BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 19 2005, pages 3797–3800
doi:10.1093/bioinformatics/bti607

Phylogenetics

An automated genotyping system for analysis of HIV-1 and other microbial sequences

Tulio de Oliveira[1,*], Koen Deforche[2], Sharon Cassol[3], Mika Salminen[4], Dimitris Paraskevis[2], Chris Seebregts[5], Joe Snoeck[2], Estrelita Janse van Rensburg[3], Annemarie M. J. Wensing[6,7], David A. van de Vijver[6], Charles A. Boucher[6], Ricardo Camacho[8] and Anne-Mieke Vandamme[2]

[1]Evolution Group at the Zoology Department, University of Oxford, UK, [2]Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, [3]HIV-1 Immunopathogenesis and Therapeutics Research Program, Department of Medical Virology, University of Pretoria, Pretoria, South Africa, [4]Department of Infectious Disease Epidemiology, HIV-Laboratory, National Public Health Institute, Helsinki, Finland, [5]Research Information Systems Division, South African Medical Research Council, Cape Town, South Africa, [6]Department of Medical Microbiology, Sub department of Virology and Eijkman Winkler Institute, [7]Division of Internal Medicine and Dermatology, Department of Internal Medicine and Infectious Diseases, University Medical Center Utrecht, Utrecht, The Netherlands and [8]Virology Laboratory, Hospital Egas Moniz, Lisbon, Portugal

Received on March 28, 2005; revised on June 23, 2005; accepted on July 28, 2005
Advance Access publication August 2, 2005

ABSTRACT

Motivation: Genetic analysis of HIV-1 is important not only for vaccine development, but also to guide treatment strategies, track the emergence of new viral variants and ensure that diagnostic assays are contemporary and fully optimized. However, most genotyping methods are laborious and complex, and involve the use of multiple software applications. Here, we describe the development of an automated genotyping system that can be easily applied to HIV-1 and other rapidly evolving viral pathogens.

Results: The new REGA subtyping tool, developed using Java programming and PERL scripts, combines phylogenetic analyses with bootscanning methods for the genetic subtyping of full-length and subgenomic fragments of HIV-1. When used to investigate the subtype of previously published reference datasets that were analysed using manual phylogenetic methods, the automated method correctly identified 97.5–100% of non-recombinant and circulating recombinant forms of HIV-1, including 108 full-length, 108 *gag* and 221 *env* sequences downloaded from the Los Alamos database.

Availability: The tool, which can be easily downloaded and installed on either a UNIX or Linux-based computer system, is available at http://www.bioafrica.net/subtypetool/html/

Contact: tulio.deoliveira@zoology.oxford.ac.uk

INTRODUCTION

A high level of commitment to AIDS research combined with recent advances in automated sequencing have lead to the rapid accumulation of large amounts of HIV-1 and microbial sequence data. Between September 2000 and 2004 the HIV-1 database increased from approximately 42 000 to 115 000 sequences. There has been a similar rapid growth in both the hepatitis B (HBV) and C (HCV) databases. Although these large datasets contain a wealth of information needed to design effective treatment and prevention strategies, it is difficult to manipulate them using stand-alone computer applications. Effective utilization of these databases will depend, in large measure, on the development of high-throughput software applications for the analysis of both nucleotide and amino acid sequence data.

A number of software applications addressing this issue are either available or under development. However, many of these programmes are highly specialized. In the setting of HIV-1, these software tools include methods for the detection of drug-induced resistance mutations (Shafer *et al*., 1999), the identification of immunodominant epitopes for inclusion in an AIDS vaccine (Korber *et al*., 2002) and, more recently, the development of an integrated interface for studying adaptive evolution in the HIV-1 genome (De Oliveira *et al*., 2003). Despite these efforts, determining the genetic subtype of HIV-1, one of the most basic analyses, is often performed manually. A rapid, accurate and reliable subtyping tool that eliminates the complexity of phylogenetic analysis, and that can be widely applied to a variety of different datasets, would be highly beneficial.

