
Segmentation of DNA sequences using Normalized Maximum Likelihood models for uncovering gene duplications

Ioan Täbuş

Department of Signal Processing
Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland
email: ioan.tabus@tut.fi
web: www.cs.tut.fi/~tabus

Abstract

The normalized maximum likelihood (NML) model [2]-[4] for a class of Markov sources [6] was recently used for the compression of full genomes, obtaining for the human genome the best existing compression results [1]. We show that one of the underlying biological features that the compression algorithm implicitly uncovers is the existence of approximate gene duplication. We proposed a refined method based on the same NML models for the segmentation of DNA sequences for uncovering gene duplications [5]. Several analysis tasks in genomic sequences involve preliminary segmentation or clustering of the data, which can be performed by a number of techniques, based on various similarity measures. Here we review and further pursue the application of MDL techniques for genomic sequence analysis. The process of sequence matching will be used for solving the problem of uncovering gene duplications with the help of a preliminary segmentation of a complex DNA locus, known to have evolved through a series of duplications.

References

- [1] G. Korodi, I. Tabus, "Normalized maximum likelihood model of order-1 for the compression of DNA sequences", in Proc. IEEE Data Compression Conference, DCC'07, pp:33 - 42, Snowbird, 27-29 March 2007.
- [2] J. Rissanen, "Fisher information and stochastic complexity", *IEEE Transactions on Information Theory*, vol. IT-42, pp. 40-47, Jan. 1996.
- [3] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data", vol. IT-47 (5), pp. 1712-1717, 2001.
- [4] Y.M. Shtarkov, "Universal sequential coding of single messages". Translated from *Problems of Information Transmission*, Vol. 23, No. 3, 3-17, July-September 1987.
- [5] I. Tabus, Y. Yang, J. Astola, "Universal models with memory for genomic sequence analysis", 3rd International Symposium on Communications, Control and Signal Processing, ISCCSP 2008, March 12-14, St. Julians, Malta, 2008.
- [6] I. Tabus, G. Korodi, "Genome compression using normalized maximum likelihood models for constrained Markov sources", *IEEE Information Theory Workshop*, Porto, Portugal, May 5-9, 2008.