Viral-KB: an integrated Knowledge-Base to support Viral molecular biodiversity studies

Di Tota F[1], Balech B[1], Lobefaro A[1], Falcone A[1], Vaccina A[1], Scioscia G[1] ,Leo P[1]

Motivation

The present work is positioned in the study of the molecular biodiversity of viruses by coupling genotype and phenotype data analyses. In this context, viruses' characterization, evolution and epidemiological studies are located beneath close relationships, combination and discovery of new events that can give clear answers on virus behaviour. Data mining and bioinformatic tools can help to describe define and highlight important biological concepts in different types of complex experimental data and enable these understandings. A difficulty, in performing such combined analysis in this domain, is concerned with the distribution of relevant data in various information sources, although cross-linked, and in the heterogeneous representation of them. In this community, the Universal Virus Database, authorized by ICTV (International Committee on Taxonomy of Viruses), offers an international taxonomy of all known viruses ever discovered, as well as a detailed morphological description, data about strains and variants belonging to a definite genus and/or family, information on viral isolate properties such as its habitat, host, geographical origin and so on. Only references to genes and genomes NCBI-accession numbers of a type-species and sometimes a group of sub-species are provided. Evidently, ICTV db is not suited to directly apply data mining and/or bioinformatics analysis that aim to support biodiversity studies. With our work we aimed to reduce such shortcomings and build a new operational information source to better support combined viral genotype and phenotype analyses. We called this resource Viral-KB. Viral-KB is core component of an integrated knowledge-base, consistent with ICTV nomenclature, and designed to support general-purpose data mining analysis. Currently, its content is ready in datalog format and provides a series of biological data, sequence and statistical approaches in order to build up data mining analysis highlighting important motifs between viral genotypes and phenotypes and their correlation with isolates properties. Typical usage scenarios we are exploring to exploit Viral-KB are concerned with the ability to support advanced semantic search as well as provide the support to data mining tasks.

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

Methods

We elaborated an extraction and analysis tool that acquired 3550 files recovered from ICTVdb and data were analyzed at species, sub-species, isolate, strain, serotype and sero-group levels. We defined a knowledge representation schema based on 47 fields of the ICTV record type and organized them in five, high-level, biological criteria: Biological-Morphological, Taxonomic-Phylogenetic, Geo-ecological, Patho-physiological, and Functional. Each field's value has been accurately revised from a group of biology curators before importing data. To ensure some data localized on specific genus of viruses we have set the eligibility threshold field to 4%. Currently the Viral-KB is stored on a relational database as well as exported in datalog to support data mining and reasoning tasks.

Results

We developed a strategy that efficiently produces a knowledge representation of the ICTV DB content on biological user-defined criteria. This knowledge base will be used for data mining analysis in order to find different and important biological patterns. 80% of our fields imported were found to be represented more than 4%, in which 14 fields out of 47 are represented up to 40% and only 9 fields have fell under 4%. The present approach of pattern discovery analysis faithfully reproduces results from several viruses and this may be an important method to study association and co-association of phenotype and genotype properties. Preliminary results on a definite case study regarding the identification of a relevant gene discriminating between viral isolates and their geographical origin correlation seem plausible and time-saving and could allow an in vivo experimental design to assure the hypothesis designed in-silico.

Contact e-mail

francesco_ditota@it.ibm.com