Published online 14 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W647–W651***
doi:10.1093/nar/gkp371

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] and Mario Stanke[1,]*

[1]Institut für Mikrobiologie und Genetik, Abteilung für Bioinformatik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077, Göttingen, Germany, [2]T-6, Los Alamos National Laboratory, [3]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545 and [4]The Santa Fe Institute, Santa Fe, NM 87501, USA

Received February 14, 2009; Revised April 17, 2009; Accepted April 24, 2009

ABSTRACT

Previously, we developed jumping profile hidden Markov model (jpHMM), a new method to detect recombinations in HIV-1 genomes. The jpHMM predicts recombination breakpoints in a query sequence and assigns to each position of the sequence one of the major HIV-1 subtypes. Since incorrect subtype assignment or recombination prediction may lead to wrong conclusions in epidemiological or vaccine research, information about the reliability of the predicted parental subtypes and breakpoint positions is valuable. For this reason, we extended the output of jpHMM to include such information in terms of 'uncertainty' regions in the recombination prediction and an interval estimate of the breakpoint. Both types of information are computed based on the posterior probabilities of the subtypes at each query sequence position. Our results show that this extension strongly improves the reliability of the jpHMM recombination prediction. The jpHMM is available online at http://jphmm.gobics.de/.

INTRODUCTION

Viruses of the so-called M (Major) 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) (1). Among these subtypes, recombination is extremely common (2). Recombinants that have been epidemiologically successful are called 'circulating recombinant forms' (CRF). Up to now, >40 CRFs and many 'unique recombinant forms' (URF) have been identified and the number is increasing (http://hiv.lanl.gov/). The accurate classification of HIV-1 genomes and the identification of recombinants, including precise breakpoint definitions, is important in many aspects, such as the design of potential vaccines and treatment strategies against HIV, as well as for epidemiological monitoring of HIV-1. For this challenging task, a wide variety of recombination detection tools has been developed. The most widely used HIV subtyping tool is Simplot (3), which has also been applied to many other viruses. For a query sequence it provides a graph reflecting the similarity of the sequence to a panel of reference sequences and predicts recombination breakpoints. RIP 3.0 (http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) also identifies recombination in a query sequence by calculating its similarity to a background alignment of HIV-1 sequences of different subtypes in a sliding window. Depending on how significantly better the 'best matching' background sequence is than the second best match, 'uncertainty regions' in the recombination prediction can be defined. The REGA HIV-1 subtyping tool (4) uses phylogenetic methods to identify the subtype of a query sequence and further analyses the sequence for recombination using bootscanning methods. Exact recombination breakpoint positions are not predicted, but the assignment to known CRFs is possible.

Previously, we developed jpHMM (jumping profile hidden Markov model), a method to detect genomic recombinations in HIV-1 (5,6) and to accurately locate recombination breakpoints. The jpHMM is a probabilistic generalization of the jumping alignment algorithm proposed by Spang *et al.* (7). For an HIV-1 genomic sequence, jpHMM predicts whether it is a recombinant of different subtypes. If so, it estimates the recombination breakpoint positions and assigns to each segment in between two breakpoints a parental subtype among the major HIV-1 subtypes. The predicted recombination pattern is represented graphically in addition to a list of fragment coordinates and their HIV-1 subtypes. The jpHMM was previously tested on a large set of real and simulated

*To whom correspondence should be addressed. Tel: +49 551 3914926; Fax: +49 551 3914929; Email: mstanke@gwdg.de, mario.stanke@gmail.com

HIV-1 data (5). The evaluation of its prediction accuracy showed that jpHMM is more accurate than competing methods for phylogenetic breakpoint detection.

Nevertheless, it is indispensable to know how reliable the predicted recombination breakpoints and parental subtypes in a particular sequence or a particular region of a sequence are. For this reason, we extended the output of jpHMM to include a tagging of regions where the model is 'uncertain' about the predicted parental subtype and provide an 'interval' estimate for each predicted breakpoint in addition to predicting its precise position. Similar approaches to assess the robustness of predicted breakpoint positions and parental subtypes (or sequences) have been developed in other recombination detection tools: TOPALi v2 (8,9) is a tool for the evolutionary analysis of multiple sequence alignments. It comprises three recombination detection tools, that look for changes in phylogenetic tree topologies moving along an alignment. Statistical significance is assessed by posterior probabilities assigned to each topology for each position in the alignment. The cBrother (10) estimates the recombinant structure of a query sequence and provides posterior support for each genotype at each query sequence position and each breakpoint position. Recco (11) provides a very good visualization tool for locating recombination breakpoints (or breakpoint intervals) in a query sequence. It identifies the parental sequences within a given set of sequences and indicates robust sequence positions.

