BioCloud Search EnGene: Surfing Biological Data on the Cloud

Nicoletta Dessì, Emanuele Pascariello, Gabriele Milia, and Barbara Pes

Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,
Via Ospedale 72, 09124 Cagliari, Italy
dessi@unica.it, emanuele.pascariello@gmail.com,
milia.ga@unica.it, pes@unica.it

Abstract. The massive production and spread of biomedical data around the web introduces new challenges related to identify computational approaches for providing quality search and browsing of web resources. This papers presents BioCloud Search EnGene (BSE), a cloud application that facilitates searching and integration of the many layers of biological information offered by public large-scale genomic repositories. Grounding on the concept of dataspace, BSE is built on top of a cloud platform that severely curtails issues associated with scalability and performance. Like popular online gene portals, BSE adopts a gene-centric approach: researchers can find their information of interest by means of a simple "Google-like" query interface that accepts standard gene identification as keywords. We present BSE architecture and functionality and discuss how our strategies contribute to successfully tackle big data problems in querying gene-based web resources. BSE is publically available at: http://biocloud-unica.appspot.com/.

Keywords: Biomedical data exploration, Cloud computing, Data searching, Data integration, Dataspaces, Pay-as-you-go data querying.

1 Introduction

The massive production and spread of biomedical data around the web introduces new challenges related to identify computational approaches for their management and exploitation. These challenges mainly result from three issues:

- *Biomedical data are typical of the category of "big data"* [1]. The term "big data" refers to "the ever increasing amount of information that organizations are storing, processing and analyzing, owning the growing number of information sources in use" [2].

- *Biomedical data relay with a wide range of types and sources.* As biomedical research became interdisciplinary, information searching often requires the integration of information with multiple levels of granularities and relates data that pertain to different disciplines. Hence, the user search is not limited to a single source, but it is carried out through separate web resources in which information is represented in a different way.

- *Biomedical data must be accessed quickly* to determine which information to show to a user on a webpage. To do global analysis, biological researchers often need to access data from multiple archival databases.

It has been observed [3] that gene sequencing technologies have become more and more affordable but the challenge of integrating disparate resources of biologic information remains difficult and more implicit or automatic ways of joining information are needed to improve the usability of gene annotation resources where searching is often unwieldy.

The development of efficient, optimized, and highly scalable search tools is a particularly challenging task as data are reaching tsunami proportions [4] and related clinical applications are seen as a "slowly rising tide" [5].

In this work we focus on genomics, a key area of biology which places greater stress on trying to solve the problem of collecting and processing large volumes of biological information, due to the fact that biological data accumulate at an ever-faster pace.

Specifically, we envision searching genetic information in databases and web resources to be like searching information in the web: we search for the information we exactly need and capture a lot of information in a short time from different websites. To face the challenge of supporting scientists in searching genetic information, we stop thinking in terms of capabilities of individual web resources and instead think of the computational functionalities needed.

In order to avoid browsing web resources and data locked to specific infrastructures, we propose advanced search functionalities on many resources via high quality, interoperable services offered in a "neutral" territory. As it happens for web engines which are designed to search for information on the World Wide Web, these services act as specialists which mine data available in many databases or open directories and return real-time information. They are a mean of organizing and integrating information from different web sources and making them manageable and satisfactory for the user.

In this article, we present BioCloud Search EnGene (BSE), a comprehensive searching environment which facilitates the versatile integration of existing genetic and genomic information from multiple heterogeneous resources. It proposes a new operational framework in which genetic information and computing technologies are reshaping each other. Like popular online gene portals, BSE adopts a gene-centric approach: researchers can find their information of interest by means of a simple "Google-like" query interface that accepts standard gene identification as keywords. Moreover, by using advanced searching and tools, users are allowed to extend their possibilities of standard data searching on popular genetic databases. BSE heavily relies on the following key design features.

First, BSE is grounded on the concept of dataspace [6] [7], a new paradigm for data integration characterized by a very loosely structured data model and intended for the management of heterogeneous data coming from a diverse set of sources regardless their format and location.

Second, to handle important coordination tasks, BSE is built on top of a cloud platform which is the physical infrastructure for hosting the dataspace. This severely curtails issues associated with scalability and performance, especially during information retrieval as searching expands across multiple server nodes. Finally, BSE is built into an integrated cloud environment that allows a close integration with web servers and standard protocols and facilitates rapid development and updates.

The paper is organized as follows. Section 2 provides background concepts and motivates the adoption of dataspace and cloud paradigms. Sections 3 details the architectural aspects of BSE. The system functionalities are described in section 4. Finally, section 5 presents conclusions.

2 Background and Motivations

Dozens of gene annotation resources and databases exist which serve prominent roles in the genetics and genomics communities, each presenting a particular aspect of available gene notations. For example, the 20th annual Database Issue of Nucleic Acids Research (NAR) includes 176 articles half of which describe new online molecular biology databases and the other half provide updates on the databases previously featured in NAR and other journals.

As notable example, Entrez [8] [9] is the most popular system for searching and retrieving information from databases that are maintained by the NCBI (National Center for Biotechnology Information) [10]. Entrez is constantly being developed and improved. It indexes records in NCBI databases by means of nodes that correspond to specific databases including GenBank [11] [12], Protein database [13] and also scientific abstracts from the PubMed database [14] [15]. Access to these resources is provided by the graphical user interface of the NCBI Entrez system or by using NCBI Web services.

