A Support Vector Machine for the Discrimination of Coding and Non-Coding Sequences that are Conserved Between Related Genomes.

ID - 216

Re' Matteo[1], Nasi Chiara[1], Pesole Graziano[2,3], Horner David[1]

[1]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Via Celoria 26, 20133 Milano

[2]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, via Amendola 122/D, 70125 Bari, Italy

[3]Department of Biochemistry and Molecular Biology, University of Bari, Via Orabona 4, 70126 Bari, Italy

Motivation

The conservation of sequence elements between related genomes has long been recognised as an indication of functional significance and indeed recognition of homology to characterized sequences is one of the principal approaches used in the annotation of both protein coding and non-coding genes in newly sequenced genomes.

Discrimination between conserved coding and non-coding sequences is thus a topic of considerable interest, not least in the context of recent findings that the number non-coding transcripts in higher organisms is likely to be much higher than previously imagined. Additionally, it should be considered desirable to discriminate between coding and non-coding conserved sequences without recourse to the use of sequence similarity searches of protein databases as such approaches exclude the identification of novel conserved proteins that lack characterized homologs and may be influenced by the presence in databases of sequences which are erroneously annotated as coding.

Methods

Here we present a machine learning-based approach to discriminate between conserved coding and non-coding sequences. Our method calculates various statistics related to the evolutionary dynamics of two aligned sequences. These features are considered by a Support Vector Machine which designates the alignment coding or non-coding with an associated probability score.

Results

Our approach is both sensitive and accurate with respect to comparable methods and may be applicable in the discrimination between conserved coding and non-coding regions in complete genome sequences, the validation of ab-initio exon predictions and the discrimination of coding from non-coding cDNA sequences.

Email: matteo.re@unimi.it