METHODS

jpHMM

The recombination prediction of jpHMM is based on a pre-calculated multiple sequence alignment of the major HIV-1 subtypes. Each subtype in the alignment is modeled as a profile HMM (12). In addition to the usual state transitions within these profile HMMs, transitions, called 'jumps', between the different profile HMMs are allowed at almost any position in the alignment. Thus, the model can jump between states corresponding to different subtypes, depending on which subtype is locally most similar to the query sequence. The recombination prediction for a query sequence is then defined by a most probable path through the model that generates the sequence, the so-called Viterbi path. Since each state of the model only belongs to one profile HMM and each sequence position is generated by one state of the model, each position of the sequence is assigned to exactly one parental subtype. Positions of jumps between different subtypes define recombination breakpoints.

Uncertainty regions and breakpoint intervals

The new version of jpHMM presented here additionally calculates the so-called 'posterior probability' for each base of the query sequence and each subtype in the given alignment. This quantity denotes the probability that the base belongs to the subtype in our probabilistic recombinant model. The posterior probabilities are calculated using the well-known Forward and Backward algorithms (13). Based on these probabilities, 'uncertainty regions' in the recombination prediction and interval estimates of breakpoints, i.e. intervals where breakpoints can be expected to be located, are defined.

Uncertainty region. If at a certain position of the query sequence the posterior probability of the parental subtype, that was predicted by jpHMM for this position, is lower than a certain threshold $0 \ll t_{\mathrm{UR}} < 1$, this position is marked as 'uncertain' (Figure 1). This classification accounts for the fact that there is a significant ($\geq 1 - t_{\mathrm{UR}}$) probability that the predicted subtype is wrong according to the probabilistic model.

For uncertainty regions, no parental strain can confidently be determined. However, by examining the graph of the posterior probabilities, the user can see which subtypes are closest related in these regions. In the case that an uncertainty region is equally close to two subtypes, the user cannot distinguish whether the uncertainty region is close to both subtypes or far away from them. In this case, we recommend to use the branching index method (14), which quantifies how closely a query sequence clusters with a subtype clade.

Breakpoint interval. An interval estimate of a breakpoint, called 'breakpoint interval', is defined as an interval around a predicted breakpoint position where the posterior probabilities of the predicted subtypes to the left and to the right of the breakpoint, are lower than a certain threshold $0 \ll t_{\mathrm{BPI}} < 1$, but higher than the posterior probabilities of all other subtypes (Figure 1). If the posterior probability of a third subtype is higher than the posterior probability of one of the two predicted subtypes in this region, the whole region is marked as 'uncertain', since this indicates the possibility of an undetected recombination segment.

The length of a predicted breakpoint interval depends on how precisely the breakpoint can be located reliably. A large interval is the consequence of the uncertainty of the model to locate the exact breakpoint position between two subtypes. Thus, the user can see which breakpoints can be located relative precisely or which breakpoints are approximate.

Web server

The jpHMM is available online at http://jphmm.gobics.de/. The user can paste or upload up to five full-length HIV-1 genomic sequences or fragments at a time in FASTA format. A hyperlink to the results of the program run, which are stored on the server for 2 days, is returned to the user by e-mail. The result contains for each sequence the predicted recombination, including uncertainty regions and breakpoint intervals, in text format as well as a graphical representation of the predicted recombinant fragments within the HIV-1 genome. Additionally, the posterior probabilities of the subtypes for each sequence position are plotted. For uncertainty regions the originally predicted parental subtype is also provided. As thresholds for uncertainty regions and breakpoint intervals we use $t_{\mathrm{BPI}} = t_{\mathrm{UR}} = 0.99$. For each query sequence, the predicted recombination with precise breakpoint positions as well as the predicted recombination