In exploring a database, researchers are not interested in exploiting the resource full content, but they just distil a huge amount of data to obtain succinct, key information about a concept. As biology encompasses many domains of knowledge, the success of their search depends on their ability in browsing large-scale information that is stored in several databases and web sites, each having its own organization, terminology and data formats. Unearthing specialized information can also be complex, time consuming and daunting as the researcher is also involved in learning and remembering the navigation paths of each specific web site. Finally, different web portals implement the same basic functionality and are often concerned with overlapping information.

For effective searching biomedical databases in the face of the growing number of bio-resources available worldwide, we have to answer three fundamental questions.

First, how to integrate structured, semi-structured and unstructured available data with diverse and sparse schemas?

Second, how to retrieve meaningful information in an easy and efficient way?

Finally, how implementing a searching infrastructure which has to scale, hence change, to meet new requirements stemming from the growth of its searching domain?

Computational solutions ranging from database to data warehouse poorly adapt to facing the above questions as:

a) Many resources are large in size, dynamic, and physically distributed. Consequently, there is the need for mechanisms that can efficiently extract the relevant information from disparate sources on demand.

b) The resources of interest are autonomously owned and operated. Consequently, searching strategies must be devised for obtaining the necessary information within the operational constraints imposed by the data source.

c) Being heterogeneous in structure and content, information resources represent data according to their own schema which, implicitly or explicitly, defines its own concepts and relationships among concepts.

d) Searching happens in different contexts and from different user perspectives. Hence, it is necessary to implement mechanisms for extracting context-dependent information.

The research community has recently proposed the concept of dataspace [6] [7] as a new scenario for structuring information relevant to a particular organization, regardless of its format and location. The elements of a dataspace are a set of participants (i.e., individual data sources) and a set of relationships between them [7]. In this sense, a dataspace is an abstraction of a database that does not require data to be structured and has a minimal "off-the-shelf" set of search functions based on keywords. The key idea is to enhance the quality of data integration and the semantic meaning of information without an a priori schema for the data sources [16] [17] [18].

In sharp contrast to the traditional approaches, a dataspace is based on a ‘‘data co-existence’’ approach as it integrates data according to a very loosely structured data model which is intended for the management of heterogeneous data coming from a diverse set of sources. A fundamental part of a dataspace is the catalogue that contains information about participants and their relationships with associate mechanisms for its gradually extension. Advanced DBMS-like functions, queries and mappings are provided over time by different components, each defining relationships among data when required. Integrated views over a set of data sources are provided following the so-called pay-as-you-go principle that is currently emerging on the web [19] [20].

In this work, we choose the dataspace paradigm as a data integration architecture for reconciling data from heterogeneous sources and providing users with a unified view of these data.

Our key idea is to conceive BSE as a cloud based application which essentially rents its capacity from a cloud computing platform.

Recent research [21] has proposed cloud computing as an innovative computational environment for searching large-scale data in a more efficient way. Specifically, cloud computing refers to a flexible and scalable internet infrastructure where processing and storage capability are dynamically provided. A cloud infrastructure abstracts the underlying hardware (i.e. servers, networking, storage etc.) and enables on-demand network access to a shared pool of computing resources that can be readily provisioned and released.

The next section will present how the BSE architecture benefits from both the dataspace and the cloud paradigms.

3 Architectural Aspects

Grounded on the dataspace paradigm, BSE undertakes the responsibility of coordinating and organizing the search across different web resources that are assumed to be the dataspace participants.

Data integration expects no data transfer to any central repository, except for the data stored in BSE catalogue which is initially built and gradually updated. In some way, this catalogue has the same role of the table of facts in a data warehouse where the dimension tables are distributed across many web resources. However, it differs from a data warehouse schema because:

1) It contains information about various participants instead of relational tables.

2) Besides storing and indexing participants, the catalogue contains mechanisms for creating new relationships by modifying the existing ones.

3) It avoids the definition of an a priori matching schema.

From a logical point of view, the catalogue is a multi-level index that specifies how genetic information from various web resources is captured and linked together. Physically, it is implemented by an object-oriented database, specifically a key-value NoSQL database [22], which stores gene annotations, acquires and combines information from external resources that participate in the dataspace.

The current version of BSE implements a dataspace with the contribution of 34 participants. According to their role in supplying data, these participants are categorized as:

- *Local participants*, i.e. resources from which some useful content is captured and permanently stored into the catalogue.
- *Service-based participants*, i.e. resources whose content is captured at running time by specific BSE services in a pay-as-you-go fashion according to the user request.
- *External participants*, i.e. resources whose web links are dynamically built and activated when it is required.

The catalogue organizes objects in classes, each corresponding to one local participant. Table 1 shows the list of local participants and the corresponding catalogue content.

BSE also relies on the following external services:

- NCBI Entrez Programming Utilities (E-utilities) [23]
- UniChem RESTful Web Service API [24]
- Database identifier mapping [25]
- STRING API [26]
- WikiPathways Webservice/API [27]
- REST-style version of KEGG API [28]
- mygene.info REST web services [29]
- RESTful web service Europe PMC [30]

Conversely, a local participant is viewed as a repository of objects associated with the catalogue. Relationships between participants are expressed by means of key-values [31] that store gene identifiers as defined by international scientific standards.

Table 1. Local participants and corresponding catalogue content.

