Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W463–W465***
doi:10.1093/nar/gkl255

jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1

Ming Zhang[1,2], Anne-Kathrin Schultz[1], Charles Calef[2], Carla Kuiken[2], Thomas Leitner[2], Bette Korber[2,3], Burkhard Morgenstern[1] and Mario Stanke[1,*]

[1]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Goldschmidtstraße 1, 37077 Göttingen, Germany, [2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and [3]The Santa Fe Institute, Santa Fe, NM 87501, USA

Received February 14, 2006; Revised March 2, 2006; Accepted March 30, 2006

ABSTRACT

Detecting recombinations in the genome sequence of human immunodeficiency virus (HIV-1) is crucial for epidemiological studies and for vaccine development. Herein, we present a web server for subtyping and localization of phylogenetic breakpoints in HIV-1. Our software is based on a jumping profile Hidden Markov Model (jpHMM), a probabilistic generalization of the jumping-alignment approach proposed by Spang *et al.* The input data for our server is a partial or complete genome sequence from HIV-1; our tool assigns regions of the input sequence to known subtypes of HIV-1 and predicts phylogenetic breakpoints. jpHMM is available online at http://jphmm.gobics.de/.

INTRODUCTION

Currently, more than 150 000 partial or complete HIV genome sequences are available in the central HIV database at *Los Alamos National Laboratory* (1); these data are crucial for the development of drugs against AIDS. Analysis of HIV sequence data is challenging, however, since HIV is among the most genetically variable organisms known and recombinations of different HIV subtypes are very common (2). HIV-1 is divided into three major phylogenetic groups, one of which—the M group—is responsible for the AIDS pandemic (3,4). This group is classified into ten subtypes, some of which are further divided into sub-subtypes. Accurate classification of HIV-1 subtypes and recombinants is of crucial importance for epidemiological monitoring and drug development. Therefore, a number of software tools have been developed to classify HIV genome sequences and to identify phylogenetic breakpoints and subtypes in recombinant strains (5,6).

We recently developed a HMM-based method to compare nucleic acid sequences to a given multiple alignment A of a sequence family S for which a classification into subclasses is available (7). We called this method jumping profile Hidden Markov Model (jpHMM) since our approach is a probabilistic generalization of the jumping-alignment (JALI) algorithm proposed by Spang *et al.* (8,9). In JALI, a query sequence s is aligned to a multiple alignment A of a sequence family $S = \{s_1, \ldots, s_n\}$—but s is not aligned to the alignment A as a whole, but different parts of s can be aligned to different individual sequences s_i from A.

Within an alignment of the query s to the sequence family S, 'jumps' are allowed between different sequences from S depending on where the strongest degree of similarity is found. For a jump between two sequences s_i and s_j, a penalty is imposed, similar to the familiar gap penalty used in standard sequence alignment. This approach is particularly useful if the query sequence s is a result of phylogenetic recombinations such that different parts of s are related to different sequences from the family S. JALI has been shown to perform well if an alignment A is to be searched against a sequence database (9).

In our jpHMM approach, we assume that a partition of the sequences from the family S into subclasses is given. Each subclass is modeled as a profile Hidden Markov Model (10). *Within* a subclass, the usual transitions between match, insert and delete states are possible, as in standard profile HMM theory—but in addition, our model allows transitions between profile HMMs corresponding to different subclasses, so a path through our model can switch back and forth between different subclasses. Jumps between subclasses are associated with so-called jump probabilities. A detailed description of this approach is given in Schultz *et al.* (7).

