GC content dependency of open reading frame prediction

M. Pohl[1] ✉, G. Theißen[2], S. Schuster[1]

[1]Department of Bioinformatics, Friedrich Schiller University, Jena, Germany
[2]Department of Genetics, Friedrich Schiller University, Jena, Germany

Motivations

A frequently used approach for detecting potential coding regions is to search for stop codons. In the standard genetic code 3 out of 64 trinucleotides are stop codons. Hence, in random or non-coding DNA one can expect every 21st trinucleotide to be equivalent to a stop codon. In contrast, the open reading frames (ORFs) of most protein coding genes are considerably longer. Thus, the stop codon frequency in coding sequences deviates from the background frequency of the corresponding trinucleotides. This has been utilized for gene prediction, in particular, in detecting ORFs. Traditional methods based on stop codon frequency are based on the assumption that the GC content is about 50 %. However, many genomes show significant deviations from that value.

Methods and Results

With the presented method we can describe the effects of GC content on the selection of appropriate length thresholds of ORFs. Conversely, for a given length threshold, we can calculate the probability of observing it in a random sequence. Thus, we can derive the maximum GC content for which ORF length is practicable as a feature for gene prediction methods and the resulting false positive rates. A rough estimate for an upper limit is a GC content of 80 %. This estimate can be made more precise by including further parameters and by taking into account start codons as well. We demonstrate the feasibility of this method by applying it to the genomes of the bacteria Rickettsia prowazekii, Escherichia coli and Caulobacter crescentus, exemplifying the effect of GC content variations according to our predictions.