Dataset	**Catalogue content**
Entrez Gene homo sapiens gene info [32]	Main annotations about human genes
Entrez Gene Relations human gene relations [32]	Gene to gene relationships
M.A.T.A.D.O.R Manually Annotated Targets and Drugs Online Resource [33]	Gene drug relationships
Entrez gene ID to pathways [34]	Human genes pathways, according to Reactome
Entrez gene ID to Mendelian Phenotype [35]	Human Mendelian Phenotypes and their gene associations
Entrez gene ID to RefSeq [36]	Cumulative set of transcripts and proteins
H.A.G.R.- Human Ageing Genome Resource [37]	Genes possibly related to human ageing
Wellcome Trust Sanger Institute - Cancer genomics annotations [38]	Cancer Drug sensitivity Annotated genes

3.1 Query Contextualization

The schema free and non-rigid structure of the catalogue allows us to implement with relative ease new ways of querying and extracting information on the basis of what we can here define as a query context or a context from now on.

Specifically, a context is a logical structure that supports queries about common points of interest the users share in browsing dataspace participants. As an example, if the user is interested in searching information about genes associated with a specific disorder, he refers to the context "Mendelian genetic disorders". Contexts are the only way to query data. Each context presents a "gene centric view" where the users can easily identify the relevant resources and navigate the content of the resources to which the context relates. Contexts hide the complexity of data searching, where BSE services capture and present the information of interest.

From a technical point of view, contexts identify specific perspectives on dataspace participants that are kept in the dataspace catalogue. These perspectives resemble to views in relational databases. However, being the catalogue implemented by a NoSQL database, they do not result from joining structured relational tables, but from relationships expressed by key-values. As well, contexts take very little space to be stored as the catalogue contains only the definition of contexts without a copy of all the data that the context relates to.

The present version of BSE implements the following contexts:

1. Known gene name or gene identifications

Here, we assume that the user is able to identify genes by their standard identifier and wants to know further details.

2. Query by Human Mendelian Genetic disorders

The user is allowed to extract a list of genes using the name of a certain phenotype associated with a genetic disorder with Mendelian transmission character.

3. Query by pathway

This context allows the user to extract a list of human genes annotated in a given biological pathway. A pathway is a set of chemical reactions related to one or more processes within a cell. It results in expression products whose knowledge is very important in the study of biological phenomena.

4. Bulk queries.

This context allows to extract a list of human genes meeting the following searching criteria: gene biotype, chromosome belonging, ageing related annotation, chemotherapeutic sensitivity related to annotated genes according to their mutational status.

5.Query by Drug information

It moves the query focus from a purely genetic perspective to a context dealing with the relationships between pharmacologically active molecules and the human genome expression products.

3.2 Technical details

BSE is built on top of GAE (Google App Engine) [39], a platform as a service (PaaS) cloud computing environment which hosts web applications. Differently from other PaaS offerings, GAE benefits from the same infrastructure that supports basic Google applications and services such as Google search engine, You tube, Google Earth etc.

We stored the dataspace catalogue into the GAE datastore, a distributed data storage service that performs distribution, replication and load balancing automatically and supports operations to access objects (i.e. create, read, update, delete) by means of an SQL-like language called GQL.

We used Phyton as programming language and implemented BSE functionality using JavaScript/AJAX/jQuery and Django, a high-level Python web framework that runs within GAE.

For implementing the pay-as-you-go approach in searching data we relied on Biopython [40], a rich set of Python libraries which provides the ability to deal with "things" of interest to biologists while working on the cloud. Specifically, the Entrez Programming Utilities provided by NCBI were accessed by means of the Bio.Entrez library available in Biopython. This library was made available on the cloud just making some easy changes in the source code.

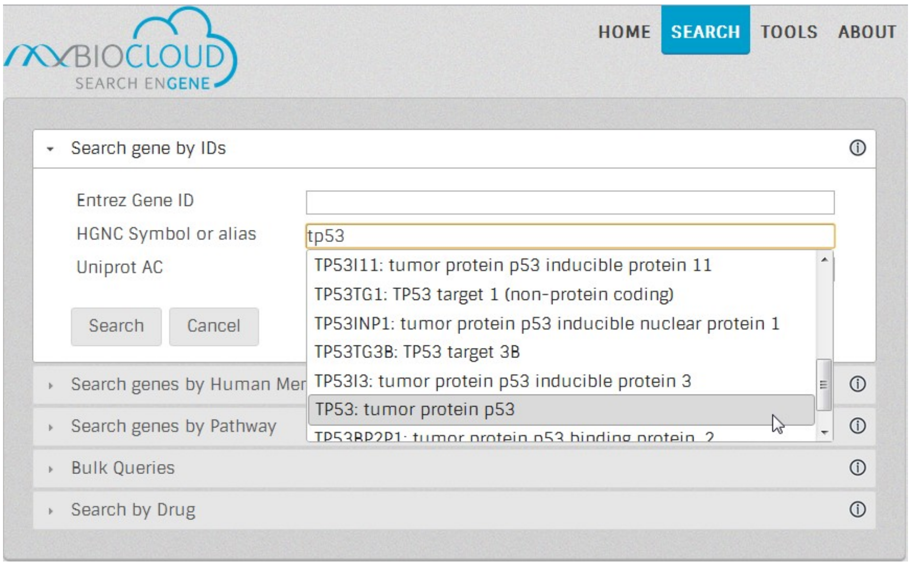
4 BSE Functionalities

BSE is publically available at: http://biocloud-unica.appspot.com/. In what follows we will present and discuss the BSE functionality.

BSE utilizes a simple graphical user interface (GUI) that takes account for the concept of usability in presenting information. Specifically, BSE GUI is implemented by an accordion i.e. a vertically stacked list of items each of ones can be “expanded” or “stretched” to reveal the content associated with that item. There can be zero or more items expanded at a time, depending on the show/hide operation users carry on.

