Comparative Studies of Neural Network Models for mRNA Analysis

Matthias Futschik, Mark Schreiber, Chris M. Brown, Nikola Kasabov

Departments of Information Science and Biochemistry
University of Otago, PO box 56, Dunedin, New Zealand
email: mfutschik@infoscience.otago.ac.nz

The cellular machinery for recognition of protein synthesis initiation codons successfully distinguishes between true and false initiation sites. This recognition problem is most marked in prokaryotes, where ribosomes must accurately discriminate between several true initiation codons (AUG, GUG, UUG) and the identical triplets within mRNAs. An understanding of this process is crucial for the construction of transgenic organisms and for gene recognition by computer. To facilitate these studies we have constructed a large database of the translation initiation regions from over 150,000 genes classified by organism [1]. Ancillary information associated with each sequence allows clustering of many of the sequences in biologically relevant ways, for example by function or codon bias. We can also extract sets of false starts from the same organism, for example, a set of 47189 non-initiation AUG triplets from *E. coli* coding regions, to compare to a set of 4270 predicted true initiation sites from the complete *E. coli* genome.

We applied different neural networks models to these data sets and compared them in terms of their performance. Among well-known algorithms such as MLP and LVQ we utilise evolving connectionist systems (ECOS) for classification of true and false initiation sites. ECOS are designed to facilitate building on-line, adaptive, knowledge-based IS and to evolve through incremental, hybrid (supervised/unsupervised), on-line learning [2]. Fuzzy input-output relations can be further accommodated by evolving fuzzy neural network structures (EFuNNs). Comparison was also carried out with previous analyses of our data sets using simple statistical models, rule-based approaches, matrices, and Hidden Markov Models.

Each of the approaches has advantages and disadvantages in terms of efficiency, accuracy and precision. Using these approaches we have identified several unusual patterns and signals in the translation initiation regions. These analyses reveal unexpected differences between organisms and subsets of genes within organisms.

References

[1] Transterm, *http://biochem.otago.ac.nz:800/Transterm/home_page.html*

[2] RICBIS, *http://divcom.otago.ac.nz/kel/CBIIS.html*