Published online 6 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W23–W27***
doi:10.1093/nar/gkp265

BioMart Central Portal—unified access to biological data

Syed Haider[1,2], Benoit Ballester[1], Damian Smedley[1], Junjun Zhang[3], Peter Rice[1] and Arek Kasprzyk[3,*]

[1]EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, [2]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK and [3]Ontario Institute for Cancer Research, MaRS Centre, 101 College Street, Toronto M5G 0A3, Canada

Received March 4, 2009; Revised and Accepted April 8, 2009

ABSTRACT

BioMart Central Portal (www.biomart.org) offers a one-stop shop solution to access a wide array of biological databases. These include major biomolecular sequence, pathway and annotation databases such as Ensembl, Uniprot, Reactome, HGNC, Wormbase and PRIDE; for a complete list, visit, http://www.biomart.org/biomart/martview. Moreover, the web server features seamless data federation making cross querying of these data sources in a user friendly and unified way. The web server not only provides access through a web interface (MartView), it also supports programmatic access through a Perl API as well as RESTful and SOAP oriented web services. The website is free and open to all users and there is no login requirement.

INTRODUCTION

The advancements in sequencing technologies and subsequent growth in the repertoire of biological information are posing serious data-management challenges. The volume of these data is expected to continue to grow exponentially. Projects such as GenBank (1), HapMap (2) and the SNP Consortium are prime examples of the high-throughput data-management challenges that we are experiencing. Querying different biological data sources in an integrated manner generally involves moving all the data into a centralized data warehouse, necessitating substantial resources for keeping it up to date with component data sources. New generation sequencing projects such as the 1000 Genomes Project and International Cancer Genome Consortium (ICGC) are expected to produce data on an unprecedented scale. Moving this type of data into a central location for integrated querying with other resources presents considerable organizational and physical transfer challenges. One solution to this challenge lies in federated databases whereby individual data providers are responsible for updates and release cycles. The federated model eliminates the need to aggregate and manage all the data in any one central location. Another dimension of this problem is the provision of fast and robust access to such large quantities of data; how do we bring this data to end-users without having to expose any of the back-end issues pertaining to discovering repository location, information retrieval and merging with other datasets to support cross querying which is often the case in biological queries. Lastly, the results to be returned from these databases must be in standard formats and where possible, semantically annotated to ensure interoperability with other databases and tools. The Distributed Annotation System (DAS) (3) as well as BioMart (4) are functional examples of such frameworks. The BioMart software system offers a generic framework for biological data storage and retrieval particularly suited for large scale 'omics data through a single point of access. The web server, BioMart Central Portal, provides access to variety of datasets that can be queried independently or in a federated way enabling users to ask complex questions over data sources that may be located at different geographical locations. These inculde Ensembl genomic, Uniprot protein, Reactome pathway, HGNC gene name, Wormbase genomic and PRIDE proteomic data (5–10). As of March 2009, BioMart Central Portal brings together an extensive range of databases (see Figure 1), serving more than 100 datasets with an average monthly usage of over 1 million server hits (see Supplementary Table S1). Furthermore, the web server provides complete access to metadata that can be used by third party client writers to emulate functionality offered by the BioMart Central Portal as per their domain requirements. We believe that this service will be of enormous benefit to many users and deployers ranging from wet-lab biologists to computer scientists working in bioinformatics setups.

