Genome Biology

POSTER PRESENTATION Open Access

The coding potential of the human genome: global compositional properties identify with statistical significance a plethora of new potential coding regions

Steven Oden*, Luciano Brocchieri

From Beyond the Genome: The true gene count, human evolution and disease genomics
Boston, MA, USA. 11-13 October 2010

Bioinformatics predictions of coding sequences rely on details models of compositional properties of genes. Such models are based on large, genome-specific training sets of known genes. Although these models are optimal for the identification of 'average' genes, they may be over-parameterized to allow recognition of genes of anomalous properties, for example genes coding for very short peptides.

We have developed two approaches to the identification of coding sequences that rely on more general compositional principles that we expect to be conserved over a wider variety of genes. The first approach is based on the observation that coding regions generally exhibit contrasting global compositional properties in the three codon positions, depending on the overall base composition of the sequence. For example, sequences rich in C and G bases have a much higher GC content in third codon position and a relatively low GC content in second codon position. General rules on the base content at the three codon positions as a function of the overall base content can be identified and exploited to score sequence regions for their coding potential. More generally, the period-three structure of coding regions imposes compositional periodicity to the sequence that, irrespective of the specific type of contrasts that we might expect to see, result in a significantly non-random distribution of bases.

Applying these principles, we have devised two algorithms to detect potential coding regions in sequences of any composition, one based on overall compositional expectations and one based on overall contrasts. We have applied our procedure to the human genome.

To our surprise, we have detected a plethora of regions, not overlapping with any of the currently annotated gene sequences, that display with high statistical significance a periodic structure often conforming to expectations for coding regions in terms of base-type composition. The frequency of these regions is far greater than the random frequency observed in corresponding scrambled sequences. Most of these regions also show levels of complexity that distinguish them from repetitive elements and that are consistent with the complexity of known genes.

Our bioinformatics results provide a rich source of information for future experimental analyses and the potential for exciting new discoveries.

Published: 11 October 2010

doi:10.1186/gb-2010-11-S1-P7
Cite this article as: Oden and Brocchieri: The coding potential of the human genome: global compositional properties identify with statistical significance a plethora of new potential coding regions. *Genome Biology* 2010 **11**(Suppl 1):P7.

Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL 32610, USA