When a web page is loaded, the accordion expands the corresponding item into a window which contains the web page and allows users to navigate through this page. Practically, an accordion is expandable whenever needed and allows to really save some space while showing a lot of information.

By default, the first item is expanded whenever an accordion appears. Each item can be open/closed by clicking on it. A new item is added dynamically on top of the accordion to present query results.



BIOCLOUD
SEARCH ENGENE
HOME
SEARCH
TOOLS
ABOUT
Search gene by IDs
Entrez Gene ID
HGNC Symbol or alias
Uniprot AC
Search
Cancel
TP53I11: tumor protein p53 inducible protein 11
TP53TG1: TP53 target 1 (non-protein coding)
TP53INP1: tumor protein p53 inducible nuclear protein 1
TP53TG3B: TP53 target 3B
TP53I3: tumor protein p53 inducible protein 3
TP53: tumor protein p53
TP53BP2P1: tumor protein p53 binding protein 2
Search genes by Human Mer
Search genes by Pathway
Bulk Queries
Search by Drug

HGNC Symbol or alias: tp53

Fig. 1. The BSE main page with the *basic accordion*. This screenshot shows a typical search by typing the HGNC official gene symbol related to tumor protein 53 gene. Within this search context the user can also search for genes by Entrez ID, or UNIPROT accession. The Alias gene identification is supported too.

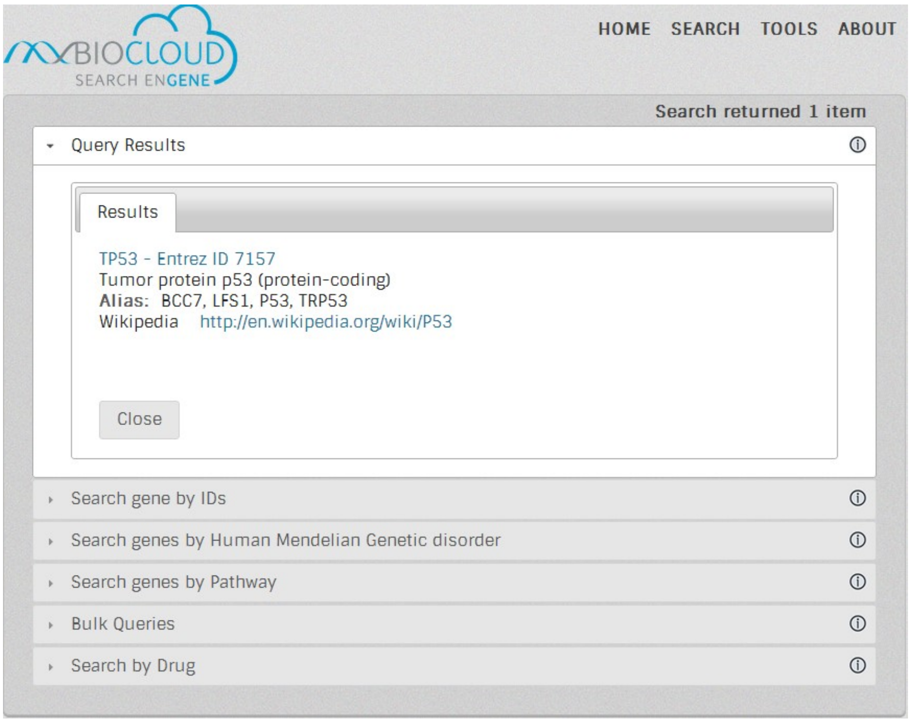
BSE implements the following accordions:

1) The *basic accordion* is the BSE main page where each item represents a query context.

2) The *gene accordion* is visualized whenever the user clicks on a gene identifier and allows to detail information about that gene.

3) The *drug accordion* is visualized whenever the user clicks on the name of a drug and allows to investigate about the drug properties and effects.

In what follows we present the structure of each accordion.



Search returned 1 item

TP53 - Entrez ID 7157
Tumor protein p53 (protein-coding)
Alias: BCC7, LFS1, P53, TRP53
Wikipedia http://en.wikipedia.org/wiki/P53

Fig. 2. Results to the query in Fig. 1.

Fig. 1 shows the BSE main page with its *basic accordion*. The first item, corresponding to the context '*Search gene by IDs'*, is expanded and shows a text field where users enter a single keyword which is the identifier of the gene they want to find. In Fig. 1 the user is typing the keyword "tp53" as standard gene identifier while BSE dynamically provides predictive suggestions by expanding the keyword "tp53" in a sliding list of its synonyms and variants. The user selects the appropriate keyword from the list, submits his query and obtains information showed in Fig. 2.

When the user clicks on the gene identifier (i.e. TP53 – Entrez ID 7157 in Fig. 2), he is redirected to the gene accordion (see Fig. 3) which details a new context to explore information about TP53. By expanding the items of this new accordion, the user can extract a series of highly detailed data and then investigate every aspect of its interest in specialized databases with a redirection that is consistent with the initial query.