HIV-1 strains are currently classified into three groups, group M, N and O (Robertson *et al*., 2000). The HIV-1 group M is the responsible for the majority of the HIV/AIDS infections worldwide. The variation in the HIV-1 group M is high, up to 30% at nucleotide level, in certain regions of the envelope gene. Group M

*To whom correspondence should be addressed.

Table 1. The current publicly available HIV-1 subtyping tools and its methods.

Bioinformatics tools: **Methods:**	**RIP** **Los Alamos**[a]	**HIVSeq.** **Stanford**[b]	HIV Genotyping **NCBI**[c]	**SUDI** **Los Alamos**[d]	**STAR**[e]	**REGA HIV-1** **Subtyping tool** (this paper)
Similarity search	*YES*	*YES*	*YES*	No	**No**	No
Amino acid matrices	No	No	No	No	YES	No
Phylogenetic analysis	No	No	No	*YES*	No	*YES*
Bootstrap support	No	No	No	No	No	*YES*
Bootscanning similarity	*YES*	No	*YES*	No	No	No
Bootscanning phylo analysis	No	No	No	No	No	*YES*
Phylogenetic signal detection	No	No	No	No	No	YES
Time for execution (1000 bp)	15 s	5 s	10 s	5 s	5 s	10 s

[a] http://www.hiv.lanl.gov/content/hiv-db/RIPPER/RIP.html
[b] http://hivdb.stanford.edu/
[c] http://www.ncbi.nih.gov/projects/genotyping/
[d] http://www.hiv.lanl.gov/content/hiv-db/SUDI/sudi.html
[e] http://pgv19.virol.ucl.ac.uk/download/star_linux.tar

strains are currently classified into 9 subtypes (labelled A–D, F–H, J and K) and 16 circulating recombinants forms (epidemic recombinant virus with more than one subtype). A high mutation rate, combined with extensive recombination, leads to the production of new recombinants and unclassified sequences on a daily basis (Rambaut *et al*., 2004). At least four web-based tools have been developed to assist researchers in the genetic classification of HIV-1. These include the Stanford HIV-Seq program for assessing the impact of RT and protease resistance mutations on phenotypic resistance (http://hivdb.Stanford.edu), the NCBI Genotyping Program (http://www.ncbi.nih.gov/projects/genotyping/), the Los Alamos Recombinant Identification Program (RIP) (http://hivweb.lanl.gov/RIP/RIPsubmit.html) and the European-based Subtype Analyzer Program (STAR) (http://pgv19.virol.ucl.ac.uk/download/star_linux.tar) for the subtyping of both recombinant and non-recombinant viruses. The first three programs utilize a similarity search tool, implemented in BLAST, to determine the genotype of a query sequence. These 'similarity'-based methods allow for the identification of recombinant viruses using boot-scanning methods, but they all require further confirmation using proper phylogenetic methods (Rozanov *et al*., 2004; Gale *et al*., 2004). The fourth software program uses amino acid matrices to create an identification index. This index is then used to determine the genetic subtype of the query sample. In this report, we describe the development of an automated testing algorithm using phylogenetic analyses, not only to determine subtype of a query sequence and to identify inter-subtype recombinants, but also to assess the quality of sequence alignments used during the analysis. The tool can be easily installed on a local computer or can be accessed on a remote server via the internet using a web-based browser interface.