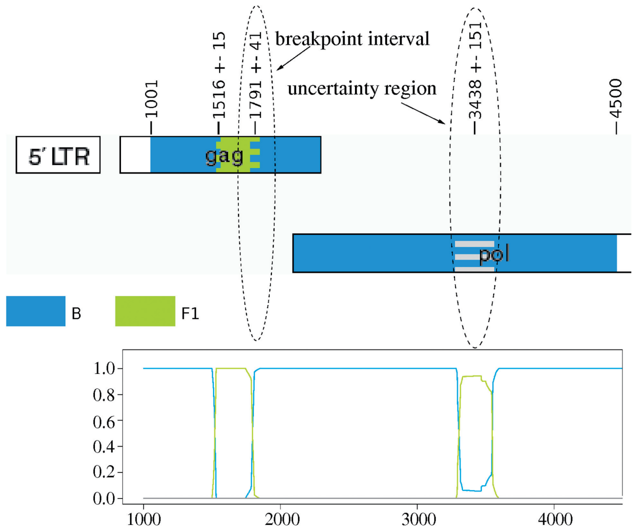


Figure 1. Part of the jpHMM web server output for an artificial recombinant containing alternating B/F1 fragments of lengths 1500/300 nt. Above the genome map of the predicted recombination is shown (drawn with the HIV Sequence Locator Tool, http://hiv.lanl.gov/), below the posterior probabilities of the subtypes. Breakpoint intervals are shown by an interfingering of the colors of the two predicted subtypes, uncertainty regions by an interfingering of grey and the color of the predicted subtype. For the uncertainty region around position 3438 the posterior probabilities give a hint to the correct subtype F1.

including uncertainty regions and breakpoint intervals, a list of the breakpoint intervals and uncertainty regions and the posterior probabilities of the subtypes can be downloaded. Additionally, the alignment of each input sequence to the HXB2 sequence (15), defined by jpHMM, is provided for download. HXB2 is the most commonly used HIV-1 reference sequence and is part of the multiple sequence alignment we use to build the model. Figure 1 shows an excerpt of the jpHMM output for an artificial recombinant HIV-1 sequence.

Evaluation

The accuracy of the new extension of jpHMM was evaluated on 40 semi-artificial near full-length inter-subtype recombinant sequences. For evaluation, we considered the accuracy of the predicted breakpoint intervals and the accuracy of the predicted parental subtypes at positions outside uncertainty regions and breakpoint intervals. As customary, and in lack of real recombinant sequences with exactly known breakpoint positions, these test sequences are real HIV-1 sequences but with artificially introduced breakpoints. Each of the test sequences is a recombination of two ‘real-world’ parental sequences from two different HIV-1 (sub-)subtypes. Hereby, we chose every possible pair of the subtypes A1, B, C, D, F1, G and CRF01 as parental subtypes. To simulate unknown sequences that also differ by mutations from the known sequences, the parental sequences of all test sequences are not contained in the multiple sequence alignment we use to build the model.

The parental sequence pairs were used in three different datasets. In the first dataset, we introduced breakpoints at every 1000th position in the sequences. So, segments of length 1000 nt of one subtype were interrupted by segments of length 1000 nt of another subtype. In the second and third datasets, we used simulated recombinants where alternating long segments (1500 nt) from one subtype are interrupted by short segments (500 and 300 nt, respectively) from another subtype. So, in total, jpHMM was tested for 120 artificial recombinant sequences, each having eight to ten recombination breakpoints.

Table 1. Comparison of the accuracy of breakpoint intervals (BPI) predicted by jpHMM and the accuracy of BPI of fixed length

Threshold t_{BPI}	BPI length		Percentage of BP found using	
	Average	Minimum/ Maximum	P_{post}	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

