

Pair HMM for genome analysis

Made possible by \LaTeX and Beamer

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ June 7, 2022

Leonardo Gori

CONTENTS

1. The problem
 - Sequence alignment
2. The solution
 - Pair HMM
 - The forward algorithm
3. The implementation
 - UML class diagram
 - OpenMP
4. Conclusions



THE PROBLEM

Sequence alignment

4

« *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 $\mathcal{P}_s(a, b)$

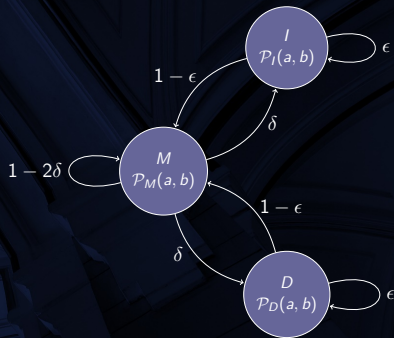


Figure 1: Naive representation of a Pair HMM, composed by its transmission states and emission probabilities

The PHMM forward algorithm

7

Algorithm 1 Pseudo-code of the Pair HMM forward algorithm

1: Initialize:

$$\begin{aligned} \blacksquare M_{i,0} &= I_{i,0} = D_{i,0} = 0, & \forall 0 \leq i \leq |\mathcal{R}| \\ \blacksquare M_{0,j} &= I_{0,j} = 0, & \forall 0 \leq j \leq |\mathcal{H}| \\ \blacksquare D_{0,j} &= 1/n, & \forall 0 \leq j \leq |\mathcal{H}| \end{aligned}$$

2: **for** $1 \leq i \leq |\mathcal{R}|$ **do**

3: **for** $1 \leq j \leq |\mathcal{H}|$ **do**

$$4: \quad M_{ij} = \mathcal{P}_M(\mathcal{R}_i, \mathcal{H}_j) \cdot (M_{i-1,j-1} T_{MM} + I_{i-1,j-1} T_{IM} + D_{i-1,j-1} T_{DM})$$

$$5: \quad I_{ij} = M_{i-1,j} T_{MI} + I_{i-1,j} T_{II}$$

$$6: \quad D_{ij} = M_{i,j-1} T_{MD} + D_{i,j-1} T_{DD}$$

7: **end for**

8: **end for**

9: Total likelihood $P(\mathcal{R}|\mathcal{H})$ is $\sum_j (M_{\mathcal{R},j} + I_{\mathcal{R},j})$.



THE IMPLEMENTATION

UML class diagram

9

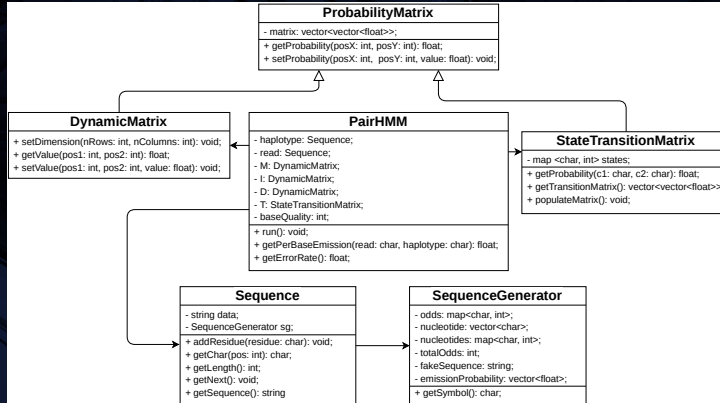


Figure 2: UML class diagram of the the project for the analysis of genomic sequences



Without parallel execution

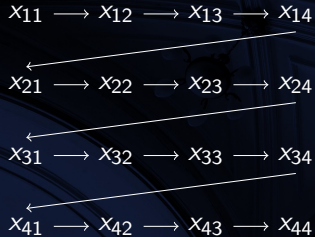


Figure 3: Flow of computation of the PHMM forward algorithm

With parallel execution

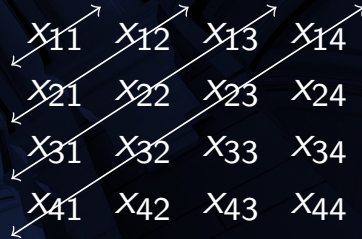


Figure 4: Flow of computation of the algorithm proposed in the implementation

The background of the slide is a low-angle, dark blue-tinted photograph of a grand, classical-style building. The building features multiple stories with large windows and ornate architectural details. A prominent green rectangular overlay covers the middle portion of the image, containing the word 'CONCLUSIONS' in white capital letters. At the bottom of the image, the Russian text 'НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ' is visible on the building's facade.

CONCLUSIONS

Further work


12


- 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!