Gene 448 (2009) 134–138

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

The phylogeny of orthoretroviral long terminal repeats (LTRs)

Farid Benachenhou, Vidar Blikstad, Jonas Blomberg *

Section of Virology, Dept. of Medical Sciences, Uppsala University, Dag Hammarskjolds v. 17, SE-75185 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 29 April 2009
Received in revised form 24 June 2009
Accepted 2 July 2009
Available online 9 July 2009

Received by N. Okada

Keywords:
Endogenous retrovirus
LTR
Phylogeny
Hidden Markov Model

ABSTRACT

LTRs are sequence elements in retroviruses and retrotransposons which are difficult to align due to their variability. One way of handling such cases is to use Hidden Markov Models (HMMs). In this work HMMs of LTRs were constructed for three groups of orthoretroviruses: the betaretroviruslike human MMTV-like (HML) endogenous retroviruses, the lentiviruses, including HIV, and gammaretroviruslike human endogenous retroviruses (HERVs). The HMM-generated LTR alignments and the phylogenetic trees constructed from them were compared with trees based on alignments of the *pol* gene at the nucleic acid level. The majority of branches in the LTR and *pol* based trees had the same order for the three retroviral genera, showing that HMM methods are successful in aligning and constructing phylogenies of LTRs. The HML LTR tree deviated somewhat from the *pol* tree for the groups HML3, HML7 and HML6. Among the gammaretroviruslike proviruses, the exogenous Mouse Leukemia Virus (MLV) was highly related to HERV-T in the *pol* based tree, but not in the LTR based tree. Aside from these differences, the similarity between the trees indicates that LTRs and *pol* coevolved in a largely monophyletic way.

1. Introduction

Endogenous retroviruses and LTR retrotransposons are present in a wide variety of organisms ranging from plants and insects to humans. For a review, see for example Blikstad et al. (2008) and Jern and Coffin (2008). One of their most characteristic features is two identical long terminal repeats (LTRs) flanking the protein coding genes. The LTRs vary considerably in length and internal structure but do have a few conserved motifs such as target site duplications (TG-CA), three A-rich regions, which occasionally encompass a TATA signal and always a polyadenylation signal (AATAAA box), see Benachenhou et al. (2009). Due to this diversity LTRs cannot be aligned with the commonly used alignment algorithms such as ClustalW (Thompson et al., 1994) and consequently are not used in e.g. phylogenetic analyses. This is despite the fact that the majority of endogenous retroviruses occur as single LTRs in many genomes (Mager and Medstrand, 2003) and completely lack the other protein coding genes such as the *pol* gene. Thus an important source of information is ignored.

In Benachenhou et al. (2009) several groups of LTRs from vertebrate exogenous and endogenous retroviruses were aligned by means of Hidden Markov Models (HMMs). The two most conserved groups were LTRs from the human MMTV-like (HML) endogenous retroviruses (Blikstad et al., 2008) and from the exogenous lentiviral retroviruses (including HIV-1 and HIV-2). The gammaretroviral HERV LTRs were on the other hand more variable. Here we explore whether phylogenies can be reconstructed from LTR Viterbi alignments for the three groups and compare them with trees obtained from *pol* gene alignments.

2. Results

Profile Hidden Markov Models were built according to the methodology of Benachenhou et al. (2009). The most important issue in the model building is to avoid overfitting by regularising the HMMs. The regularisation method in Benachenhou et al. (2009) was taken from Brand (1999). It has a parameter z that can be thought of as introducing disorder in the training set if negative.

The scoring of the sequences was performed using reverse-sequence null models (Karplus et al., 2005). This scoring method has the virtue of being insensitive to the composition bias of the sequence since it is the difference between the logarithm of the raw score of the sequence and the logarithm of the same sequence in reverse order.

For the three retroviral groups many HMMs were built with increasing number of match states (M) and with different z-values.