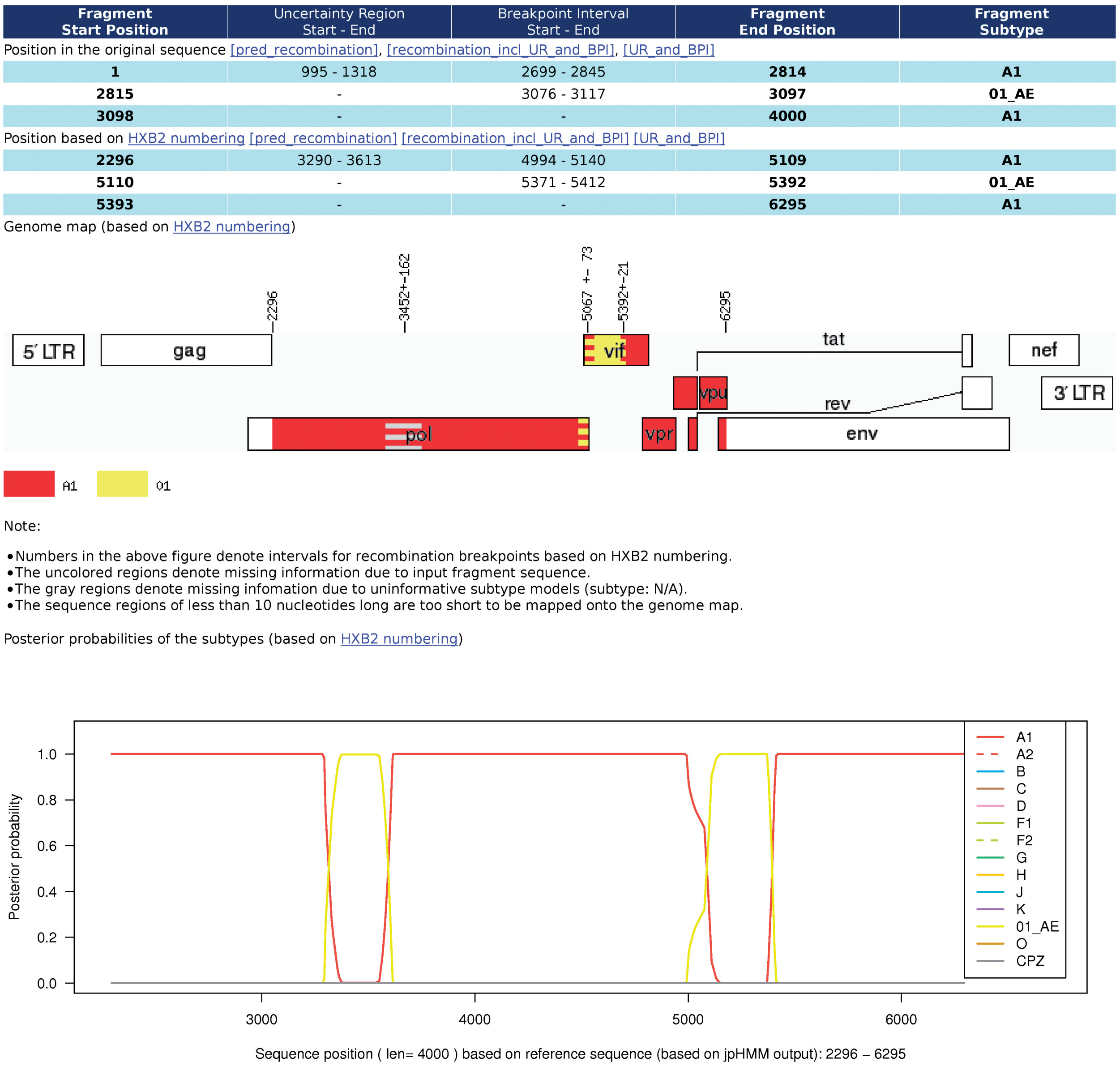
Results are shown for the first dataset. In column 1, the threshold t_{BPI} is given. The average length of the BPI defined by t_{BPI} is given in column 2, the minimal and maximal length in column 3. In column 4, the percentage of real breakpoints detected with these BPI is shown, in column 5 the percentage of breakpoints using the average length of the predicted BPI as fixed BPI length (naïve method). For each threshold the highest value is marked in bold face.

RESULTS

We determined for different thresholds t_{BPI} the number of real breakpoints detected by the predicted breakpoint intervals for each data set. A breakpoint is defined as ‘detected’, when the breakpoint interval contains the true breakpoint and the two neighboring subtypes are predicted correctly. In Table 1, the results are shown for the first dataset (1000/1000 nt fragments). For each threshold the average, the minimal and the maximal length of the predicted breakpoint intervals, and the percentage of detected breakpoints is given. For example for the default threshold $t_{BPI} = 0.99$, 92.50% of the real breakpoints could be detected (Table 1, column 4). The average length of the predicted breakpoint intervals for this threshold is 48.58 nt, the minimal length is 5 nt and the maximal length is 233 nt.

Besides the accuracy of the predicted breakpoint intervals, we were also interested in the ability of jpHMM to predict the true recombination pattern, i.e. the correct sequence of subtypes. Please note, that the recombination pattern of a sequence can be predicted correctly, even if not all breakpoints were detected, according to our definition of a detected breakpoint. For the first dataset, for 39 of the 40 test sequences the recombination pattern was predicted correctly, only in one sequence one recombinant segment was not identified.

