Showing Figures and Captions in the BioText Journal Search Engine

Marti A. Hearst*, Michael A. Wooldridge †, Jerry Ye* and Anna Divoli*

*School of Information, University of California, Berkeley, CA 94720, †California Digital Library, Oakland, CA 94612

Summary

The BioText Search Engine is a freely available web-based application that provides biologists with new ways to access the scientific literature.

One novel feature is the ability to search and browse article figures and their captions.

Motivation

All prominent literature search systems are based on information from abstracts alone when:

- Researchers in the area of text-mining have started to investigate approaches for full-text analysis.
- PubMed Central has made available a large collection of full-text articles (Open Access), overcoming licensing restrictions from the publishers.
- Figure captions and figures can be especially useful for locating experimental results.

The Views

The BioText Search Engine allows users to search in ***"Abstracts - list view"***, ***"Captions - list view"*** and ***"Captions - grid view"***. The interface is carefully designed according to usability principles and techniques.

Clicking on any figure opens a new window with a **large version** of the figure accompanied with its caption.

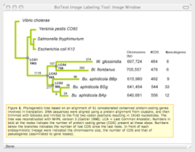
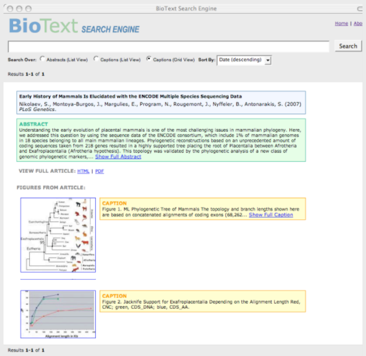

BioText Image Labeling Tool: Image Window
Vibrio cholerae
Yersinia pestis CO92
Salmonella thyphimurium
Escherichia coli K12
LCA1 1983
LCA2 865
LCA3 588
LCA4 758
1118
107
274
282
96
277
38
24
Chromosome (bp)
#CDS
#pseudogenes
W. glossinidia 697,724 484 8
Bl. floridanus 705,557 476 6
Bu. aphidicola BBp 615,980 492 9
Bu. aphidicola BSg 641,454 544 33
Bu. aphidicola BAp 640,681 556 12
Done

Figure 1. Phylogenetic tree based on an alignment of 61 concatenated conserved protein-coding genes involved in translation. DNA sequences were aligned using a protein alignment from clustalw, and then trimmed with Gblocks and limited to the first two codon positions resulting in 19143 nucleotides. The tree was reconstructed with NHML version 3 (Galtier 1998). LCA = Last Common Ancestor; Numbers in bold at the nodes indicate the number of protein coding genes (CDS) present at these steps. Numbers below the branches indicates the number of lost CDS since the last node. In front of each endosymbiotic lineage were indicated the chromosome size, the number of CDS and that of pseudogenes (assimilated to gene losses).

All views lead to the ***"Endgame view"***. *"Endgame view"* displays a summary of a paper that an abstract or caption originates. This summary comprises of the *TITLE, CITATION, ABSTRACT*, all *FIGURES and corresponding CAPTIONS* from the paper in a list and HTML & PDF links to the paper.


BioText Search Engine
BioText SEARCH ENGINE
Home | Abo
Search
Search Over: ○ Abstracts (List View) ○ Captions (List View) ● Captions (Grid View) Sort By: Date (descending)

Results 1-1 of 1

Early History of Mammals Is Elucidated with the ENCODE Multiple Species Sequencing Data
Nikolaev, S., Montoya-Burgos, J., Margulies, E., Program, N., Rougemont, J., Nyffeler, B., Antonarakis, S. (2007) *PLoS Genetics*.

ABSTRACT
Understanding the early evolution of placental mammals is one of the most challenging issues in mammalian phylogeny. Here, we addressed this question by using the sequence data of the ENCODE consortium, which include 1% of mammalian genomes in 18 species belonging to all main mammalian lineages. Phylogenetic reconstructions based on an unprecedented amount of coding sequences taken from 218 genes resulted in a highly supported tree placing the root of Placentalia between Afrotheria and Exafroplacentalia (Afrotheria hypothesis). This topology was validated by the phylogenetic analysis of a new class of genomic phylogenetic markers,... Show Full Abstract

VIEW FULL ARTICLE: HTML | PDF

FIGURES FROM ARTICLE:

CAPTION
Figure 1. ML Phylogenetic Tree of Mammals The topology and branch lengths shown here are based on concatenated alignments of coding exons (68,262... Show Full Caption

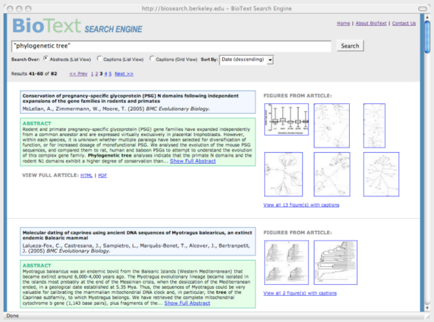
CAPTION
Figure 2. Jacknife Support for Exafroplacentalia Depending on the Alignment Length Red, CNC; green, CDS_DNA; blue, CDS_AA.