SYSTEMS AND METHODS

The subtyping algorithm consists of four sequential steps. In the initial step, the query sequence is first compared with a full genome reference alignment constructed from 27 pure subtype sequences representing group M subtypes A–D, F–H, J and K, and then trimmed to a uniform length (detailed information on the alignments can be found at http://www.bioafrica.net/subtypetool/alignment.html). The alignment, created using the profile alignment functions of Clustal W (Thompson *et al*., 1994), is then used to construct a phylogenetic tree using the HKY evolutionary model with gamma distribution of sites as implemented in the PAUP* software programme (Swofford). Sequences that form a tight cluster within a known 'pure' subtype are considered to be non-recombinant, while sequences that branch out between subtype clusters are considered to be either CRFs, recombinant viruses or unclassified viral subtypes. The reliability of the clustering is assessed using 100 bootstrap replicates, considering 70% as the cut-off value. In the second step, this entire process is repeated using a more complex reference alignment consisting of 28 previously-characterized full genome CRF strains, in addition to 22 full genome 'pure' subtype sequences. Tight clustering (again >70% of bootstrap replicates) with a known CRF provides information on the mosaic nature of recombinant sequences, while repeat clustering with the same subtype provides further confirmation of non-recombinant 'pure' strains. The user should take care that assignment to a CRF can only be done in regions where the CRF contains a recombination breakpoint. Sequences that do not segregate with any of the known reference strains are given the designation 'unclassifiable' and are further investigated as new recombinants, or as 'new' subtypes. In the third step, the query sequence is divided into small segments and a sliding window of 400 bp is moved along the sequence in 20 bp increments. Each segment in the query sequence, and the reference alignment, is then analysed for recombination using bootscanning methods, implemented in PAUP*. Finally, the alignments are examined to determine whether they contain sufficient phylogenetic signal for subtype determination using the likelihood mapping analysis implemented in the Treepuzzle software (Strimmer and von Haeseler, 1997). Upon completion of this multi-step process, a set of PERL scripts is used to read the program output files and produce an html report containing information on the genotype of the query sequence and its bootstrap support, and on the phylogenetic trees and their sequence alignments. For sequences larger than 800 bp, the bootscanning results are provided in graphical format. For sequences below 800 bp, information is also provided on the quality of the alignment (the phylogenetic signal). The final output of the analysis is a report showing details of the different phylogenetic trees (i.e. with or without CRF reference strains), the bootstrap support for each of the trees, a graphic image of the boot-scanning analyses and values for the phylogenetic signal (and noise). Characteristics of the automated (REGA) tool relative to other vailable subtyping methods is shown in Table 1. Of the two phylogeny-based methods, REGA is the only method that incorporates both bootscanning and signal analyses.

The main program for these analyses is written in Java and can be readily installed on computers running on UNIX, Linux or Mac OS X operating systems. A cgi-bin interface has also been developed to assist with the development of web interfaces. A web-based interface supporting the free implementation of this application is available on the BioAfrica website http://www.bioafrica.net/subtypetool/html

Table 2. Results of the subtyping tool

Dataset	Los Alamos reference	Los Alamos reference	Los Alamos reference	Los Alamos reference	Snoeck *et al*. (1999)	Gordon *et al*. (2004)	CATCH 2004	Portugal blood bank 2003
Number of sequences	108	108	108	221	48	78	2040	1591
Method subtyped	Los Alamos Manual phylogenetic	Los Alamos Manual phylogenetic	Los Alamos Manual phylogenetic	Los Alamos Manual Phylogenetic	Manual phylogenetic	Manual phylogenetic	Manual phylogenetic	HIVSeq.Stanford/ BLAST
Match with REGA HIV-1 subtyping tool (%)	100	100	100	99	98	100	97.5	92.8
Genetic region	Complete genome	GAG	POL	ENV	POL	ENV	POL	POL
Size (bp)	≅10.000	≅2.400	≅1.400	≅2.550	≅1.200	≅600	≅1.200	≅1.300

TESTING AND VALIDATION

As with all new methods, software programmes must be thoroughly validated against well-established 'gold standards'. As described in Table 2, a total of 4302 sequences were used in the test process. These sequences represent reference subtype sequences stored in the HIV Los Alamos Sequence Database (Korber *et al*., 2002), including both published and unpublished sequences. The reference datasets analysed in this study included 403 well-characterized group M subtypes and 142 circulating inter-subtype recombinant (CRF14_BG, etc) virus sequences from the Los Alamos HIV database, a set of *po*l sequences from Belgium (Snoeck *et al*., 2002), an African *env* dataset from KwaZulu-Natal, South Africa (Gordon *et al*., 2003) and two very large *pol* datasets, one from the Europe-wide drug resistance CATCH study with retrospectively collected *pol* sequences from therapy-naïve recently and chronically infected patients in the period 1996–2002 (Wensing *et al*., 2005) and the other from Portugal (Camacho *et al*., personal communication). The recent dramatic increase in *pol* sequences is due to the growing demand for resistance genotyping to help guide treatment programmes. Table 2 shows the high level of agreement with other methods (92.8–100%), the highest agreement being with manual phylogenetic analyses. The availability of an automated subtyping tool will further enhance the evaluation of *pol* as a region for HIV-1 subtype classification, and will help in obtaining information that is epidemiologically and geographically relevant to the global AIDS pandemic.

