SRS implementation of the IARC TP53 Mutation Database

ID - 182

Romano Paolo[1], Marra Domenico[1]

[1]Bioinfomratics and Structural Proteomics, National Cancer Research Institute, Genova

Motivation

The importance of the function of the p53 protein in the regulation of the apoptosis process in the cell is well known and recognized. This protein can be considered as a tumour suppressor, in the sense that it prevents the occurrence of tumours by promoting apoptosis in pre-cancer, altered cells. The TP53 Mutation Database [1] of the International Agency for the Research on Cancer (IARC) currently is the biggest and most detailed database of mutations of this protein. It includes somatic mutations (mutation type, tumour prevalence, prognostic value), germline mutations, polymorphisms, functional properties of mutated proteins and cell line status, together with related literature. The IARC web site provides a tool to analyse TP53 mutation patterns in cancer: the database can be queried on-line but only overall analysis, like distributions or frequencies, are returned. Moreover, some datasets (e.g., mutation and prognosis, germline mutations, polymorphisms) cannot be searched on-line and only the complete data sets can be downloaded, in a tab delimited format.

We present here an SRS [2] site focused on TP53 somatic mutations data, including all data sets provided by IARC.

Methods

All data sets of the IARC TP53 Mutation Database (Release 11) have been downloaded and inserted in a relational database. Perl scripts have been developed for extracting data from the database and creating flat files for SRS. The information has partially been reformatted for easier indexing and searching through SRS.

Reformatting was carried out through an automatic procedure, so that it can be replicated for future releases of the database with a minimal effort. It also involved a few changes and additions in the database fields and their contents (e.g., the Tumor origin has been split and a new Metastasis localization field has been created; unique identifiers were added when missing). All datasets have then been implemented as separate libraries. Data and indexes formats were defined by using Icarus language. Data fields were defined in agreement with the controlled vocabularies that were used for data input at IARC, so that SRS index keys relate to the same terminology. Links between libraries have been created whenever possible.

HTML links were defined to Medline at NCBI. SRS links have been defined so that they can be used for the creation of data views incorporating information from more libraries. Data views have been created for displaying both complete and focused data sets, the latter starting from mutation, sample or patient specific data.

All data sets have been made available on-line and can easily and effectively be queried through well known SRS query forms. Due to the careful definition of data fields, terms included in the controlled vocabularies that were used during data input at IARC can also be used from within the SRS extended query form, thus allowing for a data-driven search.

Results

The IARC TP53 Somatic Mutations Database, including a great number of annotated information on mutations of the p53 human protein and their relationships with cancer and related pathologies, is available on-line in a purpose SRS site and can be easily and effectively queried by using standard, well known SRS query forms. Due to the careful definition of data fields and indexing rules, terms included in the same vocabularies that were used during data input at IARC can also be used from within the SRS extended query form, thus allowing for a data-driven search. SRS links allow to retrieve data from one database by also imposing restrictions on the other. HTML links from the literature references database to PubMed site at the National Center for Biotechnology Information (NCBI) allow a direct access to available information, either abstract or full text. Purpose views can be used to focus attention on data subsets, hiding not pertinent information, thus achieving a more compact output. By using SRS views manager, personal data views can also be created. The configuration files can be downloaded for implementation in further SRS sites.

Currently, more than 23,500 somatic mutations and 1,700 germline mutations, whose data was taken from about 2,150 papers, are present in the database. More than 5,600 unique mutations are described together with their effects and functional properties.

We welcome collaborations with all researchers that would be willing to contribute to our effort by submitting mutation databases to our SRS implementation.

Availability: http://srs.o2i.it/srs71bin/cgi-bin/wgetz?-page+top

Email: paolo.romano@istge.it