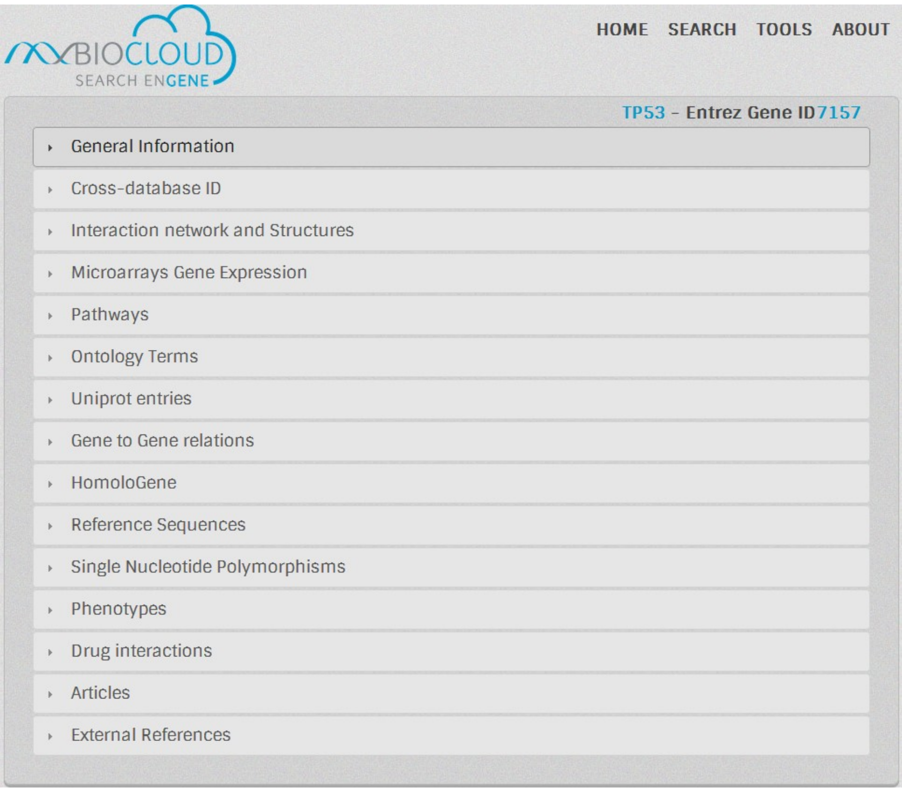
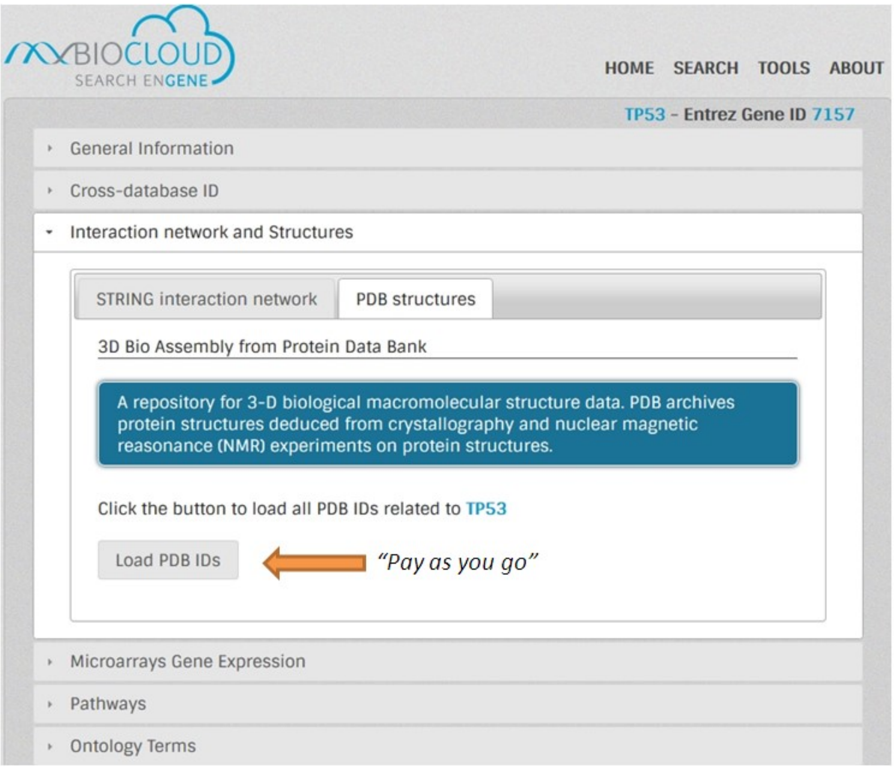


Fig. 3. The *gene accordion* and its items.

For example, Fig. 4 shows the effects of expanding the item “Interaction network and Structures”. Here, in order to limit the number of query results, the searching process follows a pay-as-you-go approach: the user is invited to load additional information if necessary. In this case, he interactively triggers the capture of data by clicking the “Load PDB IDs” button. Captured information is permanently stored in a buffer during 24 hours and then released.

Fig. 5 shows the results of this capture including PDB IDs, the images about related 3D structures from Protein Data bank, and FASTA Sequence of the corresponding structure. As showed on the left of Fig. 5, images can be expanded. Clicking on the sky blue arrows which have a wavy tail (see Fig. 5 on the right), the user is redirected to an external web site providing more detailed information.

The same design logic features the organization of the other search contexts of the basic accordion, i.e. *Search genes by Human Mendelian Genetic disorder*, *Search genes by Pathway*, *Bulk Queries*, *Search by Drug*.


BIOCLOUD
SEARCH ENGENE
HOME
SEARCH
TOOLS
ABOUT
TP53 - Entrez Gene ID 7157
General Information
Cross-database ID
Interaction network and Structures
STRING interaction network
PDB structures
3D Bio Assembly from Protein Data Bank
Load PDB IDs
"Pay as you go"
Microarrays Gene Expression
Pathways
Ontology Terms

A repository for 3-D biological macromolecular structure data. PDB archives protein structures deduced from crystallography and nuclear magnetic reasonance (NMR) experiments on protein structures.

