GAMIV: A Genetic Algorithm for Identifying Variable-length Motifs in Noncoding DNA

David J. Gagne
University of Southern Maine
Portland, ME 04104, USA
david.gagne1@maine.edu

ABSTRACT

GAMI uses a genetic algorithm to identify putatively conserved motifs of a pre-selected length in noncoding DNA from diverse species. In this work, I present an extension to the system, GAMIV, that identifies putatively conserved motifs of variable length. The system begins with an initial set of very short motifs and allows them to grow through a pair of custom operators. A fitness function that rewards both motif conservation and motif length is used to evolve a population of conserved motifs of variable length. This paper describes the motivation for GAMIV, discusses the design of the system, and presents initial results for the system. Based on these initial results, GAMIV is a promising tool for the inference of variable-length motifs in noncoding DNA.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics; I.2.8 [**Artificial Intelligence**]: Problem solving, Control methods, and Search

General Terms

Algorithms

Keywords

Evolutionary computation, genetic algorithms, DNA motif inference

1. INTRODUCTION

Only an estimated 5% of the human genome is known to code for proteins [8]; the remaining (noncoding) regions contain relatively short functional patterns of DNA that help regulate the expression of nearby genes by providing binding sites for regulatory proteins (transcription factors). GAMI [4], [3] (Genetic Algorithms for Motif Inference) is a system that identifies candidate functional elements using a genetic algorithms (GA) search. Following the notion that functional elements are more likely to remain conserved across evolutionary time than the surrounding background sequences, GAMI searches for DNA patterns (motifs) conserved across a set of orthologous (related to the same gene) genetic sequences from diverse species. In a single run, GAMI identifies candidate motifs of only a single length; since functional elements vary in length, the user must guess at the length of the elements prior to a run. In this work, I introduce GAMIV, an extension of GAMI that identifies conserved motifs of arbitrary length, so that the expected length of the motifs need not be known in advance.

In the Background section of this paper, I provide a brief overview of existing approaches to motif inference and describe the GAMI approach in further detail. In the System Design section, I describe the changes made to the GAMI system in order to evolve motifs of variable-length, including a fitness metric that considers both motif conservation and motif length, as well as additional operators for varying the lengths of motifs. In the Methodology section, I describe the methods used to conduct experiments comparing GAMI with GAMIV. In the Results and Conclusion sections, I describe and analyze the outcome of the experiments, and present conclusions drawn from the results. Finally, I consider future directions for the work presented in this paper.

2. BACKGROUND

The degree to which a gene is expressed in a higher organism may vary across different tissues in the organism, over time, and in response to changes in environment. Gene expression is primarily regulated by proteins known as transcription factors, which promote or inhibit the expression of a gene by binding to a host site in the noncoding regions near the gene. The identification of these binding sites is an important step in understanding the complex expression profiles of many genes, yet many regulatory elements remain unidentified in the regions far upstream or far downstream of the gene they regulate. Traditional laboratory methods for the detection of binding sites, such as DNASE I hypersensitivity and gel shifts, are expensive and time consuming, and thus are not well-suited to the exploration of the large noncoding regions found in the human genome. Given the limitations of experimental methods for binding site discovery, the computational inference of regulatory elements plays an important role in narrowing the field of candidates for laboratory assay.

Many computational systems exist for identifying candidate functional elements in noncoding DNA; a review of several such systems is given in [14]. One common approach

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.

is to identify nucleotide patterns (motifs) conserved across a set of sequences from co-expressed genes, such as in [12], since genes that are expressed concurrently are more likely to share common transcription factor binding sites [1]. Another approach is to identify nucleotide patterns conserved across a set of sequences from orthologous genes (genes that serve equivalent functions in different species).

