

Frame-Cluster Mapping: Genomewide Analysis for Detection of Localized Signal Regions

Kou Amano^{1,2} **Hidemitsu Nakamura**² **Hisataka Numa**²
amano@slis.tsukuba.ac.jp hide7087@affrc.go.jp hisataka@nias.affrc.go.jp
Yoshiaki Nagamura² **Hiroaki Ichikawa**² **Kaoru Fukami-Kobayashi**³
nagamura@nias.affrc.go.jp hichkw@affrc.go.jp kfukami@brc.riken.jp

¹ University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan

² National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

³ RIKEN BioResource Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Keywords: clustering, genomewide sequence analysis, oligonucleotide

1 Introduction

Short characteristic sequence patterns found on a genome are often called “signals” whether their functions are known or not. Not only the presence of such signals but also their localizations on the chromosomes should be redefect chromosomal structure and genomic functions.

Existing methods can map sequence patterns to visualize their genomic positions[1], if the patterns are specified *a priori*. When the target signals are not known, however, various sequence patterns can be candidates for signals, and therefore it is difficult to visualize all of them.

We propose a new method, Frame-cluster mapping (FCM), to evade such problems, and show the results of pilot tests. In our method, genome sequence is fragmented into “*frames*”, which are further “*clustered*” into several groups based on their sequence characteristics and “*mapped*” to their genomic positions. If the clustered frames are localized in the genome, the frames should have some meaningful signals in their sequences. By detecting such clusters localized in a genome, our method can find any signal candidates whether they are known or not.

2 Materials and Methods

FCM clusters oligonucleotide patterns of frames and plots their distributions in the genome sequence in the form of such as a dot plot and a histogram. Basic protocol of FCM is:

1. Fragmenting whole chromosome sequence into small frames with the same size.
2. Discarding unsuitable frames (i.e. those containing a lot of character 'N's).
3. Calculating oligonucleotide frequencies of all the remaining frames.
4. Clustering the frames based on their oligonucleotide frequencies.
5. Mapping individual frames in each cluster to their genomic positions.

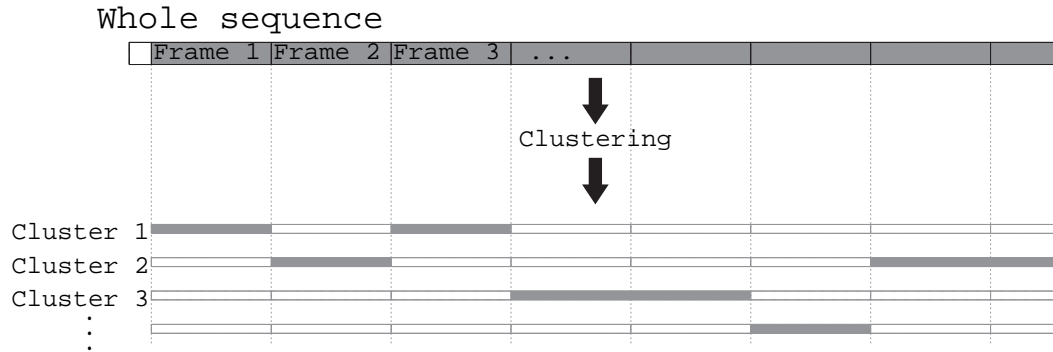


Figure 1: Concept of FCM.

A genome sequence is fragmented into frames, which are the same size and have no overlaps. All the frames of each cluster are separately mapped to their genomic positions.

Then, a dot plot or a histogram is obtained from each cluster. To cluster the frames, Self-organizing clustering (SOC)[3] was used. To make graphic images, Mathematica(R) was used. Figure 1 shows the concept of FCM.

Using FCM, we tested several data sets (species) of whole genome sequences from DDBJ/EMBL/GenBank or TIGR database. The number of selected species was 15. These species were: *Apis mellifera*, *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Oryza sativa*, *Pan troglodytes*, *Plasmodium falciparum*, *Rattus norvegicus*, and *Tribolium castaneum*.

3 Results and Discussion

We obtained some remarkable results as follows.

- In the analysis of 94,460 frames with 1 kbp in length from all the six chromosomes of *C. elegans* using tetranucleotide frequencies, FCM detected the TTAGGC repeat (CeRep26)[2] region, although the presence of the repeat was not known *a priori*.
- In the analysis of 118,962 frames with 1 kbp in length from all the five chromosomes of *A. thaliana* using pentanucleotide frequencies, FCM detected the centromere region of each chromosome, although the presence of the centromeric repeat was not known *a priori*.

We also obtained interesting results from analyses of the other 13 species. Thus, we conclude that Frame-cluster mapping method is effective for detecting localized signals in genome sequences.

References

- [1] Berger, J.A., Mitra, S.K., Carli, M., and Neri, A., *Proc. GENSIPS*, 2002.
- [2] Wicky, C., Villeneuve, A. M., Lauper, N., Codourey, L., Tobler, H. and Müller, F., Telomeric repeats (TTAGGC)_n are sufficient for chromosome capping function in *Caenorhabditis elegans*, *Proc. Natl. Acad. Sci. USA*, 93(17):8983–8988, 1996.
- [3] Amano, K., Nakamura, H., and Ichikawa, H., Self-Organizing Clustering: A novel non-hierarchical method for clustering large amount of DNA sequences, *Genome Informatics*, 14:575–576, 2003.