Classification of genomic regions by chaos game representation images and fractal dimension

Caetano Jimenez Carezzato
DCC-IME-USP, University of São Paulo – Brazil – caetano@vision.ime.usp.br
Junior Barrera
DCC-IME-USP, University of São Paulo – Brazil – jb@ime.usp.br
Sandro José de Souza
ILPC – Brazil – sandro@compbio.ludwig.org.br
Luciano da Fontoura Costa
IFSC-USP, University of São Paulo – Brazil – luciano@if.sc.usp.br

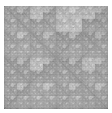
Abstract

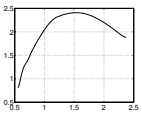
This work describes a new approach to classify genomic regions by applying the multiscale fractal dimension [1] over images generated by chaos game representation (CGR) of sequences. Since the introduction of chaos game representation of sequences [4], it has been found evidences that it is possible to obtain discriminative measures from images produced by this methodology and to feed such features into classifiers in order to identify the origin of gene fragments. Also, it has been showed [2] that it is possible to reconstruct filogenetic trees just using this representation.

In this project, we propose a new feature extractor of sequences that can help the classification of genomic regions. The feature extraction consists of determining the CGR of a sequence and estimating the fractal dimension of the generated image. Such measurements are organized into a feature vector.

For certain genomic regions, as the GC-content (mean percentage of guanine and cytosine) changes, the CGR also changes. We are going to verify if CGR contains more meaningful information than GC-content that can be easily and fastly exploited for genomic regions classification. Our preliminary results based on cluster techniques show that methods based on this approach should be better than GC-content based ones. We compared, for each approach, the ability to distinguish genic regions of the human chromossome 22. Currently, we are finishing the validation of the tests and developing a way to classify genic regions.


DNA sequence


CGR image



fractal dimension

References

[1] Costa, L. da F., Manoel, E.T.M., Faucereau F., Chelly J., van Pelt J., and Ramakers G. A shape analysis framework for neuromorphometry. Network, 13, 283-310, 2002.

[2] Deschavanne P.J., Giron A., Vilain J., Fagot G., and Fertil B. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. Molecular Biology Evolution, 16(10):1391–1399, 1999.

[3] Duda, R.O., Hart P.E., and Stork D.G. Pattern Classification. John Wiley and Sons, 2000.

[4] Oliver J.L., Bernaola-Galván P., Guerrero-García J., and Román-Roldán R. Entropic profile of DNA sequences trought chaos-game-derived images. Journal of Theorical Biology, 160:457–470, 1993.