Silvestre-Ryan, Jordi; Holmes, Ian

Consensus decoding of recurrent neural network basecallers. (English) Zbl 1392.92008

Jansson, Jesper (ed.) et al., Algorithms for computational biology. 5th international conference, ALCoB 2018, Hong Kong, China, June 25–26, 2018. Proceedings. Cham: Springer (ISBN 978-3-319-91937-9/pbk; 978-3-319-91938-6/ebook). Lecture Notes in Computer Science 10849. Lecture Notes in Bioinformatics, 128-139 (2018).

Summary: There is an extensive literature using probabilistic models, such as hidden Markov models, for the analysis of biological sequences. These models have a clear theoretical basis, and many heuristics have been developed to reduce the time and memory requirements of the dynamic programming algorithms used for their inference. Nevertheless, mirroring the shift in natural language processing, bioinformatics is increasingly seeing higher accuracy predictions made by recurrent neural networks (RNN). This shift is exemplified by basecalling on the Oxford nanopore technologies' sequencing platform, in which a continuous time series of current measurements is mapped to a string of nucleotides. Current basecallers have applied connectionist temporal classification (CTC), a method originally developed for speech recognition, and focused on the task of decoding RNN output from a single read. We wish to extend this method for the more general task of consensus basecalling from multiple reads, and in doing so, exploit the gains in both accelerated algorithms for sequence analysis and recurrent neural networks, areas that have advanced in parallel over the past decade. To this end, we develop a dynamic programming algorithm for consensus decoding from a pair of RNNs, and show that it can be readily optimized with the use of an alignment envelope. We express this decoding in the notation of finite state automata, and show that pair RNN decoding can be compactly represented using automata operations. We additionally introduce a set of Markov chain Monte Carlo moves for consensus basecalling multiple reads.

For the entire collection see [Zbl 1391.92003].

MSC:

92B20	Neural networks for/in biological studies, artificial life and related topics
92C40	Biochemistry, molecular biology
68Q45	Formal languages and automata
68T05	Learning and adaptive systems in artificial intelligence

Keywords:

nanopore sequencing; deep learning; dynamic programming; alignment envelope; finite state automata

Full Text: DOI