Molecular phylogenetic analysis of bovine viral diarrhoea virus: A Bayesian approach

Hongyan Xia\textsuperscript{a,b}, Lihong Liu\textsuperscript{a,b,\ast}, Niklas Wahlberg\textsuperscript{c,d}, Claudia Baule\textsuperscript{a}, Sándor Belák\textsuperscript{a,b}

\textsuperscript{a} Joint R&D Division in Virology, The National Veterinary Institute & The Swedish University of Agricultural Sciences, SE-751 89 Uppsala, Sweden
\textsuperscript{b} Department of Biomedical Sciences and Veterinary Public Health, Swedish University of Agricultural Sciences, SE-751 89 Uppsala, Sweden
\textsuperscript{c} Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden
\textsuperscript{d} Laboratory of Genetics, Department of Biology, 20014 University of Turku, Finland

Received 27 November 2006; received in revised form 15 May 2007; accepted 18 May 2007

Abstract

Genetic typing of bovine viral diarrhoea virus (BVDV) is important for precise classification of viruses. Traditionally, inferring BVDV phylogeny has been performed by distance-based method, i.e. neighbor-joining for single genes. In this study, a Bayesian approach was exploited to analyze five genetic regions of BVDV genome (5’UTR, Npro, E2a, E2b, and NS3) for 68 taxa retrieved from GenBank. The results showed that all taxa in the consensus tree of E2a have been assigned correctly to corresponding groups, i.e. type-2 BVDV, and BVDV-1a, -1b, -1c, -1e, and -1g, supported by a high posterior probability. In contrast, subgroup 1a formed polytomies in the consensus trees of 5’UTR and NS3. Polytomies also appeared among the subgroup 1b in the consensus trees of Npro and E2b. Analysis of a combined dataset produced an unambiguous, well-supported phylogenetic hypothesis. The topologies found for each genetic region separately and combined were different, but the differences were mainly weakly supported by the data. Combining the data allowed the identification of well-supported clades of strains that correspond to some of the previously defined subgroups. Only a combined approach will allow the confident placement of new strains in the current classification of viruses into genotype and subgenotype.

Keywords: Bayesian inferring; Bovine viral diarrhoea virus; Phylogeny

1. Introduction

The genus \textit{Pestivirus} of the family \textit{Flaviviridae} contains four approved species: bovine viral diarrhoea virus (BVDV) genotypes 1 and 2, border disease virus (BDV), classical swine fever virus (CSFV), and a fifth tentative species represented by a single strain (H138) isolated from a giraffe in Kenya (Becher and Thiel, 2002; Heinz et al., 2000). BVDV is subdivided into two major genotypes: BVDV genotype 1 (BVDV-1) has been reported worldwide, whereas genotype 2 (BVDV-2) has been observed mainly in North America. BVDV has emerged as one of the most important pathogens in cattle population. BVDV infection causes substantial economic losses in both dairy and beef industries worldwide. In addition, cows infected with BVDV may give birth to persistently infected (PI) calves. These PI animals become a large virus reservoir and are at risk of developing mucosal disease (Nettleton and Entrican, 1995; Thiel et al., 1996).

The viral genome is a single-stranded, positive sense RNA molecule of 12.3 kb in size. It comprises two untranslated regions (UTR) at the 5’ and 3’ ends, and one open reading frame (ORF) encoding a polyprotein. The 5’UTR contains secondary structures for translation initiation of viral genome; therefore it is highly conserved during the evolution. The polyprotein is co- and post-translationally processed into 12 polypeptides in the following order: N-terminal autoprotease (N\textsuperscript{pro}), capsid protein (C), envelope proteins (E\textsuperscript{ns}, E1, and E2), p7, and non-structural (NS) proteins (NS2, NS3, NS4A, NS4B, NS5A, and NS5B) (reviewed by Thiel et al., 1996). Of all these proteins, N\textsuperscript{pro} is a small protease responsible for cleavage of the N-terminal polyprotein. It also plays a role in blocking IFN-alpha/beta induction during virus-host interaction. As a
surface glycoprotein, E2 is the major antigen and therefore is highly variable: direct interactions between host immune system and the virus favor viruses with mutations in the antigenic sites. Thus, these mutated viruses may escape host detection. NS proteins are conserved although recombination can occur within NS genes, which may lead to a non-cytopathogenic virus switching to a cytopathogenic type and cause animal death.