Multiple studies suggest that functional elements are more likely to remain conserved across evolutionary time than the background noncoding regions [5], [11]. Further evidence suggests that conservation across species separated by greater evolutionary distance provides stronger support that a motif is functional, since this helps distinguish between regions that are functional and regions that have not had sufficient evolutionary time to diverge [13]. In [10] Lones and Tyrell provide an excellent review of computational approaches to motif inference, with a focus on evolutionary computation as an approach. As the authors note, many tools rely on a global sequence alignment as the first step in motif elicitation. But global sequence alignment is computationally expensive, especially for multiple long sequences, and may be problematic for sequences with poor overall conservation [7], such as those from diverse species. Some approaches to motif inference, such as MEME [2] and [6], search for sets of window-locations that are highly conserved across the sequences. While this approach avoids the problems associated with a global sequence alignment, the search space becomes computationally intractable for long sequences or large sets of sequences.

Few systems identify motifs of variable length. In [9], Kaya proposes the use of a multi-objective genetic algorithm for motif discovery (MOGAMOD) to identify conserved motifs of varying length. Because the MOGAMOD approach searches for sets of highly conserved window-locations of variable length, this approach suffers from the same limitations in sequence length as MEME and other window-based searches.

2.1 GAMI

Rather than search the space of possible window-locations, GAMI uses a genetic algorithms approach to search the space of possible motifs. Because the search space does not increase with the lengths of the sequences and because this approach does not rely upon a global sequence alignment, GAMI is well suited to working with a large number of potentially long sequences from evolutionarily diverse species. In [4], the authors demonstrate GAMI's ability to find highly conserved motifs in sets of orthologous sequences. In a given run, GAMI only searches for motifs of a single, user-specified length; yet the lengths of functional elements vary. In this project, I present an extension to the GAMI system, GAMIV, that identifies putatively conserved motifs of variable length.

GAMI represents a motif as a pattern of nucleotides, and a motif is assigned a fitness score based on how well it is represented across the set of sequences being searched. For example, the GAMI population might include the following pattern of 8 nucleotides (an 8-mer): AACTGCAA. The fitness score for a GAMI motif is based on the sum of the best matches for the motif from each sequence, with imperfect matches allowed. For instance, if the best match for the above motif in a given sequence is AGCTGCAA, then the motif would receive a match score of 7 for the sequence, and the maximum possible match count for the motif across a set of five sequences would be 40. Since the match count score is dependent on both the length of the motif and the number of sequences in the data set, the percentage of the maximum possible match score is used to measure the conservation of a given motif. If our example motif had a match count of 36 across a set of 5 sequences, then the motif would receive a fitness score of 90%. In [3], the authors found that this simple match-based scoring function led to the identification of motifs better conserved than those identified when GAMI used a more traditional information content scoring function.

3. SYSTEM DESIGN

Several changes were necessary in order to extend GAMI to search for motifs of variable length. First, it was necessary to modify the internal structure of the GAMI system to accommodate motifs of variable length. Several approaches are possible for searching the space of motifs of varying length. One method would be to generate a population of random motifs of varying length and to evolve the population using the traditional crossover and mutation operators. An alternative is to begin the search with a population of very short motifs and to use operators that allow the motifs to increase in length. GAMIV uses the latter approach, since a series of small changes to a short, well-conserved motif is more likely to produce a well-conserved long motif than would the generation of a long motif at random. I introduce two new genetic operators in order to allow motifs to increase length. In addition, GAMI's fitness function does not consider the length of a motif; in order to compare motifs of variable length, it was necessary to design a fitness function that incorporates both the length and the conservation of a motif.

3.1 Evolutionary Operators

GAMIV implements two custom operators in order to allow motifs in its population to change length. The first is a mutation that increases the length of a motif by a single nucleotide. A randomly selected base is appended to a motif selected for this operation, with an equal probability of adding the base to the start or end of the motif. The second operator merges together overlapping motifs. At the start of a run, the user specifies a sequence to be used in identifying overlap between motifs. When a motif is selected to undergo the merge operation, the system identifies the best-matching location(s) for the selected motif in the user-specified sequence (more than one such location may exist). The operation searches the population for another motif whose best-matching location in the sequence overlaps that of the motif in question, and a new motif is formed by merging the two. Since this operation uses co-location in a sequence as the criteria for merging, it allows motifs to be merged even when the nucleotides in the overlapping regions do not match exactly. If nucleotides in the overlapping regions differ between the motifs, then the nucleotides from the higher scoring motif are used. This merge operation produces a longer motif that is supported in at least one of the sequences.

