SOFTWARE TOOL ARTICLE

Biobtree: A tool to search and map bioinformatics identifiers and special keywords [version 2; peer review: 2 approved with reservations]

Previously titled: Biobtree: A tool to search, map and visualize bioinformatics identifiers and special keywords

Tamer Gur
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD Cambridge, UK

Abstract
Biobtree is a bioinformatics tool to search and map bioinformatics datasets via identifiers or special keywords such as species name. It processes large bioinformatics datasets using a specialized MapReduce-based solution with optimum computational and storage resource usage. It provides uniform and B+ tree-based database output, web interface, web services and allows performing chain mapping queries between datasets. It is managed by a single binary file without requiring installation, specific technical knowledge or extra maintenance. Biobtree is open source and available at GitHub.

Keywords
bioinformatics, identifiers, search, mapping, visualization

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the EMBL-EBI collection.
Corresponding author: Tamer Gur (tgur@ebi.ac.uk)

Author roles: Gur T. Conceptualization, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2019 Gur T. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Gur T. Biobtree: A tool to search and map bioinformatics identifiers and special keywords [version 2; peer review: 2 approved with reservations] F1000Research 2019, 8:145 (https://doi.org/10.12688/f1000research.17927.2)

First published: 04 Feb 2019, 8:145 (https://doi.org/10.12688/f1000research.17927.1)
Amendments from Version 1

In this new version, based on reviewers’ comments article has been reconstructed and tool has been improved by integrating new datasets and adding new major feature for performing mapping queries via web interface and web services. No new figures were added, and many were removed leaving only figure one.

Any further responses from the reviewers can be found at the end of the article

Introduction

Mapping bioinformatics datasets through a web interface or programmatically via identifiers or specific keywords and attributes such as gene name, gene location, protein accessions and species name is a common need during genomics research. These mappings play an essential role in molecular data integration (Huang et al., 2011) and allow the gathering of maximum biological insight (Mudunuri et al., 2009) for these diverse bioinformatics datasets.

There are several existing tools for these mapping needs; these tools are gene-centric, protein-centric or can provide both gene- and protein-centric solutions. One of the common gene-centric tools is BioMart (Zhang et al., 2011)-based tools such as Ensembl BioMarts (Kinsella et al., 2011) which covers Ensembl (Zerbino et al., 2018) and Ensembl Genomes (Kersey et al., 2018) datasets. The R programming language package biomartRt (Durinck et al., 2009) is also widely used via performing queries with BioMart-based tools. Other common gene-centric tools are MyGene.info (Xin et al., 2016), DAVID (Huang da et al., 2009) and g:Profiler (Raudvere et al., 2019). Uniprot ID mapping service (Huang et al., 2011) provides a protein-centric solution. bioDBnet (Mudunuri et al., 2009) and BridgeDb (van Iersel et al., 2010) provide services for both gene- and protein-centric solutions.

On the other hand, genomics data size is increasing continuously (Langmead & Nellore, 2018) especially via high throughput sequencing, so performing these mapping on these expanding data sizes in local computers, cloud computing or existing computing environments in a rapid and effective way via tools with easy installation and requiring minimum maintenance is a challenge (Marx, 2013).

The referenced existing gene-centric tools currently do not support large Ensembl Bacteria genomes. Existing tools either provide only online service or require specific technical knowledge such as a particular database or specific programming language to install, use and adapt to different computational environments such as a local computer. Another limitation of the referenced tools is that they provide one-dimensional filtering capability in a single mapping query.

Biobtree address these problems of existing tools. First, it can be used via a single executable file without requiring installation, specific technical knowledge or extra maintenance such as database administration. To process large datasets, it uses a specialized MapReduce-based solution which is discussed in the next section. MapReduce is an effective way to deal with large datasets (Langmead & Nellore, 2018). After processing data, Biobtree provides a web interface, web services and chain mapping and filtering query capability in a single query with its intuitive query syntax which is demonstrated in the use cases section. Biobtree covers a range of bioinformatics datasets including Ensembl Bacteria genomes. The data resources currently used are ChEBI (Hastings et al., 2016), HGNC (Braschi et al., 2019), HMDB (Wishart et al., 2018), InterPro (Mitchell et al., 2019), Europe PMC (Europe PMC Consortium, 2015), UniProt (UniProt Consortium, 2019), Chembl (Gaulton et al., 2017), Gene Ontology (The Gene Ontology Consortium, 2019), EFO (Malone et al., 2010), ECO (Giglio et al., 2019), Ensembl (Zerbino et al., 2018) and Ensembl Genomes (Kersey et al., 2018). Table 1 shows details of these datasets.