The precise classification of BVDV is dependent on genetic typing, which in turn is important for the development of molecular epidemiology (Vilcek et al., 2005). Initially only two subgroups of BVDV-1 were recognized: BVDV-1a (NADL-like) and BVDV-1b (Osloss-like), based on the analysis of a fragment of 5′ UTR (Pellerin et al., 1994; Ridpath et al., 1994). Later studies revealed additional two to three subgroups (Baule et al., 1997, 1999). Furthermore, 11 complete genome sequences of BVDV as well as several representative sequences of BVDV subtypes were also exploited for this purpose (Nagai et al., 2004).

More than studies inferring BVDV phylogeny have been using distance-based methods, with a neighbor-joining algorithm. This method is extremely fast and has been advocated for analysis of large datasets (Tamura et al., 2004). However, recovery of the “true tree” is guaranteed only if the distance matrix is correct, and calculation of genetic distances is complicated by biological processes such as rate heterogeneity (Felsenstein, 2004). It is thus not recommended for use in finding a final tree (McCormack and Clewley, 2002). Maximum likelihood (ML), one of the character-based methods, has also been used for phylogenetic analysis of BVDV. Under an evolutionary model, the most probable tree is found by an optimality criterion based on the character (nucleotide) at each position of a set of sequences. Disadvantages using ML are that it is computationally intensive when dealing with many taxa, and may yield unreliable results with regard to complex parameter-rich model (Holder and Lewis, 2003). In practice, these two methods are often combined when inferring a phylogenetic tree, i.e. a starting tree is generated using neighbor-joining algorithm, and the best tree is found by branch swapping on the starting tree using maximum likelihood criterion. The robustness of the so-called “best tree” can be estimated statistically by bootstrapping (e.g. 1000 replicates) the original dataset and a value of more than 70% is thought to indicate support for a group on the tree.

The Bayesian approach has been recently developed for inferring phylogeny (Yang and Rannala, 1997; Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2001). In contrast to the traditional ML method that only gives the topology of a tree, the Bayesian analysis produces both a tree estimate and measures of uncertainty for the groups on the tree, and thus it provides a measure of support faster than ML bootstrapping. By using a Markov chain Monte Carlo (MCMC) algorithm, Bayesian phylogenetic inference allows implementation of complex parameter-rich evolution models, in which different genes can be analyzed together under different models.

Despite its wide application in molecular phylogenetics and evolution, this relatively new method has not been used in the phylogenetic analysis of BVDV. Considering that different genes may evolve at different rates and may display different relationships among viral sequences, the purpose of this study is to use the Bayesian method to infer BVDV phylogenies based on each of the five genetic regions and a combined dataset of these regions, and particularly to investigate the monophyly of proposed subgroups of BVDV-1 viral sequences. We retrieved sequences of 48 Japanese BVDV isolates described previously by Nagai et al. (2004) from GenBank. The choice of this dataset was based on: (a) five different genetic regions of each virus isolate are available (except the NS3 region of 190CP); (b) the dataset represents both BVDV-1 including several subgroups and BVDV-2; (c) a neighbor-joining tree is presented in the previous paper; (d) basic clinical data are also available. Furthermore, 11 complete genome sequences of BVDV as well as several representative sequences of BVDV subtypes were also retrieved from GenBank.

2. Materials and methods

2.1. Nucleotide sequences retrieved from GenBank

Five genetic regions of BVDV were retrieved from GenBank, including 20 reference viruses and 48 field viruses (Nagai et al., 2004). The genetic regions used in this study are 5′ UTR (position 135–354), Npro (position 391–766), E2a (position 2462–2712), E2b (position 2879–3298), and NS3 (position 5864–6305). All positions are referred to BVDV strain SD-1 genome (GenBank accession no. M96751). Subgroup reference sequences are given in Table 1.

