Hidden Markov Models for Genome Analysis

Leonardo Gori

May 2022

Abstract

//TODO ...

1 Summary of the activities performed

Before the implementation of the project, an analysis of the problem has been performed. In particular, the attempt made was understanding the concepts and the components used for the correct execution of the algorithm.

In the very beginning, the concepts of pairwise alignment, Markov chains and hidden Markov models, the Viterbi algorithm and the forward algorithm have been gathered from chapter 2 and 3 of [DEKM98].

Once gained confidence with the theory behind these models and algorithms, as a second stage of the analysis, the attempt of understanding the Pairwise alignment using HMMs in section 4 of [DEKM98] was made with less difficulties.

Subsequently, following the extract by Benjamin ([Ben17]) and the precious information by [RBAA18], the actual implementation of the algorithm has been made.

2 Language and APIs

The code is entirely written in C++ programming language, with the use of the following libraries and APIs (omitting the standard ones):

- random: used for the random generation of sequences and the random definition of state transition probabilities
- algorithm: used for the shuffling of sequences, used for randomization purposes

Since the computation of the components that build up the belief state of the alignment is independent, the execution of the algorithm has been developed in a parallel fashion, by making use of the OpenMP APIs.

3 Class diagram

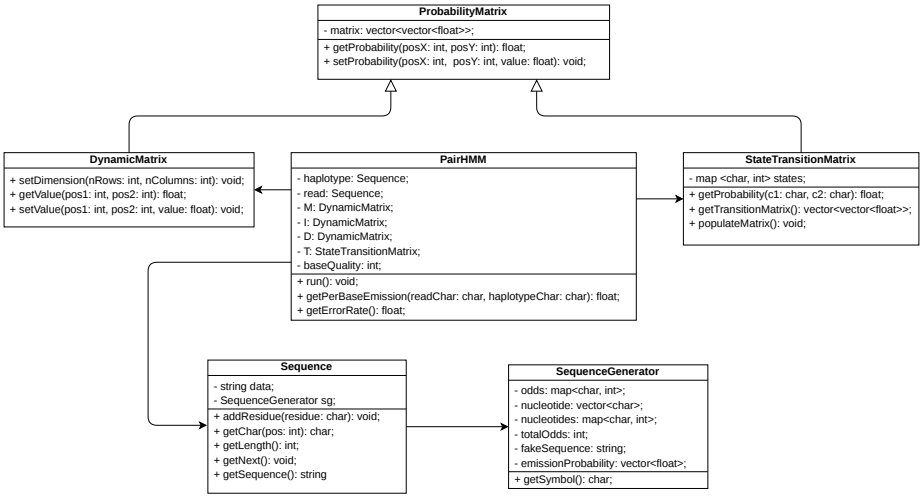

Figure 1: High level UML diagram

4 Results

...

5 Conclusions

...

References

[Ben17] David Benjamin. Pair HMM probabilistic realignment in HaplotypeCaller and Mutect. `https://github.com/broadinstitute/gatk/blob/master/docs/pair_hmm.pdf`, August 2017.

[DEKM98] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[HMR+17] Sitao Huang, Gowthami Jayashri Manikandan, Anand Ramachandran, Kyle Rupnow, Wen-mei W. Hwu, and Deming Chen. Hardware acceleration of the pair-hmm algorithm for dna variant calling. In *Proceedings of the 2017 ACM/SIGDA International Symposium*

on Field-Programmable Gate Arrays, FPGA '17, page 275–284, New York, NY, USA, 2017. Association for Computing Machinery.

[RBAA18] Shanshan Ren, Koen Bertels, and Zaid Al-Ars. Efficient acceleration of the pair-hmms forward algorithm for gatk haplotypecaller on graphics processing units. *Evolutionary Bioinformatics*, 14:1176934318760543, 2018. PMID: 29568218.