Determining *coding* CpG islands by identifying regions significant for pattern statistics on Markov chains

Meromit Singer[1*], Alexander Engström[2*], Alexander Schönhuth[3†], Lior Pachter[1,2,4†]

[1] Department of Computer Science, University of California, Berkeley, USA
[2] Department of Mathematics, University of California, Berkeley, USA
[3] Centrum Wiskunde & Informatica, Amsterdam, Netherlands
[4] Department of Molecular and Cell Biology, University of California, Berkeley, USA
[*] Joint first authorship
[†] Joint last, corresponding authors

alexander.schoenhuth@cwi.nl
lpachter@math.berkeley.edu

Abstract

Recent experimental and computational work confirms that CpGs can be unmethylated inside coding exons, thereby showing that codons may be subjected to both genomic and epigenomic constraint. It is therefore of interest to identify *coding* CpG islands (CCGIs) that are regions inside exons enriched for CpGs. The difficulty in identifying such islands is that coding exons exhibit sequence biases determined by codon usage and constraint that must be taken into account.

We present a method for finding CCGIs that showcases a novel approach we have developed for identifying regions of interest that are significant (with respect to a Markov chain) for the counts of any pattern. Our method begins with the exact computation of tail probabilities for the number of CpGs in all regions contained in coding exons, and then applies a greedy algorithm for selecting islands from among the regions. We show that the greedy algorithm provably optimizes a biologically motivated criterion for selecting islands while controlling the false discovery rate.

The statistical criterion we apply to evaluating islands greatly reduces the number of false positives in existing annotations, and our approach to defining islands reveals significant numbers of undiscovered CCGIs in coding exons. Many of these appear to be examples of functional epigenetic overloading in coding exons.