Advanced Genomic Data Mining

Xosé M. Fernández-Suárez*, Ewan Birney

European Bioinformatics Institute, Ensembl Group, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Introduction

As data banks increase in size, one of the current challenges in bioinformatics is to be able to query them in a sensible way. Information is contained in different databases, with various data representations or formats, making it very difficult to use a single query tool to search more than a single data source.

Data mining is vital to bioinformatics as it allows users to go beyond simple browsing of genome browsers, such as Ensembl [1,2] or the UCSC Genome Browser [3], to address questions; for example, the biological meaning of the results obtained with a microarray platform, or how to identify a short motif upstream of a gene, amongst others. There are a number of integrated approaches available, some of which are described below (Figure 1).

The Table Browser at UCSC [4] supports text-based batch queries to the UCSC Genome Browser, limiting the output to entries meeting the selected criteria. A disadvantage of this tool is that users need to be familiar with the underlying database schema in order to know where their data is stored. Similarly, performing complex queries might require multiple steps that can be burdensome with this tool. Galaxy [5] provides a set of tools that can retrieve data from the Table Browser (Table Browser and BioMart will be explained below), facilitating complex queries that require multiple joins (Figure 2).

BioMart provides a query-oriented data management system to interact with different datasets (Ensembl [2], RGD [6,7], and WormBase [8], among many others). This data “warehouse” was originally developed for Ensembl, creating EnsMart [9,10]. From there, it was first deployed across the European Bioinformatics Institute (EBI), and now it has become a joint project between EBI and Cold Spring Harbor Laboratory (CSHL). The generic query system has shifted toward a federated approach that has been deployed for several biological databases, and has become a component of the Generic Model Organism Database (GMOD) project.

In this contribution, we provide some solutions for data mining; we focus on advanced ways of interacting with BioMart using other applications to retrieve information through different platforms such as Galaxy [3] and the biomaRt package of BioConductor [11,12]. Many of these tools also interact with the UCSC Table Browser and have similar approaches using the UCSC system. We also address programmatic access using BioMart’s own implementation of Web services (MartService). For local deployment of BioMart, see Table 1.

BioMart Web Interface

First we will focus on BioMart’s Web interface (http://www.biomaart.org) to illustrate how to join two different datasets: Reactome [13], a database of metabolic pathways, and UniProt [14], a catalogue of protein entries meeting the selected criteria. In this example, we need to obtain a catalogue of enzymes involved in carbohydrate metabolism in humans, as we are interested in a congric disorder in this pathway. To ask this question without an integrated data mining tool, one would have to start with Reactome to find enzymes involved in reaction pathways in human and then compare those enzymes to a list of entries in UniProt. However, BioMart allows us to join the two databases.

We can start our query by clicking on ‘MartView’ from the Web interface at http://www.biomaart.org, and selecting the Reactome database. Next, we can enrich our search for enzymes…GO ID(s)’ in the secondary dataset. Also select, under ‘External references’: ‘Entries with EC ID(s)’, to limit our query to enzymes only, and ‘eukaryota’ along with ‘Homo sapiens’ under ‘SPECIES’ (Species and Proteome Name, respectively). This will give a count of 257 in the secondary dataset. The output by choosing the following Attributes: “Species” “Gene Name” “Gene ID”. At this stage, 2,432 entries meet our criteria (i.e. we have asked for all human reaction pathways in the Reactome database).

In this contribution, we provide some solutions for data mining; we focus on advanced ways of interacting with BioMart using other applications to retrieve information through different platforms such as Galaxy [3] and the biomaRt package of BioConductor [11,12]. Many of these tools also interact with the UCSC Table Browser and have similar approaches using the UCSC system. We also address programmatic access using BioMart’s own implementation of Web services (MartService). For local deployment of BioMart, see Table 1.

BioMart Web Interface

First we will focus on BioMart’s Web interface (http://www.biomaart.org) to illustrate how to join two different datasets: Reactome [13], a database of metabolic pathways, and UniProt [14], a catalogue of protein entries meeting the selected criteria. In this example, we need to obtain a catalogue of enzymes involved in carbohydrate metabolism in humans, as we are interested in a congenic disorder in this pathway. To ask this question without an integrated data mining tool, one would have to start with Reactome to find enzymes involved in reaction pathways in human and then compare those enzymes to a list of entries in UniProt. However, BioMart allows us to join the two databases.

