Pair HMM for genome analysis

Made possible by LaTeX and Beamer

June 7, 2022

Leonardo Gori

CONTENTS

THE PROBLEM

Sequence alignment

« Nature is a tinkerer and not an inventor » Jacob 1977

- Finding new methods for comparing the similarity of two genome sequences gives vital information on evolution and development
- New sequences are adapted from pre-existing sequences rather than invented *de novo*
- Evolving sequences accumulate insertions and deletions as well as substitutions
- Before comparing them, it's necessary finding a plausible alignment between them

THE SOLUTION

Pair Hidden Markov Models

Special types of Hidden Markov models for the generation of a pair of sequences

They are composed by 2 main properties:

- State transition distribution (between the states M,I,D)
- Pair emission distribution $\mathcal{P}_s(a, b)$

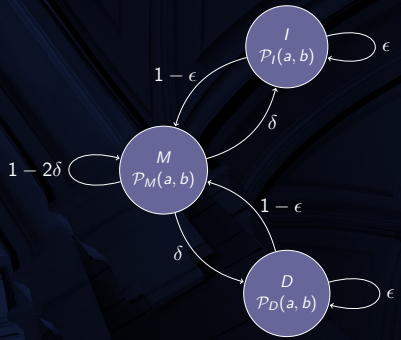

Figure 1: Naive representation of a Pair HMM, composed by its transmission states and emission probabilities

The PHMM forward algorithm

Algorithm 1 Pseudo-code of the Pair HMM forward algorithm

Initialize:

- $M_{i,0} = I_{i,0} = D_{i,0} = 0, \qquad \forall\ 0 \leq i \leq |\mathcal{R}|$
- $M_{0,j} = I_{0,j} = 0, \qquad \forall\ 0 \leq j \leq |\mathcal{H}|$
- $D_{0,j} = 1/n, \qquad \forall\ 0 \leq j \leq |\mathcal{H}|$

for $1 \leq i \leq |\mathcal{R}|$ **do**
  for $1 \leq j \leq |\mathcal{H}|$ **do**
    $M_{ij} = \mathcal{P}_M(\mathcal{R}_i, \mathcal{H}_j) \cdot (M_{i-1,j-1} T_{MM} + I_{i-1,j-1} T_{IM} + D_{i-1,j-1} T_{DM})$
    $I_{ij} = M_{i-1,j} T_{MI} + I_{i-1,j} T_{II}$
    $D_{ij} = M_{i,j-1} T_{MD} + D_{i,j-1} T_{DD}$
  end for
end for
Total likelihood $P(\mathcal{R}|\mathcal{H})$ is $\sum_j (M_{\mathcal{R},j} + I_{\mathcal{R},j})$.

THE IMPLEMENTATION

UML class diagram

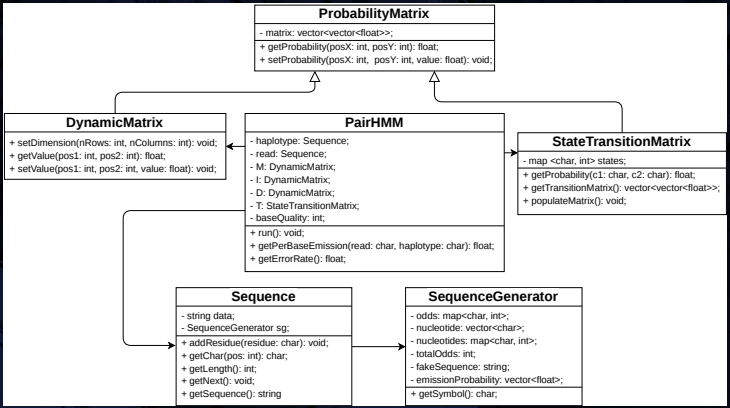

Figure 2: UML class diagram of the the project for the analysis of genomic sequences

OpenMP APIs

Without parallel execution

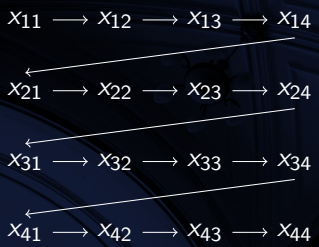

Figure 3: Flow of computation of the PHMM forward algorithm

With parallel execution

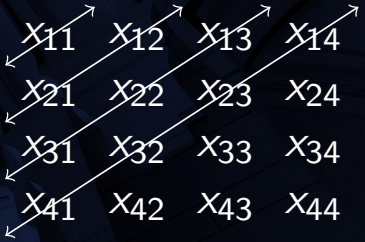

Figure 4: Flow of computation of the algorithm proposed in the implementation

CONCLUSIONS

Further work

- Performance comparison between the standard algorithm and the one proposed in the project
- Metamorphic malware's mitigation through Profile Hidden Markov Models

Contacts

Leonardo Gori

xleonardogori@gmail.com

/LeoGori

/leonardo-gori

Thank you for your attention!