Methods

Implementation

The Biobtree implementation process starts by retrieving selected datasets as shown in Table 1 and retrieving data entries belonging to these datasets with their attributes and mapping information from their public locations, which are also shown in Table 1. During this data retrieval, the whole of the data do not get stored and uncompressed on the disc, instead data are retrieved and uncompressed in a streaming manner in the memory, which allows avoiding excessive disc space usage. Necessary data, which are these mapping and attributes, are compactly stored as chunks on the disc. During these data retrievals, all the idle CPUs have been utilized to merge and sort these chunks recursively with each other. It is essential that the produced files are sorted to make fast batch inserts to the LMDB database which Biobtree uses as a database to store its result data. Once the data retrieval process is completed, result chunk files are globally merged using the patience sort technique and inserted into the LMDB database as key and values. Keys consist of identifiers and special keywords such as gene names or species name, and values are attributes and mapped datasets information. In these processes, data retrieval and creation of sorted chunks represent the map phase, global merge of the chunks and database creation represent the reduce phase of the MapReduce solution. Once the database is created, the Biobtree web module provides a web interface and web services to perform both searching for identifiers and mapping queries. Mapping queries has been done with a query syntax which allows chains of mapping and filtering between datasets. An example use case with this syntax is demonstrated in the next section. Biobtree uses a B+ tree data structure-based LMDB key-value store. LMDB provides fast batch inserts and reads which fits the bioinformatics datasets update cycle well where datasets are often updated periodically, and then only intensive read operations are performed. LMDB is embedded into Biobtree’s executable binary code so it does not require a separate installation or special maintenance.

Use cases

This section covers three use cases. Each use case consists of two or three inputs, a command for how to run them on a local
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Location</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEBI</td>
<td>ChEBI reference accession data</td>
<td>ftp://ebi.ac.uk/chebi/Flat_file_delimited/</td>
<td>TSV</td>
</tr>
<tr>
<td>HGNC</td>
<td>Human gene nomenclature</td>
<td>ftp://ftp.ebi.ac.uk/genenames/new/json/</td>
<td>JSON</td>
</tr>
<tr>
<td>InterPro</td>
<td>Protein Families</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/interpro/current/</td>
<td>XML</td>
</tr>
<tr>
<td>Literature</td>
<td>Literature mappings</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/literature/</td>
<td>CSV</td>
</tr>
<tr>
<td>UniProt</td>
<td>UniProt Knowledgebase reviewed</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/uniref90/</td>
<td>XML</td>
</tr>
<tr>
<td>Uniref50</td>
<td>UniProt sequence clusters</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/uniref50/</td>
<td>XML</td>
</tr>
<tr>
<td>Uniref90</td>
<td>UniProt sequence clusters</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/uniref90/</td>
<td>XML</td>
</tr>
<tr>
<td>Uniref100</td>
<td>UniProt sequence clusters</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/uniref100/</td>
<td>XML</td>
</tr>
<tr>
<td>GO</td>
<td>GO</td>
<td><a href="http://www.geneontology.org/GO.owl">http://www.geneontology.org/GO.owl</a></td>
<td>RDF/XML</td>
</tr>
<tr>
<td>ECO</td>
<td>ECO</td>
<td><a href="http://purl.obolibrary.org/obo/eco.owl">http://purl.obolibrary.org/obo/eco.owl</a></td>
<td>RDF/XML</td>
</tr>
<tr>
<td>EFO</td>
<td>EFO</td>
<td><a href="http://www.ebi.ac.uk/efo.owl">http://www.ebi.ac.uk/efo.owl</a></td>
<td>RDF/XML</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>Chemical database of bioactive molecules</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBL-RDF/latest/</td>
<td>RDF/XML</td>
</tr>
</tbody>
</table>
computer and the expected output after performing the use case. The first input is called ‘terms’, which contains a set of identifiers or special keywords such as species name separated by commas. The mapping process starts based on records belonging to these terms. The second input is a mapping query which consists of chains of map and filter function queries. The map function query takes a single dataset name as an argument and the filter function gets a Boolean query based on given source datasets attributes to filter out mappings. Datasets’ attributes are mostly available for source datasets shown in Table 1 and can be explored via the web interface by searching identifiers. The third optional input is dataset to filter out terms which occurred in different dataset with same value. To perform use cases, a given command needs to be run from a terminal via a Biobtree binary file which can be downloaded from the GitHub page. Once a command runs, it processes related data and the Biobtree web interface opens and the use case can be performed via the web interface. For use cases 1 and 2, because it is the same dataset, it is enough to run the command once. For use case 3, the commands need to be run again to process specific bacteria datasets instead of the default human genome dataset.