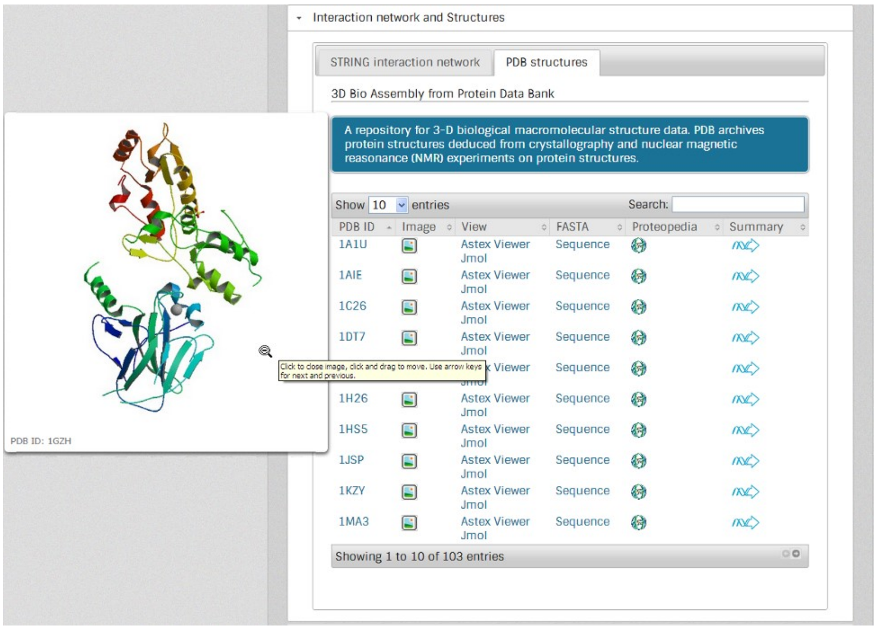
Click the button to load all PDB IDs related to TP53

Fig. 4. Expansion of the item "Interaction network and Structures" in the gene accordion. The arrow shows the button to catch data in "pay as you go" fashion.

In the *Search by Drug* context, as a further example, the user specifies a drug name and BSE auto-completes the user input using M.A.T.A.D.O.R. [33], a public repository which annotates relationships between human genes and drugs. Fig. 6 depicts results of searching for the drug "Aspirin".

Finally, the *drug accordion* occurs whenever the user clicks on a drug name. For example, in Fig. 6, when the user clicks on "Aspirin - Pubchem ID 2244" in the window "Results", he is redirected to the drug accordion (Fig. 7) to obtain additional information.

In Fig. 7, the item "General Information" is expanded and shows details about the drug "Aspirin" and the related 2D structure. The drug accordion enables searching for specific molecular information about drugs. For example, the "Protein Interactions" item shows the relationships among drugs and Gene expression products as annotated in the M.A.T.A.D.O.R. dataset [33].



Interaction network and Structures
STRING interaction network
PDB structures
PDB ID: 1GZH
Click to close image, click and drag to move. Use arrow keys for next and previous.

3D Bio Assembly from Protein Data Bank

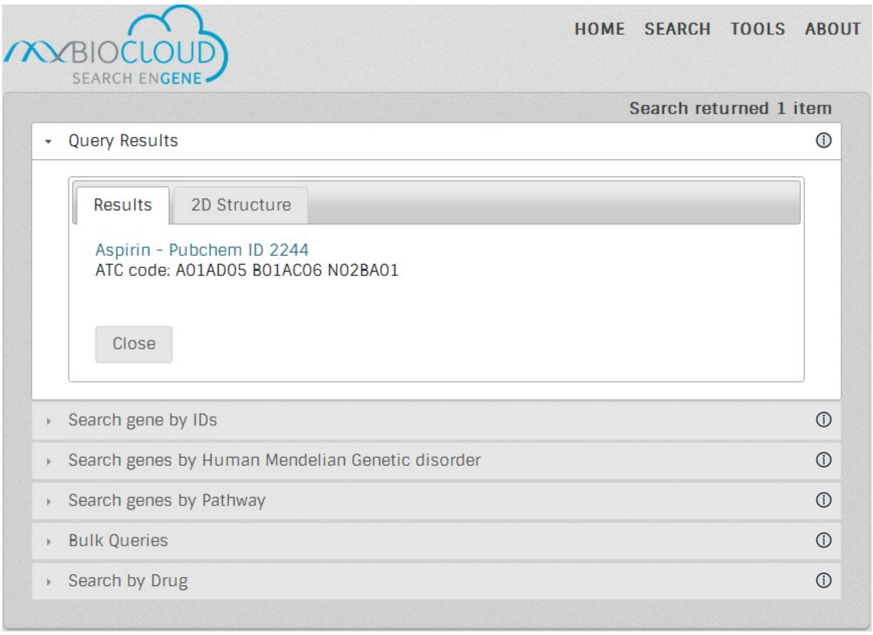
A repository for 3-D biological macromolecular structure data. PDB archives protein structures deduced from crystallography and nuclear magnetic reasonance (NMR) experiments on protein structures.

Show 10 entries   Search:

PDB ID	Image	View	FASTA	Proteopedia	Summary
1A1U	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1AIE	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1C26	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1DT7	🖼	Astex Viewer Jmol	Sequence	🌐	➪
[illegible]	[illegible]	k Viewer Jmol	Sequence	🌐	➪
1H26	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1HS5	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1JSP	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1KZY	🖼	Astex Viewer Jmol	Sequence	🌐	➪
1MA3	🖼	Astex Viewer Jmol	Sequence	🌐	➪

Showing 1 to 10 of 103 entries

Fig. 5. Data captured in pay-as-you-go fashion.