For the test sequences containing segments of length 1500/500 nt, 82.72 % of all breakpoints could be detected (with an average breakpoint interval length of 43.73 nt),



Fragment Start Position	Uncertainty Region Start - End	Breakpoint Interval Start - End	Fragment End Position	Fragment Subtype
Position in the original sequence [pred_recombination], [recombination_incl_UR_and_BPI], [UR_and_BPI]				
1	995 - 1318	2699 - 2845	2814	A1
2815	-	3076 - 3117	3097	01_AE
3098	-	-	4000	A1
Position based on HXB2 numbering [pred_recombination] [recombination_incl_UR_and_BPI] [UR_and_BPI]				
2296	3290 - 3613	4994 - 5140	5109	A1
5110	-	5371 - 5412	5392	01_AE
5393	-	-	6295	A1

Genome map (based on HXB2 numbering)

Note:

- Numbers in the above figure denote intervals for recombination breakpoints based on HXB2 numbering.
- The uncolored regions denote missing information due to input fragment sequence.
- The gray regions denote missing infomation due to uninformative subtype models (subtype: N/A).
- The sequence regions of less than 10 nucleotides long are too short to be mapped onto the genome map.

Posterior probabilities of the subtypes (based on HXB2 numbering)

Posterior probability
1.0
0.8
0.6
0.4
0.2
0.0
3000
4000
5000
6000
Sequence position (len= 4000) based on reference sequence (based on jpHMM output): 2296 – 6295
A1
A2
B
C
D
F1
F2
G
H
J
K
01_AE
O
CPZ

Figure 2. Extract of the jpHMM web server output for an artificial recombinant. The output contains a list of fragments from the input sequence that are assigned to different HIV-1 subtypes, including predicted breakpoint intervals and uncertainty regions. In the center, a graphical representation of the predicted recombinant fragments within the HIV-1 genome is given. At the bottom, the posterior probabilities of all HIV-1 subtypes are plotted.

and for the sequences containing segments of length 1500/300 nt, 87.50% (average breakpoint interval length is 41.24 nt). The near full-length sequences used in the input multiple alignment are not always complete at the end regions, and therefore, the multiple sequence alignment we use is 'frayed' and less informative at the sequence ends. jpHMM is thus often not able to assign any subtype to positions located near the ends of the genome. For 33 of the 40 sequences containing segments of length 1500/500 nt, one short segment (500 nt) at the sequence end was not assigned to the correct subtype, since jpHMM was not able to assign any subtype in this region. So, 9.35% (33 out of 353) of the real breakpoints could not be detected, only because they were located within an unclassified region. Apart from these 33 breakpoints, the predicted recombination pattern was correct. In contrast, in eight sequences of the third dataset (1500/300 nt), in fact one short segment (300 nt) was not assigned to the correct subtype. Five of these eight segments were classified as uncertainty region, the other three segments were predicted to have the same subtype as their neighbors, i.e. they could not be identified as a recombinant segment. For the remaining 32 sequences the recombination pattern was also predicted correctly.

Using $t_{UR} = t_{BPI}$ as threshold for the uncertainty regions, for all given thresholds (0.75–0.9999), 92.44–92.68% of the positions outside breakpoint intervals and uncertainty regions were assigned a subtype and classified correctly. Additionally, 6.74% of the positions at the sequence ends were not assigned to any subtype, so the total percentage of positions outside breakpoint intervals and uncertainty regions that were classified incorrectly is only 0.58–0.82%. For precise recombination prediction with jpHMM, i.e. including precise breakpoint estimates and no uncertainty regions, this is 1.51%. For the other two datasets, 0.66–0.99% (1500/500 nt) and 0.75–1.08% (1500/300 nt) of those positions were predicted incorrectly, compared with 1.55% and 1.97 %, respectively, for precise jpHMM recombination prediction.

Comparison to a naïve approach

The accuracy of the predicted breakpoint intervals was compared with the accuracy of a naïve method, that predicts breakpoints in a symmetric interval of fixed length, centered around the predicted breakpoint position. This naïve approach is the most obvious method to define breakpoint intervals around predicted breakpoint positions, if no further information is provided. For a direct comparison, we used the average length of the predicted breakpoint intervals as the fixed interval length in the naïve method, rounded to the nearest even number.