**usecase-1 Map Affymetrix identifiers to Ensembl human genome identifiers and then map these to the molecular function type GO terms**

**Command**  
biobtree start  
**Terms**  
202763_at,209310_s_at  
**Mapping query**  
map(transcript).map(ensembl).filter(ensembl.genome=="homo_sapiens").map(go).filter(go.type=="molecular_function")  
**Dataset**  
affy_hg_u133_plus_2  
**Output**  
Query results 21 GO terms. The terms which maps to identifier 202763_at are GO:00020200, GO:0004190, GO:0004197, GO:0004861, GO:0005123, GO:0005515, GO:0008233, GO:0008234, GO:0016005, GO:0016787, GO:0044877, GO:097153, GO:097199 and GO:097200. And the terms which maps to identifier 209310_s_at are GO:0004197, GO:0005155, GO:0008233, GO:0008234, GO:0016787, GO:0050700 and GO:097199.

**usecase-2 Map human Ensembl identifiers with given genome location to the reviewed Uniprot identifiers**

**Command**  
biobtree start  
**Term**  
homo_sapiens  
**Mapping query**  
map(ensembl).filter(ensembl.start>10000000 && ensembl.end<10100000 && ensembl.seq_region_name=="*X" ).map(uniprot).filter(uniprot.reviewed)  
**Output**  
Query results 9 Uniprot reviewed proteins. These protein identifiers are Q43657, Q9H2S6, P33240, O66087, Q96C24, Q8TAB3, Q5H913, Q6PP77 and Q9Y5S8.

**usecase-3 Map all taxonomic children of given bacteria and then map these children to Ensembl with given genome location and contains a given word**

**Command**  
biobtree -d +ensembl_bacteria -sp "serovar_virchow" start  
**Term**  
Salmonella enterica subsp. enterica  
**Mapping query**  
map(txchild).map(ensembl).filter(ensembl.start<10000 && ensembl.description.contains("SopD"))  
**Output**  
Query results 7 Ensembl Genomes Bacteria genes. These gene identifiers are AEW14_05145, AEW14_15935, ACH54_23895, ACH56_04205, DE27_21250, DE87_06330 and LPMST02_21800.

**Discussion**

A mapping between bioinformatics datasets via identifiers or special keywords such as species names is often performed during genomic analyses and plays an essential role in molecular data integration and getting maximum biological insight from these datasets. There are several gene-centric, protein-centric and both protein- and gene-centric tools for addressing these mapping needs. These tools currently do not support the large Ensembl Genomes Bacteria dataset. In addition, these tools provide either only online services or require specific technical knowledge to install and adapt to new computing environments. Existing tools also provide one-dimensional filtering in a single mapping query. Biobtree addresses these problems by managing a tool with a single executable file without requiring specific technical knowledge and processing large datasets with its specialized MapReduce-based solution. Based on processed data, it creates a uniform database and allows searching identifiers and chain mappings and filtering queries with its web interface and web services.

**Future work**

More datasets can be integrated into the existing system such as gene expression. In addition, following and experimenting with the advancements in large data processing techniques, databases and data structures fields to improve the tool further.

**Data availability**

All data underlying the results are available as part of the article and no additional source data are required.

**Software availability**

All source codes and binaries available at: https://www.github.com/tamerh/biobtree.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2547047

License: BSD 3-Clause “New” or “Revised” license.
References


Open Peer Review

Current Peer Review Status: ?