We can start our query by clicking on ‘MartView’ from the Web interface at http://www.biomaart.org, and selecting the Reactome database. Now, select the reaction dataset. Filters applied will be simply ‘Limit to Species’ Homo sapiens. Attributes can be selected as “Reaction name” and “Gene ENSEMBL ID”. At this stage, 2,432 entries meet our criteria (i.e. we have asked for all human reaction pathways in the Reactome database).

Click on the ‘count’ button at the top to obtain this number.

Next, we can enrich our search for enzymes in the UniProt database. This will require the ‘linked’ or secondary dataset. Follow this description, or view the tutorials for use of the linked database at http://www.ensembl.org/common/Workshops_Online?id = 117.

Click on the second ‘Dataset’ option at the left of the page. Select ‘UniProt proteomes’ as the database. In this instance, we will add as a filter the Gene Ontology (GO) [15] term GO:0005975 (associated with carbohydrate metabolic processes); this will be under ‘EXTERNAL IDENTIFIERS’, ‘Limit to proteins…GO ID(s)’ in the secondary dataset. Also select, under ‘External references’: ‘Entries with EC ID(s)’, to limit our query to enzymes only, and ‘eukaryota’ along with ‘Homo sapiens’ under ‘SPECIES’ (Species and Proteome Name, respectively). This will give a count of 257 in the secondary dataset. The genome location can be displayed in the output by choosing the following Attributes: “Chromosome” “Start Position” and “End Position” for the coordinates. Click ‘Results’ for the table in Figure 2.

Now you have a list of enzymes in UniProt involved in carbohydrate metabolism in humans.

BioConductor

BioConductor is open source software for the analysis of genomic data. It is based...
on the R language [16] (which is an implementation of the S language, a statistical programming language originally developed at Bell Laboratories to support research and data analysis of large statistical projects [17]).

R is an integrated software environment for data manipulation, which can be used as a statistics system (throughout many different packages). There are a large number of biologically relevant modules in BioConductor, some of which are described in [12] and at http://www.bioconductor.org/packages/release/Software.html. The biomart package provides an API (Application Programming Interface) in the scripting language R, allowing interaction with biomart databases. These include Ensembl, which produces and maintains automatic annotation on selected eukaryotic genomes; VEGA [18], the manually annotated Vertebrate Genome Annotation database; dbSNP [19], the Single Nucleotide Polymorphism database of NCBI; Gramene [20], a resource for comparative grass genomics; WormBase [9], the canonical database for *Caenorhabditis elegans* and related nematodes RGD [6,7]; and Reactome [13], a curated knowledgebase of biological pathways, amongst others.

R can be installed on different platforms; there are binaries available for Unix, Windows, and Macintosh. For a list of the Comprehensive R Archive Network (CRAN), go to http://cran.r-project.org/mirrors.html. Once obtained, the source should be unpacked and installed following the instructions provided. There is a built-in help facility invoked with help,
for instance ‘help(debug)’ will provide documentation about the debug function. Furthermore, following the installation of a package, a pre-built help search index is created. To know what commands are available in biomaRt use ‘help.search(‘biomaRt’):’.

Once R is installed and compiled, the default set of BioConductor packages is easily installed using the biocLite.R installation script as follows:

```r
> source(‘http://bioconductor.org/biocLite.R’) > biocLite(‘biomaRt’)```

With biomaRt installed, load the relevant library with the `library(biomaRt)` command, and then connect to any public BioMart database. The listMarts function will show which BioMart services are available. BioMart is structured in tables with attributes (the information you want to know) and filters (the information you know). You need to select a dataset (e.g., `rnorvegicus_gene_ensembl`), if you are interested in rat gene annotation from Ensembl. Issuing the following command: `rat = useMart(‘ensembl’, dataset = ‘rnorvegicus_gene_ensembl’)`, would set the dataset queried to be the Ensembl rat genes.

Below are two commands to query the library to see the currently available marts and datasets on the central server (Figure 3).