Table 1
Reference sequences for BVDV genotyping

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Name</th>
<th>Accession no.</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>NADL</td>
<td>NC_001461</td>
<td>Pellerin et al. (1994)</td>
</tr>
<tr>
<td></td>
<td>Osloss</td>
<td>M96687</td>
<td>Pellerin et al. (1994)</td>
</tr>
<tr>
<td>Ib</td>
<td>Deer-NZ1</td>
<td>U80903 (Npro), AF144614 (E2)</td>
<td>Becher et al. (1997)</td>
</tr>
<tr>
<td></td>
<td>Bega</td>
<td>AF052303 (NS3)</td>
<td>Becher et al. (1997)</td>
</tr>
<tr>
<td>Ic</td>
<td>Deer-GB1</td>
<td>U80902 (Npro), AF144615 (E2)</td>
<td>Becher et al. (1997, 1999)</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Xia, H. et al., Molecular phylogenetic analysis of bovine viral diarrhoea virus: A Bayesian approach, Virus Res. (2007), doi:10.1016/j.virusres.2007.05.017
2.2. Sequence alignment and genetic divergence

Multiple sequence alignment of each genetic region was done with CLUSTAL W (Thompson et al., 1994). Sequence divergences were estimated using the program MEGA V.3.1 (Kumar et al., 2004). For each genetic region, the following numbers of sites were calculated: total sites (including gaps), conserved sites, and variable sites.

2.3. Model test and Bayesian inference of phylogeny

MrModelTest V.2.2 (Nylander, 2004) was used to estimate best-fit model by hierarchical likelihood ratio tests (hLRTs) and approximate Akaike information criterion (AIC). The best-fit model was selected and used for phylogenetic analysis of each genetic region. Bayesian inference analysis was performed with the software MrBayes 3.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The model settings were a Dirichlet prior for both substitution rates of sites invariable (rates = invgamma). The MCMC search was run with four chains for 5,000,000 generations, sampling the Markov chain every 1000 generations. The first 25% trees were discarded as “burn-in”. Each analysis was performed three times and a representative consensus tree is shown in this paper.

3. Results

3.1. Model selection

The most complex model (GTR + I + G model of substitution) was selected by both hLRTs and AIC in MrModeltest 2.2 for four genetic regions (Npro, E2a, E2b, and NS3). For 5' UTR, a GTR + G model selected by AIC was used for Bayesian inference, although a K80 + G model was suggested by hLRTs. Base frequencies and nucleotide substitution rates of each region are summarized in Table 2. The proportions of invariable sites among-site rate variation were 0.2723 for Npro, 0.2068 for E2a, 0.0477 for E2b, and 0.5202 for NS3. For each genetic region, the shape parameter of the $\gamma$-distribution was selected by both hLRTs and AIC in MrModeltest 2.2. The shape parameter for 5' UTR, a GTR + G model selected by AIC was used for Bayesian inference, although a K80 + G model was suggested by hLRTs. Base frequencies and nucleotide substitution rates of each region are summarized in Table 2. The proportions of invariable sites among-site rate variation were 0.2723 for Npro, 0.2068 for E2a, 0.0477 for E2b, and 0.5202 for NS3. For each genetic region, the shape parameter of the $\gamma$-distribution was selected by both hLRTs and AIC in MrModeltest 2.2.

3.2. General phylogenetic results

In all analyses BVDV-1 sequences were found to form a strongly supported monophyletic group (posterior probability 1.0 regardless of which genetic region was analysed). The overall relationships among all Japanese BVDV-2 isolates were same in the consensus trees of each genetic region and the combined dataset. Therefore the BVDV-2 isolates were removed to simplify the consensus trees, and replaced with words “Outgroup BVDV-2”.

3.3. Phylogenetic analysis of 5' UTR sequences

Within BVDV-1, the 5' UTR sequences identified three major clades with low posterior probability values, corresponding to 1a, 1b and an unnamed subgroup. The latter subgroup consisted of the Japanese isolates IS25CP/01, IS26NCP/01 that clustered with a Chinese isolate ZM-95, and grouped with the unique isolate So CP/75 (Fig. 1). The relationship of the three clades was unresolved. Within clade 1a, two subgroups were also well supported by a high posterior probability (1.00). These were a subgroup composed of IS7NCP/97, IS8NCP/97 and IS14NCP/99, and a subgroup containing 190CP, 190NCP, and KS86-Incp. The other Japanese isolates were either clustered with BVDV strain NADL (1a) or Osloss (1b). Relationships of most of the isolates within the major clades were largely unresolved.

