences Using a Genetic Algorithms Approach" *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5.1** (2008): 1 – 14.
[5] Olivier Elemento, Noam Slonim, & Saeed Tavazoie. "A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types" *Molecular Cell* **28** (2007): 337 – 350.
[6] P. Karran, M. Bignami. "DNA damage tolerance, mismatch repair and genome instability" *BioEssays* **16.11** (1994):833 – 839.
[7] In the context of this paper, incomplete motifs refer to motifs that do not match the "ideal" motif, which is the motif that is being used as the template for other motifs found in the non-coding DNA.