Comparative performance of the REGA subtyping tool version 2 versus version 1

Ana B. Abecasis a,b,1, Yunpeng Wang a,1, Pieter Libin a, Stijn Imbrechts a, Tulio de Oliveira c, Ricardo J. Camacho b,d, Anne-Mieke Vandamme a,*

Laboratory for Clinical and Epidemiological Virology, REGA Institute for Medical Research, Katholieke Universiteit Leuven, Minderbroedersstraat 10, Leuven, Belgium
 Centro de Malária e Outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Portugal
 Africa Centre for Health and Population Studies, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, South Africa
 Laboratory for Clinical and Epidemiological Virology, Hospital de Egas Moniz, Centro Hospitalar de Lisboa Ocidental, Portugal

ARTICLE INFO

Article history:
Received 22 June 2009
Received in revised form 29 September 2009
Accepted 30 September 2009
Available online xxxx

Keywords:
HIV-1
Subtyping

Abstract

The REGA HIV-1 subtyping tool is a phylogenetic-based method for subtyping HIV-1 genomic sequences that was published in 2005. The subtyping tool combines phylogenetic approaches with recombination detection methods. Recently, version 2 was released (http://www.bioafrica.net/rega-genotype/html/index.html) as an improvement of version 1. Version 2 implements a Decision-Tree-based algorithm that was not implemented in version 1. We wanted to compare the two versions on a large sequence dataset to assess the improvements of version 2 and to verify whether features lost during updating the tool needed to be recovered. We analysed the results of the two versions in the genotyping of 4676 HIV-1 pol sequences. We compared those results to a manual approach, used in previous studies. Our results show that version 2 has an overall better sensitivity but especially for the detection of subtypes A, B, D, F, G and CRF14_BG and CRF06_CPX. For the other subtypes, no significant differences were observed in the sensitivity of versions 1 and 2. The overall increase in sensitivity was however accompanied by a decrease in the specificity for the detection of subtype B. This is the main limitation of version 2. However, while the number of false negatives decreased by 53 samples, the number of false positives increased only by 5 samples from version 1 to 2. The performance of the REGA HIV-1 subtyping tool was considerably improved from one version to the other. Our results are very valuable and allow us to make suggestions for further improvement of the tool for a version 3 release.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Human immunodeficiency virus (HIV) is a retrovirus that causes the acquired immunodeficiency syndrome (AIDS). Due to high rates of mutation and replication, the fast accumulation of proviral variants during the course of infection and the high rate of recombination, HIV exhibits an extraordinary genetic diversity. Historically, HIV-1 has been classified into three major groups: M (major), O (outlier) and N (non-M, non-O). Group M is responsible for the global pandemic, and based on phylogenetic analysis it was further divided into 9 subtypes, A–D, F, G, H, J and a large number of inter-subtype recombinants. Recombinant viruses that are identified in at least three epidemiologically unrelated individuals and characterized by full genome sequencing are designated as circulating recombinant forms (CRFs) (Robertson et al., 2000). Up until now, 43 CRFs have been identified (Los Alamos database). The remaining recombinant forms, which are found in isolated or small groups of epidemiologically related individuals, are called unique recombinant forms (URFs).

Different subtypes and CRFs have distinct global distribution patterns (Osmanov et al., 2002). On a global scale, the most prevalent HIV-1 genetic forms are subtypes A, B, C, D and G, accounting for 12%, 10%, 50%, 3% and 6%, respectively, and CRF01_AE, CRF02_AG, each accounting for 5%, of all HIV-1 infections worldwide. In particular, subtype B is responsible for 67% of the overall infections in newly diagnosed patients in Western Europe (Abecasis et al., 2008), while subtype A accounts for 79% in Eastern Europe (Buonaguro et al., 2007; Hemelaar et al., 2006).

The high level of genetic variability of HIV-1 may have important implications for HIV pathogenesis, transmission, diagnosis, treatment and vaccine development. It is plausible that different subtypes have different biological properties resulting in differences in transmissibility and pathogenicity, but this issue is still a matter of debate (Hemelaar et al., 2006). A few studies have shown that subtype D may lead to more rapid disease progression than other subtypes (Baeten et al., 2007). It was also reported that...
Fig. 1. Decision-Tree algorithm of the REGA subtyping tool used to subtype query sequences longer than 800 bp. A brief description of the algorithm is described in Section 1.

