A set of simple workflows simplifying access to TP53 mutation database

Domenico Marra and Paolo Romano
National Cancer Research Institute IST, Largo R. Benzi 10, I-16132, Genova, Italy

Introduction

The correlation between genomic mutations and the development of cancer pathologies is a keystone for oncology researchers; in fact, many studies have been oriented to the retrieval of information on genomic mutations of cancer genes and their effect to the protein activity. Mutations might lead, for instance, to a complete loss of function, a partial inactivation, or even to a change in the function of the protein.

Consistent with its role in regulating both cell cycle progression and apoptosis, the human tumor suppressor p53 protein (reference sequence is SwissProt P04637) is found mutated in nearly 50% of human tumors being in fact one of the proteins with the highest number of known and studied mutation variants, with a large majority of single aminoacid substitutions. Consequently, specific databases have been created and made available to the scientific community through the Internet. The TP53 Mutation Database (Olivier et al, 2002) of the International Agency for the Research on Cancer (IARC) currently is the biggest and most detailed database. Release 10 includes 21,587 somatic mutations whose description has been derived from 1,876 papers, 1,839 of which are included in the Medline database. Information on somatic mutations includes data on the mutation, the sample, the patient and his/her life style. Reference vocabularies and standardized annotations are used extensively for the description of the mutation, tumour site, type and origin and for literature references. Examples of the former are ICD-O (International Classification of Diseases – Oncology) and SMOMED nomenclatures. Accessory databases have also been created at IARC. They refer to somatic mutations in sporadic cancers, germline mutations in familial cancers, polymorphisms identified in human populations, functional properties of mutant proteins and gene status in human cell lines.

Researchers can choose to browse and query this vast amount of data starting from two different websites: the IARC TP53 Mutation Database web site (http://www-p53.iarc.fr/index.html), that provides a purpose query interface, and an SRS site where all IARC mutation databases have been implemented (http://srs.o2i.it/srs71/).

In both cases, the researcher must have some experience concerning the query system environment as well as the structure and meaning of each data field; in other words we can say that the above described query methods are useful to users who have a deep knowledge of the scientific issues (p53 protein and its variants) and who are proficient in accessing Internet, know database structures and are able to use involved software tools.

We developed a new query option, based on the use of some specifically created workflows, that can be useful not only to the experienced researcher but, particularly, to all users approaching for the first time this scientific domain.

Methods

Our aim was to provide the researcher with very simple tools which could be used to retrieve significant data subsets of the tp53 databases starting even from the simplest queries.

In practise we adopted the following rules:

- Reduction of number of option queries, usually not more than two.
- Use of simple option queries: only fields like Exon, Intron, Codon Number etc were taken into account as possible query parameters
- Simplification of query outputs

Eight workflows have been developed to make queries to the SRS distribution of the IARC TP53 databases. They can be divided into two groups on the basis of the type of query and of its output.

The first group, comprising six workflows, can be described as a set of queries that interrogates the tp53 somatic mutation database; according to input parameters, the output is one or more sets of somatic mutation entries. The output doesn't allow any further elaboration. It can only be used for a very basic level of analysis, regarding one or more single somatic mutations.
The second group of two workflows is used to make queries, either in pipeline or in parallel, involving two or more databases; the output can be one or more set of entries, which can be integrated together, from any of the tp53 databases. A deeper level of analysis of the output is possible, since the retrieval of data from more specialized databases offer the researcher a specific view of queried mutations. The latter workflows are described here below.

Workflow GetTP53MutationFunctionEntriesAndTP53CellLineEntriesByMutAaAndCodonNumber
Aim: Identify coherent subsets of cell lines status and mutation functions databases whose records have the same mutation, that is determined on the basis of the location of the mutation and the mutated aminoacid, in order to study effects of a specific mutation.
Input data: Mutant Amino Acid and Codon Number at which the mutation is located
Main workflow steps:

- Retrieval of the complete list of ids of tp53 cell line status database which present mutations for the specified Mutant Amino Acid
- Retrieval of the complete list of ids of tp53 cell line status database which present mutations for the specified Codon Number
- Comparison of the two above lists and identification of the common subset
- Retrieval of tp53 cell line status database records using as input the common subset identified
- Retrieval of the complete list of ids of p53 mutant function database which present mutations for the specified Mutant Amino Acid
- Retrieval of the complete list of ids of p53 mutant function database which present mutations for the specified Codon Number
- Comparison of the two above lists and identification of the common subset
- Retrieval of p53 mutant function database records using as input the common subset identified

Output: Two distinct records sets: one set of tp53 cell line status entries and another of p53 mutant function entries that have in common the same Mutated Amino Acid on the same Codon Number

Workflow GetTP53MutationFunctionEntriesByExon
Aim:
Input data: Exon of the tp53 gene
Main workflow steps:

- Retrieval of the complete list of ids of tp53 somatic mutation database which present mutations for the specified Exon
- Retrieval of the list of Codon Numbers of the tp53 somatic mutation database using as input the above list of identifiers
- Retrieval of the list of Mutant Amino Acids of the tp53 somatic mutation database using as input the above list of identifiers
- Retrieval of p53 mutant function database records using as input the two list of Codon Numbers and of Mutant Amino Acids just retrieved

Output: a list of records of the p53 mutant function database.

Some other actions occur during the enactment of the workflows that are not included in these lists; they carry out some string elaborations which are needed for the correct passage of data between the various queries

Conclusions

Workflows have been tested and they proved to be useful for a first level approach to the tp53 field. In the next future, in order to satisfy the feedbacks we received from some users, we plan to implement a more compact output (to be flanked to the standard one) for some workflows, thus giving a better readability for a specific subset of record fields.

In order to offer a deeper level of investigation, new workflows must be implemented which must be able to integrate different specialized databases in order to produce a better and wider view of all available information.

Acknowledgements

This work has been partially carried out with the contribution of the Laboratory of Experimental Oncology B, National Cancer Research Institute (IST), in particular our thanks go to Dr. Fronza Gilberto, Dr. Inga Alberto and Dr Menichini Paola who tested the workflows and provided us with some corrections for the workflows' better functioning