*To whom correspondence should be addressed. Email: arek.kasprzyk@oicr.on.ca

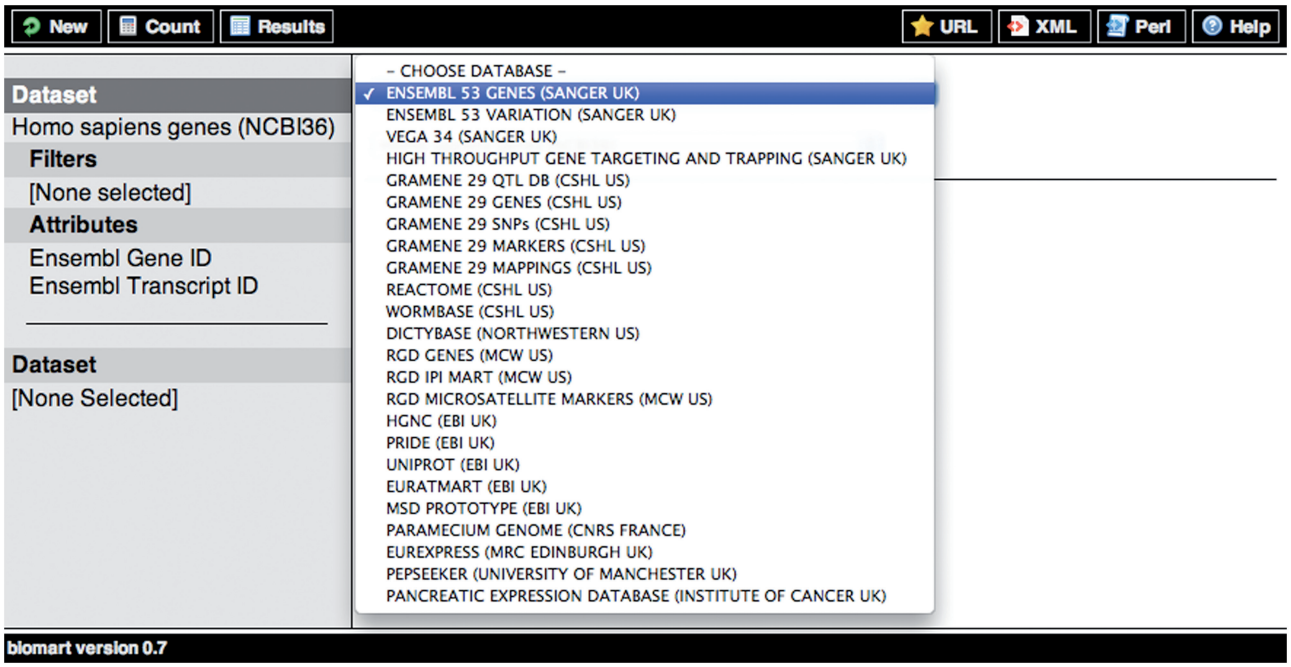

Figure 1. List of databases available through BioMart Central Portal (March 2009).

BIOMART CENTRAL PORTAL

The BioMart Central Portal is a web server interface of BioMart software and provides a unified view over disparate data sources that enable bioscientists to retrieve data from one or multiple sources in a simple and efficient way. The library behind the web server handles user request and takes over the responsibility of fetching data from respective locations, aggregating results and subsequent formatting in the specified format. Figure 2 describes the high-level system architecture and the data flow. A query to the BioMart Central Portal primarily consists of three simple abstractions (Dataset, Filters and Attributes). *Dataset* being the logical boundary of the query, *Filters* (optional) are the inputs and *Attributes* are the user specified outputs. The BioMart Central Portal handles queries from several interfaces, all utilizing these three abstractions in a coherent way across all interfaces. These interfaces are:

- Perl API
- Web interface (MartView)
- URL based access
- RESTful web service (MartService)
- SOAP web service (MartServiceSoap)
- DAS server

All the query interfaces are written in Perl. A detailed description of usage and query formulation is explained in (11) and the project docs available at www.biomart.org/install.html.

In the sections to follow, we will describe the access to BioMart Central Portal through its web service end-point, MartServiceSoap. The BioMart queries can be fundamentally categorized into two types; metadata and data access. A machine readable XML based description of inputs and outputs of these queries are published in Web Service Definition Language (WSDL) and XML Schema Definition (XSD) files available at http://www.biomart.org/biomart/martwsdl and http://www.biomart.org/biomart/martxsd.

