DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID bar049, doi:10.1093/database/bar049

Editorial

BioMart: driving a paradigm change in biological data management

Arek Kasprzyk*

Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800 Toronto, Ontario, M5G 0A3, Canada.

***Corresponding author:** Tel: 647 258 4321; Fax: 647-258-4321; Email: arek.kasprzyk@gmail.com

Biological data management is a challenging undertaking. It is challenging for database designers, because biological concepts are complex and not always well defined, and therefore the data models that are used to represent them are constantly changing as new techniques are developed and new information becomes available. It is challenging for collaborating groups based in different geographical locations who wish to have unified access to their distributed data sources, because combining and presenting their data creates logistical difficulties. Finally, it is challenging for users of biological databases, because in order to correctly interpret the experimental data located in one database, additional information from other databases is frequently needed, requiring the user to learn multiple systems.

The BioMart project (www.biomart.org) was initiated to address these challenges. The BioMart software is based on two fundamental concepts: data agnostic modelling and data federation. Data agnostic modelling simplifies the difficult and time-consuming task of data modelling. In BioMart, this is achieved by using a predefined, query-optimized relational schema that can be used to represent any kind of data (1). Data federation makes it possible to organize multiple, disparate and distributed database systems into what appears to be a single integrated virtual database. It therefore allows users to access and cross reference data from these data sources using a single user interface, without the need for database administrators to physically collate the data in one location.

Using these fundamental concepts, the BioMart project has driven a change in the biological data management paradigm, where individual biological databases are managed by different custom built systems. To give more control to both the users and the data providers, a new, innovative solution was required. BioMart started by adapting data warehousing ideas to create one universal software system for biological data management and empower biologists with the ability to create complex, customized datasets through a web interface without the need for bioinformatics support (1). It subsequently introduced a new innovative way of creating large multi-database repositories that avoid the need to store all the data in a single location (2), and finally it proved that large-scale projects involving next generation sequencing data can be managed efficiently in a distributed environment (3).

BioMart has successfully adapted data warehousing ideas such as data marts, dimensional modelling (4), and query optimization into the world of biological databases (5–13). BioMart's ability to quickly deploy a website hosting any type of data, user-friendly graphical user interface, several programmatic interfaces and support for third party tools contributed to its success and adoption by many different types of projects around the world as their data management platform (14). During the 10 years of its existence, BioMart has grown from humble beginnings as a 'data mining extension' for the Ensembl website (1), to become an international collaboration involving large number of different organizations located on five continents: Asia, Australia, Europe, North America and South America (3,15). It has a large community of users and developers and it has been successfully used in both academia and industry. The latest version of the BioMart software has been significantly enhanced with numerous graphical user interfaces that are tailored to different user groups. In addition, it has been further improved by parallel query processing, it is now extensible with different analysis tools

Database, Vol. 2011, Article ID bar049, doi:10.1093/database/bar049

and the installation process can be effortlessly completed with just a few mouse clicks (16).

Building on the wealth of information that has become accessible through the BioMart interface, the BioMart Central Portal (15) has introduced an innovative alternative to the large data stores maintained by specialized organizations such as The European Bioinformatics Institute (EBI) or The National Center for Biotechnology Information (NCBI). BioMart Central Portal is a first-of-its-kind, community-driven effort to provide unified access to dozens of biological databases. All development and maintenance of individual databases is left to the individual data providers, making it a very cost-effective approach. The groups that maintain individual sources can do so at their own location without the necessity of any data exchange procedures. In addition, they can draw on the wealth of information available through the portal to expose their data in the context of third party annotations. The BioMart Central Portal approach is very democratic: everyone can join or remove their data source at any time. BioMart Central Portal is effectively a 'Virtual Bioinformatics Institute' with no onsite personnel, minimal administration, and a very 'green' footprint.

More recently, the International Cancer Genome Consortium (ICGC) Data Portal has demonstrated how BioMart can scale to manage large collaborative projects involving next generation sequencing data (3). The ICGC is generating data on an unprecedented scale by sequencing 500 cancer genomes and matched normal control genomes for 50 different cancer types (17). The effort is distributed between multiple participating countries and sequencing centres. Given the scale of the effort, moving all of the data to a single location is impractical. Instead, the ICGC Data Portal relies on BioMart data federation. By replicating and distributing the data model across different centres that produce the same type of data according to the same recipe, the scalability of the effort is greatly improved. Each centre is only responsible for managing their own data while data access to all of the consortium data is managed by the BioMart software. This presents a scalable approach, not only in the traditional sense of parallelizing data processing and storage, but also in a more general sense of outsourcing the external annotation expertise by federating annotations from additional, independently-maintained databases that are available in the BioMart Central Portal.

