A Bioinformatic Workflow for Grapevine Viral Diseases Analysis with Reference to Grapevine Leafroll Complex

Balech B[1], Creanza TM[1], Di Tota F[1], Scioscia G[1], Leo P[1]

Motivation

Grapevines (Vitis spp.) are affected by many viral diseases. The most harmful and widespread ones are fanleaf degeneration, Leafroll complex, rugose wood, and fleck. Leafroll disease occurs in all major grape-growing areas worldwide and is one of the most destructive viral diseases of grapevines. Grapevine Leafroll associated Viruses (GLRaVs) are a complex of viruses in the genus Ampelovirus, family Closteroviridae, where GLRaV-3 is the predominant species in the world. At least nine serologically distinct viruses are associated with Leafroll disease. This disease impacts both vine health and grape quality where yield losses may reach as much as 40%. The international committee on taxonomy of viruses describes in its database, the Universal Virus Database (ICTVdb), GLRaV-3 as type-species of Leafroll complex, and provides morphological descriptions and general properties of this virus complex, as well as records of genomic and protein sequences located in GenBank (NCBI). The present study illustrates a bioinformatic workflow for studying one of the most represented genes of GLRaVs complex, namely the Heat shock Protein (HSP70), and a preliminary analysis of comparative genomics of the available complete genome sequences. The aim of these analyses was mainly the identification of tip association traits between a phylogenetic inference HSP70-based and categorical characters, namely isolate geographical origin and host-virus adaptation.

Methods

A bioinformatic workflow shape has been adopted to conduct this study. The first analyses concerned the investigations of ICTVdb files to underline the main properties of Leafroll complex viruses and their instructive genes in their taxonomy. The links to NCBI taxonomy database were exploited to get complete and partial genomic and gene nucleotide sequences. A nucleotide dataset corresponding to HSP70 gene sequences (complete and/or partial) was constructed from all Leafroll viruses (1-11) excluding GLRaV-2, -7 and -8 due to their lack of sequence records for this gene. Nucleotide sequences were multiple aligned using MUSCLE algorithm and than translated into amino acid sequences using Transeq (Emboss package) in order to refine the multiple aligned nucleotides through amino acid sequences alignment. The multiple alignment of both nucleotide and amino acid

[1] MBLab.IBM GBS Business Analytics and Optimization.IBMItalia S.p.A. Via Pietro Leonida Laforgia, 14 - 70125 Bari (ITALY)

sequences showed that the sequences of some of these viruses were located in different gene regions, the fact that led to split this dataset into three new datasets, the first has included GLRaV-4, -5, -6, -9, -10 and -11 (27 taxa), the second GLRaV-3 (23 taxa) and the third GLRaV-1 (42 taxa). In order to evaluate the discrimination capacity of HSP70 gene between and within GLRaVs, a Bayesian phylogenetic inference approach was adopted and a consensus phylogenetic tree was obtained for each created dataset. Phylogeny-trait correlation analyses was performed between genetic clusters, obtained from Bayesian phylogeny, and categorical characters, namely geographical origin and isolate host origin as represented in NCBI features for each sequence. In details, three statistical tests, namely Association Index statistics (AI), Parsimony Score Statistics (PS) and Monophyletic Clade size statistic for a particular trait value (MC) were implemented in BaTS_beta_build2 software which computes p-values by randomization procedures. Finally, we conducted a comparative genomic experiment on the available complete genomic sequences of GLRaV-2, GLRaV-3 and GLRaV-10 using Mauve Genome Alignment software in the aim to obtain other datasets based on homologous viral genes and to observe rearrangement events between Leafroll viruses complex.

Results

Regarding GLRaV-4, -5, -6, -9, -10 and -11 dataset, HSP70 gene showed a clear discrimination capacity between these five viruses by clustering each one in a single genetic group; furthermore, MC correlation analyses, performed on this dataset, with viruses' geographical origin showed that one genetic group of GLRaV-4 variants is highly correlated to its geographical origin while for the other variants this correlation was not significant. Regarding GLRaV-3 dataset we obtained a well supported discrimination capacity between variants within this group where the correlation cultivar-virus was not significant showing by that a poor capacity of this gene and for this dataset in giving host-virus adaptation information. The same discrimination capacity results were obtained for GLRaV-1 group but correlation analyses were not performed due to the small number of observations for each single trait. Preliminary results of comparative genomics showed clear genomic rearrangements of these viruses but it could highlight so many other homologous genomic regions on which the same bioinformatic workflow can be conducted, either singularly or combined, in order to obtain viruses genotype-trait and genotype-phenotype correlations.

Contact e-mail

bachir_balech@it.ibm.com