D780–D785 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 15 November 2006*
doi:10.1093/nar/gkl781

The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System—a one-stop gateway to online bioinformatics databases and software tools

Yi-Bu Chen*, Ansuman Chattopadhyay, Phillip Bergen, Cynthia Gadd[1] and Nancy Tannery

Health Sciences Library System, University of Pittsburgh, 200 Scaife Hall, 3550 Terrace Street, Pittsburgh, PA 15261, USA and [1]Department of Biomedical Informatics, Vanderbilt University, 2209 Garland Avenue, Nashville, TN 37232-8340, USA

Received August 14, 2006; Revised September 13, 2006; Accepted October 1, 2006

ABSTRACT

To bridge the gap between the rising information needs of biological and medical researchers and the rapidly growing number of online bioinformatics resources, we have created the Online Bioinformatics Resources Collection (OBRC) at the Health Sciences Library System (HSLS) at the University of Pittsburgh. The OBRC, containing 1542 major online bioinformatics databases and software tools, was constructed using the HSLS content management system built on the Zope® Web application server. To enhance the output of search results, we further implemented the Vivísimo Clustering Engine®, which automatically organizes the search results into categories created dynamically based on the textual information of the retrieved records. As the largest online collection of its kind and the only one with advanced search results clustering, OBRC is aimed at becoming a one-stop guided information gateway to the major bioinformatics databases and software tools on the Web. OBRC is available at the University of Pittsburgh's HSLS Web site (http://www.hsls.pitt.edu/guides/genetics/obrc).

INTRODUCTION

In the past decade, the emergence and rapid advance of genomic and proteomic technologies have generated never-before-seen amounts of genomic and proteomic data. As the genomes of 294 model organisms have been sequenced with 1206 more on the way (1), the amount of nucleotide sequence data alone nearly doubles every year. Such explosive growth of data has spawned hundreds of Web-based, publicly available bioinformatics resources, including databases and software tools, in various fields of biological sciences. The number of the online databases listed in the *Nucleic Acids Research* (*NAR*) Molecular Biology Database Collection alone has increased more than 14-fold from 58 in 1996 to 858 in 2006 (2). The majority of these newly emerged online resources are specialized databases and Web servers that provide not only sequence information, but also data on gene expression, macromolecular structures, genotype and phenotype of model organisms, as well as computational tools for analyzing macromolecular sequences/structures and global gene expression. Representing the best state of knowledge in the corresponding fields, these expert curated databases and specialized software tools may greatly assist researchers in designing their own experiments, as well as interpreting and validating their results.

Although the proliferation of bioinformatics databases is a manifestation of collective efforts by the life science community to help individual researchers coping with the phenomenal growth of biological data and information, many researchers find themselves struggling to keep up-to-date with the research in their fields (3,4). The situation is further exacerbated by the fact that locating such large numbers of online resources is anything but an easy task (5). The problem stems from the fact that the information about these online resources is scattered in various life science journals and around the Web, and that few web sites currently provide a guided access point with searchable links to a majority of these resources. Studies suggested that locating bioinformatics resources through literature searches is often very difficult (6–8). One study reported that >50% of the participating researchers use the Web to search for bioinformatics resources (9). However, searches using popular Web search engines, such as Google, are often ineffective. This is because Web search engines rank web sites by popularity rather than their relevance, and that Web search engines do not discriminate between reliable and unreliable web sites. The lack of

*To whom correspondence should be addressed. Tel: +1 412 383 6887; Fax: +1 412 648 8819; Email: yic18@pitt.edu

The authors wish it to be known that, in their opinion the first two authors should be regarded as joint First Authors standard search terms and the fact that Web search engines lump all hits together regardless of the nature of each hit, as long as they all contain the searched terms, further reduces the usefulness of the Web search engines as a mean to locate bioinformatics resources (5).