```r
> library(biomaRt) > listMarts()```

Additionally, the useMart function allows you to select the relevant BioMart dataset, using the name provided by listMarts.

```r
> ensml = useMart(‘ensembl’) > listDatasets(ensml)```

### Table 1. URLs for additional information.

| BioMart documentation | http://www.biocmart.org/user-docs.pdf |
|-----------------------|---------------------------------------|
| BioMart tutorials     | http://www.ensembl.org/info/using/website/tutorials/ |
|                       | http://www.ensembl.org/common/Workshops_Online |
| BioMart Central Server| http://www.biocmart.org/ |
| BioConductor          | http://www.bioconductor.org/ |
|                       | http://www.bioconductor.org/packages/release/Software.html |
| R                     | http://www.r-project.org/ |
| Installing R          | http://cran.r-project.org/doc/manuals/R-admin.html |
| R Archive Network     | http://cran.r-project.org/mirrors.html |
| biomaRt documentation | http://www.bioconductor.org/packages/1.8/bioc/vignettes/biomaRt/inst/doc/biomaRt.pdf |
| Galaxy                | http://main.g2.bx.psu.edu/ |

![Figure 3. BioMart libraries available in the Central Server.](http://www.biomart.org/user-docs.pdf)
Codelink from Applied Microarrays which features Whole Genome Bioarrays (a platform finish the session, use martDiscon

Functions available allow extraction of identifiers from different sources including Ensembl IDs, several microarray platforms, UniProt, RefSeq [21], and EntrezGene [22]. Genome sequences can be retrieved by specific chromosomal coordinates for a given species, allowing a user to mine regions they define. For example, a user could view all annotations upstream of a differentially expressed gene in order to investigate putative regulatory elements. Similarly, compare_mart_homology_47 supports queries across different species in order to identify homologous genes.

To illustrate how to use this tool, we provide an example: if you were interested in all mouse protein coding genes on Chromosome 10 along with their Ensembl and MGI identifiers, the following series of commands would carry out this query:

```r
>library(biomaRt)
>listMarts()
>ensembl=useMart("ensembl")
>listDatasets(ensembl)
>mouse=useMart("ensembl", dataset="mmusculus_gene_ensembl")
>getBM(attributes = c("codelink","ensembl_transcript_id","mgi_symbol"), filters="
 chromosome_name",values=10,mart=mouse)
```

The output could be saved to a file gene.ids, which can be invoked by simply typing `"".gene.ids`.

Researchers use DNA microarrays to establish the expression profiles of thousands of genes in a single experiment. Microarrays in their different incarnations have been used in a wide range of applications, e.g., disease characterization [23,24] and identification of novel genes or gene regulatory networks [25].

A more specific use of biomaRt involving the recently developed CodeLink Rat Whole Genome Bioarrays (a platform from Applied Microarrays which features approximately 34,000 transcript and EST targets) will show how to further analyse data obtained from these bioarrays. The CodeLink [26] R package can be installed (biocLite("codelink")) and used to obtain plotting functionality in the R statistical computing environment (e.g., plotMA, plotCorrelations, plotDensities, etc.). CodeLink bioarrays have been successfully applied to the identification of molecular signatures in colon cancer development [27]. In this paper, we will use some of the up-regulated probes from this study to illustrate how annotation could be retrieved using biomaRt, focusing on Ensembl genes associated to the probes. Data was obtained from the matrix plot of gene expression values (the authors provide supplementary data online at http://dnguyen.ucdavis.edu/~hum/datasets/3/main.html).

```r
>library(biomaRt)

Loading required package: XML

Loading required package: RCurl

>rat=useMart("ensembl", dataset="rnorvegicus_gene_ensembl")

Checking attributes and filters ...

Character strings are inserted using double (") or single (') quotes, while the c() function is used to concatenate arrays.

```r
>upregulated=c("AF003944","U67136","J04597","AF281635","BC090354","U12309","M4105","M4104","BC088159","D9450","B036719","D90404","M6854","L00320","L00313","L00314","L00315","L00316","L00317","L00318","M12151","J00719","M6854","M6854","J00720","J00721","J00722","J00723","J00724","J00725","J00726","M04452","K00966","K01626","K01721","K00250","M13241","M13650","M68853","M19972","X63545","X12355","D63378","BC062393")
```
Table 2. R output (using biomaRt library) providing CodeLink bioarray IDs and their mappings to Ensembl transcripts (chr: start-end position), as explained in the text.

