RNNScan: Eukaryotic Gene Finding via Hybrid Recurrent Neural Networks

Srinivasan, Anand

Today's vast reserves of genetic data necessitates a computational tool which is capable of automatically predicting boundaries of protein-coding regions in the genome. In eukaryotic sequences, this remains an open problem due to gene interruption by non-coding introns. Popular probabilistic tools such as GENSCAN employ Hidden Markov models, but cannot take into account the locality of genetic signals which describe the initiation or termination sites of coding sequences – an automatic segmentation task made difficult by erratically distributed intergenic sequences. We propose a recurrent-neural-network based tool, RNNScan, which incorporates specialized “memory units” to force gradient retention, as well as a gated information-flow architecture to allow convergence on terminal gene boundaries following arbitrary-length noncoding sequences. A novel “sluice gate” unit regulates error flow between memory blocks. The network is "hybridized" with an auxiliary probabilistic feature, which is a function of differences in information content (relative entropy) between regions of the genome. These features, called the nucleotide scores, are computed over a linear Bayesian model and marginalized via a custom message-passing algorithm. This takes into account the influence of certain k-mer distributions around splice-sites, a known determinant of snRNP activity. On Burset and Guigo’s standardized dataset of 570 vertebrate sequences, RNNScan performs on par with GENSCAN at nucleotide- and exon-level accuracy but outperforms it on the exact whole-gene identification test (71% vs. 43% sensitivity). We show how RNNScan is able to achieve this performance by solving the localization problem, and discuss applications in genome-tailed pharmaceuticals and cancer diagnostics.

Awards Won:

European Organization for Nuclear Research-CERN: All expense paid trip to tour CERN