The urgent need of organizing the bioinformatics resources has recently been raised (5,10). Among the existing efforts to solve the problem are the Molecular Biology Database Collection compiled by the *NAR* (2), the Bioinformatics Links Directory (11,12), the Expasy Life Sciences Directory (http://www.expasy.org/links.html), the DBcat (13), the Database of Databases (14) and the Pathguide (15). Although these projects are highly valuable, their sole reliance on categorical content structure, limitations in annotation and coverage, and the lack of sophisticated search features may affect their usability and appeal to a wide audiences. For example, the output of search results from the Bioinformatics Links Directory is pages of a scrollable list, which may require users to examine the entire list in order to find the results relevant to their queries. There are also no ranking of the results or indications of any relationships that may exist among the results. Such limitations may pose even bigger problems as the number of the bioinformatics resources is expected to continuously grow at a rapid pace. Different approaches, such as using document clustering techniques (16) to organize search results, may enable users to quickly navigate through a large number of search results (17,18).

In order to help biomedical researchers to quickly find the most relevant bioinformatics resources for their specific information needs, we sought to develop a concrete and innovative search strategy as a part of a fledging library-based molecular biology information service at the Universtiy of Pittsburgh (19). For this purpose, we constructed the Online Bioinformatics Resources Collection (OBRC) at the Health Sciences Library System (HSLS), University of Pittsburgh. This collection currently includes 1542 online bioinformatics databases and software tools, most of which have been published by *NAR* or listed in its Molecular Biology Database Collection (2). In addition, we implemented the Vivísimo Clustering Engine® to OBRC to help users navigate through their search results.

METHODOLOGY

The new search strategy consists of two major components: a centralized collection of the curated information on major online bioinformatics databases and software tools, and the implementation of the Vivísimo Clustering Engine® to enhance the output of search results.

Source materials

The primary sources of OBRC are the databases and software tools published by the *NAR* (http://nar.oxfordjournals.org/). Specifically, the source materials were mainly the databases published in the *NAR* Annual Database Issues from 2001 to 2006, and the software tools published in the *NAR* Annual Web Server Issues from 2004 to 2006. Other databases listed in the *NAR* Molecular Biology Database Collection, including those published by *NAR* before 2001 and those not published by the *NAR*, were also selected. Selected databases and software tools described in other peer-reviewed journals, such as *Bioinformatics* and *BMC Bioinformatics*, were included in the collections. In addition, a number of unpublished but popular online software tools were also entered.

Collection construction, organization and maintenance

Information on each resource was entered using the HSLS content management system built on the Zope® Web application server. For each entry, the information for the following fields was entered: URL to the resource; name of the resource; a one-sentence description of the major functions; URL to the relevant PubMed abstract(s); last modification date of the entry; highlights of the resource; and keywords. The title, description and highlights for each entry were generated based on the PubMed abstract(s), as well as the content and scope of the resource. Together with the keywords, the textual information in these fields are automatically indexed by the Zope® Zcatalog and subsequently processed by the Zope®-based search engine.

As a major part of curation efforts, keywords were generated based on the information in the PubMed abstract(s), the MESH terms of the abstract(s), the information posted on corresponding web site, as well as the domain knowledge in molecular biology. Standard terminologies, commonly used by researchers in their publications, were used. The main types of keywords include biological concepts, entities, organism names, widely studied gene and protein names, and common molecular biology tasks. Whenever possible, common synonyms of the most important keywords were included as a conscious effort to improve the recall.