3.2 Fitness Measurement

A long motif that has been conserved across evolutionary time is a stronger candidate for functionality than a shorter motif of equal conservation, since a long motif is less likely to remain conserved by pure chance. The match percentage metric used by GAMI does not consider the length of a motif, while the match count metric gives too much weight to longer motifs, without a penalty for a loss of conservation. For example, consider adding a base to the example motif mentioned in the previous section to make it AACTGCAAC. If the final nucleotide matches in just one of the five sequences, the motif is considered more fit by the match count metric, but less fit by the match percentage metric. In order to balance the trade-off between length and conservation, we introduce a parameterized scoring metric that allows adjustments to the weight given motif length versus the weight given conservation. A linear weighting scheme only preserves a balance between length and conservation over a small range of lengths, but when the length of the motif reaches a critical point, it dominates the score. I found that applying the weights as exponents of the match percentage and motif length, and then taking the product of the values resulted in a more stable scoring function. In a series of preliminary trials, I found that multiplying the cube of the match percentage by the motif length resulted in a fitness function that promoted the growth of motifs without dominating conservation.

4. METHODOLOGY

In order to assess GAMIV's ability to identify motifs of variable length, I compared the performance of the extension with the performance of the original system using data sets containing motifs of variable length. As a preliminary assessment of the system, I used fifteen sets of artificial sequences with variable-length motifs implanted, which were produced in the manner described below. GAMIV was run twenty times on each data set, using a different seed to the random number generator for each run. Rather than run GAMI once for each motif length ranging from 5-mers to 35-mers, system was run twenty times on each data set using motif lengths of 9, 12, 18, and 27 bases.

4.1 Data

Each artificial data set consisted of ten background sequences of 10k randomly selected bases (a uniform distribution of nucleotides was used). Twenty motifs varying in length from 5-mers to 35-mers were implanted at random, non-overlapping locations in each of the sequences, so that each of the motifs would be fully-conserved across the set of sequences.

4.2 Parameter Settings

A GA run for either system used a population of 1000 motifs and a maximum of 50,000 trials (calls to the fitness functions). In both cases, a mutation rate of 2% was used, with 25% of the population preserved by elitism in each generation. GAMI uses two non-traditional mutation operators: a slide mutation that removes a base from one end of the motif and adds a randomly selected base to the opposite end, and a directed mutation that replaces one base in the motif with the nucleotide that produces the best score for the motif. For both systems, a directed mutation rate of 0.4 and a slide mutation rate of 0.2 were used. GAMI was run using a crossover rate of 60% and using the match percentage of a motif as the scoring metric in four separate runs with motif lengths of 9, 12, 18, and 27. GAMIV was run using the scoring metric described previously in the system design section, with an exponential weight of 3 given to the match percent score. In each GAMIV run, the initial motif length was set to 5 bases, and a length mutation rate of 0.3 was used, along with a merge mutation rate of 0.5.

Table 1: Implanted Motifs Identified By GAMIV

Data Set	Location ID(%)	Exact Match ID(%)
A	100	40
B	100	65
C	100	45
D	100	65
E	100	45
F	100	75
G	100	75
H	100	70
I	100	50
J	100	65
K	100	65
L	100	55
M	100	50
N	100	45
O	100	55
Average	100	58
Standard Deviation	0	12

5. RESULTS

For each data set, I determined whether the location of the implanted motif was represented in the results. An implanted motif was considered identified by location if one of the motifs identified by the system was co-located with the implanted motif in one of the sequences. To test this, I compared the best-matching locations of each motif to the location of the implanted motif. If the result motif overlapped the implanted motif in one of the sequences, then the location of the implanted motif was considered identified. Both systems were able to identify the locations of all 20 of the implanted motifs in each of the 15 data sets.

