Alvira: comparative genomics of viral strains

Francois Enault[1,*,†], Romain Fremez[1], Eric Baranowski[2] and Thomas Faraut[1] ¸

[1]Laboratoire de Genetique Cellulaire INRA UMR444, Chemin de Borde Rouge BP52627 31326 Castanet Tolosan ´´
Cedex and [2]Laboratoire des Interactions Hotes-Agents Pathogenes ENVT UMR1225, 23 chemin des Capelles `
31000 Toulouse, France

ABSTRACT

Motivation: The Alvira tool is a general purpose multiple sequence alignment viewer with a special emphasis on the comparative analysis of viral genomes. This new tool has been devised specifically to address the problem of the simultaneous analysis of a large number of viral strains. The multiple alignment is embedded in a graph that can be explored at different levels of resolution.

Availability: The Alvira software is available at: http://bioinfo.genopole-toulouse.prd.fr/Alvira

Contact: fenault@toulouse.inra.fr

Supplementary information: A tutorial is available at Alvira's homepage.

1 INTRODUCTION

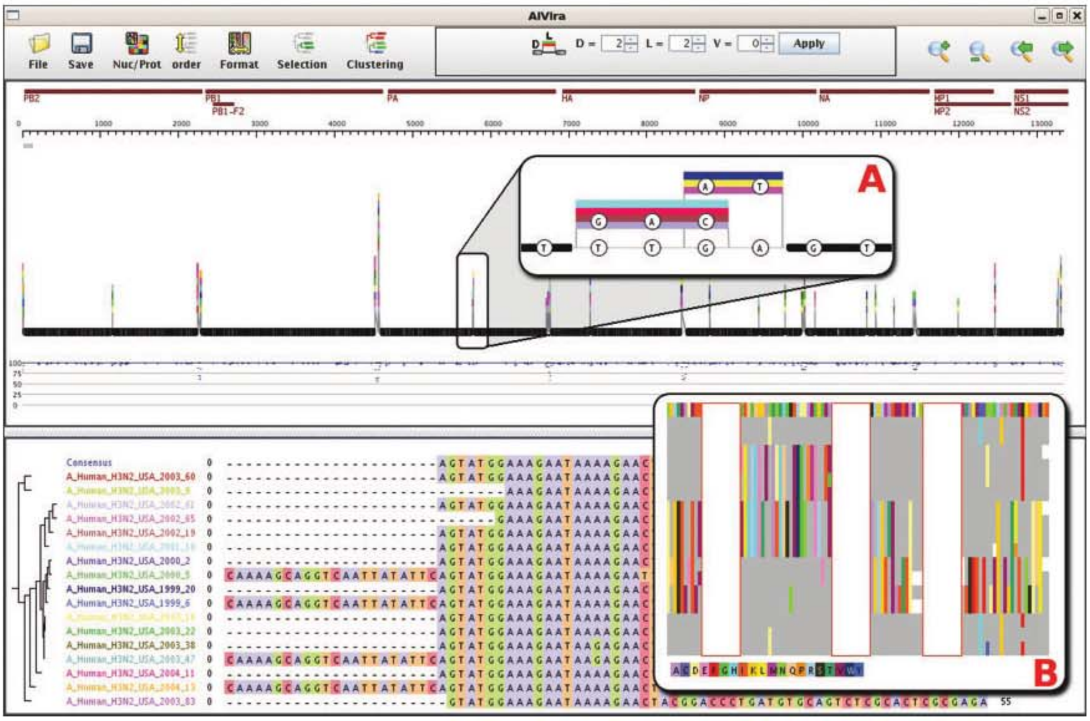
Because of their limited replication fidelity, populations of RNA viruses are extremely heterogeneous and evolve as dynamic mutant swarms composed of related but non-identical genomes. Recent large-scale sequencing efforts offer the opportunity to address, on a sequence analysis basis, the understanding of the evolutionary mechanisms and the population dynamics of these major pathogens (Carrillo et al., 2006; Ghedin et al., 2005; Obenauer et al., 2006). Analysing such large collections of viral genomes requires adapted multiple alignment (MA) tools. While the NCBI viral projects or the Influenza Viral database provide an access to a large amount of viral genome sequences and related information, they are not focused on analytical tools. In contrast, general MA tools and viewers like kalign, clustalW or Jalview (Chenna et al., 2003; Clamp et al., 2004; Lassmann and Sonnhammer, 2006) are not well adapted for the analysis of very long sequences. Moreover, the only MA tool developed for viral genome analysis, Base-by-base (Brodie et al., 2004), is rather centered on the editing of pairwise alignments and is not appropriate for large-scale data sets. In contrast, Alvira has been specifically designed for the comparison of collections of RNA viral genomes. To achieve this goal, two complementary linked views of MAs are simultaneously displayed in the Alvira window (Fig. 1): (i) a global and stylized view where the whole alignment is represented by the consensus sequence and shared divergences, (ii) a local view where alignments are shown in a customizable row-column format. The global view can be zoomed and selected parts are displayed in the local view. This combination provides a very powerful and flexible tool. For example, regions with peculiar variability can be easily located on the global view and further scrutinized at a local scale in both views. Another powerful functionality is the combination of nucleic and protein alignments as an MA of annotated coding sequences can be linked to its corresponding genome's MA. In addition, to help in the analysis of large data sets, MAs can be reduced in both dimensions. First, columns can be removed on the basis of the pattern of dissimilarities they exhibit. Second, subsets of sequences (i.e. rows) can either be selected or created by grouping similar ones into a single consensus. All these features ease the investigation of characteristics that define a viral family as well as the variable positions that contribute to phenotypic traits like virulence. Alvira can also lead to the identification of even more complex biological characteristics of virus population dynamics, like segment reassortment in the influenza population.

