LA-UR- 08-7959

*Approved for public release;
distribution is unlimited.*

Title: Improving the reliability of the jpHMM recombination prediction in HIV

Author(s): M. Zhang, Z#: 188651, T-6/T-Division
T. Leitner, Z#: 120084, T-6/T-Division
B. Korber, Z#: 108817, T-6/T-Division

Intended for: Journal: Bioinformatics Journal

Los Alamos
NATIONAL LABORATORY
EST. 1943

Improving the reliability of the jpHMM recombination prediction in HIV

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

[1]Department of Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, Germany; [2]Los Alamos National Laboratory, Los Alamos, NM 87545, USA; [3]The Santa Fe Institute, Santa Fe, NM 87501, USA 4 CNLS, Los Alamos, NM 87544

ABSTRACT

Summary: Accurate classification of HIV and the identification of recombinants, including precise breakpoint definitions, is of crucial importance for epidemiological monitoring and the design of potential drugs. Recently we developed jpHMM, a new method to detect recombinations in HIV-1 genomes. jpHMM predicts phylogenetic recombination breakpoints in a query sequence and assigns to each segment of the sequence one of the major HIV-1 subtypes. For the user the reliability of the predicted breakpoint positions and parental subtypes is most important. For this reason we extended the output of jpHMM to include the information on regions where the model is 'uncertain' about the parental subtype and an interval estimate of the breakpoint. This information is determined using the posterior probabilities of the subtypes at each query sequence position.

Availability: jpHMM is available online at http://jphmm.gobics.de.

Contact: {anne, mario}@gobics.de

Viruses of the so-called M(ajor) Group of HIV-1 are mainly responsible for the HIV pandemic. This clade has been divided into nine genetic subtypes, A - D, F - H, J, K, and four sub-subtypes (A1, A2, F1, F2). Among these subtypes recombination is extremely common. Recombinants that have been epidemiologically successful are called *circulating recombinant forms* (CRF). Up to now 43 CRFs have been identified and the number is increasing.

Recently we developed jpHMM, a jumping profile hidden Markov model to detect recombinations in HIV-1 genomes (Schultz *et al.*, 2006; Zhang *et al.*, 2006). jpHMM aligns a query sequence to a pre-calculated multiple sequence alignment of pure-subtype HIV-1 sequences, predicts phylogenetic recombination breakpoints and assigns to each segment of the sequence one of these subtypes. Each subtype in the alignment is modeled as a profile HMM (Eddy, 1998). In addition to the usual state transitions within these profile HMMs, transitions, called *jumps*, between the different profile HMMs are allowed. Thus the model can jump between states corresponding to the different subtypes, depending on which subtype is locally most similar to the database sequence. The recombination prediction for a query sequence is then defined by the most probable path through the jpHMM that generates the query sequence, the so-called Viterbi path. Since each state of the jpHMM only belongs to one profile HMM and each sequence position is generated by one state of the model, each position of the query sequence is assigned to exactly one parental subtype. Positions of jumps between different subtypes define recombination breakpoints.

jpHMM was evaluated on a large set of real and simulated HIV-1 data (Schultz *et al.*, 2006). A comparison of its prediction accuracy to competing methods such as Simplot (Lole *et al.*, 1999) and RDP (Martin *et al.*, 2005) showed that jpHMM is far more accurate than existing methods for phylogenetic breakpoint detection.

Nevertheless it is very useful and important for the user to a get a hint about the reliability of the predicted recombination pattern in a particular region of the query sequence. For this reason we extended the output of jpHMM to include the information on regions where the model is 'uncertain' about the parental subtype and, as opposed to a point estimate, an interval estimate of the breakpoint, called 'breakpoint interval' here. For each query sequence position the so-called posterior probability for each subtype is calculated. This is the probability that the respective sequence position belongs to the considered subtype given that the whole sequence is generated by the model. These probabilities can be calculated using the well-known Forward and Backward algorithms (Durbin *et al.*, 1998). The posterior probabilities are used to define uncertainty regions in the recombination prediction and breakpoint intervals: If at a certain position of the query sequence the posterior probability of the predicted subtype is lower than a certain threshold t_1, this position is marked as uncertain (e.g. Fig. 1, position 3434 ± 149). A breakpoint interval is defined by an interval around a predicted breakpoint position where the posterior probabilities of the two successive predicted subtypes is lower than a certain threshold t_2 but higher than the posterior probabilities of all other subtypes (e.g. Fig. 1 position 5063 ± 41).

This extension of jpHMM was evaluated on a large set of simulated full-length inter-subtype recombinant sequences with known breakpoints. Each of the sequences is a recombination of two 'real-world' parental sequences from two different HIV-1 subtypes. As parental subtypes we chose every possible pair of the (sub-) subtypes A1, A2, B, C, D, F1, F2, G and CRF01, except A1-A2, F1-F2 and B-D. For each pair of subtypes we created two different artificial recombinants, e.g. A1-B and B-A1, so, in total $2 * 33 = 66$ artificial recombinants were evaluated. In a first test run we introduced breakpoints at every 1000th position, results for this data set are shown in Table 1. Additionally we tested two more data sets where we introduced segments of length 500nt and 300nt respectively at every 1500th position.