As shown in Table 2, the REGA subtyping tool is fully concordant with the phylogenetic analysis of full-length and *gag* sequences from the Los Alamos databank, including both 'pure' -subtype, and CRF sequences. Complete (100%) concordance was also observed between the REGA subtyping results and previously-determined subtype classifications for *env* sequences from South Africa. This high level of performance, across a spectrum of different HIV-1 group M subtypes and CRFs and across full-length *gag* and *env* sequences, suggests that the new REGA subtyping tool will be applicable to wide range of different databases. Overall, for both 'pure' subtypes and known CRFs, our subtyping results matched the published data for $>95\%$ of sequences when compared with manual phylogenetic analysis. The highest level of disagreement was between the BLAST-based Stanford HIVSeq and the phylogeny-based REGA subtyping tools, with 7.2% of 1591 *pol* sequences giving discrepant subtype results.

DISCUSSION

The most robust way of assigning a subtype to an unknown sequence is by using phylogenetic analysis. Our tool is the first one of its kind to incorporate both phylogenetic and bootscanning methods in an automated process. We, therefore predict that this tool will be more reliable and accurate than tools that rely on similarity methodologies to define the subtype of an HIV-1 sequence. A significant difficulty with all automated (similarity and phylogenetic) tools is that current subtyping methods are not well defined. The inclusion of large numbers of recombinant sequences has largely invalidated the historically defined HIV-1 subtyping systems that are used today (Rambaut *et al*., 2004). Our automated subtyping tool overcomes many of these difficulties, and it is designed to accept different HIV-1 datasets, including *gag, pol* and *env* subgenomic sequences. The same approach, used to develop the HIV-1 interface, can be easily extended to the analysis of other organisms that can be sub-classified based on their genetic diversity. Subtyping is particularly important when it leads to the identification of organisms that are refractory to current treatment strategies, or that have increased transmissibility and virulence. We are currently developing subtyping interfaces for HCV, HBV and HHV8. In the future, we plan to develop and introduce additional microbial genotyping programs to the automated REGA subtyping tool, based on reference alignments constructed by experts and on the classification system of the International Committee of Virus Taxonomy. The REGA tool is designed to use profile alignment, phylogenetic inference and boot-scanning methods in a UNIX or Linux based computer system. These computer systems increase the speed and stability of the phylogenetic analysis process. For example, a sequence of 1000 bp is subtyped in $\sim$40 s using automated subtyping, whereas the same analysis takes at least 5 min on a stand-alone system. Batch sequence submissions are also available, allowing the subtyping of large numbers of sequences at any one time. Interfaces are freely available for remote web access and are simple and easy to use. In summary, the availability of the REGA subtyping tool will greatly facilitate the process of genotyping sequences from HIV-1 and other organisms, especially from large databases, and in settings where phylogenetic expertize is limited.

ACKNOWLEDGEMENTS

The CATCH-investigators: Gioacchino Angarano, University of Foggia, Foggia, Italy. Birgitta Åsjö, University of Bergen, Bergen,

