Improving the capacity of the CSTMiner algorithm to correctly classify conserved sequences

Horner D (1), Re M (1), Nasi C (1), Pesole G (2)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita degli Studi di Milano
(2) Dipartimento di Biochimica e Biologia Molecolare, Universita di Bari

Motivation

The CSTminer algorithm identifies sequences conserved between genomes (CSTs) through the use of a BLAST-like similarity search. A simple algorithm describing the evolutionary dynamics expected of codingsequences (a predominance of synonymous subsitutions and conservative amino acid changes) is used to ascribe a Coding Potential Score (CPS) to conserved elements. Such elements are classified as coding or non-coding through reference to results obtained from coding and non-coding "training sets".While in general this approach has proved extremely effective, the currently implemented methodology fails to unambiguously classify a significant number of (predominantly short) CSTs.

Methods

While the original method to evaluate the coding potential of CSTs relied on a simple threshold value derived from CPS values obtained empirically from known coding and non-coding sequences, we have developed a statistical measure of CPS scores derived from a simple data randomizationprocedure. Our method also incorporates observations of the "shadow effect" whereby a reverse reading frame tends, for coding csts, to exhibit a high coding potential. We thus calculate a P-value which indicates whether an observed CPS score can be explained by chance, given the composition, and level of conservation of a CST.

Results

Here we show that a simple bootstrap-like data randomization procedure can improve the accuracy of CST classification. Furthermore, we demonstrate that this approach is also effective in the classification of short conserved stretches, suggesting that it may also be of use in the detection of novel short exons.

Contact email: david.horner@unimi.it