2 METHODS

In order to facilitate the dynamic exploration of alignments, shared divergences are located in a preprocessing stage. These divergences are regions comprising more than x sequences and longer than y positions that differ from the consensus sequence. The search for such divergences can quickly become a computationally intensive task. To accelerate this search, MAs are represented as directed acyclic graphs. Each node contains a substring shared by one or more sequences and the edges correspond to the succession of substrings in the different sequences. Within this representation, each sequence is represented by a single path on the graph. Although this graph representation is not new (Lee et al., 2002), using nodes that can contain substrings and not just single letters allows a considerable reduction of the encoded MA. This graph structure is particularly well suited for a quick identification of shared divergences and proves to be fast enough even on large alignments.

3 DISPLAY AND FEATURES

Three main sections are available on the Alvira web site: (i) the user can launch Alvira on any MA (multiple fasta format) or can give raw sequences that will first be aligned by Muscle



D = 2 L = 2 V = 0

```
Consensus                  0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2003_60   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2003_5    0 --------------------------------AAAGAATAAAAGAAC
A_Human_H3N2_USA_2002_61   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2002_65   0 -------------------------------GAAAGAATAAAAGAAC
A_Human_H3N2_USA_2002_19   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2001_16   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2000_2    0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2000_5    0 CAAAAGCAGGTCAATTATATTCAGTATGGAAAGAATAAAAGAAT
A_Human_H3N2_USA_1999_20   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_1999_6    0 CAAAAGCAGGTCAATTATATTCAGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2003_19   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2003_22   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2003_38   0 --------------------------AGTATGGAAAGAATAAGAGAAC
A_Human_H3N2_USA_2003_47   0 CAAAAGCAGGTCAATTATATTCAGTATGGAAAGAATAAGAGAAC
A_Human_H3N2_USA_2004_11   0 --------------------------AGTATGGAAAGAATAAAAGAAC
A_Human_H3N2_USA_2004_13   0 CAAAAGCAGGTCAATTATATTCAGTATGGAAAGAATAAAAGAACT
A_Human_H3N2_USA_2003_83   0 ---------------------------GTATGGAAAGAATAAAAGAACTACGGACCCTGATGTGCAGTCTCGCACTCGCGAGA 55
```

Fig. 1. The Alvira graphical interface is divided into two panels: a global view showing the consensus sequence and shared divergences and a local view in row-column format. (A) zoom of a region of interest in the global view, (B) the customized format of the local view applied to the corresponding protein alignment helps in finding genome fragment reassortment.

(Edgar et al., 2004) on our server, (ii) pre-computed alignments are available for RNA viruses with more than three sequenced strains or isolates available in GenBank and (iii) a selection can be made among Influenza sequences. If available, information on coding sequences found in GenBank files is displayed and, for pre-computed viruses, a protein mode is also available. In this protein mode, alignments of known coding sequences of the viruses are concatenated into a single MA (Fig. 1B). The Alvira window is divided into two parts: the global view in the upper half and the local view in row-column format in the lower half. 'Zoom in/out' and 'left'/'right' buttons allows navigation within the whole alignment. Three parameters can be changed by the user: the depth x and the length y of the divergences and the parameter z that selects only columns exhibiting one or more variability shared by at least z sequences. When applied to protein alignments with z ¼ 3, we obtain the proteotype described in Obenauer et al. (2006), which is not available in any other software to our knowledge. This view is of great interest, e.g. in finding genome segment reassortment (Fig. 1B). At any step, the selection of a subset of sequences can be made using the phylogenetic tree derived from the alignment or using a text list. Reduced alignments can also be obtained by grouping together sequences using a clickable phylogenetic tree or an automatic procedure. Phylogenetic trees are generated using the Java code from Zmasek and Eddy (2001). The number of viral genomes is at the moment limited to 300, a threshold under which Alvira proves to be fast enough on a standard PC. Alvira is written in Java 1.5 and runs as an application throughout the Java Web Start Technology.

ACKNOWLEDGEMENTS

This work is supported by the EADGENE Network Of Excellence and the Genopole Toulouse Midi-Pyrenees.

Conflict of Interest: none declared.

REFERENCES

Brodie,R. et al. (2004) Base-By-Base: single nucleotide-level analysis of whole viral genome alignments. BMC Bioinformatics, 5, 96.

Chenna,R. et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res., 31, 3497-3500.

Clamp,M. et al. (2004) The Jalview Java alignment editor. Bioinformatics, 20, 426-427.

Carrillo,C. et al. (2006) High throughput sequencing and comparative genomics of foot-and-mouth disease virus. Dev. Biol. (Basel), 126, 23-30.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792-1797.

Ghedin,E. et al. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature, 437, 1162-1166.

Lassmann,T. and Sonnhammer,E.L. (2006) Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. Nucleic Acids Res., 34, W596-W599. Lee,C. et al. (2002) Multiple sequence alignment using partial order graphs. Bioinformatics, 18, 452-464.

Obenauer,J.C. et al. (2006) Large-scale sequence analysis of avian influenza isolates. Science, 311, 1576-1580.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics, 17, 383-384.