Version 2

Reviewer Report 27 September 2019

https://doi.org/10.5256/f1000research.22620.r53984

© 2019 Shokhirev M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Maxim N. Shokhirev
Razavi Newman Integrative Genomics and Bioinformatics Core, Salk Institute for Biological Studies, La Jolla, CA, USA

The author now includes references to previously published annotation tools and services such as BioMart and MyGene.info and points out several limitations of the extant tools that are addressed with Biobtree. In addition, the author now includes an online portal for exploring data with a structured query language along with a few examples of specific use cases and a table of databases implemented.

Given the improvements, the work can now be beneficial for biologists interested in mapping and annotation of specific genes or terms, which can serve as an alternative to already established annotation databases and services.

However, there are still several limitations to this work. First, supposedly one of the main advantages of Biobtree is that it can be run as a standalone executable tool. This is great in theory but there doesn't seem to be any executables in the referenced repository or an explanation of how to install the Biobtree in various environments. The manuscript goes into some detail about the underlying algorithms and datastructures, but there is no explanation of what is required for you to install the program or get it working in either the manuscript or the repository. Therefore, without documentation or description the only way to use the tool currently is through the online portal. The second major limitation is that while there are three use-cases described, there is no documentation provided for how one would build arbitrary queries. A manual or easier to use interface is needed before the full power of the tool can be leveraged. Finally, I couldn't find a way to download the results from the online portal. There is often hundreds of pages of results but it seems there is no way to obtain a table of the results for use in downstream tools or analyses. Given the fact that there is no description of how to install the tool or which prerequisites are needed, and the fact that you can't download the results from the online portal, this tool is currently only useful for manual exploration of the mappings and not suitable for inclusion in bioinformatics pipelines. Since it seems the author intended for the latter, instructions for how to install the tool, a detailed description of the API or commands, and a way to save the output to a table of some kind is still needed. I look forward to seeing the revised manuscript.

Competing Interests: No competing interests were disclosed.
Reviewer Expertise: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 15 April 2019

https://doi.org/10.5256/f1000research.19605.r46335

© 2019 Lampa S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Samuel Lampa
Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

The article describes a commandline tool, Biobtree, that is claimed to allow to process relations between bioinformatics datasets based on various characteristics such as identifiers and keywords.

The manuscript describes the tool in a clear way technically, making it quite clear what it does in technical terms, and how it is supposed to be used.

Also, I was able to install and run the tool in a simple way on my laptop (i5 CPU, 8GB RAM and 10-15 GB free hard drive, Xubuntu 16.04 64 bit) without problems. It provides a simple but good looking and easy to use web interface.

I'm seeing at least two major issues with the tool and manuscript though, that needs being thoroughly addressed to make them acceptable.

Main problem 1: Visualization?

Firstly, the title claims that the tool does visualization of the database produced by the tool. Perhaps I'm missing something, but I have not found any visualization in the tool apart from a form of search hit result listings. I don't think this is enough to be called "visualization". Especially as it is unclear how the current form of output is supposed to be used in a concrete biological usecase. With the current wording, I would expect something more graphical, like a graphviz-like graph view of dataset relations.

Suggested edits to make the tool and paper acceptable:

- Provide graphical visualization beyond results listings (or explain how to show them, if I have missed them), or else remove "visualization" from the title and other places.
- Use this/these visualizations in the use cases/demonstrators discussed above, to explain how they contribute to solving concrete biological problems.
Main problem 2: Lack of context and discussion of biological relevance

The first and main problem with the manuscript is that it does not provide a clear enough description of what biological problem it is solving. Nor does it provide an overview of existing tools and solutions in this field. Right now, the manuscript only states what the tool can do in technical terms. It somehow reads like a (well written) user guide or README file, but not yet a scientific paper. To help potential new users understand why they might need this tool, it needs to be put in context and compared with other existing tools.