Rule 1A: pure—subtype assigned based on sequence >800 bp, clustering with a pure subtype with bootstrap >70% without recombination in the bootscan, and do not clustering with a CRF with bootstrap >70%.

Rule 1B: pure—subtype assigned based on sequence >800 bp, clustering with a pure subtype with bootstrap >70% without recombination in the bootscan, clustering with a CRF with bootstrap >70% however not clustering inside the CRF cluster.

some subtype G samples are less susceptible to protease inhibitors (Abecasis et al., 2006). Group O viruses are known to be naturally resistant to non-nucleoside reverse transcriptase inhibitors. The M group viruses are shown to have similar susceptibility to currently used drugs, at least in vitro (Abecasis et al., 2006; Palmer et al., 1998) but some groups reported that subtype B is different from other subtypes in the generation of drug resistance mutations under treatment selective pressure (Abecasis et al., 2005; Grossman et al., 2004; Pieniazek et al., 2000).

Recently, classification of HIV-1 sequences became more based on online subtyping tools than on manual phylogenetic analysis. At present, several web-applications are available: the NCBI genotyping program (http://www.ncbi.nlm.nih.gov/projects/genotyping/form/ page.cgi), the Los Alamos RIP program (http://hivweb.lanl.gov/RIP/ RIPSubmit.html), the Stanford HIV-seq program (http://hivdb-Stanford.edu), the STAR subtype analyser (http://www.biochem-ucl.ac.uk/bsm/virus_database/) and the REGA subtyping tool (http://www.bioafrica.net/subtypetool/html). The NCBI genotyping program is based on a BLAST-based sliding window approach. The Los Alamos RIP program uses a sliding window approach based on similarity distance measurements. With the Stanford HIV-seq software, the subtype of the most similar reference sequence, in protease (PR) and reverse transcriptase (RT) separately, is assigned to the query sequence. The STAR subtype analyser uses position-specific scoring matrix (PSSM)-based genotyping. Finally, the REGA HIV-1 subtyping tool is the only one to use a phylogeny-based subtyping method (de Oliveira et al., 2005; Gale et al., 2004; Korber et al., 2002; Rozanov et al., 2004).

The first version of the REGA subtyping tool was based on a streamline scheme. Two Neighbour-Joining (NJ) trees of the query with two sets of pre-selected reference sequences, one only with the pure subtype reference sequences and the other containing both pure subtypes and CRFs, were constructed sequentially. After building each tree, bootstrap testing (100 replicates) was used to test the reliability of the tree clustering. Finally, bootscanning analysis and likelihood mapping analysis were used to test recombination and phylogenetic signal (Salminen et al., 1995; Strimmer and von Haeseler, 1997). A bootstrap value of 70% was used as cut-off value in the preceding two trees, for the assignment of the query sequence either to a particular pure subtype reference or to a CRF reference (de Oliveira et al., 2005).

Version 2, on the other hand, follows a Decision-Tree-based algorithm, as presented in Fig. 1. At the first step, a pure NJ tree is built, containing the query and reference sequences of the so called ‘pure’ subtypes. Depending on the bootstrap value, a different branch of the tree is followed. A bootstrap value of 70% is used as threshold for the split decision between branches of the Decision-Tree. Then, the bootscan method is applied and the bootscan support (defined as the fraction of windows in which the sequence clusters with the more frequently supported subtype with a bootscan support above 70; threshold = 0.9) is now used as split decision criterion. If the bootscanning procedure with only pure subtypes has a bootscan support >0.9, a NJ tree with only CRFs is made. Subsequently the CRF clustering with the query in the CRFs tree is added to the ‘pure’ subtype references and the bootscan method is applied (see Fig. 1). The final result provides the genome subtype pattern schema, the phylogenetic signal based on likelihood mapping analysis, the alignment(s), the tree(s) and the bootscan plot(s). The philosophy of the tool is to assign only when confident. This results in the majority of sequences being assigned a subtype, and where not possible, being flagged for further verification by manual phylogenetic procedures. In this context, some sequences remain unassigned, compared to other subtyping tools, however we see this as an advantage rather than a disadvantage as reported by Holguin et al. (2008), who treated such sequences as wrong assignments, thereby claiming the REGA tool to be unreliable.

In this paper we wanted to compare the two versions of the REGA subtyping tool on a dataset of 4676 pol sequences to assess the improvements of version 2 and to verify whether good features of version 1 were lost during updating the tool that need to be recovered for the next version. We compared those results to a manual approach, used in previous studies. Such comparisons are very valuable to further improve the tool for a version 3 release.

