HCCDB: DATA MINING SYSTEM FOR HUMAN CELL CYCLE GENE

Milanesi L.*, Alfieri R., Merelli I.
CNR Institute of Biomedical Technologies, Milan, Italy
*
Corresponding author: e-mail: luciano.milanesi@itb.cnr.it

Key words: cell cycle, systems biology, database, bioinformatics, data mining

SUMMARY

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 the problem of finding as quickly as possible the information related to all the genes that are involved in this cellular process. This work aims to implement a new database containing information about the human cell cycle to support studies on genetic diseases related to this crucial biological process.

Results: HCCdb is a data mining system that integrates information related to genes and proteins involved in human cell cycle.

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

INTRODUCTION

Several resources that collect many biological pathways, such as the cell cycle, are available for different organisms, but in the state of the 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.

METHODS

For the models of the gene networks introduced in Likhoshvai et al. (2001), we consider here the corresponding 3-dimensional dynamical systems.

The HCCdb database “Human Cell Cycle Database” is a resource, which relies on the data taken from Kegg and Reactome. To integrate the data, we query many resources to collect the information related to each gene and protein previously selected. The database infrastructure is designed to make possible an automated data integration and to query a set of selected biological databases retrieving information about genes and proteins. Moreover, we create a database automatic updating through a pipeline that queries public databases to integrate new data in our resource. The database administrator can access a specific page where he can insert a gene name and perform the pipeline for data integration. As a result, an updating of all tables of the database occurs: in this way the
resource can be maintained up to date. For example, the genes involved in the complete cell cycle pathway, in apoptosis pathway and in MAP kinase signaling pathway are taken from Kegg, while the genes involved in mitotic and checkpoint pathways are taken from Reactome. The main goal of this work is the integration of the data related to each gene or protein. For this reason, users can query the database contents both inserting the gene/protein name and using the IDs of public databases. The query results page is a complete report, and users can browse the data using direct links to the different biological related database. Users can also query the database using keywords: the result is a list of genes related to the query. HCCdb data are stored in a relational database based on MySQL DBMS. HCCdb data mining system is made up using a “snowflakes” schema, which present the important information about genes and proteins and auxiliary data. In particular, for each gene we give the promoter information, validated experimental data about gene’s expression and quantitative PCR primers. The HCCdb database can be used to find useful information about each protein, such as complexes, protein–protein interactions, protein structures and surfaces, as show in Fig. 1 for BCL2 gene structure. The HCCdb database is accessible through a web interface made up of a set of HTML pages dynamically generated from PHP scripts.

**Figure 1.** An example of an HCCdb entry for BCL2 protein structures and surfaces visualizations.

**RESULTS AND DISCUSSION**

We have developed a data integration system by integrating data from many biological resources: NCBI, Ensemble, Kegg, Reactome, dbSNP, MGC, DBTSS, Unigene, OPPD, TRANSFAC, UniProt, InterPro, PDB, TRANSPATH, BIND, MINT, and IntAct. HCCdb database use the data warehouse approach that allows for bringing all the related data in a single database and experimental data. This approach is used in order

- To develop a unified data model that can accommodate all the information stored in various source databases;
- To develop programs that will take data from the source databases, transform them to match the unified data model and load them into the warehouse;
- To increase the efficiency of retrieving the specific information related to a specific query;
- To perform data mining of different kinds of information through a single query; and
- To increase the information accuracy and better control over the information standard.
The HCCDB is a MySQL relational database with a “snowflake” schema aimed to collect principal information about genes and different kind of auxiliary data related to genes and proteins. An example of the “Gene report” from HCCdb by using BCL2 gene as input query is shown in Fig. 2.

Figure 2. An example of a “Gene report” from HCCdb by using BCL2 gene as an input query. The report collects the information integrating data from different resources.

ACKNOWLEDGEMENTS

The work was supported by the grant No. 03-51-5218 INTAS, European SSH BioinfoGRID and MIUR FIRB LITBIO projects. The authors are grateful to J. Hatton and C. Bishop, for helpful discussions and Web interface.