3.4. Phylogenetic analysis of Npro sequences

The Japanese isolates IS25CP/01, IS26NCP/01 grouped with the reference viruses A and L (Fig. 2), named BVDV-1g by Vileck et al. (2001). Three isolates IS7NCP/97, IS8NCP/97 and IS14NCP/99 clustered with isolates Deer-NZ1 and Bega, termed BVDV-1c by Becher et al. (1999). Both subgroups were strongly supported. The 190CP, 190NCP, and KS86-Incp were clustered with Deer-GB1 identified as BVDV-1e (Becher et al., 1999). The isolate So CP/75 formed an independent branch.

3.5. Phylogenetic analysis of E2a sequences

The exciting result using the Bayesian approach was achieved from inferring E2a region. In the consensus tree of E2a sequences (Fig. 3), each subgroup formed a clade, and the relationships of the subgroups were fully resolved and well
Fig. 1. Phylogenetic tree of 5′ UTR region inferred by Bayesian approach. The Japanese isolates were analysed together with nine reference strains (GenBank accession numbers are shown in bracket). The posterior probability value is indicated on the top left part of a branch. Genotype and subgroup are shown in the right part of the corresponding grouping. The bar shows number of changes per site.

### 3.6. Phylogenetic analysis of E2b sequences

Except BVDV-1a and -1c, all other groups were well resolved and had strong support (Fig. 4). However, the topology was different from the one of E2a. For instance, the single lineage (So CP/75) was placed in the same branch as BVDV-1a whereas it was placed in a separate branch.

### 3.7. Phylogenetic analysis of NS3 sequences

The topology of the consensus tree of NS3 (Fig. 5) was similar to the results of E2a and Npro regions. The BVDV-1a
Fig. 2. Phylogenetic tree of a 377-bp fragment of the N gene inferred by Bayesian approach. The Japanese isolates were analysed together with 14 reference strains (GenBank accession numbers are shown in bracket). The posterior probability value is indicated on the top left part of a branch. Genotype and subgroup are shown in the right part of the corresponding grouping. The bar shows number of changes per site.

formed polytomies comprising two major Japanese clades, as well as reference strains. Two references (ILLC and ILLNC) also formed a clade and were included in this subgroup, which is in good agreement with the previous report (Nagai et al., 2004). BVDV-1b formed a monophyletic group, which was supported by a posterior probability value of 1.00. Subgroup 1c was also fully resolved and supported by a high posterior probability value.

3.8. Phylogenetic analysis of the combined dataset

Analysis of the combined dataset yielded a fully resolved, well-supported consensus tree (Fig. 6). Three subgroups BVDV-1a, -1c, and -1e shared almost the same topology as inferred from E2a region. The recombinant strains (ILLC and ILLNC) were clustered within subgroup BVDV-1b. The relationships of each subgroup were consistent with an analysis of an even larger dataset (145 taxa) of pestiviruses using different methods (unpublished).

4. Discussion

Dichotomously branching trees are used to depict two fundamentally different concepts: phenetic approaches attempt to cluster samples which are most similar to each other, regardless of evolutionary history, while phylogenetic approaches attempt to cluster samples which share an evolutionary history. Phenetic methods, such as neighbor-joining, are quick heuristic methods, which can be useful for identifying strains of viruses. However, trees inferred using such methods should not be confused with phylogenetic trees, which depict the evolutionary relationships of lineages.
It is now commonly recognized that to confidently infer the evolutionary relationships of lineages, one needs to use all available data (the so-called total evidence approach; Kluge, 1989). The total evidence approach acknowledges that single datasets may have systematic biases due to homoplasy, and that analysing different datasets together can bring out the hidden phylogenetic signal in them (Gatesy et al., 1999; Wahlberg et al., 2005). Such synergistic effects can lead to increased support for relationships which are otherwise unsupported or weakly supported in separate analyses.