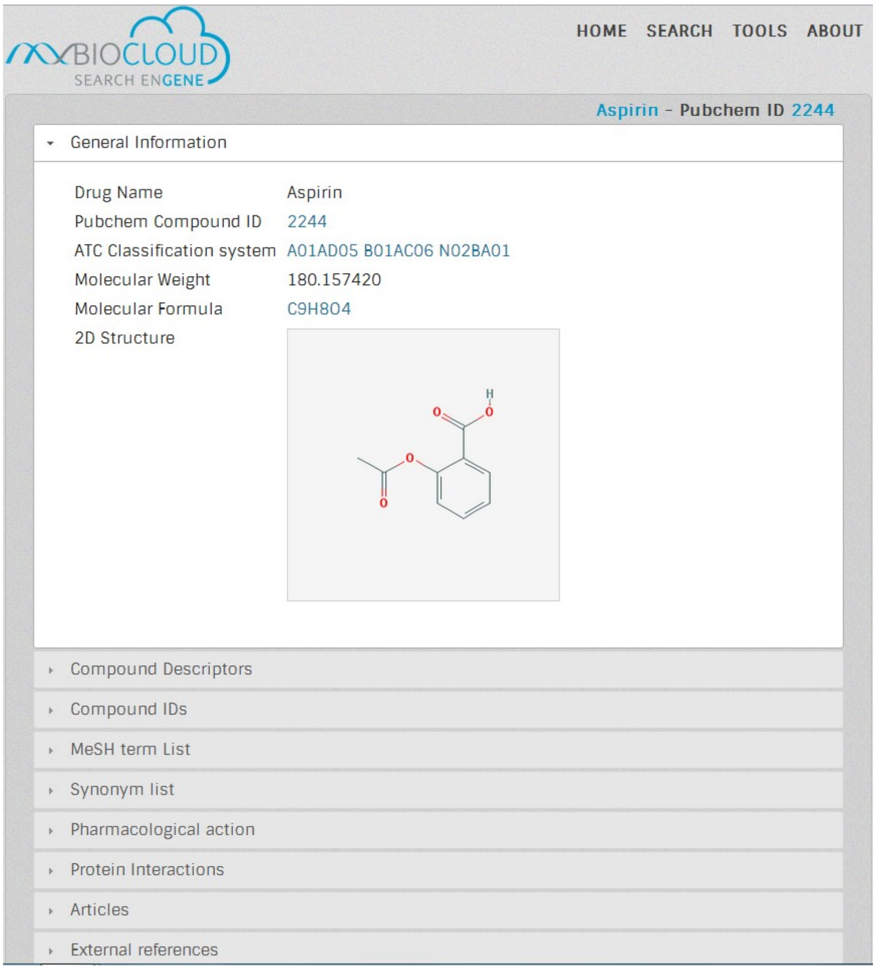


Search returned 1 item

Aspirin - Pubchem ID 2244

ATC code: A01AD05 B01AC06 N02BA01

Fig. 6. Search by drug context in the basic accordion: query results.



Aspirin - Pubchem ID 2244

General Information

Drug Name	Aspirin
Pubchem Compound ID	2244
ATC Classification system	A01AD05 B01AC06 N02BA01
Molecular Weight	180.157420
Molecular Formula	C9H8O4
2D Structure	

Fig. 7. The *drug accordion* which shows the contexts related to the drug "Aspirin".

The current functionalities of BSE could be extended to incorporate scalable tools for appropriate use cases in order to facilitate rapid large scale analysis of genetic information. In this direction, we are starting to implement tools which are made available easily through BSE and benefit from BSE searching capabilities.

We believe that this combination (searching-plus-tools) will allow for easy, user-friendly and transparent analysis of genetic data without requiring the user to know anything about the technical specifications of different systems (i.e. job submission, localization of web resources etc.).

Finally, the BSE user interface is unique in its focus on aggregating distributed web content in a flexible menu, a model that is highly amenable to future extension and customization by adding additional gene annotation resources, and by customizing the accordion menu to suit specific user needs.

5 Conclusions

Designed for people involved in the analysis of biological data (i.e. molecular biologists, biochemists, medical doctors, molecular pathologists etc.), BSE is a suitable and scalable cloud application that allows simple and advanced data searching in different databases. Going further from simply integrating content within genetic databases, as data warehousing systems do, BSE considers cloud and dataspaces the basic paradigms for effective searching big data from genomic resources. This is a unique feature of BSE.

Specifically, our work has explored how the convergence of cloud computing and dataspaces can offer both added-value service components and flexibility, making this convergence an attractive combination also for any scientific domain. BSE meets some important requirements, such as high performance, fault handling and compensation, scalability, elasticity, trust and security support, multi-tenancy, quality of service, and so on.

Most importantly, we tried to identify the nature of the technology we need in order to address big data searching issues in bioinformatics field in that complementing the capabilities of genetic portals. Albeit relatively new, cloud computing and dataspace paradigms seem to offer a prospect of new insights in bioinformatics. Finally, we are confident that, even implemented for data searching in genetic databases, our approach might reveal new directions for improving web based exploration of big data in life science.

Acknowledgments. We thank Francesco Masulli for his appreciations and precious suggestions and the CIBB 2013 general chairs Enrico Formenti, Roberto Tagliaferri and Ernst Wit for hosting the presentation of BSE as special event at CIBB 2013.