Results 1-1 of 1

Abstracts - list view

Searches in *TITLES, ABSTRACTS, AUTHOR NAMES*

Returns in a list *TITLE, CITATION, ABSTRACT*, thumbnails of *FIGURES* from the paper, HTML & PDF links to the paper and link *to the "Endgame view"*


http://biosearch.berkeley.edu – BioText Search Engine
BioText SEARCH ENGINE
Home | About BioText | Contact Us
"phylogenetic tree"
Search
Search Over: ● Abstracts (List View) ○ Captions (List View) ○ Captions (Grid View) Sort By: Date (descending)
<< Prev 1 2 3 4 5 Next >>
FIGURES FROM ARTICLE:
View all 13 figure(s) with captions
FIGURES FROM ARTICLE:
View all 2 figure(s) with captions
Done

Results 41-60 of 82

Conservation of pregnancy-specific glycoprotein (PSG) N domains following independent expansions of the gene families in rodents and primates
McLellan, A., Zimmermann, W., Moore, T. (2005) *BMC Evolutionary Biology*.

ABSTRACT
Rodent and primate pregnancy-specific glycoprotein (PSG) gene families have expanded independently from a common ancestor and are expressed virtually exclusively in placental trophoblasts. However, within each species, it is unknown whether multiple paralogs have been selected for diversification of function, or for increased dosage of monofunctional PSG. We analysed the evolution of the mouse PSG sequences, and compared them to rat, human and baboon PSGs to attempt to understand the evolution of this complex gene family. **Phylogenetic tree** analyses indicate that the primate N domains and the rodent N1 domains exhibit a higher degree of conservation than... Show Full Abstract

VIEW FULL ARTICLE: HTML | PDF

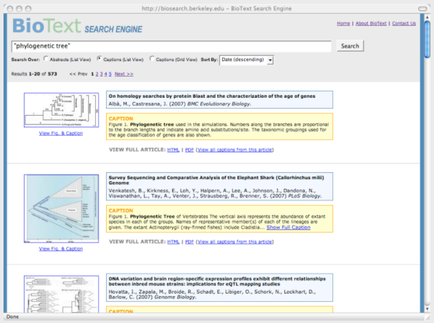
Molecular dating of caprines using ancient DNA sequences of Myotragus balearicus, an extinct endemic Balearic mammal
Lalueza-Fox, C., Castresana, J., Sampietro, L., Marquès-Bonet, T., Alcover, J., Bertranpetit, J. (2005) *BMC Evolutionary Biology*.

ABSTRACT
Myotragus balearicus was an endemic bovid from the Balearic Islands (Western Mediterranean) that became extinct around 6,000-4,000 years ago. The Myotragus evolutionary lineage became isolated in the islands most probably at the end of the Messinian crisis, when the desiccation of the Mediterranean ended, in a geological date established at 5.35 Mya. Thus, the sequences of Myotragus could be very valuable for calibrating the mammalian mitochondrial DNA clock and, in particular, the **tree** of the Caprinae subfamily, to which Myotragus belongs. We have retrieved the complete mitochondrial cytochrome b gene (1,143 base pairs), plus fragments of the... Show Full Abstract

Captions - list view

Searches in *CAPTIONS*

Returns in a list *TITLE, CITATION, CAPTION*, corresponding *FIGURE*, HTML & PDF links to the paper and link *to the "Endgame view"*


http://biosearch.berkeley.edu – BioText Search Engine
BioText SEARCH ENGINE
Home | About BioText | Contact Us
"phylogenetic tree"
Search
Search Over: ○ Abstracts (List View) ● Captions (List View) ○ Captions (Grid View) Sort By: Date (descending)
<< Prev 1 2 3 4 5 Next >>
View Fig. & Caption
View Fig. & Caption
Done

Results 1-20 of 573

On homology searches by protein Blast and the characterization of the age of genes
Albà, M., Castresana, J. (2007) *BMC Evolutionary Biology*.

CAPTION
Figure 1. **Phylogenetic tree** used in the simulations. Numbers along the branches are proportional to the branch lengths and indicate amino acid substitutions/site. The taxonomic groupings used for the age classification of genes are also shown.

VIEW FULL ARTICLE: HTML | PDF (View all captions from this article)

Survey Sequencing and Comparative Analysis of the Elephant Shark (Callorhinchus milii) Genome
Venkatesh, B., Kirkness, E., Loh, Y., Halpern, A., Lee, A., Johnson, J., Dandona, N., Viswanathan, L., Tay, A., Venter, J., Strausberg, R., Brenner, S. (2007) *PLoS Biology*.