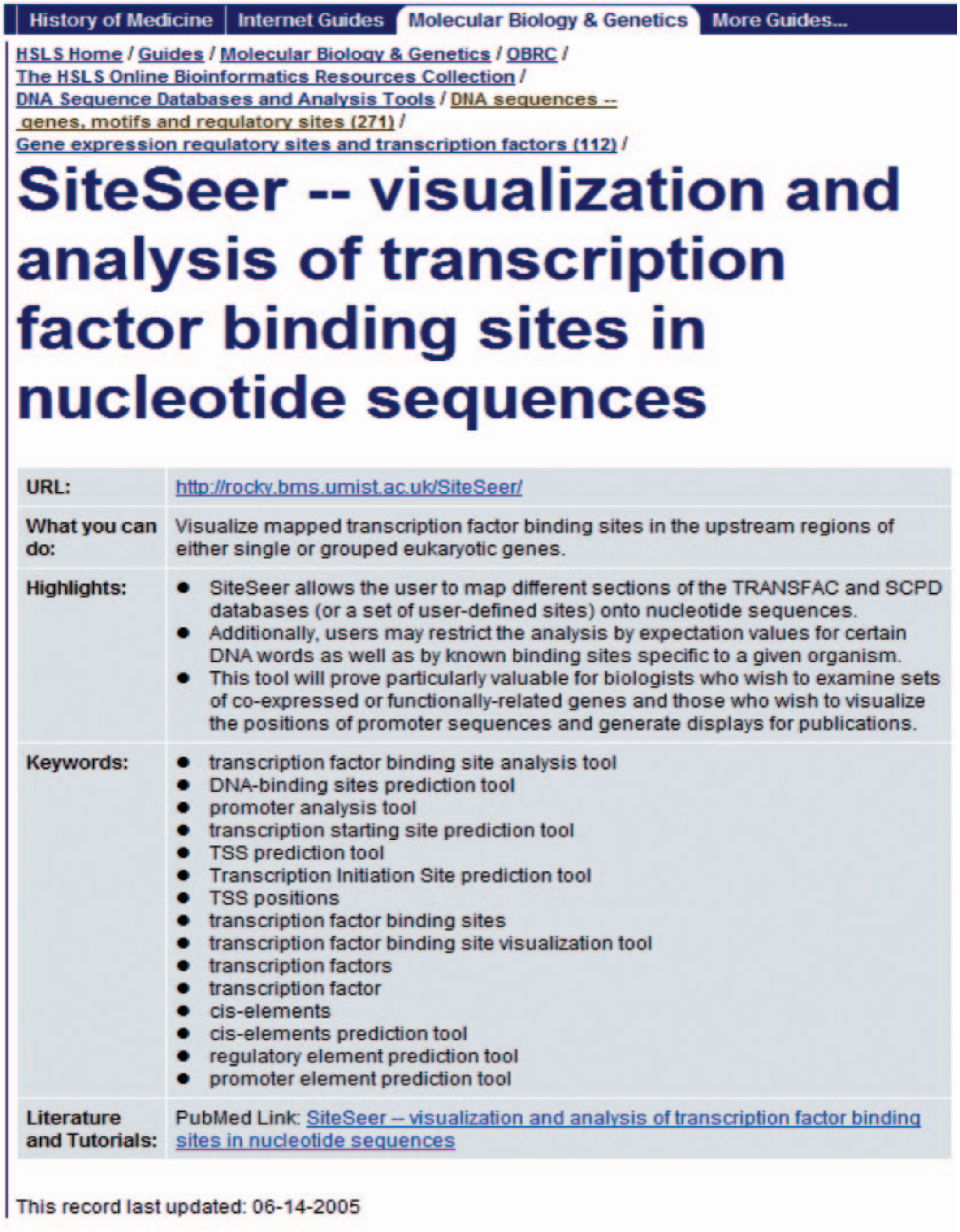
We implemented a categorical structure and basic classification theme that were derived from those used in the *NAR* Molecular Biology Database Collection (2). To facilitate users to browse OBRC, we consolidated the category structure and limited it to three levels. We also expanded the category names to make them more self-evident.

To ensure the up-to-dateness and running status of each entry, we perform link analysis and content verification at least every 6 months. The results are used to update the URLs and remove the entries that are no longer available.

Vivísimo Clustering Engine® implementation

The Vivísimo Clustering Engine® is based on a novel, intricate three-pass algorithm that is augmented with hundreds of special processing heuristics and endowed with thousands of specific facts and general patterns of English and other languages (http://Vivisimo.com/). It automatically organizes large number of search results into different groups and enables users to quickly survey and identify relevant groups. The Vivísimo Clustering Engine® has been successfully applied on the Web by search engines such as the Clusty (http://clusty.com) and ClusterMed™ (http://www.clustermed.info).

