Use of an innovative clustering algorithm to summarize the outcome of genome-wide motif searches

ID - 205

Fu Limin[1], Isella Claudio[1], Cora' Davide[2], Caselle Michele[2], Medico Enzo[1]
[1]Department of Oncological Sciences, University of Torino, Candiolo - Torino
[2]Department of Theoretical Phisics, University of Torino, Torino

Motivation

In the past years, a number of motif searching algorithms have been proposed to identify putative transcription binding site motifs (TFBS) from genomic data. Some of them directly work on Position Weight Matrix (PWM), while the others yield a set of hits of over-represented motifs. These hits very often are strongly correlated, being different variants of the same motif or overlapping portions of the same binding sequence.

Methods

Here we use a novel clustering algorithm named FLAME [1] to cluster motifs that are similar or overlapping to each other. Unlike other clustering algorithms, in FLAME cluster assignments of each object are solely based the memberships of its neighbors, that means, the object has high membership degrees in certain clusters if and only if its neighbors have high membership degrees in these clusters. This feature of FLAME is appealing in clustering motifs for our aim, since this will lead to better identification of overlapping motifs. To cluster motifs, a similarity matrix is generated first based on simple pairwise alignments of motifs. This matrix is then fed into GEDAS software (http://www.sourceforge.net/projects/gedas) to perform FLAME clustering, to generate clusters of motifs that are regarded as belonging to the same TFBS. At last a PWM for a putative motif can be built from each motif cluster with proper alignments.

Results

The ab-initio motif identification method proposed by Cora et al [2] was used to search for over-represented motifs in groups of genes concordantly regulated by hypoxia in an in vitro cell model. This searching yielded a total of 1562 significantly enriched short motifs. These motifs were clustered using FLAME, with clustering parameter settings optimized to generate clusters of reasonable size and diversity. For each cluster, the motifs were subsequently aligned to generate a PWM for each cluster. Scanning of the PWMs obtained through this process against TFBS databases led to the identification of binding sites for transcription factors known to mediate the hypoxia response, which confirmed the validity of our approach. Motif clustering by FLAME is currently being implemented in the GEDAS software.

[1] Fu L, Medico E: FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 2007, 8:3.
[2] Cora D, Di Cunto F, Provero P, Silengo L and Caselle M: Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrep-resented upstream motifs. BMC Bioinformatics 2004, 5:57.

Acknowledgements: This work was supported by grants from AIRC, FIRB-MIUR and Regione Piemonte. LF is the recipient of a PhD fellowship from Fondazione CRT - Progetto Lagrange.

Email: limin.fu@ircc.it