PREDICTION OF PHYLOGENETIC RECOMBINATION POINTS IN HIV-1 AT GOBICS

In (7), we found that jpHMM is a useful tool to predict phylogenetic breakpoints and subtypes in recombinant HIV and hepatitis C sequences (11). For HIV subtyping, we start with a pre-calculated multiple alignment of HIV-1 genome sequences consisting of all major subtypes and sub-subtypes; these (sub-)subtypes are modeled as profile HMMs in our

*To whom correspondence should be addressed. Tel: +1 831 459 5232; Fax: +1 831 459 1809

jpHMM approach. It turned out that 'jumps' between these (sub-)subtypes correspond quite well to known phylogenetic breakpoints and (sub-)subtypes to which a query sequence *s* is aligned, reliably indicate the real (sub-)subtypes in recombinant HIV sequences. To evaluate our tool and to compare its prediction accuracy to competing methods such as Simplot (12) and *RDP* (13), we used a large set of real and simulated data from HIV-1 and hepatitis C. These test runs demonstrated that jpHMM is far more accurate than existing tools for phylogenetic breakpoint detection. Details of this program evaluation are described in (7).

To make jpHMM available to the HIV research community, we set up an easy-to-use WWW interface at *Göttingen Bioinformatics Compute Server (GOBICS)*: http://jphmm.gobics.de/ At our server, the user can paste or upload up to 5 full-length HIV-1 genome sequences that is to be searched for phylogenetic breakpoints and subtypes. Our server uses a pre-calculated multiple alignment of 309 HIV sequences from the major HIV (sub-)subtypes obtained from the HIV database at http://hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html. These sequences include nine subtypes *A–D, F, G, H, J, K*, and a persumed recombinant 01_*AE*. Subtype *A* has two sub-subtypes, *A*1 and *A*2; similarly *F* has two sub-subtypes, *F*1 and *F*2. *B* and *D* could be regarded as sub-subtypes because their relative distance and relation are similar to *A*1 and *A*2, *F*1 and *F*2, respectively. But we still consider *B* and *D* as subtypes, not sub-subtypes because of historical reasons (14).

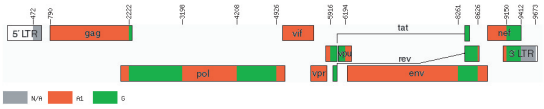
jpHMM result:

Sequence #1: 02_AG.NG.x.IBNG

This sequence is related to subtype: **A1 G** .

Fragment Start Position	Fragment End Position	Fragment Subtype
Position in the original sequence [text]		
1	329	N/A
330	1743	A1
1744	2722	G
2723	3732	A1
3733	4450	G
4451	5439	A1
5440	5726	G
5727	7766	A1
7767	8152	G
8153	8676	A1
8677	8938	G
8939	9201	N/A
Position based on HXB2 numbering [text]		
472	789	N/A
790	2221	A1
2222	3197	G
3198	4207	A1
4208	4925	G
4926	5915	A1
5916	6193	G
6194	8260	A1
8261	8625	G
8626	9149	A1
9150	9411	G
9412	9673	N/A

Genome map (based on HXB2 numbering)



Note:

- Numbers in the figure are breakpoint positions 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.

Figure 1. Sample output from our jpHMM web server. The output file contains a list of fragments from the input HIV-1 sequences that are assigned to different HIV subtypes, including predicted breakpoints. At the bottom of the file, a graphical representation of the input sequence is given where recombinant subtypes are color coded. Gray regions denote missing subtype information due to uninformative subtype models.

01_*AE*, though being called recombinant, contains the only information of subtype *E*. Thus we include 01_*AE* in the alignment. The alignment of these sequences has been carried out using HMMER (15) and subsequent manual improvement.

A hyperlink to the results of the program run is returned to the user by e-mail. The result file contains a list of fragments of the input sequence that are assigned to different subtypes and sub-subtypes, including predicted breakpoints between these fragments. In addition, the output file contains a graphical representation of the predicted recombinant fragments within the HIV-1 genome. A sample output file is shown in Figure 1. The predicted breakpoint positions are provided in two ways. One is based on the original sequence position, and the other is based on HXB2 numbering. HXB2 (GenBank accession number K03455) is the most commonly used reference strain for many different kinds of HIV-1 functional studies. The HXB2 numbering provided for the output breakpoints is especially useful to facilitate the identification of the precise location of interest in HIV sequences.

