Developement of a data mining system for human cell cycle data analysis

Milanesi L (1), Alfieri R (2), Merelli I (1)

(1) Institute of Biomedical Technologies, CNR, Milano (Italy)
(2) Department of Biotechnology and Bioscience, University of Milano Bicocca, Milano (Italy)

Motivation

The cell cycle is a complex biological process that implies the interaction of a large number of genes. Disease studies on tumour proliferation and de-regulation of human cell cycle have to face with the problem of finding as quickly as possible information related to all the genes that are involved in this cellular process. This work aims to implement a new resource which collects useful information about the human cell cycle to support studies on genetic diseases related to this crucial biological process. Some resources that collect many biological pathways, such as cell cycle, are available for different organisms, but in the state of art there are no specific resources for human cell cycle data integration. The most important resources are Kegg Pathway Database (http://www.genome.ad.jp/kegg/pathway.html) and Reactome (http://www.reactome.org/). Kegg acts in a larger range because it is a collection of pathway maps for metabolic processes, genetic and environmental data such as signal transductions and human diseases. Reactome is a resource for human biological processes which relies on information about single reactions grouped into pathways. Another resource is Cyclonet (http://cyclonet.biouml.org/index.html), a database specifically focused on the regulation of eukaryotic cell cycle. It is less integrated with other biological databases and it is less user-friendly than others.

Methods

"HCCdb" the "Human Cell Cycle Database" is a resource which relies on data taken from Kegg and Reactome. In particular genes involved in the complete cell cycle pathway, in apoptosis pathway and in MAP kinase signalling pathway are taken from Kegg, while genes involved in mitotic and checkpoint pathways are taken from Reactome. To integrate data, we query many resources to collect information related to each gene and protein previously selected. The database infrastructure is designed to make possible an automated data integration: by using a set of Perl libraries it is possible to query a set of selected biological databases retrieving information about genes and proteins. Moreover, we created a database automatic updating system through a pipeline that queries public databases to integrate new data in our resource. The database administrator can access to a specific page where he can insert a gene name and perform the pipeline for data integration. As result it occurs an updating of all tables of the database: in this way the resource can maintained up to date. The main goal of this work is the integration of data related to each gene or protein. For this reason users can query the database contents both inserting the gene/protein name or using the IDs of public databases. The query results page is a complete report and users can browse data using direct links to the different biological database from which data are taken. Users can also query the database using key-words: the results is a list of genes related to the query. HCCDb data are stored in a relational database and a MySQL server is used for this purpose. HCCDb has a "snowflakes" schema, which present the important information about genes and proteins in the inside tables, while collects auxiliary data in the outside tables. The HCCdb database is accessible through a web interface made up of a set of HTML pages dynamically generated from PHP scripts.

Results

HCCdb is a resource that integrates as much as possible information related to genes and proteins involved in human cell cycle. The use of HCCdb has been tested while studying the Cyclin D1 genes, a regulator of the transition from G1 to S phase, which plays an important role in tumourgenesis.

While investigating this gene, HCCdb has demonstrated its importance in retrieving

information about experimental data, promoter and PCR primers that will be used to re-sequencing this cell cycle regulator gene. This database has been realized in the frame of MIUR - LITBIO Project.

Availability: http://cellcycle.itb.cnr.it/

Contact email: luciano.milanesi@itb.cnr.it

Supplementary informations
The LITBIO web site is available at http://www.litbio.org