| codelink | ensembl_transcript_id | enml | chromosome_name | start_position | end_position |
|----------|-----------------------|------|-----------------|----------------|-------------|
| GE13154  | ENSRNOT00000014152    | AF003944 | 1 | 125280974 | 125293051 |
| GE13549  | ENSRNOT0000018050    | AF281635 | 4 | 152975181 | 152979366 |
| GE20053  | ENSRNOT0000015476    | BC061719 | 8 | 101405877 | 101437760 |
| GE20496  | ENSRNOT0000020478    | BC062393 | 3 | 108216369 | 108240138 |
| GE19851  | ENSRNOT0000015723    | BC081159 | 16| 17591820 | 17597859 |
| GE13549  | ENSRNOT0000018050    | BC090354 | 4 | 152975181 | 152979366 |
| GE1195465 | ENSRNOT0000028196  | D00250 | 1 | 81345425 | 81359415 |
| GE20496  | ENSRNOT0000020478    | D63378  | 3 | 108216369 | 108240138 |
| GE20053  | ENSRNOT0000015476    | D84450  | 8 | 101405877 | 101437760 |
| GE20338  | ENSRNOT0000022342    | D90404  | 1 | 146629802 | 146661183 |
| GE21631  | ENSRNOT0000022342    | D90404  | 1 | 146629802 | 146661183 |
| ENSRNOT0000047540 | J00719 | 1 | 81266845 | 81290470 |
| GE1195465 | ENSRNOT0000028196  | J00720  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00721  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00722  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00724  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00725  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00726  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | J00728  | 1 | 81345425 | 81359415 |
| GE21002  | ENSRNOT0000088416    | J04597  | 6 | 98923111 | 98926505 |
| GE1195465 | ENSRNOT0000028196  | K00996  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | K01626  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | K01721  | 1 | 81345425 | 81359415 |
| ENSRNOT0000047540 | L00313 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00314 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00315 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00316 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00317 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00318 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00319 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | L00320 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | M11251 | 1 | 81266845 | 81290470 |
| GE1195465 | ENSRNOT0000028196  | M13234  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | M13650  | 1 | 81345425 | 81359415 |
| GE19851  | ENSRNOT0000015723    | M14104  | 16| 17591820 | 17597859 |
| GE19851  | ENSRNOT0000015723    | M14105  | 16| 17591820 | 17597859 |
| GE1195465 | ENSRNOT0000028196  | M19972  | 1 | 81345425 | 81359415 |
| GE1195465 | ENSRNOT0000028196  | M26853  | 1 | 81345425 | 81359415 |
| ENSRNOT0000034845 | M26854 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | M26854 | 1 | 81266845 | 81290470 |
| ENSRNOT0000034845 | M26855 | 1 | 81266845 | 81290470 |
| ENSRNOT0000047540 | M26855 | 1 | 81266845 | 81290470 |
| GE1195465 | ENSRNOT0000028196  | M34452  | 1 | 81345425 | 81359415 |
| ENSRNOT0000047540 | M37134 | 1 | 81266845 | 81290470 |
| GE20281  | ENSRNOT0000014785    | U12309  | 19| 58601188 | 58628500 |
| GE21915  | ENSRNOT0000014785    | U12309  | 19| 58601188 | 58628500 |
| GE21002  | ENSRNOT0000088416    | U67136  | 6 | 98923111 | 98926505 |
| GE20496  | ENSRNOT0000020478    | X12355  | 3 | 108216369 | 108240138 |
| GE20381  | ENSRNOT0000041580    | X63545  | 1 | 81780088 | 81853249 |
| GE22156  | ENSRNOT0000041580    | X63545  | 1 | 81780088 | 81853249 |

doi:10.1371/journal.pcbi.1000121.t002
select: Get Data > BioMart (Central Server) to query the ‘Ensembl genes’ database and retrieve 300 bp upstream (Attributes > Sequences > Flank-coding region) to ENSG00000100985. Following this, we can run fuzznuc (from EMBOSS) to find out if the Runx2 motif is upstream of the human gene. Such short motifs (length <10) cannot be identified by BLAST, SSAHA, or BLAT; that’s why we will use EMBOSS.

This example shows how Galaxy adds a layer of analysis to the genomic sequences retrieved from the UCSC Table Browser, or BioMart in this case (Figure 4). From the output of fuzznuc, we could expect a similar regulation for human MMP9, as we can find the Runx2 binding motif upstream of the human gene.

Galaxy offers users without large compute capacity the possibility of undertaking the analysis of multiple alignments (whole genome alignments are stored locally at the Galaxy site, compressed and indexed). One of Galaxy’s strengths is the ease with which new tools can be integrated—new suites of tools for massively parallel sequence data, metagenomics, and statistical genetics are growing rapidly.

