pubmed2ensembl: Linking Publications and Genes

Joachim Baran[1], Martin Gerner[1], Maximilian Haeussler[1], Goran Nenadic[2] and Casey M. Bergman[1]
[1] Faculty of Life Sciences, University of Manchester, Manchester, UK
[2] School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, Manchester, UK

Introduction

To overcome the lack of integration between genomic data and biological literature, we have developed an extension to the BioMart data-management system called pubmed2ensembl that links over 2,000,000 articles in Pubmed to Ensembl genes for 50 species. We exploit several sources of curated (e.g. Entrez Gene) and automatically generated (e.g. gene name recognition in MEDLINE records, BLAST of EMBL records) gene-publication information, allowing users to filter and combine different data sources to suit their needs for information extraction and biological discovery. In addition to extending the Ensembl BioMart database to include information on publications, we also implemented a novel BioMart interface that allows text-based search queries on Pubmed abstracts to be performed in conjunction with queries on genomic features (Figure 2). By allowing biologists to find the relevant literature on specific regions of the genome or set of functionally related genes more easily, pubmed2ensembl offers a much-needed genome informatics inspired solution to accessing the ever-increasing biomedical literature.

Data Source	Articles	Genes
Entrez Gene	469,872	102,415
MEDLINE	1,867,773	36,310
PMC	102,406	26,008
EMBL BLAST	69,764	64,335
EMBL XREF	28,982	82,940
text2genome	9,128	11,560
	2,093,067	148,019

Table 1. Number of mapped articles and genes per data source

Method

pubmed2ensembl is build on top of a customised version of the data-mining system BioMart 0.7, [1], that is tailored to fit the demands of researches who wish to query our database for links between genes and articles. BioMart provides tools to create and build an optimised data-mining system out of one or more propriety databases. The resulting system's database, a.k.a. "mart", can then be queried through the various interfaces of BioMart, such as BioMart's

- MartView: an interactive web-interface for constructing database queries and to display query results
- MartExplorer: an application for accessing marts
- MartShell: a command line interface to query marts
- MartService: a RESTful interface for programmatic mart access
- build-in DAS server
- etc.

We have extended MartView and DAS server to permit full-text searches within Pubmed/PubMed Central via Entrez Utilities (eUtils) and to export gene-related publication information as DAS-track, respectively. Even though MartExplorer, MartShell, MartService and BioMart's other interfaces are not tailored or customised for improved accessibility, it is still possible to query our imported publication data-sources via these interfaces.

Mart Creation

We have created and implemented a scripting language, named MartScript, that allows for the automated creation of marts based on the Ensembl database structure. MartScript is an imperative programming language without branching code, i.e. if-then-else constructs are intentionally not supported, and its syntax resembles written English to further simplify the scripting language.

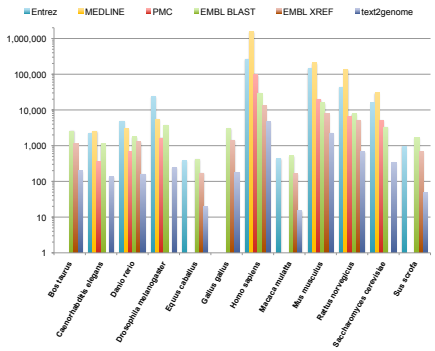

Figure 1. Number of mapped articles per data source for 12 model organisms of the 50 available organisms

Data Sources

We extended WTSI/EBI's Ensembl mart (version 56) that contains gene-centric genomic information for 50 species with 6 additional data sources that link Ensembl genes and PMIDs of articles that are obtained by importing Entrez Gene mappings between genes and articles, named entity recognition runs over MEDLINE and PMC, information extraction from EMBL records with subsequent blasting of relevant sequences against Ensembl genes, importing of mappings from the text2genome project.

Entrez Gene

Entrez Gene is an information source about genes that appear as part of the Reference Sequence collection (RefSeq), genes that are part of certain recognised genome databases that annotate genes, genes that are reported by the NCBI Genome Annotation Pipeline and genes that are identified as known genes when sequencing new model organisms [2].