2. Materials and methods

The dataset we used contained 4676 pol sequences derived from patients at the Egas Moniz Hospital, Lisbon, submitted to resistance testing either for therapy failure or for baseline genotyping of drug naïve patients. Data was retrieved from the Egas Moniz RegaDB instance. The sequences were obtained by population sequencing using the ViroSeq 2.0 toolkit (Abbott Laboratories, Abbott Park, IL, USA). The sequences were in general approximately 1300 bp long (Min = 993 bp; Max = 1311 bp; Average = 1295 bp).

HIV-1 pol subtyping was first performed using REGA HIV subtyping tool version 1 and version 2, separately. These results for version 2 were obtained by running this program 10 times and combining the outcome from each run by occurrence. This is due to the random nature of the bootstrap procedure and the high similarity between these subtypes, the use of a bootscan value of 90 as split criteria does not generate consistent assignments of these subtypes. Version 1 on the other hand does not use bootscan value as subtyping criterion, therefore the results are highly consistent.

We then performed manual analysis as follows:

(a) The sequences that were classified as ‘unassigned’ by or discordant between one of the two versions, were subtyped manually by phylogenetic analysis. 423 sequences were unassigned by both, and 68 were discordant between both (Table 1). The programs CLUSTALW (Thompson et al., 1994), Simplot (Lole et al., 1999) and PAUP (Swoford, 1998) were used for alignment, recombination detection and construction of the maximum likelihood tree, respectively. For measuring
the reliability of the clustering. 1000 bootstrap replicates were run in each version. In the situation that the query was clearly assigned by phylogenetic analysis and not assigned by the subtyping tools, we recorded these results as false negatives for both tools.

(b) The concordant assignments of sequences by both tools, in total 4195 sequences, were verified first by bootstrapping analysis with our new reference set, a set of carefully chosen reference sequences downloaded from the Los Alamos HIV sequence database (Los Alamos National Laboratory, 2005). If the bootscan plots showed an apparent crossover between two or more references, a manual phylogenetic analysis was performed. The assignment was considered as false positive if the manual analysis indicated either that the sequence was unassignable to any subtype or CRF or a different assignment than the one indicated by the subtyping tool. Otherwise, if the result of the manual analysis was consistent with the subtyping tool results, we considered the assignment as a true positive and no further manual analyses was performed.

The final results were summarized in terms of specificity, sensitivity and reproducibility. True positive and true negative for each version are defined as the number of queries that get concordant results from manual analysis and the tools, either assigning the sequence to one subtype (true positive) or leaving it unassigned (true negative). False positives and false negatives for each version are the number of queries that get discordant results from manual analysis and each version, respectively. A false positive is a sequence that is incorrectly assigned to one subtype by the tool, while a false negative is a sequence that is incorrectly left as unassigned by the tool. Sensitivity is defined as: number of true positives/(number of true positives + number of false negatives); specificity defined as: number of true negative/(number of true negatives + number of false positives). These measures were summarized both for all subtypes together and in a subtype-specific way. By doing this, we expect to discriminate between subtypes/CRFs that the tools are more/less efficient to genotype.

### 3. Results

In this study, we aimed to compare the performance of the version 2.0 to the original one, so we used reference sequences from the same pure subtypes and CRFs: A–D, F, G, H, J, and K and CRF01-14. We detected more than a hundred new URFs and other CRFs, which have already been published but were not included in the reference set, in which case we consider the lack of assignment by the subtyping tool as true negative.

The discordant results between these two versions are mainly because version 1 failed to classify 58 sequences, whereas version 2 assigned most of the sequences. Upon manually verifying the assignment from version 2, we found that most of the unassigned sequences by version 1 were subtype B, in total 24 sequences. Among these discordant results, one sequence was false positively assigned to subtype B by version 2. Furthermore, 10, 2, 14, 4, 3 and 1 sequences were correctly assigned to subtype G, CRF02_AG, CRF06_CPX, subtype D, subtype A and subtype F, respectively. Finally, 2 sequences were misassigned to CRF06_CPX and CRF02_AG by version 2. The result shows that version 1 is more conservative than version 2, presenting less false positives, but on the other hand much more false negatives than version 2.

