

## **Reviewer's report**

**Title:** Partitioning clustering algorithms for protein sequence data sets

**Version:** 2 **Date:** 3 February 2009

**Reviewer:** Zied Elouedi

### **Reviewer's report:**

1. Is the question posed by the authors new and well defined?

Authors propose four partitioning clustering approaches using Smith-Waterman local alignment algorithms in order to find pair-wise similarities of sequences. Several methods are used in clustering of protein sequences. However, most of them are either hierarchical or graph-based approaches, whereas few applications have been found in the field of protein sequence clustering based on partitioning like what is proposed in this paper. So, the question handled by authors is original and well defined.

2. Are the methods appropriate and well described, and are sufficient details provided to replicate the work?

In general, what is presented by the authors in this paper is appropriate and well described. So they describe thoroughly the four clustering methods namely Pro-kmeans, Pro-LEADER, Pro-CLARA, Pro-CLARANS. Algorithms' codes are also presented facilitating consequently the replication of this work.

3. Are the data sound and well controlled?

Data used in experimentations are various. So, four datasets are experimented where each one presented a different number of sequences and also different classes. These data are sound and well controlled.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?

The manuscript starts by an abstract, the methods and results. Then, the background is presented, the developed methods are detailed. Next, results and discussions are proposed and finally authors end their paper by a conclusion summarizing the different developments.

Hence, the manuscript adheres to the relevant standards for reporting and data deposition.

5. Are the discussion and conclusions well balanced and adequately supported by the data?

The section relative to discussions and conclusions is interesting and well balanced and adequately supported by the data.

6. Do the title and abstract accurately convey what has been found?

The title and the abstract accurately convey the content of the paper dealing partitioning clustering algorithms for protein sequence

7. Is the writing acceptable?

Except some minor error mentioned below, the paper is well written.

#### Minor Essential Revisions

1- There are some minor mistakes to correct

- Page 4

Replace "they can not be applied" by "they cannot be applied"

Replace "Methods presented bellows" by "Methods presented below"

Replace "Methods defined bellows" by "Methods defined below"

- Page 6:

Replace "The Pro-Kmeans algorithm proposed here, start by a random partition of the data set

D into K clusters and then use the Smith Waterman algorithm" by "The Pro-Kmeans algorithm proposed here, starts by a random partition of the data set D into K clusters and then uses the Smith Waterman algorithm"

- Page 7:

Replace "The algorithm detect the nearest leader" by "The algorithm detects the nearest leader"

Replace "and compare the score" by "and compares the score"

- Page 9

Replace "Pro-CLARA use the optimal set of medoids" by "Pro-CLARA uses the optimal set of medoids"

- There two equations number 7 so the number of these equations should be corrected

#### Major Compulsory Revisions

1- I agree that there are few clustering methods based on partitioning techniques applied in protein sequence classification. However, this choice should be more motivated in other words why it is interesting to do such developments since there already exist some developments within hierarchical, graph-based clustering methods dealing with the clustering of protein sequence.

2- Leader is presented as an incremental partitioning clustering method whereas for example k-means is considered as a non-incremental partitioning clustering. So how to explain the use of these two different manners of clustering and basically the incremental one i.e what is its interest in your case of protein

sequence. Besides, it is not clear that in your pro-LEADER algorithm you have made it as incremental or non-incremental

3- Equation (7) in page 9 should be more explained especially the different used variables and values

4- In experimentations, it is not clear how you have obtained the testing set.

#### Discretionary Revisions

1- what about the computational complexity of the developed methods in this paper and more exactly their time complexity.

2- Comparisons of partitioning clustering methods for protein sequence with hierarchical and graph-based clustering methods may be done in future works

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

I declare that I have no competing interests