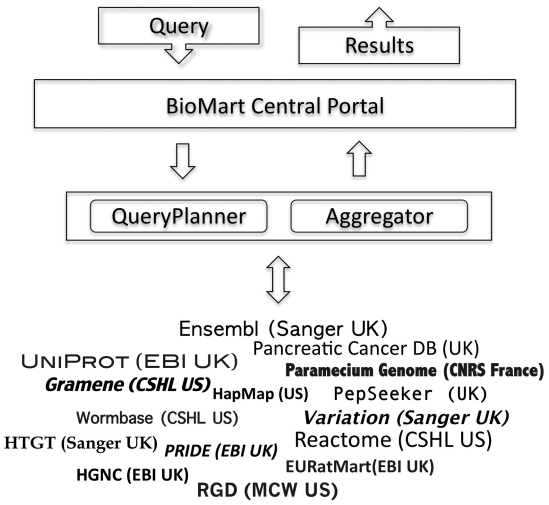

Figure 2. The schematic representation of BioMart Central Portal.

Metadata Access

These requests are used to retrieve information about which databases, datasets, filters, attributes and associated formatters are made available by BioMart Central Portal. These queries support not only programmatic access, they also return additional information which may be used to write domain specific specialized clients to access BioMart

Central Portal remotely. These requests are described as follows:

getRegistry. This request retrieves information contents such as name, location, host, port etc about all the databases/marts available at BioMart Central Portal. The output is equivalent to the list displayed by MartView, see Figure 1.

getDatasets. This request retrieves a list of datasets available under each mart, mart name being the input of the request.

getFilters and getAttributes. These two requests retrieve a list of all the filters and attributes available given a dataset. Additional information about hierarchy, limitations and output formatters is also returned. Most importantly, the W3C suggested property *'modelReference'* in the output, if configured by the data publisher, provides the Uniform Resource Identifier (URI) of the concept in an ontology that contains description of the output attribute/s. This feature offers a framework for semantic annotation of terms in BioMart databases. This feature will improve interoperability of BioMart results with non-BioMart data sources and analysis tools.

Data Access

In order to access biological content of the marts available through the BioMart web server, a *query* request is used. Figure 3a illustrates an example query in MartSoapService format that spans two datasets (Ensembl Homo Sapiens & Reactome Pathways) residing at different locations (Sanger & CSHL). The query finds the alleles in genes involved in the regulation of DNA replication. A user can specify the attributes of interest along with any possible limitations (filters) from a given dataset/s and in return gets results as shown in Figure 3b. Users are neither expected to ascertain the database specific access protocol, nor its physical location. From a user's point of view, all datasets appear to be residing at BioMart Central Portal that takes care of all underlying federation logic.

Query processing

The BioMart server-side software constitutes of a *QueryPlanner* and an *Aggregator*. The QueryPlanner consumes data access queries and formulates an execution plan. If BioMart Central Portal has direct access credentials to the database server, then SQL statements are compiled, otherwise XML-based web service requests are sent to the remote BioMart web server over HTTP stream and results are retrieved over the same connection. The execution scheme consists of ANSI SQL statements (to ensure compatibility across MySQL, Oracle and PostgreSQL) or web service requests or combination of both if a query involves one or more datasets providing direct database access and others proving only web service access. To minimize database or HTTP time-outs and slow response times, the query engine uses a sophisticated batching system that performs the job over several iterations. The results are piped back to the user as soon as the first batch in finished. The Aggregator component enables merging of data coming from different sources on a common concept. This is achieved by extending the afore-mentioned abstractions, Attributes and Filters, to *Exportables* and *Importables*. A dataset that exposes an attribute as exportable is able to integrate data from all those sources whereby a filter with similar name is tagged as importable. The exportables and importables are columns with similar contents in a database table. The aggregation of results is an in-memory operation that does not prove to be very costly given the batching model described above.