Abbreviations: LTR, long terminal repeat; HMM, Hidden Markov Model; MMTV, Mouse Mammary Tumor Virus; HML, human MMTV-like; HERV, human endogenous retrovirus; *pol*, polymerase gene nucleotide sequence; Pol, polymerase gene amino-acid sequence; Gag, group antigen amino-acid sequence; ERV, human endogenous retrovirus; PBS, primer binding site; HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; LST, lhoest's monkey; MND, mandrill; SUN, sun-tailed macaque; DRL, drill monkey; RCM, red-capped mangabey; CPZ, chimpanzee; O.BE, O. CM, HIV-1 type O; SAB, African green monkey, *sabaeus* subspecies; SIV-VER, vervet monkey; SYK, Sykes' monkey; MON, Mona's monkey; MUS, moustached monkey; GSN, greater spot-nosed monkey; DEB, DeBrazza's monkey; DEN, Dent's Mona monkey; COL, *guereza colobus*; BIV, bovine immunodeficiency virus; Visna, ovine maedi-visna virus; FIV, feline immunodeficiency virus; EIAV, equine infectious anemia virus; MLV, murine leukemia virus; GaLV, gibbon ape leukemia virus; FLV, feline leukemia virus.

* Corresponding author. Fax: +46 18 55 10 12.
E-mail address: Jonas.Blomberg@medsci.uu.se (J. Blomberg).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.07.002

The score of the training set plotted against the number of match states showed a characteristic linear rise followed by a plateau where the score stayed constant (see Supplementary materials 1, 2 and 3).

For lentiviral LTRs it was found necessary to remove the part of the LTR which codes for the *nef* protein in order to obtain good alignments. Otherwise, this part interfered with the non-coding part of the LTR during HMM training. For the HML LTRs the long insertion mentioned in Benachenhou et al. (2009) (which sometimes also contained an open reading frame) had a similar effect and was therefore also removed.

Each HMM yielded a Viterbi alignment (Rabiner, 1989) of the training set. The Viterbi alignment with its insert states removed was used to construct a phylogenetic tree. Individual trees from different models varied to some extent. Ten to fifteen trees from models with *M*-values in the plateau and a fixed *z* were therefore combined to yield a 50% majority rule consensus tree. This proved especially useful for the broader gammaretroviral group because some groupings appeared consistently but not in the same individual tree. Negative *z* yielded consensus trees with somewhat higher bootstrap support. In Benachenhou et al. (2009) it was found that the HMMs trained with negative *z*-parameters were the most sensitively detecting ones and this is in line with other approaches to regularisation such as simulated annealing (see Eddy, 1995).

The resulting LTR trees were compared with trees based on ClustalX (Thompson et al., 1997) alignments of the *pol* gene at the nucleic acid level (see Figs. 1, 2 and 3).

For HML endogenous retroviruses the LTR tree did not group hml-7 with hml-8 which seems to be the correct grouping according to both ClustalW (Thompson et al., 1994) alignments of LTRs (data not shown) and the *pol* tree (see Fig. 1). However it is well known that HMM methods perform less efficiently than ClustalW when aligning closely related sequences within subgroups (Edgar and Sjolander, 2003). HMMs are on the other hand superior in aligning the different subgroups. The other difference between the LTR and the *pol* tree is the branching order of HML3 and HML6. If the LTR coevolved with the *pol* gene, this discrepancy could be explained as a long branch attraction between HML5 and HML6 in the *pol* tree, since they are both relatively distant from the other HML groups (see Fig. 1). However, there may also be other explanations (see below). In the LTR tree the bootstrap support for MER9 (HML3) is admittedly weaker but this could be due to the misplacement of MER11D (HML7). In Benachenhou et al. (2009) HML6 was detected in human chromosome 19 even though it was absent from the LTR training set. This was not the case for HML3 when it was absent from the training set, confirming the branching order of the LTR tree, i.e. that the HML6 LTR is closer to the HML1-2/4/8-10 LTRs than is the HML3 LTR. In addition, as described in Lavie et al. (2004), both HML-5 and HML-3 use different primer binding sites in comparison to the other HMLs. HML-5 uses methionine or isoleucine tRNA while HML-3 uses arginine or asparagine tRNA instead of lysine tRNA (which gave the alternative name HERV-K). This opens the possibility that the true LTR and *pol* trees are not identical, i.e. the evolution of the LTR and the *pol* gene may not have been monophyletic for all HML groups.

For lentiviruses the LTR and *pol* trees (Fig. 2) can be compared to the robust phylogenetic tree in Gifford et al. (2008). This tree was based on the Gag and Pol proteins at the amino-acid level. Both trees correctly group the non-primate lentiviruses BIV, Visna, FIV and EIAV outside the primate lentiviruses but their branching orders do not completely agree with Gifford et al. (2008). In the *pol* tree the SAB lentiviral sequence (African green monkey, *sabaeus* subspecies) branches differently. On the other hand the LTR tree has generally lower bootstrap support than the *pol* tree.