In my view, the manuscript needs the following points thoroughly addressing to be acceptable:

1. In the introduction: Elaborate on the field of mapping/visualising dataset relations, mentioning relevant existing similar tools, what are the typical problems, and what particular problem Biobtree solves.
2. Explain a few examples of biological problems that can be solved with this tool, or type of tool.
3. E.g. in the results: Provide at least one, and optimally two or three potentially simple, but relevant, biological demonstrators or use cases, that can be addressed with the tool. Provide complete instructions on how to re-run this or these demo(s) and provide outputs for this/these in terms of figures or diagrams and how these were produced. In this way, both reviewers and users can make sure that they understand how to operate the tool.
4. In the discussion: Connect back to the explained problem the tool is addressing, and explain how the problem was solved, again reinstating the relevance of this specific tool compared to other existing tools, and what improvement it provides to the end user trying to solve biological problems, exemplified by the demonstrators or use cases.

Language issues

The manuscript also contains quite a number of language issues. I'm listing a few language suggestions below as examples, but further language proofing or editing is highly recommended, to make sure there are not more of these:

1. **Methods** section:
   "in GO programming language" --> "in the Go programming language"
   (Note the "the" and that only G is uppercase in "Go").
2. **Update phase** section:
   "to LMDB" --> "to the LMDB"
3. **Update phase** section:
   "Updating reads selected datasets as a stream"
   I don't understand this sentence. Please language-check it.
4. "is used in next" --> "is used in the next"
5. **Generate phase** section:
   "the project github page" --> "the project's GitHub page"
6. **Web phase** section:
   "starts web phase" --> "starts the web phase"
7. **Web interface** section:
   "The Web interface allow user" --> "The web interface allows the user"
8. **Operation** section:
   "Biobtree executable" --> "The biobtree executable".

Is the rationale for developing the new software tool clearly explained?
Partly
Is the description of the software tool technically sound?  
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?  
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?  
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?  
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Scientific workflow tools, Cheminformatics, Semantic web.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 07 March 2019

https://doi.org/10.5256/f1000research.19605.r45074

© 2019 Shokhirev M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Maxim N. Shokhirev  
Razavi Newman Integrative Genomics and Bioinformatics Core, Salk Institute for Biological Studies, La Jolla, CA, USA

While it is important to create a consistent and queryable database of biological identifiers, it is unclear what advances this tool brings to the field. For example, how does this tool compare to other queryable database tools such as mygene.info, or BioMart? The paper will greatly benefit from a comparison to these and other such tools.

I downloaded and ran the tool but it seems I can't get through the update phase when I run ./biobtree update (It seems to hang after uniprot_reviewed finishes) without any other messages. When I rerun using biobtree --d uniprot_reviewed update it finishes but there is an error:

Error while reading file-> ./out/index/0_13.938476000.gz
panic: gzip: invalid header
I tried running generate and web after that regardless, but couldn't get it to work:

```
panic: mdb_txn_commit: MDB_BAD_TXN: Transaction must abort, has a child, or is invalid
```

Error while reading meta information file which should be produced with generate command. Please make sure you did previous steps correctly.

The author needs to debug/test their code to ensure that it can be used by others.

**Is the rationale for developing the new software tool clearly explained?**
Partly

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
No

*Competing Interests*: No competing interests were disclosed.

**Reviewer Expertise**: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

---

**Author Response 10 Mar 2019**

**Tamer Gur**, EMBL European Bioinformatics Institute, UK

Thank you for reviewing the article. I agree that there are several similar tools exist with different dataset and functionalities such as Biomart and mygene.info. However, this tool can still complement them for following main reasons.

- Biobtree can work in local machine. This can be especially useful when large number of requests needs to be performed. For instance currently similar Uniprot tool documentation suggests either split the requests or download underlying data when number of requests are above 50K. These types of limitations for bulk requests are sensible for fair usage of a public service and can be more suitable with Biobtree type locally runnable tool.
- Users custom dataset can be integrated.
- Tool provides new intuitive web interface.
In relation to reported errors, I have added demo of tool in case such errors happen again. I have also added integration test which runs periodically on Linux, MacOS and Windows operating systems via Azure DevOps platform. These tests can be accessed publicly and test and demo links can be found at github page. Based on these tests, it seems that tool is working as expected. I believe that hanged process is a specific issue or bug which I am happy to resolve if I have more information. The rest of problems which have been reported are most probably due to the prematurely exited hanged process. Either starting in a new folder or passing --clean parameter can solve the issue.

**Competing Interests:** No competing interests were disclosed.