An experimental comparison of two different paradigms in Evolutionary Computation

Giovanna Martínez Arellano and Carlos A. Brizuela
Computer Sciences Department
CICESE Research Center
Km 107, Carr. Tijuana - Ensenada, Ensenada, B.C., México
Phone: +52-646-1750500
Email: {gimartin, cbrizuel}@cicese.mx

Abstract

The DNA motif finding problem is of great relevance in molecular biology. Motifs play an important role in all biological processes since they control the production of certain proteins by turning on and off the genes that codify them. These motifs consist of a short string of unknown length that can be located anywhere throughout the genome. This fact turns the problem much more difficult, so to find these sites, a set of DNA sequences, of orthologous genes or genes that are known to be controlled by the same motif, is taken and the strings that are more alike in all sequences are considered ocurrences of the same motif. Therefore, the main idea of this problem is to discover short, conserved sites in genomic DNA without knowing, *a priori*, the length nor the chemical composition of the site, turning the original problem into a combinatorial one, where computational tools can be applied to find the solution.

Pevzner and Sze [1] studied a precise combinatorial formulation of this problem, called *the planted motif problem*, which is of particular interest because it is intractable for commonly used motif-finding algorithms [2]. This algorithmic challenge consists of finding twenty planted occurrences of a motif of length fifteen in roughly twelve kilobases of genomic sequence, where each occurrence of the motif differs from its consensus in four randomly chosen positions. Pevzner and Sze [1] introduced new algorithms to solve their (15,4)-motif challenge, but these methods do not scale efficiently to more difficult problems in the same family, such as the (14,4)-, (16,5)-, and (18,6)-motif problems. Due to this results, many other approaches have been sugested, among them, Random Projections [2] and Pattern Branching [3] seem to be the most effective ones.

On the other hand, evolutionary computation is known to be a good strategy to solve combinatorial problems. For this reason, in this work we study the performance of this kind of strategy on the planted motif finding problem. To do this, we design two algorithms. The first one is based on the basic paradigm of genetic algorithms, while the second

one, and since Pattern Branching performs well regarding computation time and solution's quality, we decide to borrow some ideas from this approach to codify the individuals and to measure their fitness. We test the performance of both algorithms in planted motif instances such as (10,2),(15,4), (16,5), and (18,6).

Preliminary experimental results show a superior performance of the algorithm based on the Pattern Branching ideas over the standard one. This will motivate a comparative study of this new approach with the state-of-the-art methods.

References

[1] Pevzner, P., and Sze, S.-H. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology*, 269-78.

[2] Buhler, J., and Tompa, M. 2002. Finding Motifs Using Random Projections. *Journal of computational Biology*, Volume 9, Number 2, 225-242.

[3] Price, A., Ramabhadran, S., and Pevzner, P. 2003. Finding Subtle Motifs by Branching from Sample Strings. *Bioinformatics.* Volume 1, Number 1, 1-7.