Detecting non-coding RNAs in vertebrate genomes

Gayle McEwen & Wally Gilks*

MRC Biostatistics Unit & MRC-RFCGR

Our aim is to develop a statistical method to detect non-coding RNA (ncRNA) genes within vertebrate genomes. ncRNA genes are very difficult to identify computationally and so new methods for mining the wealth of readily available genomic sequence data are of immediate interest.

RNA is an important mediator of cellular processes, which acts in a variety of structural, catalytic and regulatory roles within the cell. In every cell, RNA is involved in the production of proteins from the instructions encoded in DNA. Regions of DNA are transcribed into *messenger* RNA (mRNA) which then serves as a template for protein synthesis. These DNA regions are called *coding regions* and the mRNA is *coding* RNA.

In recent years, many new species of so-called *non-coding RNAs* (ncRNAs) have been discovered. Transfer RNA (tRNA) and ribosomal RNA (rRNA) have been long known and act as functional RNA molecules in the process of protein synthesis, but many new small RNA molecules have been discovered; these RNAs do not encode proteins but function directly as RNA in processes such as transcriptional regulation, RNA processing and modification, protein degradation and chromosome replication (for a review see Eddy, 2001; Stortz, 2002).

Detecting ncRNA genes computationally within the genome is more difficult than detecting protein coding genes because ncRNA genes carry a much smaller amount of statistical information than protein coding genes and do not have such characteristic signals, such as hexamer bias etc.

In most cases the structure of a ncRNA is important for its function. RNA *tertiary structure* (its full 3-dimensional shape) is highly complex and very difficult to predict; however, RNA *secondary structure* prediction is more tractable. For this, only the base-pairing within the molecule is predicted. Secondary structure can be predicted with reasonable success from thermodynamic principles, *i.e.* by minimising free energy. Unfortunatley, predicted minimum free energy turns out to be a poor discriminator between false and genuine ncRNA sequences, especially in the context of a whole genome scan. However, such a method may be of use when examining smaller sets of sequences that have some other evidence to suggest they may be ncRNAs, provided a sensitive statistical test is employed.

We develop an approach similar to that of Rivas and Eddy (2000), in which a query sequence is compared to those of random sequences of the same dinucleotide content, using a sensitive statistical measure of ncRNA potential based on folding free energy. We apply it to experimentally validated ncRNA sequences, and to a set of well-conserved non-coding sequences obtained from a comparative whole-genome analysis of human and the pufferfish, *fugu rubripes*.

References

Coward E. (1999) Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics*, **15**, 1058-9.

Eddy, S.R. (2001). Noncoding RNA genes and the modern RNA world. *Nat Rev Gen*, **2**, 919-929.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, **125**, 167-188,

Rivas, E. & Eddy, S.R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **7**, 583-605.

Storz G. (2002) An expanding universe of noncoding RNAs. *Science* (Related Articles, Links Abstract), **296**, 1260-3.

Workman C, Krogh A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, *27*, 4816-22.