PROGRAM LIMITATIONS AND FUTURE WORK

It should be mentioned that our tool is sometimes not sensitive to detect HIV-1 subtypes H, J, K, as only few full-length genome sequences of these subtypes are available to train our model. For these subtypes, we recommend to compare the results of jpHMM with those of other HIV-1 subtyping tools, for example, RIP (http://hiv-web.lanl.gov/content/hiv-db/RIPPER/RIP.html).

As shown in (7), the overall prediction accuracy of our method is high compared with alternative approaches. Nevertheless, it would be useful for the user to assess the relative reliability of individual predicted breakpoints. In principle, this is possible by using posterior probabilities that can be calculated using the Forward and Backward algorithms as explained in (16). We are currently implementing these algorithms to estimate the (local) reliability of our predictions. This feature will be available on our web site in the near future.

For predicted recombinants, users of our software may want to know putative parental sequences. Our method cannot provide this information directly, since jpHMM compares input sequences to a model derived from a pre-calculated alignment of representative sequences. It is possible, however, to search predicted recombinant segments of input sequences against the HIV-1 database to retrieve potential parent sequences. We are planning to add this functionality to our web server soon.

ACKNOWLEDGEMENTS

This project was funded in part by grant NIH Y1-AI-1500-01, the NIH-DOE interagency agreement, the HIV Immunology and Sequence Database and by BMBF grant 01AK803G (Medigrid) and DFG grant 1048/1-1 to BM. We thank Rasmus Steinkamp and Maike Tech for helping us with the web server at GOBICS. Two anonymous referees made useful comments on the manuscript. Funding to pay the Open Access publication charges for this article was provided by the annual budget of BM's research group.

Conflict of interest statement. None declared.

REFERENCES

1. Leitner,T., Foley,B., Hahn,B., Marx,P., ,McCutchan,F., Mellors,J., Wolinsky,S. and Korber,B. (eds) (2005) *HIV Sequence Compendium 2005*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM.
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. Sharp,P.M., Shaw,G.M. and Hahn,B.H. (2005) Simian immunodeficiency virus infection of chimpanzees. *J. Virol.*, **79**, 3891–3902.
4. 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**, 55–57.
5. Martin,D. and Rybicki,E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, **16**, 3535–3540.
6. Siepel,A.C., Halpern,A.L., Macken,C. and Korber,B.T. (1995) A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses*, **11**, 1413–1416.
7. Schultz,A.-K., Zhang,M., Leitner,T., Kuiken,C., Korber,B., Morgenstern,B. and Stanke,M. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*, **7**, 265.
8. Spang,R., Rehmsmeier,M. and Stoye,J. (2000) Sequence database search using jumping alignments. In Bourne,P., Gribskov,M., Altman,R., Jensen,N., Hope,D., Lengauer,T., Mitchell,J., Scheeff,E., Smith,C., Strande,S. and Weissig,H. (eds), *Proceedings of Intelligent Systems for Molecular Biology 2000*. AAAI Press.
9. Spang,R., Rehmsmeier,M. and Stoye,J. (2002) A novel approach to remote homology detection: Jumping alignments. *J. Comput. Biol.*, **9**, 747–760.
10. Krogh,A., Brown,M., Mian,I., Sjolander,K. and Haussler,D. (1994) Hidden markov models in computational biology: applications to protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.
11. Zhang,M., Schultz,A.-K., Morgenstern,B., Stanke,M., Korber,B. and Leitner,T. (2005) Greater HIV genome diversities inferred from re-subtyping of HIV database sequences. In *Proceedings of German Conference on Bioinformatics (GCB'05), Discovery Notes, Poster Abstracts*, pp. 5–7.
12. 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. Virol.*, **73**, 152–160.
13. Martin,D.P., Williamson,C. and Posada,D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.
14. Kuiken,C. and Leitner,T. (2005) Chapter HIV-1 Subtyping. *Computational and Evolutionary Analysis of HIV Molecular Sequences*. Kluwer Academic Publishers, pp. 27–53.
15. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, **28**, 405–420.
16. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.