Registry

The BioMart Central Portal does not store any data locally except meta information of all the datasets. The server maintains a registry containing references to remote BioMart web servers. To add a new mart to this registry, we only require the URL of the BioMart server hosting the databases or read access to the database server. This information is added to the registry file of the web server and following a configuration rerun, the whole bioinformatics community can benefit from the data through BioMart Central Portal as well as several third party softwares, see www.biomart.org for a complete list. The web server stays in sync with any of the data updates carried out on various databases. However, updates relating to metadata are made available shortly after the stable release of such updates upon reconfiguration of the web server.

FUTURE DIRECTIONS

We are working on extending the system to support multiple and more specialized web GUIs. This includes integration of analysis and visualization plugins with special focus on cancer research. We also envisage substantial development towards semantic annotation of *attributes* and *filters* by data publishers that would enhance the interoperability of mart datasets with analysis tools and non-BioMart databases. MartServiceSoap provides a complete framework to define ontology references for the annotation of these terms and we would like to collaborate with data providers to achieve this goal.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are very thankful to Dr Paul Flicek (EMBL-EBI) for his feedback on this manuscript.

FUNDING

Ontario Institute for Cancer Research; the Wellcome Trust, EMBL; the European Commission within its FP6 Programme under the thematic area 'Life sciences, genomics and biotechnology for health', contract number

```
<Soapenv:Body>
    <mart:query>
        <virtualSchemaName>default</virtualSchemaName>
        <count>0</count>
        <Dataset>
            <name>hsapiens_gene_ensembl</name>
            <Attribute>
                <name>allele</name>
            </Attribute>
        </Dataset>
        <Dataset>
            <name>pathway</name>
            <Filter>
                <name>_displayname</name>
                <value>DNA replication initiation</value>
            </Filter>
            <Attribute>
                <name>_displayname</name>
            </Attribute>
            <Attribute>
                <name>pathway_db_id</name>
            </Attribute>
        </Dataset>
    </mart:query>
</Soapenv:Body>
```

```
<Soapenv:Body>
    <queryResponse xmlns="http://www.biomart.org:80/MartServiceSoap">
    <resultsRow>
        <item attribute="allele" xsi:type="xsd:string" modelReference="">TG/-</item>
        <item attribute="_displayname" xsi:type="xsd:string" modelReference="">DNA replication initiation</item>
        <item attribute="pathway_db_id" xsi:type="xsd:string" modelReference="">68952</item>
    </resultsRow>
    <resultsRow>
        <item attribute="allele" xsi:type="xsd:string" modelReference="">T/G</item>
        <item attribute="_displayname" xsi:type="xsd:string" modelReference="">DNA replication initiation</item>
        <item attribute="pathway_db_id" xsi:type="xsd:string" modelReference="">68952</item>
    </resultsRow>
    ..........
    ..........
    </queryResponse>
</Soapenv:Body>
```

Figure 3. (**a**) SOAP request envelope representing data federation between Ensembl Homo Sapiens (Sanger-UK) and Reactome pathway (CSHL-US) datasets. The query finds the alleles in genes involved in the regulation of DNA replication (**b**) SOAP response envelope for the query shown in figure 3a.

LHSG-CT-2004-512092. Funding for open access charge: Ontario Government and Ministry of Research and Innovation.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res*., **37**, D26–D31.
2. The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 851–861.
3. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
4. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*., **14**, 160–169.
5. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,K., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–D697.
6. The UniProt Consortium., (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **36**, D190–D195.
7. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al*. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*., **8**, R39.
8. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res*., **36**, D445–D448.
9. Bieri,T., Blasiar,D., Ozersky,P., Antoshechkin,I., Bastiani,C., Canaran,P., Chan,J., Chen,N., Chen,W.J., Davis,P. *et al*. (2007) WormBase: new content and better access. *Nucleic Acids Res*., **35**, D506–D510.
10. Jones,P., Côté,R.G., Cho,S.Y., Klie,S., Martens,L., Quinn,A.F., Thorneycroft,D. and Hermjakob,H. (2008) PRIDE: new developments and new datasets. *Nucleic Acids Res*., **36**, D878–D883.
11. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart—biological queries made easy. *BMC Genomics*, **10**, 22.