For the concordant assignments, we found that the misassignments concentrate on complex recombinant CRF06_CPX, in which case the results of the phylogenetic analysis show that the query sequences seem to be new recombinant forms between CRF02_AG and CRF06_CPX. In this situation, both versions have the same problem.

Table 2 shows the details of the performance of each version for each subtypes and CRFs. For subtype B, there are 9 wrong
assignments for version 1 and 10 for version 2. The frequent crossover between subtypes B and D, due to the close relatedness of these subtypes occurring at the middle region of the pol sequences, is a source of difficulty in assigning sequences to subtype B. There were no misassignments for subtypes A, C, G, H and CRF13_CPX. But both versions conversantly gave 2, 2, 2 and 10 wrong assignments to sequences of subtype D, subtype F, CRF02_AG and CRF06_CPX, respectively. Version 2 also gave 2 extra misassignments to the latter 2 CRFs resulting in the assignment of 58 discordant sequences. Phylogenetic analysis showed that these two sequences wrongly assigned to ‘subtype D’ were actually recombinants between subtypes B and D, and those misassigned to subtype F were recombinants between subtypes B and F. Again, the wrong assignments for CRF02_AG and CRF06_CPX resulted from the fact that the queries were recombinants of CRF02_AG and CRF06_CPX.

For sequences which were not assigned by both tools, manual analysis further assigned 23 subtype Bs, 20 subtype Gs, 2 subtype Ds, 1 subtype A, 2 subtype Fs, 1 subtype H, and 8 CRF06_CPXs. All of the 23 subtype Bs have the same pattern of crossover between subtypes B and D at the middle region of the pol sequences. We found that, in general, if more than two reference subtypes are very closely related, then during bootscan analysis they will compete with one another in the bootstrap procedure. Thus, the query could not get a consistent support to any of them.

Our analysis shows that version 2 performs generally better than version 1, especially in terms of sensitivity (see Table 2). Furthermore, we recorded the running time for both versions, and version 2 is faster that version 1. Subtyping 4676 sequences by both versions concordantly gave 2, 2, 2 and 10 wrong assignments to CRFs resulting in the assignment of 58 discordant sequences. Phylogenetic analysis showed that these two sequences wrongly assigned to ‘subtype D’ were actually recombinants between subtypes B and D, and those misassigned to subtype F were recombinants between subtypes B and F. Again, the wrong assignments for CRF02_AG and CRF06_CPX resulted from the fact that the queries were recombinants of CRF02_AG and CRF06_CPX.

A summary of the advantages and disadvantages of the two versions of the subtyping tool is presented in Table 3.

### 4. Discussion

The REGA subtyping tool version 1, based on a streamline procedure, is more computationally intensive and time consuming, because for any query sequence all the same procedures have to be done before reaching the final result. This takes some unnecessary time on one simple case and means a lot more time when datasets of thousands of sequences are analysed. Whereas subtyping tool version 2, based on a Decision-Tree model, allows to subtype some sequences early in the Decision-Tree process, making the procedure faster in these cases. This feature makes this version faster than version 1.

While a low false positive rate is assured in version 1, this is achieved at the price of a high number of false negative results. During the verification step of version 1, using the bootscan method, any kind of conflict signal will be considered as uncertainty thus leaving the query unassigned. On the other hand, version 2 uses bootscan value to check whether that conflict signal is strong enough to prevent assigning the query. This procedure resulted in 481 sequences that were unassigned in version 1, while this number was decreased to 423 sequences in version 2, only at the cost of 5 extra false positive results in version 2 (Table 1).

Concerning the assignment of “pure” subtype sequences, the main difficulty of the tool, in both versions, was to assign subtype B sequences. However, this can be explained by the high similarity between subtypes B and D and to the frequent occurrence of crossover between these two subtypes in the beginning of RT. This also accounts for the high number of false negatives in subtype G, due to its history of recombination with subtype A (Abecasis et al., 2007). These problems will have to be resolved in the next version of the subtyping tool.

The classification of CRF sequences might be difficult, especially if no recombination breakpoints are included in the analysed sequence. For example, our dataset only contained pol sequences and CRF01_AE and CRF14_BG do not have recombination breakpoints in this region. However, rule 1C is expected to take this into account, by verifying the inside/outside clustering of putative sequences of these CRFs with CRF reference sequences. This procedure assigns a CRF even in the absence of breakpoints, if it clusters significantly within the cluster of reference CRF sequences. Since these reference sequences were chosen only when a full genome is available thus assuring the CRF assignment. In the case of CRF14_BG sequences of our dataset, no sequence was assigned by rule 1C but by rules 4 and 8 that do not verify the inside/outside clustering of the query. Therefore, the correctness of this assignment is dubious and this is why we considered sequences assigned to CRF14_BG as subtype G.

