jpHMM: Improving the reliability of recombination prediction in HIV-1

Anne-Kathrin Schultz[1], Ming Zhang[2,3], Ingo Bulla[1], Thomas Leitner[2], Bette Korber[2,4], Burkhard Morgenstern[1], Mario Stanke[1]

[1]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Germany; [2]T-6, Los Alamos National Laboratory; [3]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; [4]The Santa Fe Institute, Santa Fe, NM 87501, USA

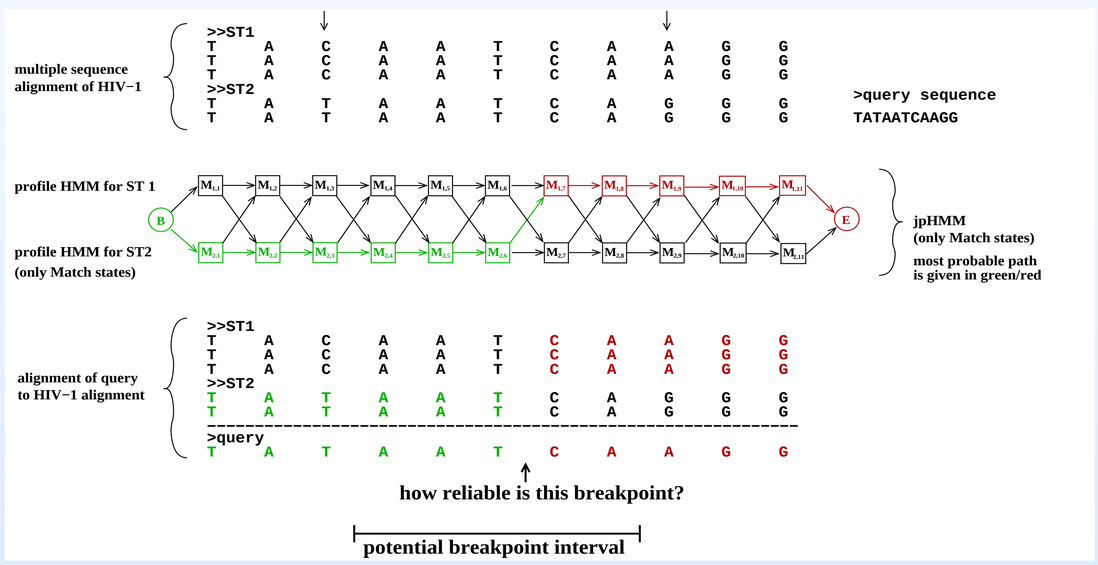
contact: anne@gobics.de

Introduction

- For an HIV-1 genomic sequence **jpHMM** [Schultz *et al.*, 2006] predicts whether it is a recombinant of different subtypes.
  If so, jpHMM **estimates the recombination breakpoint positions** and **assigns to each segment** in between two breakpoints **a parental subtype** among the major HIV-1 subtypes.
- Now, the output of jpHMM includes **information about 'uncertainty' regions in the recombination prediction and an interval estimate of the breakpoint**.
- jpHMM is available online at `http://jphmm.gobics.de/` [Zhang *et al.*, 2006].

jpHMM

- **given:** multiple sequence alignment of HIV-1 sequences subdivided into (sub)subtypes
- **input:** a query HIV-1 sequence (partial or complete genome)
- **jpHMM:** each subtype is modeled as a profile HMM; jumps between different subtypes are allowed
- **recombination prediction:** defined by the most probable path through the jpHMM generating the query sequence; jumps between different subtypes define recombination breakpoints

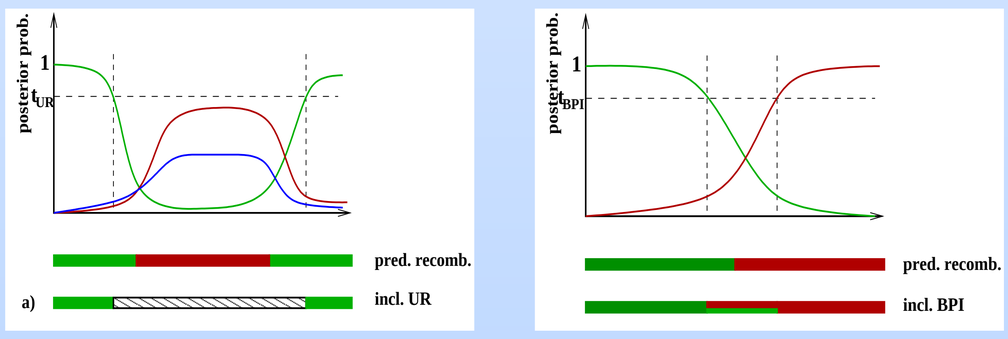


Uncertainty regions and breakpoint intervals

Posterior probability of a subtype at a certain base of the query sequence
= probability that the base belongs to the considered subtype in our probabilistic recombinant model

Uncertainty regions in the recombination prediction and interval estimates of breakpoints:

- **Uncertainty region** (UR): region in the query sequence where the posterior probability of the predicted subtype is lower than a certain threshold t_{UR}
- **Interval estimate of a breakpoint** (breakpoint interval, BPI): region around a predicted breakpoint where the posterior probabilities of the two predicted subtypes are lower than a certain threshold t_{BPI} but higher than the posterior probabilities of all other subtypes



jpHMM output

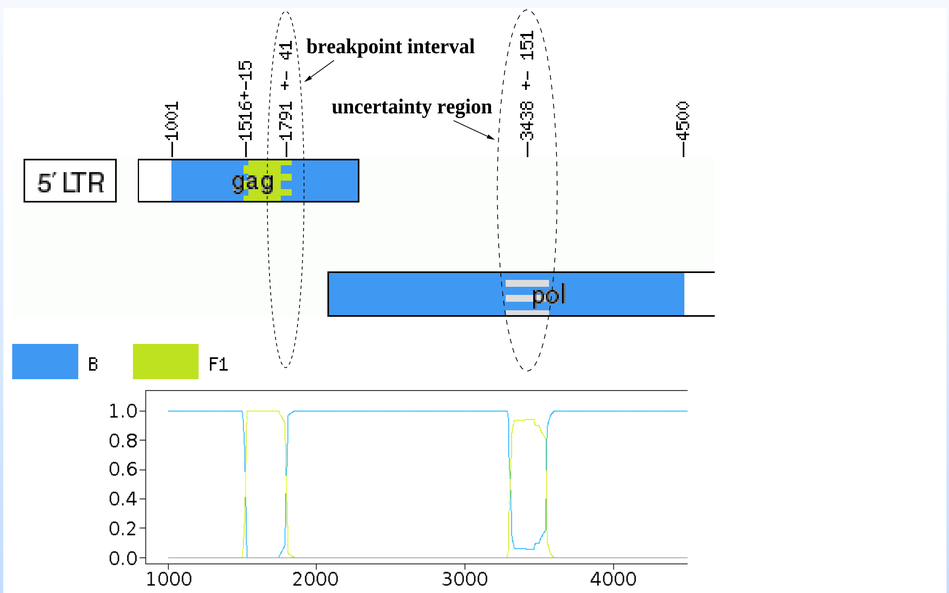


FIGURE 1: Extract of the new jpHMM output for an artificial recombinant sequence

Evaluation and Results

Test data:

40 semi-artificial near full-length HIV-1 inter-subtype recombinant sequences

- of two real-world parental sequences from two different subtypes (A1, B, C, D, F1, G and CRF01).
- with artificially introduced breakpoints
  1. at every 1000th position
  2. where alternating long segments (1500 nt) from one subtype are interrupted by short segments (500 nt (2.1.) and 300 nt (2.2.), respectively) from another subtype

Accuracy of the predicted breakpoint intervals: (results are shown for 1.)

threshold t_{BPI}	BPI length average	min / max	% BP found using posterior probs	fixed BPI length
0.75	16.12	0 / 113	54.17	50.28
0.85	22.46	0 / 121	68.06	56.39
0.90	26.89	2 / 135	74.72	59.44
0.95	34.05	2 / 202	81.11	65.00
0.99	48.58	5 / 233	92.50	71.94
0.9999	84.77	11 / 492	98.06	81.39

E.g. for $t_{\mathrm{BPI}} = 0.99$ (df), 92.50 % of the real breakpoints could be detected (column 4)
(Average length of predicted BPI = 48.58 nt, minimal length = 5 nt maximal length = 233 nt)
(see "Naïve method" for column 5)

Accuracy of the predicted parental subtypes: (results are shown for 1.)

- 0.58 % to 0.82 % of positions outside UR and BPI classified incorrectly, compared to
- 1.51% for precise recombination prediction with jpHMM (precise BP estimates, no UR)

($t_{\mathrm{UR}} = t_{\mathrm{BPI}} \in \{0.75, \ldots, 0.9999\}$)

(The results for dataset 2.1. and 2.2. are similar)

Comparison to a naïve approach

Naïve method: predicts breakpoints in a symmetric interval of fixed length, centered around the predicted breakpoint position
(= most obvious method to define BPI around predicted BPs, if no further information is provided)

fixed interval length = average length of the predicted BPI, rounded to the nearest even number

- 92.50 % of all breakpoints could be detected with jpHMM using posterior probabilities compared to
- 71.94 % with the naïve method (for the default threshold $t_{\mathrm{BPI}} = 0.99$)
  ⇒ sensitivity of our method is up to 20 percentage points higher than that of the naïve method

Conclusions

The new extension strongly improves the reliability of the jpHMM recombination prediction:

- BPI defined by our method are far more accurate than BPI of fixed length
- Outside UR and BPI the user can now be more confident in predicted parental subtypes
- Definition of UR helps to avoid drawing wrong conclusions based on doubtful, uninformative regions (e.g. postulation of a new CRF)

Additional information given by the posterior probabilities:

- Varying length of BPI gives information about which BP can be located relative precisely/approximately
- In a UR the graph of posterior probabilities shows which subtypes are closest related in these regions.

References

[Schultz *et al.*, 2006] Anne-Kathrin Schultz, Ming Zhang, Thomas Leitner, Carla Kuiken, Bette Korber, Burkhard Morgenstern, and Mario Stanke. A Jumping Profile Hidden Markov Model and Applications to Recombination Sites in HIV and HCV Genomes. *BMC Bioinformatics*, 7:265, 2006.

[Zhang *et al.*, 2006] Ming Zhang, Anne-Kathrin Schultz, Charles Calef, Carla Kuiken, Thomas Leitner, Bette Korber, Burkhard Morgenstern, and Mario Stanke. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res.*, 34:W463–W465, 2006.