Table 1 shows that for all tested thresholds t_{BPI}, especially for high thresholds, the number of breakpoints detected with breakpoint intervals defined by the posterior probabilities is much higher than when using breakpoint intervals of fixed length (Table 1, column 5). For example, for the default threshold, only 71.94% of all breakpoints could be detected with the naïve method, compared with 92.50%. So, the sensitivity of our method is up to 20 percentage points higher than that of the naïve method. For the other two datasets, the results are similar. For the sequences containing segments of length 500 nt, for the default threshold only 62.04% of all breakpoints could be detected using breakpoint intervals of fixed length, compared with 82.72% with our method. For segments of length 300 nt this is 73.75% compared with 87.5%.

CONCLUSIONS

We extended the jpHMM output to include information about the reliability of the predicted recombination breakpoints and parental subtypes. Our results show that breakpoint intervals defined by the posterior probabilities of the subtypes are far more accurate than breakpoint intervals of fixed length as used in naïve approaches. Additionally, <1% of all positions outside uncertainty regions and breakpoint intervals were classified incorrectly, so the user can now be more confident in the predicted parental subtypes outside these regions. The definition of uncertainty regions helps researchers to avoid drawing wrong conclusions based on doubtful, uninformative regions, such as the postulation of a new CRF.

ACKNOWLEDGEMENTS

We thank Tilman Sauer for helpful discussions and comments on the user interface and the anonymous referees for their suggestions.

FUNDING

DFG (grants STA 1009/4-1 and STA 1009/5-1); NIH-DOE Interagency Agreement Y1-AI-8309. Funding for open access charge: Deptartment of Bioinformatics, Georg-August-Universität Göttingen.

Conflict of interest statement. None declared.

REFERENCES

1. Robertson,D.L., Anderson,J.P., Bradac,J.A., Carr,J.K., Foley,B., Funkhouser,R.K., Gao,F., Hahn,B.H., Kalish,M.L., Kuiken,C. *et al*. (2000) HIV-1 nomenclature proposal. *Science*, **288**, 55d.
2. Hoelscher,M., Dowling,W.E., Sanders-Buell,E., Carr,J.K., Harris,M.E., Thomschke,A., Robb,M.L., Birx,D.L. and McCutchan,F.E. (2002) Detection of HIV-1 subtypes, recombinants, and dual infections in East Africa by a multi-region hybridization assay. *AIDS*, **16**, 2055–2064.
3. Lole,K.S., Bollinger,R.C., Paranjape,R.S., Gadkari,D., Kulkarni,S.S., Novak,N.G., Ingersoll,R., Sheppard,H.W. and Ray,S.C. (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.
4. de Oliveira,T., Deforche,K., Cassol,S., Salminen,M., Paraskevis,D., See-bregts,C., Snoeck,J., van Rensburg,E.J., Wensing,A.M.J., van de Vijver,D.A. *et al*. (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.
5. Schultz,A.-K., Zhang,M., Leitner,T., Kuiken,C., Korber,B., Morgenstern,B. and Stanke,M. (2006) A jumping profile hidden markov model and applications to recombination sites in HIV and HCV Genomes. *BMC Bioinformatics*, **7**, 265.
6. Zhang,M., Schultz,A.-K., Calef,C., Kuiken,C., Leitner,T., Korber,B., Morgenstern,B. and Stanke,M. (2006) jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res*., **34**, W463–W465.
7. Spang,R., Rehmsmeier,M. and Stoye,J. (2002) A novel approach to remote homology detection: jumping alignments. *J. Comp. Biol*., **9**, 747–760.
8. Milne,I., Wright,F., Rowe,G., Marshall,D.F., Husmeier,D. and McGuire,G. (2004) TOPALi: software for automatic identication of recombinant sequences within DNA multiple alignments. *Bioinformatics*, **20**, 1806–1807.
9. Milne,I., Lindner,D., Bayer,M., Husmeier,D., McGuire,G., Marshall,D.F. and Wright,F. (2009) TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*, **25**, 126–127.
10. Fang,F., Ding,J., Minin,V.N., Suchard,M.A. and Dorman,K.S. (2007) cBrother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics*, **23**, 507–508.
11. Maydt,J. and Lengauer,T. (2006) Recco: recombination analysis using cost optimization. *Bioinformatics*, **22**, 1064–1071.
12. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
13. 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.
14. Hraber,P., Kuiken,C., Waugh,M., Geer,S., Bruno,W.J. and Leitner,T. (2008) Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index. *J. Gen. Virol*, **89**, 2098–2107.
15. Korber,B., Foley,B., Kuiken,C., Pillai,S. and Sodroski,J. (1998) Numbering positions in HIV Relative to HXB2CG. *Human Retroviruses and AIDS 1998*, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, pp. 102–111.