Queries can be formed with basic Boolean operators. Queries are first processed by the Zope®-based search engine that leverages on Zope® search tools. The results are then processed by the Vivísimo Clustering Engine® on-the-fly using the textual information from a set of fields selected from the following fields: title, descriptions, highlights and



HSLS Home / Guides / Molecular Biology & Genetics / OBRC / The HSLS Online Bioinformatics Resources Collection / DNA Sequence Databases and Analysis Tools / DNA sequences -- genes, motifs and regulatory sites (271) / Gene expression regulatory sites and transcription factors (112) /

SiteSeer -- visualization and analysis of transcription factor binding sites in nucleotide sequences

URL:	http://rocky.bms.umist.ac.uk/SiteSeer/
What you can do:	Visualize mapped transcription factor binding sites in the upstream regions of either single or grouped eukaryotic genes.
Highlights:	● SiteSeer allows the user to map different sections of the TRANSFAC and SCPD databases (or a set of user-defined sites) onto nucleotide sequences. ● Additionally, users may restrict the analysis by expectation values for certain DNA words as well as by known binding sites specific to a given organism. ● This tool will prove particularly valuable for biologists who wish to examine sets of co-expressed or functionally-related genes and those who wish to visualize the positions of promoter sequences and generate displays for publications.
Keywords:	● transcription factor binding site analysis tool ● DNA-binding sites prediction tool ● promoter analysis tool ● transcription starting site prediction tool ● TSS prediction tool ● Transcription Initiation Site prediction tool ● TSS positions ● transcription factor binding sites ● transcription factor binding site visualization tool ● transcription factors ● transcription factor ● cis-elements ● cis-elements prediction tool ● regulatory element prediction tool ● promoter element prediction tool
Literature and Tutorials:	PubMed Link: SiteSeer -- visualization and analysis of transcription factor binding sites in nucleotide sequences

This record last updated: 06-14-2005

Figure 1. The screenshot of a sample record display of OBRC.

