

# Development 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](mailto:luciano.milanesi@itb.cnr.it)

### Supplementary informations

The LITBIO web site is available at <http://www.litbio.org>


Cell Cycle Database

- Home page
- Gene name search
- Protein name search
- Text search
- Links
- Acknowledgements

### Gene report: CCND1

**Alternative names:**

- BCL1
- PRAD1

**Description:** cyclin D1

**Pathway:**

- KEGG
- REACTOME

**Protein name:** CCND1

**Sequence Information:**

<a href="#">gene length</a>	<a href="#">gene sequence</a>
888 bp	<a href="#">view sequence</a>

**SNP List:** [view](#)

**Full-length cDNA Informations:**

locus	Clone Library	MGC id	Image id
BC000076	10691	2316	3508088
BC001501	10691	2233	3507598
BC014078	10691	20169	4124540
BC023620	7641	23386	4650919
BC025302	22072	39267	5457963

**Isoform and transcript:**

isoform	transcript
NM_053056	ENST00000227507

**Links to other genomic databases:**

REFSEQ	ENTREZ CODE	AC	ENSEMBL	genecard	genomebrowser
NM_053056	595	X59798	ENSG00000110092	GC11P069165	Z23022

**Promoter region:**

promoter sequence	localization	start	strand	Transcription start site position
<a href="#">view sequence</a>	chr11	69229012	reverse	69228803

**Transcription factors:** [view](#)

**Experimental data:**

- **Stanford University Data:** [view data](#)
- **Unigene Expression Profile:** [view profile](#)
- **GEO profiles:** [view data](#)

**Quantitative PCR Primer Info:** [view](#)

### Protein report: CCND1

**G1/S-specific cyclin D1**

**Alternative names:**

- PRAD1 oncogene
- BCL-1 oncogene

**CCND1 participates in following processes:**

Cell Cycle, Mitotic  
G1 Phase; Cyclin D associated events in G1; Formation of Cyclin D:Cdk4/6 complexes;  
1. Formation of Cyclin D1:Cdk4 complexes [Homo sapiens]  
2. Formation of Cyclin D1:Cdk6 complexes [Homo sapiens]

**Belongs to Cyclin family; Cyclin D subfamily**

**Gene name:** CCND1

**Sequence Information:**

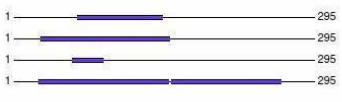
<a href="#">protein length</a>	<a href="#">protein sequence</a>
295 aa	<a href="#">view sequence</a>

**Protein links to other biological databases:**

UNIPROT	ENTREZ PROTEIN	ENTREZ OMM	GO
P24385	NP_444284	168461	GO:0005515

**InterPro Domains:**

domain	domain database	position
IPR006670	SM00385	62-146
IPR006671	PF00134	26-153
IPR006671	PS00292	57-88
IPR011028	SSF47954	24-152 154-262



**Protein Interactions:**

- Bind
- Mint
- IntAct

**Protein complexes:**

- [Transpath molecule](#)