The future developments for BioMart involve specialized 'pre-packaged' and reusable data portals. One example already in development is the OncoPortal, aimed at researchers managing cancer data. It will include preconfigured access to sources of annotations that are useful for cancer research such as Ensembl (5), Reactome (12), COSMIC (9), Pancreatic Expression Database (10) and others. It will also include a set of tools that are specifically designed for cancer data analysis. There are plans to build other preconfigured portals for different research areas, such as a mouse portal and a model organism portal. It is an ambition of the BioMart community that the BioMart project remains at the forefront of innovative solutions for biological data management in the years to come. By creating these specialized solutions and further reducing the barriers to entry, the aim is to encourage more groups to share their data through BioMart, thereby further enhancing the entire BioMart community.

References

1. Kasprzyk,A., Keefe,D., Smedley,D. *et al*. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
2. Haider,S., Ballester,B., Smedley,D. *et al*. (2009) BioMart Central Portal—unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.
3. Zhang,J., Baran,J., Cros,A. *et al*. (2011) International Cancer Genome Consortium Data Portal: a ane stop-shop for cancer genomics data. *Database*, **2011**, DOI: 10.1093/database/bar026.
4. Kimball,R. (1998) *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, New York: Wiley. xviii, p. 771.
5. Kinsella,R., Kähäri,A., Haider,S. *et al*. (2011) Ensembl BioMarts: a hub for data retrieval across the taxonomic space. *Database*, **2011**, DOI: 10:1093/database/bar030.
6. Jones,P., Binns,D., McMenamin,C. *et al*. (2011) The InterPro BioMart: Powerful, federated query and web-service access to the InterPro Resource. *Database*, **2011**, DOI: 10.1093/database/bar033.
7. Moreews,F., Klopp,C., Rauffet,G. *et al*. (2011) SigReannot-mart: a query environment for expression microarray probe re-annotations. *Database*, **2011**, DOI: 10.1093/database/bar025.
8. Stevenson,P., Richardson,L., Venkataraman,S. *et al*. (2011) The BioMart interface to the eMouseAtlas gene expression database EMAGE. *Database*, **2011**, DOI: 10.1093/database/bar029.
9. Shepherd,R., Forbes,S.A., Beare,D. *et al*. (2011) Data mining using the Catalogue of Somatic Mutations in Cancer BioMart. *Database*, **2011**, DOI: 10.1093/database/bar018.
10. Cutts,R.J., Gadaleta,E., Lemoine,N.R. *et al*. (2011) Using BioMart as a framework to manage and query pancreatic cancer data. *Database*, **2011**, DOI: 10.1093/database/bar024.
11. Perez-Llamas,C., Gundem,G. and Lopez-Bigas,N. (2011) Integrative Cancer Genomics (IntOGen) in BioMart. *Database*, **2011**, DOI: 10.1093/database/bar039.
12. Haw,R., Croft,D., Yung,C.K. *et al*. (2011) The Reactome BioMart. *Database*, **2011**, (This issue), doi:10.1093/database/bar031.
13. Ndegwa,N., Coté,R.G., Ovelleiro,D. *et al*. (2011) Critical amino acid residues in proteins: a BioMart integration of Reactome protein annotations with PRIDE mass spectrometry data and COSMIC somatic mutations. *Database*, **2011**, (This issue), doi:10.1093/database/bar047.

14. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
15. Guberman,J.M., Arnaiz,O., Baran,J. *et al*. (2011) BioMart Central Portal: an open database network for the biological community. *Database*, **2011**, DOI: 10.1093/database/bar041.
16. Zhang,J., Haider,S., Baran,J. *et al*. (2011) BioMart: a data federation framework for large collaborative projects. *Database*, **2011**, DOI: 10.1093/database/bar038.
17. Hudson,T.J., Anderson,W., Artez,A. *et al*. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.