The accuracy of predicted breakpoint positions is usually measured

[*] to whom correspondence should be addressed

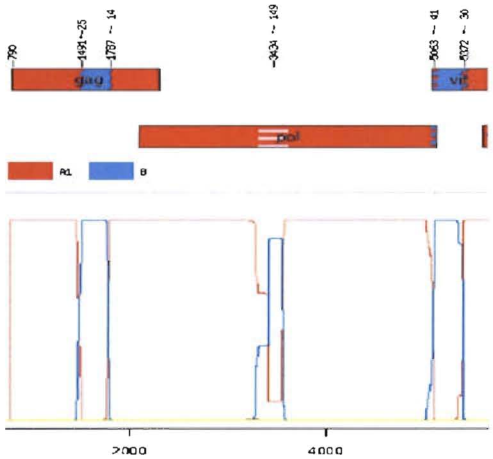

Figure 1. Excerpt of the jpHMM web server output for an artificial recombinant sequence of subtype A1 containing segments of length 300nt from subtype B at every 1500th position. Breakpoint intervals are shown by an interfingering of the colors of the two predicted subtypes, uncertainty regions by an interfingering of grey color and the color of the predicted subtype.

by the distance of the predicted to the correct breakpoint positions. This measure corresponds to testing whether an interval of fixed length around each predicted breakpoint contains the true breakpoint. If, for example, the median of the distances of the predicted to the real breakpoint is 8, then 50% of all real breakpoints are located within a breakpoint interval of length 16 around a predicted breakpoint (with the predicted breakpoint as center).

To assess the accuracy of the breakpoint intervals defined by the posterior probabilities for a certain threshold t_2, we compared the number of real breakpoints located within these breakpoint intervals to the number of breakpoints found using breakpoint intervals of fixed length. As fixed breakpoint interval length we chose the average length of the breakpoint intervals, rounded to the nearest even number, defined by the posterior probabilities for threshold t_2. In Table 1 the results are given for different thresholds t_2. At the top of the table the results are given for data set A with breakpoints at every 1000th position, at the bottom the results for data set B. In the first column the threshold t_2 is given (e.g. 0.95 in row 4 for data set A). The average length of the breakpoint intervals defined by this threshold is given in the second column (e.g. 33.83). In column 4, the percentage of real breakpoints detected with these breakpoint intervals is given. For example for a threshold of 0.95 87.54% of the real breakpoints were located within one of the predicted breakpoint intervals. The percentage of breakpoints found using the average length of the breakpoint intervals defined by the posterior probabilities for this threshold as fixed breakpoint interval length (e.g. 33.83, rounded to the nearest even number = 34) is given in column 5 (e.g. 69.02%). The table shows that, especially for higher thresholds, the number of breakpoints found using the posterior probabilities is much higher than the number of breakpoints found using breakpoint intervals of fixed length. For example for a threshold of 0.99 only 5.4% of the real breakpoints were not detected using the posterior probabilities whereas for breakpoint intervals of fixed length we can observe a fivefold increase of this percentage (26.6%).

data set A				
threshold t_2	BPI length		% BP found using	
for P_{post}	average	min/max	P_{post}	fixed BPI length
0.75	16.25	1 / 269	56.23	50.67
0.85	22.52	1 / 308	69.53	58.08
0.90	27.31	1 / 329	76.94	65.49
0.95	33.83	1 / 351	87.54	69.02
0.99	46.63	3 / 388	94.61	73.40
0.9999	83.66	11 / 587	97.14	80.64

Table 1. Comparison of the accuracy of breakpoint intervals (BPI) defined by the posterior probabilities (P_{post}) for threshold t_2 to the accuracy of BPI of fixed length. For each threshold the highest value is marked in red. Details are given in the text.

(INSERT RESULTS FOR 300 and 500)

The length of a predicted breakpoint interval depends on how clear the breakpoint position is. A large breakpoint interval is the consequence of the uncertainty of the model to locate the exact breakpoint position between two subtypes. This means that the user can now see which breakpoint can be ocated relative precisely or which breakpoints are approximate. Further she can be more confident in the subtype predicted at positions outside breakpoint intervals and uncertainty regions. Using $t_1 = t_2$ for all given thresholds t_2 as threshold for the uncertainty regions for data set A 94.94 – 95.3% of those positions were predicted correctly. Due to the model architecture 4.32% of all positions were not assigned to any subtype (these positions are located at both ends of the sequence) so the total percentage of positions not located within an uncertainty region or a breakpoint interval that were classified incorrectly is less than 1%. For uncertainty regions no parental strain can confidently be determined. This helps to avoid drawing wrong conclusions based on doubtful, uninformative regions, such as the postulation of a new CRF. However, by examining the graph of the posterior probabilities the user can see which subtypes are closest related in these regions.

The program is available online as a web interface at `http://jphmm.gobics.de`. The source code can also be downloaded from this web page.

ACKNOWLEDGEMENT

This work was funded in part by the DFG grants STA 1009/5-1 and STA 1009/4-1.

REFERENCES

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Lole, K. S., *et al.* (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virology*, **73**, 152–160.

Martin, D. P., Williamson, C., and Posada, D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.

Schultz, A.-K., *et al.* (2006) A Jumping Profile Hidden Markov Model and Applications to Recombination Sites in HIV and HCV Genomes. *BMC Bioinformatics*, **7**.

Zhang, M., *et al.* (2006) jpHMM at GOBICS: a web server to detect recombinations in HIV-1. *Nucleic Acids Res.*, **34**, 463–465.