MEDLINE and PubMed Central

We used a named entity recognition approach to find gene mentions in MEDLINE abstracts and full-text PMC articles, where the gene-recognition tool GNAT, [3], was used in conjunction with the species-recognition tool LINNAEUS, [4], to determine gene mentions within the corpus. GNAT/LINNAEUS produced mappings between genes and articles for only 8 species within Ensembl, but it is our largest source of article references.

EMBL Nucleotide Sequence Database

EMBL comprises of records of annotated nucleotide sequences, with each record typically addressing the origin of the sequence whilst providing several link-outs to external databases that contain further information about the record. We utilise EMBL to establish links between publications and genes in two ways: first, explicitly mentioned articles and Ensembl gene link-outs are extracted from those records where they are present, and second, we align the sequences of records with article mentionings against Ensembl genes to establish a gene/article mapping computationally.

We improve the quality of the sequence-alignment data in two steps: first, we remove articles that appear in 100 or more EMBL records in order to increase the specificity between and gene matchings, and second, we remove articles whose EMBL annotation refers to the regular expressions 'mir-*', 'mitochond*', 'tRNA', 'ribosom*' or 'plasmid*'.

text2genome

We also established links between articles and genes by integrating information generated by the text2genome project, which localises and identifies nucleotide sequences in articles [5].

Results

In total we have mapped 2,093,067 articles to 148,019 Ensembl genes by forging six additional data sources into a customised Ensembl mart. Table 1 summaries the number of linked articles and genes for all our data sources that we provide respectively. Figure 1 shows the number of mapped articles per data source for a dozen selected species.

The largest data source in terms of linked articles is MEDLINE, for which 1,867,773 articles have been mapped to 36,310 Ensembl genes. These are more articles than all our other data sources provide alltogether.

In terms of most genes that are linked to articles now, the data sources Entrez, EMBL XREF and EMBL BLAST perform best, with 102,415; 82,940 and 64,335 linked genes respectively. In total, 148,019 genes out of the 1,090,540 genes in Ensembl were linked to articles (13.6%).

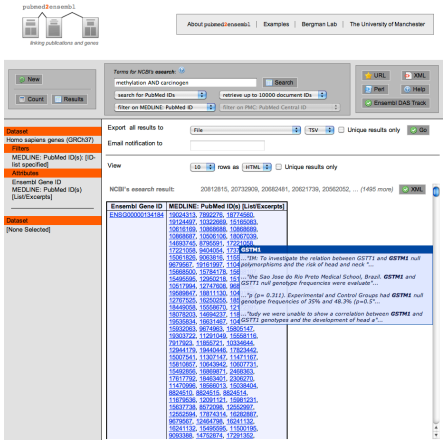

Figure 2. Customised BioMart interface for browsing and querying our extended Ensembl mart

References

1. D. Smedley, *et.al.*, **BioMart – biological queries made easy**. BMC Genomics, volume 10, 2009
2. D.Maglott, *et.al.*, **Entrez Gene: gene-centered information at NCBI**. Nucleic Acids Research, volume 35, 2007
3. J. Hakenberg, *et.al.*, **Inter-species normaliza-tion of gene mentions with GNAT**. Bioinforma-tics, volume 24, 2008
4. M. Gerner, *et.al.*, **LINNAEUS: A species name identification system for biomedical literature**. BMC Informatics, volume 11, 2010
5. M. Haeussler, *et.al.*, **Annotating genes and genomes with DNA sequences extracted from biomedical articles**, submitted

Acknowledgements

We like to thank Arek Kasprzyk and Syed Haider of the Ontario Institute for Cancer Research for their support and attention received regarding the BioMart software. We are also grateful that Rhoda Kinsella at EBI was so kind to provide us with the BioMart-XML configuration files for building an Ensembl mart as well as helping us to resolve problems that occasionally occurred with mart builds.

Links

pubmed2ensembl
`www.pubmed2ensembl.org`

BioMart 0.7 plus extras
`github.com/joejimbo/biomart-plus-extras`

MartScript
`github.com/joejimbo/MartScript`