Web services provide an alternative way of integrating databases and tools (e.g., Taverna [32] and the biomaRt package of BioConductor), but users require some programming awareness. Galaxy removes this requirement; it relies on Web services to interact with external data sources such as BioMart and the UCSC Table Browser, providing a structured Web interface. Behind the scenes, when dealing with user data, jobs are wrapped and run in an abstract interface, to ensure reproducibility and avoid any problems associated with changes in the underlying Web services. For a more detailed description of Web services, see [33].

Conclusions

We have seen how to go beyond simple browsing of data with data mining tools leveraging the BioMart system from different platforms, e.g., BioConductor (biomaRt), to find the association between microarray probe and Ensembl gene sets. Galaxy allowed us to use BioMart to extract information from Ensembl to identify some short motifs (beyond the threshold of BLAST detection) in the promoter region of a gene (MMP9). The Web interface of BioMart supports complex queries joining different datasets.

Accession Numbers Used in the Text

GO IDs: GO:0005975
EMBL: AF003944, U67136, J04597, AF281635, BC090354, AF281635, BC090354, U12309, M14105, M14104, BC089159, D84450, BC611719, D90404, M26855, M26854, L00320, L00313, L00314, L00315, L00316, L00317, L00318, L00319, M11251, J00719, M37134, M26855, M26854, J00728, J00729, J00720, J00721, J00722, J00723, J00724, J00725, J00726, M34452, K00996, K01626, K01721, D00250, M13234, M13650, M26853, M19972, X63545, X12355, D63378, BC062393.

Ensembl Transcript IDs: ENSRNOT00000014152, ENSRNOT00000018050, ENSRNOT00000015476, ENSRNOT00000015723, ENSRNOT00000028196, ENSRNOT00000022342, ENSRNOT00000047540, ENSRNOT00000008416, ENSRNOT00000034845, ENSRNOT00000014785, ENSRNOT000001580
References

1. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An overview of Ensembl. Genome Research 14: 1000–1010.

2. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. Nucleic Acids Res 35: D610–D617.

3. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, et al. (2007) The UCSC Genome Browser database: Update 2007. Nucleic Acids Res 35: D668–D673.

4. Karolchik D, Hinrichs AS, Jaerisch Y, R doomed BJ, Caw Ji- et al. (2004) Galaxy: A platform for interactive large-scale genome analysis. Genome Research 15: 1451–1454.

5. de la Cruz N, Bromberg S, Pasko D, de la Cruz N, Bromberg S, Pasko D, et al. (2005) The rat genome database (RGD): Developments towards a phenotype database. Nucleic Acids Research 33: D463–D491.

6. Twigger SN, Shinoyama M, Bromberg S, Kwitek AE, Jacob HJ, et al. (2007) The Rat Genome Database, update 2007—Easing the path from disease to data and back again. Nucleic Acids Research 35: D26–D31.

7. Chen N, Harris TW, Antoshechkin I, Bastiani C, Twigger SN, Shimoyama M, Bromberg S, et al. (2005) The vertebrate genome annotation (VEGA) database. Nucleic Acids Res 33: D459–D465.

8. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: The NCBI database of single nucleotide polymorphisms. Nucleic Acids Research 29: 308–311.

9. Jiaiwa H, Gentleman RC, Carey V, Huber W, Irizarry R, eds (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor Supplier.

10. Vaistrak I, D’Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) Reactome: A knowledge base of biologic pathways and processes. Genome Biology 8: R39.

11. Gentleman RC, Carey V, Huber W, Irizarry R, eds (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor Supplier.

12. Kasprzyk A, Keefe D, Smedley D, London D, Trumbower H, Thomas DJ, et al. (2007) Ensembl gene IDs: ENSRNOG00000017737, ENSG0000000100985.

13. Twigger SN, Shimoyama M, Bromberg S, Twigger SN, Shimoyama M, Bromberg S, et al. (2007) The vertebrate genome annotation (VEGA) database. Nucleic Acids Res 35: D668–D673.

Acknowledgments

We gratefully acknowledge Arak Kasprzyk, Damian Smedley, and Giulietta Spudich for their useful comments on the manuscript, as well as James Taylor (Galaxy) and Steffen Durinck (BioConductor) for their help. We also want to thank the Ensembl and BioMart teams.