Poster H-14

Markov Models of Genome Segmentation

Authors:
Vivek Thakur (*School of Information Technology, Jawaharlal Nehru University, India*)
R. K. Azad (*Dept. of Biological Sciences, University of Pittsburgh, USA*)
R. Ramaswamy (*School of Physical Sciences, Jawaharlal Nehru University, India*)

Short Abstract: Entropic segmentation dissects the DNA sequences into compositionally homogeneous sequences and finds application in locating the boundaries of CpG islands, isochores etc. It however has limitations in resolving boundaries when tested for biological chimeric DNA sequences of different origin. To improve its performance we implemented Markovian approach and observed substantial rise in sensitivity and specificity.

Long Abstract:

Availability of increasing number of complete genomes of prokaryotes and eukaryotes has triggered the analysis of genomic sequences for various biological and statistical features. Among the statistical features the heterogeneity of nucleotide composition within genomes has been of interest for quite long. The method of sliding window has been used extensively for studying the heterogeneity, in terms of G+C content, of DNA sequences. This, however, has methodological limitations in selecting window size as well as the sliding size, which in turn affects the results. Genome segmentation methods, on the other hand, partition the genomes into compositionally homogeneous segments and are free from the above limitations. In order to obtain optimal segments several strategies, based either on multiple change-point problem or Hidden Markov Models, have been applied [1]. These methods vary considerably in explaining the biological relevance to the large set of segments produced. One of these methods, the entropic segmentation, is based on multiple change point problem and generates segments which are homogeneous with respect to a chosen criteria but heterogeneous to its immediate neighbors [2]. It uses Shannon-entropy to describe the distribution of putative segments and computes the difference between entropies of a selected pair of neighboring segments by Jensen-Shannon divergence. Several applications of this method for biologically well charecterized features have been reported [3]; it identifies the boundaries of CpG islands and isochores with higher precision, locates the replication origin, helps in charecterizing the degree of complexity in the (compositional) organization of prokaryotic genomes etc. In an in-silico experiment it successfully identified regions of a genome artificially inserted into another evolutionary distant genome [4].

The genomes have very complex organization and are presumed to have many more underlying homogeneous strucures than obtained using criteria such as G+C content or ATGC content. We proposed consideration of DNA sequences to have originated from a Markov source as an attempt to improve upon the existing model. For the validation of our proposal we generated an ensemble of chimeric sequences as a set of heterogeneous sequences and these were subjected to different segmentation models. We observe that the higher order segmentation model performs better than others. Their performance was more

pronounced when the G+C content of segments forming chimera was equal or closer. This finding reveals ability of the higher order model in distinguishing the genomic fragments of different evolutionary origin accurately. This property can be used in locating the DNA fragments which have been horizontally inserted. Such distinction being local in nature can also help in finding the events of chromosomal rearrangements.

Refrences:

1. J. V. Brown and H.-G. Muller, Ststist. Sci. 13, 2 (1998).
2. P. Bernaola Galvan, R. Roman Roldan, and J. E. Oliver, Phys. Rev. E 53,5181 (1996).
3. W. Li, P Bernaola Galvan, F. Haghighi, and I. Grosse, Comput. Chem. 26, 491 (2002).
4. R. K. Azad, J. Subba Rao, W. Li, and R. Ramaswamy, Phys. Rev. E 66, 031913 (2002).