Overview

During the last years the advances of high-throughput technologies in biology have produced an unprecedented growth of repositories and databases storing relevant biological data. Today there is more biological information than ever but unfortunately the current status of many of these repositories is far from being optimal many times. Some of the most common problems are: a) information is spread out in many small repositories and databases, b) lack of standards between different repositories, c) unsupported databases, d) specific and unconnected information, etc.

All these problems make very difficult: a) to integrate or join many different sources into only one database to work or analyze experiments; b) to access and query this information in programmatically way.

To cope with all these problems we have designed and developed a NoSQL database that integrates the most relevant biological information about genomic features and proteins, gene expression regulation, functional annotation, genomic variation and systems biology information. We use the most relevant repositories such as Ensembl, Uniprot, ClinVar, COSMIC or IntAct among many others (you can browse them Data sources and species). The information integrated covers:

- Core features: genes, transcripts, exons, proteins, genome sequence, etc.
- Regulatory: Ensembl regulatory, TFBS, miRNA targets, CTCF, Open chromatin, etc.
- Functional annotation: OBO ontologies (Gene Ontology, Human Disease Ontology), etc.
- Genomic variation: Ensembl Variation, ClinVar, COSMIC, etc.
- Systems biology: IntAct , Reactome, gene co-expression, etc.

To make this entire database accessible to researchers, an exhaustive RESTful Web service API has been implemented. This API contains many methods that will facilitate researchers to query and obtain different biological information from a single database saving a lot of time. Another benefit is that researchers can make easily queries about different biologTical topics and link all this information together as all information is integrated.

Currently *Homo sapiens*, *Mus musculus* and other 20 species are available and many others will be included soon. Results are offered in JSON format, making all this information accessible to both software or web applications.

Availability

CellBase is open-source and freely available at **https://github.com/opencb/cellbase**

Publications

CellBase was published at Nucleic Acids Research (2012):

http://nar.oxfordjournals.org/content/40/W1/W609.short