Both versions of the subtyping tool try to avoid the difficulties of determining the breaking point of recombinant sequences by using CRF references. However, currently only CRF01-14 reference sequences are included in the reference set of the tool, therefore all other CRFs (15–43) cannot be assigned by the tool. This is one of the reason why there is such a large number of unassigned sequences. Manual phylogenetic analysis assigned 30 BF recombinants, and some of them have already been published in the Los Alamos sequence database as a new CRF. As the number of CRFs continues to increase, it becomes infeasible to continue to include all these CRF reference sequences in the reference set. One approach could be that the subtyping tool assigns only based on pure subtype references, and after having determined the breaking point of the recombinant, a true phylogenetic relationship could be inferred for each segment. However, this procedure probably also represents much more computation time. Another approach could be to verify which CRF is a reliably assigned CRF and plays a substantial role in the epidemic, leaving out those CRFs that are exceedingly rare or are disappearing from the epidemic. This latter approach is currently being investigated and we are preparing a publication assessing the relevance of assigning particular CRFs by an automated subtyping tool.

Although the performance of especially version 2 of the subtyping tool was good as compared to manual assignment, we believe that improvements could be achieved by changing the criteria rule. In the situation of two closely related subtypes (for example subtypes B and D, or subtypes G and A), when a pure subtype query sequence is taken through the first two steps of the procedure (Fig. 1) and appears to be related to two or more subtypes, then the rule used to discriminate the true recombinant from a pure subtype with noise from the other subtype references (for example: subtype B with noise from subtype D) could be applied with a less stringent bootscan value (currently 90%). Otherwise, the query will go through the rest of the Decision-Tree and remain unassigned. The underlying reason is that when two or more references related to the queries are very closely related in one of the analysed genomic regions, the bootstrap replicates will sometimes support one of the references and other times the other.
therefore, none of these references will get a support above 70 and the bootstrap value will be lower than 90. So, this region will be labelled as unassigned. Optimally, a set of references that are as divergent as possible could reduce this difficulty to some extent. But due to the nature of some subtypes, such as B and D that are not as divergent as the others, and A and G that have a history of recombination events – subtype G being in fact a CRF with CRF02_AG as parent instead of the other way round (Abecasis et al., 2007) – it will be very difficult to fine tune the criteria without introducing too many false positive assignments. For all the reasons mentioned above, we might have to treat different subtypes differently in our Decision-Tree algorithm.

The differences implemented in version 2 compared to version 1 give the tool the strength to detect more pure subtype B sequences with about the same level of false positive rate. However, the discrimination between subtype B and B/D recombinants remains the weak part of the subtyping tool, especially since the majority of sequences submitted to the tool are still subtype B. This partly accounts for the reason why version 2 has to be run several times for a safer assignment. The other reason is that our data set contains sequences from the pol region of the HIV-1 genome, where a relatively low phylogenetic signal exists. Also, no breaking point exists in this region for CRF14_BG and CRF01_AE. CRF14_BG and CRF01_AE are therefore highly similar to the references of subtypes G and A. Even though this problem is taken into account in version 2, by using the inside/outside clustering rule.

Finally, the correct identification of CRF06_CPX is hindered by recombinants between CRFs. In our dataset, we identified 17 new recombinants between CRF02_AG and CRF06_CPX, which have not been published in Los Alamos. Currently used programs arbitrarily assign these cases to CRF02 or CRF06. This is due to the fact that the REGA subtyping tool never includes two CRF reference sequences in the same bootscanning plot, therefore recombinants between two CRFs are impossible to detect. Also this problem will have to be solved in version 3 of the tool.

In this paper, we compared the performance of the two versions of the REGA subtyping tools. Version 2 showed 98.9% and 92.6% sensitivity and specificity, respectively, versus 97.6% and 93.8% of the REGA subtyping tools. Version 2 showed 98.9% and 92.6% sensitivity and specificity, respectively, versus 97.6% and 93.8% of the REGA subtyping tools. Version 2 showed 98.9% and 92.6% sensitivity and specificity, respectively, versus 97.6% and 93.8% of

**References**