Norway. Claudia Balotta and Michela Violin, University of Milan, Milan, Italy. Enzo Boeri, Diagnostica e Ricerca San Raffaele, Milan Italy. Maire-Laure Chaix Laboratoire de virologie, Hôpital Necker Paris, France. Dominique Costagliola INSERM EMI 0214, CHU Pitié-Salpétrière, Paris, France. Andrea De Luca, Institute of Clinical Infectious Diseases, Catholic University, Italy, Rome. Inge Derdelinckx and Kristel Van Laethem, Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium. Zehava Grossman Sheba Medical Center, Tel-Hashomer, Israel. Osamah Hamouda and Claudia Kücherer, Robert Koch Institute, Berlin, Germany. Angelos Hatzakis and Dimitris Paraskevis, Athens University Medical School, Athens, Greece. Robert Hemmer, Jean-Claude Schmit and Francois Schneider, Centre Hospitalier de Luxembourg, Luxembourg. Andy Hoepelman and Rob Schuurman, University Medical Center Utrecht, Utrecht, the Netherlands. Andrzej Horban and Grzegorz Stanczak, Hospital for Infectious Diseases & AIDS Diagnosis and Therapy Center, Warsaw, Poland. Klaus Korn, University of Erlangen, Erlangen, Germany. Thomas Leitner, Los Alamos National Laboratory, Los Alamos, USA. Clive Loveday and Eilidh MacRae, International Clinical Virology Centre, Buckinghamshire, England, United Kingdom. Irina Maljkovic and Karin Wilbe, Swedish Institute for Infectious Disease Control, Solna, Sweden. Carmen de Mendoza and Vincent Soriano, Hospital Carlos III, Madrid, Spain. Laurence Meyer, INSERM U569, Kremlin-Bicêtre, France. Claus Nielsen, Statens Serum Institute, Copenhagen, Denmark. Eline L. Op de Coul National Institute for Public Health and the Environment, Bilthoven, the Netherlands, Vidar Ormaasen, Ullevaal University Hospital, Oslo, Norway. Luc Perrin and Sabine Yerly, Geneva University Hospital, Geneva, Switzerland. Elisabeth Puchhammer-Stöckl, University of Vienna, Vienna, Austria. Lidia Ruiz, Retrovirology Laboratory IRSICAIXA Foundation, Badalona, Spain. Mika Salminen, National Public Health Institute, Helsinki. Finland. Maja Stanojevic, University of Belgrade, Belgrade, Serbia-Montenegro, Maurizio Zazzi, University of Siena, Siena, Italy. Supported by grant #061238 from the Wellcome Trust, UK, by the Flanders Bi-lateral Cooperation Grant (BIL02/41), by the European Commission QLK2-CT-2001-01344 SPREAD-programme, by FWO-Vlaanderen grant G.0266.04 and by the Katholieke Universiteit Leuven through Grant OT/04/43. Koen Deforche was funded by a PhD grant from the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWT). T.O. is funded by the Marie Curie Fellowship. Funding to pay the Open Access publication charges for this article was provided by the Flanders Bi-Lateral Cooperation.

Conflict of Interest: Wensing declares that she has received travel grants from Abbott, GlaxoSmithKline, Bristol-Myers Squibb and Roche, and that she has served as a temporary advisor for GlaxoSmithKline and Bristol-Myers Squibb.

REFERENCES

De Oliveira,T. *et al*. (2003) An integrated genetic data environment (GDE)-based LINUX interface for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **19**, 153–154.

Gale,C.V. *et al*. (2004) Development of a novel human immunodeficiency virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London. *AIDS Res. Hum. Retroviruses*, **20**, 457–464.

Gordon,M. *et al*. (2003) Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu-Natal, South Africa: implications for vaccine and antiretroviral control strategies. *J. Virol.*, **77**, 2587–2599.

Korber,B.T.M. *et al*. (2002) HIV Molecular Immunology 2002. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico

Rambaut,A. *et al*. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet.*, **5**, 52–61.

Robertson,D.L. *et al*. (2000) HIV-1 nomenclature proposal. *Science*, **288**, 55–56.

Rozanov,M. *et al*. (2004) A web-based genotyping resource for viral sequences. *Nucleic Acids Res.*, **32** (Web Server Issue), W654–W69.

Shafer,R.W. *et al*. (1999) Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Res.*, **27**, 348–352.

Snoeck,J. *et al*. (2002) Prevalence and origin of HIV-1 group M subtype among patients attending a Belgian hospital in 1999. *Virus Res.*, **85**, 95–107.

Strimmer,K. and von Haeseler,A. (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA*, **94**, 6815–6819.

Swofford,D.L. PAUP* 4.0: phylogenetic analysis under parsimony (and other methods), version 4.0b2a. Sinauer Associates Inc., Sunderland, Mass.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wensing,M.J. *et al*. (2005) Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *J. Infectious Diseases*, **192**, 958–966.