Reviewer's report

**Title:** Finding Gapped Motifs by A Novel Evolutionary Algorithm

**Version:** 1  **Date:** 6 July 2010

**Reviewer:** Jesus Aguilar-Ruiz

**Reviewer's report:**

This paper proposes a method based on PSO for finding motifs in DNA sequences, where motifs can also contain gaps.

The idea presented in the paper is interesting, and the fact of allowing gaps to be present within motifs is an important limitation that the proposed approach tries to solve.

The paper is generally well written and well structures, even if some parts would need some revisions, for example the section entitle "GAPPED MOTIFS" has to be reviewed. In fact, in its current state, it is rather confused.

The paper presents, however, some limitations. The main one being the experimentation proposed.

Both synthetic and real data has been used. I think that the use of synthetic data is indeed helpful when one wants to prove an important aspect of the algorithm. I would have expected to see synthetic data where motifs contained gaps of different types and length, and to see that PSO+ could outperform other state of the art algorithms on these kind of data. Unfortunately, the experiments conducted on the synthetic data, are only used to prove that PSO+ has a lower running time. I could not find other measures for this data, e.g., precision, recall. I believe that running time is an important aspect, but definitely it is not the most important one. I prefer to obtain better results even if the algorithm takes longer to run.

The experimentation performed on the real datasets is also not very well performed. Basic information, like standard deviation, is missing. Authors do not provide information on how many runs of the two GAs have been executed.

No statistical test were used in order to check if the differences in
the results were significant. The discussion of the results is also poor. Authors also claim, in the abstract, on page 6 and in the conclusions that they propose a modification of the standard PSO in order to handle discrete values. This is not really a novelty, so it’s not something authors propose, but rather adopt. There are other PSO algorithms that can do that, e.g., "A discrete particle swarm optimization algorithm for the no-wait flowshop scheduling problem, Quan-Ke Pana M. Fatih Tasgetiren and Yun-Chia Liangd", 2007, while the first proposal in this sense was "J. Kennedy and R. Eberhart. A discrete binary version of the particle swarm algorithm. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pages 4104-4108, IEEE Press, Piscataway, NJ, 1997. " Authors should specify if there is any existing methods for handling discrete domain which is similar to the strategy adopted by PSO+. Another aspect that surprised me was how gaps are handled. Authors, in the introduction claim that one of the weakest point of other algorithm is that they allow to handle gap only if the user provide the length and position of the gap. So, one would expect that the proposed algorithm would not require the user to specify any information about length and position. However, the position is hard wired into the algorithm, and user has to provide the maximum length of gaps. So I cannot see the motivations behind the proposal of PSO+. OTHER COMMENTS *) title "Novel EA" I think the use of word "novel" is inappropriate, since the algorithm is basically a PSO algorithm. *) page 2 "Second, we use both consensus and PWM representations in our algorithm, taking advantage of the efficiency of consensus and the accuracy of PWMs." This is not really a novelty of the algorithm, it's a novel representation *) in abstract " taking advantage of the efficiency of the former and the accuracy of the later" -> taking advantage of the efficiency of the former and the accuracy of the latter *) page 1. "In general, PWM offers a more accurate description of motifs
than consensus sequences, but is more difficult to optimize" what is it more difficult to optimize?

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.