This study compared the phylogenetic signal in five different genetic regions using evolutionary models and Bayesian methods to estimate parameter values (including tree topology). The results show that inferring the evolutionary relationships of the type-1 BVDV lineages using single genes leads to very different estimates, with most relationships being weakly supported. Combining all possible data together produces the most robust estimate of the type-1 BVDV phylogeny (Fig. 6), and this is the preferred phylogenetic hypothesis for the group.

The total evidence analysis allows identifying evolutionarily significant groups of viral lineages and their relationships. It also allows identification of aberrant lineages that have not been classified earlier. A case in point is the lineage including So CP/75. The placement of this lineage differs greatly in the separate analyses of the genes (Figs. 1–5), and reliance on e.g. the E2a gene would suggest that it forms an evolutionarily independent lineage which diverged first from the ancestor of all other BVDV. However, the different placements of the So

Fig. 3. Phylogenetic tree of a 354-bp fragment of the E2 gene (5′-end; E2a) inferred by Bayesian approach. The Japanese isolates were analysed together with 12 reference strains (GenBank accession numbers are shown in bracket). The posterior probability value is indicated on the top left part of a branch. Genotype and subgroup are shown in the right part of the corresponding grouping. The bar shows number of changes per site.
CP/75 lineage receive very little support in most separate analyses. In the total evidence analysis, the So CP/75 lineage is placed well within the type-1 BVDV phylogeny as the sister lineage to isolates IS25CP/01, IS26NCP/01, and the Chinese strain ZM-95. Whether the So CP/75 lineage deserves a subgroup of its own (Nagai et al., 2001, 2004) is debatable. The total evidence analysis suggests that the lineage along with the enigmatic lineages in its sister group should be classified in a group of their own, for which the subgroup 1g may be available. However, the reference sequence for subgroup 1g is not complete, and does not allow the placement of the above-mentioned strains into the group with confidence.

Considering different results obtained from different genetic regions, it is interesting to investigate the effects of the genetic content on inferring phylogeny. Previous studies suggested that 5’ UTR might be not a good target for inferring phylogeny since it is highly conserved (Becher et al., 1997). However, to our knowledge, the characteristics of the sequences of the different genes have not been evaluated. Thus the proportions of conserved and variable sites were calculated in this study for all five genetic regions. NS3 and 5’ UTR are the most conserved regions analyzed so far, with a conserved site proportion of 58% and 53%, respectively, whereas E2b is the most variable region with only 9% conserved sites. These three regions thus do not contain...
suitable phylogenetic signal (NS3 and 5' UTR are not variable enough, and E2b is too variable) to resolve relationships of lineages. In contrast, E2a shows characteristics considered to be appropriate for Bayesian inference, with one-fourth conserved sites and three-fourth variable sites. Consequently, a robust phylogenetic hypothesis was inferred with high posterior probability for many branches. The Npro gene represents an intermediate region for Bayesian inference. For phylogenetic analysis and epidemiological studies, it is important to use as much data as possible to yield robust estimates of evolutionary relationships. This includes analyzing combined datasets that can yield robust phylogenetic hypotheses, with all groups unambiguously resolved and supported by high posterior probabilities. Such an approach is useful for the classification of viruses into genotypes and subgenotypes.

In summary, Bayesian approach was exploited to infer BVDV phylogeny. The best performance of the method was achieved when analyzing a genetic region with appropriate proportions of conserved and variable sites or a combined dataset composed of all five genetic regions. In the future, Bayesian method combined with other traditional tree-building methods can be used to estimate a more reliable viral phylogenetic tree and to study the emerging and/or occurrence new variants of BVDV.
Fig. 6. Phylogenetic tree of the combined dataset from five genetic regions of BVDV genome. The consensus tree was inferred by Bayesian approach. The Japanese isolates were analysed together with 12 reference strains (GenBank accession numbers are shown in bracket). The posterior probability value is indicated on the top left part of a branch. Genotype and subgroup are shown in the right part of the corresponding grouping. The bar shows number of changes per site.

References


H. Xia et al. / Virus Research xxx (2007) xxx–xxx


