IMAGE: a new tool for the discovery of Transcription Factors binding sites

Paparcone R (1), Casilli R (1), Melchionna S (2), Marongiu A (1,3),
Palazzari P (1,3), Rosato V (1,3)

(1) Ylichron Srl, c/o ENEA Casaccia Research Center, Via Anguillarese 301,
00060 S.Maria di Galeria (Roma)
(2) INFM-SOFT, Department of Physics, University of Roma "La Sapienza",
P.le A.Moro 5, 00186 Roma
(3) ENEA, Portici Research Center, Computing and Networks Service,
Via Vecchio Macello, 80055 Portici
(4)ENEA, Casaccia Research Center, Computing and Modelling Unit,
Via Anguillarese 301, 00060 S.Maria di Galeria

Motivation

The discovery of Transcription Factor binding sites is still an open problem, as most of the softwares available to date have low predictive character, particularly for complex DNA (such as the human DNA). A novel method is proposed which overcomes some of the limitations affecting the existing prediction tools.

Methods

IMAGE strategy for the discovery of TF binding sites is based on a novel approach inspired by a technique used for lossy image compression, known as vector quantization and by analogous methods to identify genes with similar functions and reconstruct phylogenetic trees by clustering algorithms. The central idea is to map all possible n-length substrings of a given DNA sequence into a properly defined n-dimensional space equipped with a distance measure which projects similar substrings, representing the same motif, into nearby points. Consequently, the goal of finding recurrent similar strings is shifted into the determination of highly clustered data points.

Results

A complete assessment of the IMAGE tool has been provided by using the web service available at the Washington site (http://bio.cs.washington.edu/assessment/submit.html), where most of the available tools for the TF binding sites discovery have been recently compared. Results demonstrate the ability of IMAGE in correctly predicting a large number of motifs with respect to the other tools, although with a minor sensitivity in discriminating false positives.

Availability: http://image.ylichron.it/

Contact email: rosato@casaccia.enea.it