CAPTION
Figure 1. **Phylogenetic Tree** of Vertebrates The vertical axis represents the abundance of extant species in each of the groups. Names of representative member(s) of each of the lineages are given. The extant Actinopterygii (ray-finned fishes) include Cladistia... Show Full Caption

VIEW FULL ARTICLE: HTML | PDF (View all captions from this article)

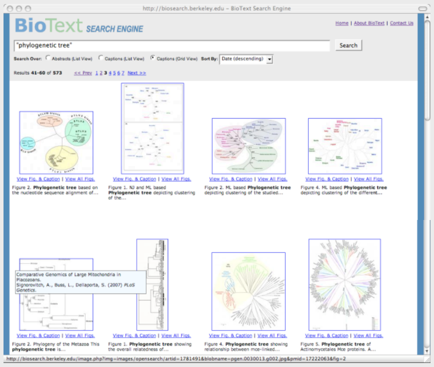
DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: implications for eQTL mapping studies
Hovatta, I., Zapala, M., Broide, R., Schadt, E., Libiger, O., Schork, N., Lockhart, D., Barlow, C. (2007) *Genome Biology*.

CAPTION

Captions - grid view

Searches in *CAPTIONS*

Returns corresponding *FIGURES* in a grid, short *CAPTION* excerpt, *CITATION* in tooltip and link *to the "Endgame view"*


http://biosearch.berkeley.edu – BioText Search Engine
BioText SEARCH ENGINE
Home | About BioText | Contact Us
"phylogenetic tree"
Search
Search Over: ○ Abstracts (List View) ○ Captions (List View) ● Captions (Grid View) Sort By: Date (descending)
<< Prev 1 2 3 4 5 6 7 Next >>
View Fig. & Caption | View All Figs.

Results 41-60 of 573

Figure 2. **Phylogenetic tree** based on the nucleotide sequence alignment of...

Figure 1. NJ and ML based **Phylogenetic tree** depicting clustering of the...

Figure 2. ML based **Phylogenetic tree** depicting clustering of the studied...

Figure 4. ML based **Phylogenetic tree** depicting clustering of the different...

Comparative Genomics of Large Mitochondria in Placozoans.
Signorovitch, A., Buss, L., Dellaporta, S. (2007) *PLoS Genetics*.

Figure 2. Phylogeny of the Metazoa This **phylogenetic tree** is...

Figure 1. **Phylogenetic tree** showing the overall relatedness of...

Figure 4. **Phylogenetic tree** showing relationship between mce-linked...

Figure 5. **Phylogenetic tree** of Actinomycetales Mce proteins. A...

http://biosearch.berkeley.edu/image.php?img=images/opensearch/artid=1781491&blobname=pgen.0030013.g002.jpg&pmid=17222063&fig=2

Technical details

The current system indexes all **Open Access** articles available at PubMed Central. This collection consists of more than 150 journals, 20,000 articles, and 80,000 figures (new articles are downloaded and indexed daily).

The **figures** are **stored locally**, in order to be able to present thumbnails quickly.

The **Lucene** open source search engine issued to index, retrieve, and rank the text (using the **default statistical ranking**).

For tokenization, the standard analysis settings for Lucene are used: words are split at punctuation characters and hyphens, unless there is an umber in the token, and uses lower-casing, simple stemming, and stop word removal.

Publication date is stored as a separate field and can also be used to **sort** the results.

The interface is **web-based** and is implemented in python and PHP. Logs and other information are stored using MySQL.

Available at: **http://biosearch.berkeley.edu**

Future work

The search engine is a work in progress. More functionality will be added over time. We plan to:

- Provide **full-text search**. Since the usability of different ranking functions for biology articles is still not well-understood, we plan to do usability testing, research how different sections should be weighted differently for different query types and investigate how best to show excerpts or summaries from full text before supporting this feature.
- Augment the caption search by indexing the **parts of the full text that refer to the caption**.
- Provide search over **table captions**.
- Incorporate topical features such as **genes/proteins** and **organisms**.
- For the grid view, we plan to provide **grouping according to categories** that are of interest to biologists, such as "sequence alignments" and "phylogenetic trees". (We are building a classifier for figures and their captions. We have developed an image annotation interface and are soliciting help with hand-labeling mated caption classifier.)

Additional future developments on the BioText Search Engine will depend on **feedback** and **requests** we receive from users, and the results of extensive **usability testing**.

Acknowledgements

This work was funded in part by NSF DBI- 0317510.

References

Marti A. Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, Preslav Nakov, Michael A. Wooldridge and Jerry Ye (2007) *"BioText Search Engine: beyond abstract search"* Bioinformatics; doi:10.1093/bioinformatics/btm301

Marti A. Hearst, Anna Divoli, Jerry Ye and Michael A. Wooldridge (2007) *"Exploring the efficacy of caption search for bioscience journal search interfaces"* ACL 2007 Workshop on BioNLP, Prague, Czech Republic

BioText SEARCH ENGINE

http://biosearch.berkeley.edu/

UC Berkeley School of Information