For each data set, I also tested whether the GAMIV results for the data set included an exact match for the implanted motif. The results of these measurements are shown in Table 1. GAMIV success rate for producing exact matches to the implanted motifs averaged 58%, with a high of 75% and a low of 40%. The standard deviation for the exact motif identification rate over the fifteen data sets was 12%.

6. CONCLUSIONS

While the GAMIV extension does not appear to improve GAMI's ability to identify putatively conserved motifs by location, it appears to perform comparably to the original without the need for multiple runs to determine the ideal motif length, leading to a reduction in run time. While GAMIV successfully identified the exact matches for the implanted motifs only 58% of the time, this is a promising initial result. I did not perform an extensive parameter sweep for GAMIV, choosing instead to base the parameter settings off of a few exploratory runs. For example, GAMIV may have produced better results without the use of GAMI's additional operators (slide mutation and directed mutation), or it may have identified more motifs exactly if I had allowed the tests to run for more trials. In addition, I believe that GAMIV would benefit from a more directed length mutation operator. Rather than add a random nucleotide to the motif, a directed length mutation would add the nucleotide that most increases the motif's fitness score. The fact that GAMIV was able to find the exact motifs more than half the time is promising, since the search space is much larger than the search space for a single motif length.

7. FUTURE WORK

The work presented here represents an early exploration of this approach to identifying variable-length motifs. I plan on conducting a set of parameter sweeps, as this is likely to improve search results. In addition, the length mutation operator adds a random base to the motif; the search might be improved by selecting the base that best improves the motif's fitness score. Similarly, the merge operation could be improved by choosing merges that result in the greatest increase in fitness. Finally, I would like to conduct farther tests of this approach, including side-by-side comparisons with other motifs inference systems.

8. ACKNOWLEDGMENTS

Thanks to Dr. Clare Bates Congdon for her continued guidance on this project. This project was supported by the Undergraduate Research Opportunities Program at USM, a National Science Foundation CAREER award (#953495), grants from the National Center for Research Resources (5P20RR024475-02) and the National Institute of General Medical Sciences (8 P20 GM103534-02) from the National Institutes of Health, and Maine Economic Improvement Funds.

9. REFERENCES

[1] D. J. Allocco, I. S. Kohane, and A. J. Butte. Quanitfying the relationship between co-expression, co-regulation, and gene function. *BMC Bioinformatics*, 5:18, Feb 2004.

[2] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. Meme: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, 34:W369–W373, Jul 2006.

[3] C. B. Congdon, J. C. Aman, G. M. Nava, H. R. Gaskins, and C. 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 Trans. Comput. Biol. Bioinform.*, 5:1–14, 2008.

[4] C. B. Congdon, C. Fizer, N. W. Smith, H. R. Gaskins, J. C. Aman, G. M. Nava, and C. J. Mattingly. Preliminary results for GAMI: A genetic algorithms approach to motif inference. In *Proc. CIBCB*, pages 97–104, 2005.

[5] I. Dubchak, M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.*, 10:1304–1306, Sep 2000.

[6] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, H. B. Harlow, J. E. Onyia, and C. Su. Discovery of sequence motifs related to co-expression of genes using evolutionary computation. *Nucleic Acids Res.*, 32(13):3826–3835, 2004.

[7] P. P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8):2433–2439, 2005.

[8] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, Oct 2004.

[9] M. Kaya. Mogamod: Multi-objective genetic algorithm for motif discovery. *Expert Systems with Applications*, 36:1039–1047, 2009.

[10] M. A. Lones and A. M. Tyrell. The evolutionary computation approach to motif discovery in biological sequences. In *Genetic and Evolutionary Computation Conference (GECCO 2005); Workshop on Biological Applications of Genetic and Evolutionary Computation.* ACM SIGEVO, 2005.

[11] L. A. Pennacchio and E. M. Rubin. Comparative genomic tools and databases: providing insights into the human genome. *J. Clin. Invest.*, 111(8):1099–1106, Apr 2003.

[12] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 30(24):5549–5560, Dec 2002.

[13] J. W. Thomas and J. W. Touchman. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.

[14] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23:137–144, Jan 2005.