This research was supported by RAS, Regione Autonoma della Sardegna (Legge regionale 7 agosto 2007, n. 7), in the project "*DENIS: Dataspaces Enhancing the Next Internet in Sardinia*".

References

1. Ranganathan, S., Schönbach, C., Kelso, J., Rost, B., et al.: Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference. BMC Bioinformatics 2011, 12 (suppl 13): S1 (2011)
2. Tankard, C.: Big data security. Network Security 2012(7), 5-8 (2012)
3. Pennisi, E.: Human genome 10th anniversary. Will computers crash genomics? Science, 331, 666-668 (2011)
4. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nat Rev Genet, 11, 647-57 (2010)
5. Marshall, E.: Human genome 10th anniversary. Waiting for the revolution. Science, 331, 526-529 (2011)
6. Franklin, M.J., Halevy, A.Y., Maier, D.: From databases to dataspaces: a new abstraction for information management. SIGMOD Record, 34(4), 27–33 (2005)
7. Halevy, A.Y., Franklin, M.J., Maier, D.: Principles of dataspace systems. In Proceedings of PODS'06, ACM, New York, 1-9 (2006)

8. Hogue, C., Ohkawa, H., Bryant, S.: A dynamic look at structures: WWW-entrez and the molecular modeling database. Trends Biochem Sci, 21, 226–229 (1996)
9. Ostell, J.: The entrez search and retrieval system. The NCBI handbook [Internet], 2002, updated 2003, http://www.ncbi.nlm.nih.gov/books/NBK21081/
10. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/
11. Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., et al.: The GenBank genetic sequence databank. Nucleic Acids Res, 14(1), 1–4 (1986)
12. Mizrachi, I.: GenBank: The nucleotide sequence database. The NCBI handbook [Internet], 2002, updated 2007, http://www.ncbi.nlm.nih.gov/books/NBK21105/
13. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., et al.: Database resources of the national center for biotechnology information. Nucleic Acids Res, 40(Database issue):D13–D25 (2012)
14. McEntyre, J., Lipman, D.: PubMed: bridging the information gap. CMAJ, 164(9), 1317–1319 (2001)
15. Canese, K., Jentsch, J., Myers, C.: PubMed: The bibliographic database. The NCBI handbook [Internet], 2002, updated 2003, http://www.ncbi.nlm.nih.gov/books/NBK21094/
16. Dong, X., Halevy, A.Y.: Indexing dataspaces. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, SIGMOD'07, ACM, New York, 43–54 (2007)
17. Howe, B., Maier, D., Rayner, N., Rucker, J.: Quarrying dataspaces: Schemaless profiling of unfamiliar information sources. In Proceedings of ICDEW'08, IEEE Computer Society, 270-277 (2008)
18. Atzori, M., Dessì, N.: Dataspaces: Where structure and schema meet. Studies in Computational Intelligence, 375, 97-119 (2011)
19. Jeffery, S.R., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go user feed-back for dataspace systems. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, SIGMOD'08, ACM, New York, 847-860 (2008)
20. Hedeler, C., Belhajjame, K., Paton, N.W., Fernandes, A.A.A., et al.: Pay-as-you-go mapping selection in dataspaces. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD'11, ACM, New York, 1279-1282 (2011)
21. Chen, J., Qian, F., Yan, W., Shen, B.: Translational Biomedical Informatics in the Cloud: Present and Future. BioMed Research International, Volume 2013, Article ID 658925, 8 pages (2013)
22. Stonebraker, M.: SQL databases v. NoSQL databases. Communications of the ACM, 53(4), 10-11 (2010)
23. Sayers, E.: E-utilities Quick Start. Entrez Programming Utilities Help [Internet], 2008, updated 2013, http://www.ncbi.nlm.nih.gov/books/NBK25500/
24. Chambers, J., Davies, M., Gaulton, A., Hersey, A., et al.: UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. Journal of Cheminformatics, 5:3 (2013)
25. The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res., 40: D71-D75 (2012)
26. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., et al.: STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res., 37(Database issue):D412-6 (2009)
27. Kelder, T., Pico, A.R., Hanspers, K., van Iersel, M.P., et al.: Mining Biological Pathways Using WikiPathways Web Services. PLoS ONE, 4(7): e6447 (2009)
28. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., et al.: KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res., 40: D109-D114 (2012)
29. Wu, C., MacLeod, I., Su, A.I.: BioGPS and MyGene.info: organizing online, gene-centric information. Nucl. Acids Res., 41(Database issue): D561-D565 (2013)
30. Europe PMC, http://europepmc.org/RestfulWebService

31. NoSQL, www.nosql-database.org
32. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res., 33(Database Issue): D54–D58 (2005)
33. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., et al.: SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res., 36(Database issue):D919-22 (2008)
34. Croft, D., O'Kelly, G., Wu, G., Haw, R., et al.: Reactome: a database of reactions, pathways and biological processes. Nucleic Acid research, 39: D691-D697 (2011)
35. McKusick, V.A.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, Johns Hopkins University Press, Baltimore (1998)
36. Pruitt, K.D., Tatusova, T., Brown, G.R., Maglott, D.R.: NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res., 40(Database issue):D130-135 (2012)
37. de Magalhaes, J.P.: The Biology of Ageing: A Primer. An Introduction to Gerontology. Cambridge University Press, Cambridge, 21-47 (2011)
38. Yang, W., Soares, J., Greninger, P., Edelman, E.J., et al.: Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucl. Acids Res., 41(Database issue): D955-D961 (2013)
39. Google App Engine, https://developers.google.com/appengine/
40. Biopython, www.biopython.org/