Transparent profile HMMs

Luis M. S. Russo, Ana T. Freitas, Arlindo L. Oliveira
lsr@kdbio.inesc-id.pt; atf@kdbio.inesc-id.pt; aml@inesc-id.pt;
INESC-ID/IST
Lisboa, Portugal

Hidden Markov Models (HMMs) have a long history in bioinformatics and genetics, with applications in modeling binding domains, gene finding and sequence alignment. Dynamic programming and BLAST can give an accurate answer for certain applications like the alignment of a small number of sequences. However, for the alignment of a family of sequences with different evolutionary distances, HMMs allow for efficiency and flexibility. A HMM model that has proven in practice to provide good compromise between flexibility and tractability was called profile HMM [1]. In order to model real sequences, the profile HMM must contain three states (the match state, the insert state, and a delete state) and two types of probabilities associated with it (the transition probability and the emission probability).

In this work we focus on the states of profile HMMs that model deletions. These states are problematic because they are silent and emit no symbols. Therefore their impact can only be verified by the performance of the remaining states. Computationally processing chains of hidden states requires some care. Naively replacing these states by transitions is problematic, since it raises the computation complexity from O(n) to $O(n^2)$, both in time and space. Moreover it is subject to overfitting, due to the large number of variables to model.

We handle these problems by using a Monge transition matrix. This means that the number of transitions if still $O(n^2)$, but we can use the SMAWK algorithm to perform the Viterbi computation, in O(n) time. The resulting model is more flexible than classical profile HMMs. The model is closer to the convex gap model than to the affine gap model. The Monge property has been useful in alignment algorithms, yielding an efficient bidirectional alignment algorithm [2]. We expect this new approach to be particularly useful with the advent of next-generation sequencing data.

References

[1] Eddy S.R. “Profile hidden Markov models”, Bioinformatics, 14(9):755-63 (1998)

[2] Luis M.S. Russo "Monge properties of sequence alignment", Theoretical Computer Science, doi:10.1016/j.tcs.2011.12.068, (2012).

Topic: learning in biological systems □Preference: oral presentation