The gammaretroviral LTR tree (Fig. 3) has as expected (because this group of LTRs is more variable) lower support and more unresolved nodes than the HML and lentiviral LTR trees. Nevertheless it generally follows the gammaretroviral *pol* tree (Fig. 3). The *pol* tree

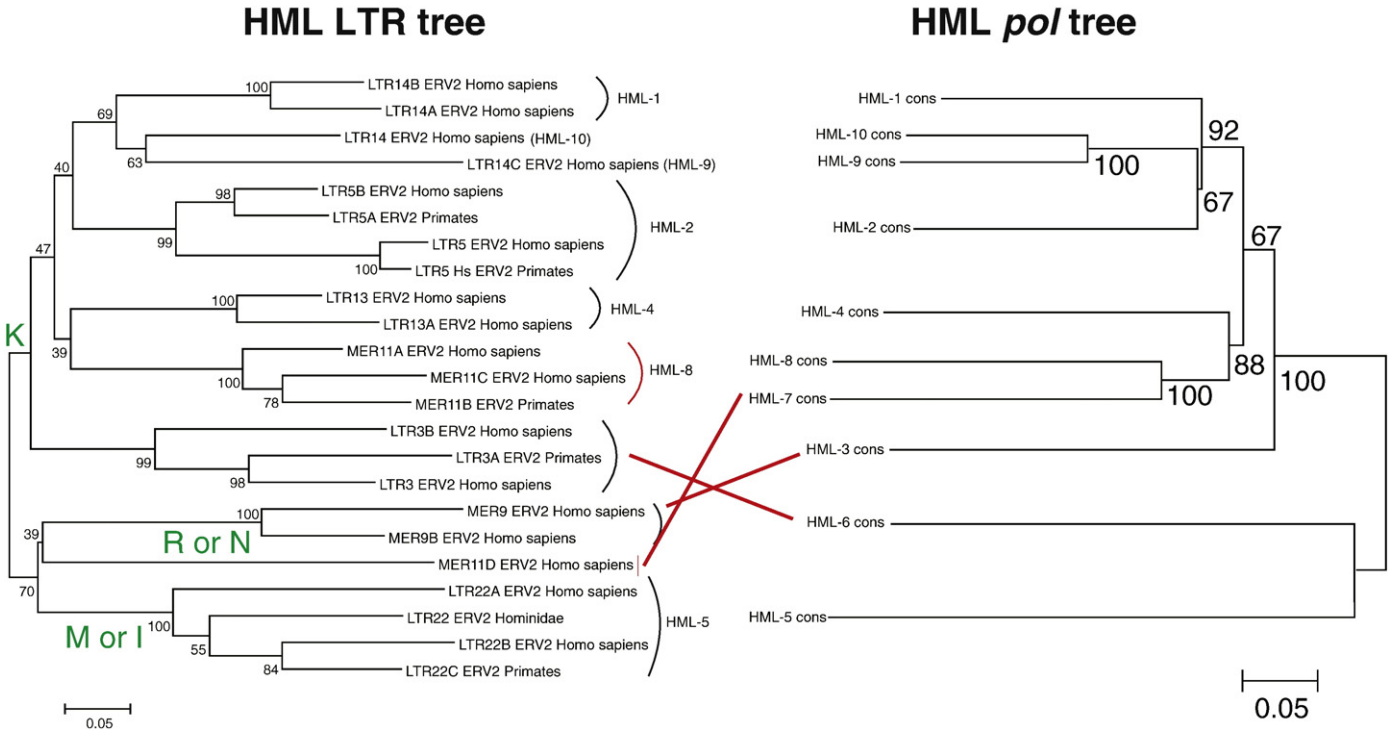

Fig. 1. HML LTR and *pol* trees. Comparison between neighbour-joining trees of HML LTR sequences and HML *pol* nucleic acid sequences. The two trees are aligned when possible; the non-congruent branches are connected with red lines. The amino acid corresponding to the primer binding site (PBS) as described in Lavie et al. (2004) is shown in the LTR tree. K: lysine, M: methionine, R: arginine, N: asparagine. The correspondence between the RepBase nomenclature and the HML names follows (Mager and Medstrand, 2003; Blikstad et al., 2008). Mega 4.1 was used with default parameters except for the pairwise deletion option. LTRs have RepBase names and *pol* sequences ERV names from the literature.