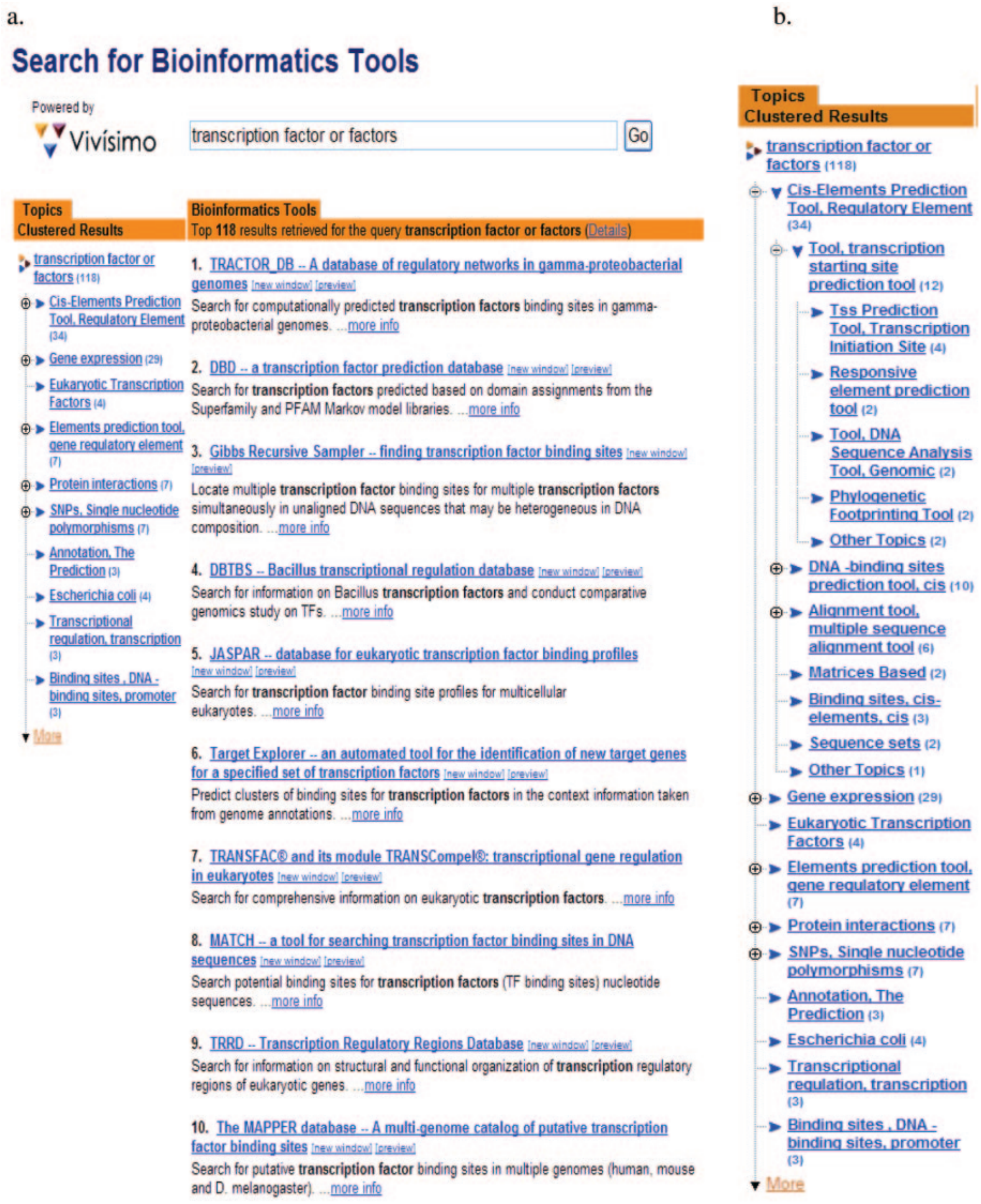
keywords. The search results organized by the Vivísimo Clustering Engine® are finally presented to the users.

RESULTS

Figure 1 shows a sample record display of OBRC.

There are a total of 1542 unique online bioinformatics resources in the current version of OBRC. The databases (475) and software tools (397) published in *NAR* Annual Database Issues (2001–2006) and Web Server Issues (2004–2006) contribute to ~30.8 and 25.7% of the total entries in OBRC, respectively. The resources published in other journals (488) contribute to ~31.6%. In addition, all the valid databases listed in the latest *NAR* Molecular Biology Database Collection (2) are included.

Organized with a three-level hierarchical category classification, OBRC was divided into 13 major categories, 40 secondary-categories and 12 tertiary-categories to assist users browsing the entire collection (Supplementary Table 1). The top five main categories are ‘DNA Sequence Databases and



Search for Bioinformatics Tools

transcription factor or factors

Topics
Clustered Results

- transcription factor or factors (118)
- Cis-Elements Prediction Tool, Regulatory Element (34)
- Gene expression (29)
- Eukaryotic Transcription Factors (4)
- Elements prediction tool, gene regulatory element (7)
- Protein interactions (7)
- SNPs, Single nucleotide polymorphisms (7)
- Annotation, The Prediction (3)
- Escherichia coli (4)
- Transcriptional regulation, transcription (3)
- Binding sites , DNA - binding sites, promoter (3)
- More

Bioinformatics Tools
Top **118** results retrieved for the query **transcription factor or factors** (Details)

1. TRACTOR_DB -- A database of regulatory networks in gamma-proteobacterial genomes [new window] [preview]
Search for computationally predicted **transcription factors** binding sites in gamma-proteobacterial genomes. ...more info

2. DBD -- a transcription factor prediction database [new window] [preview]
Search for **transcription factors** predicted based on domain assignments from the Superfamily and PFAM Markov model libraries. ...more info

3. Gibbs Recursive Sampler -- finding transcription factor binding sites [new window] [preview]
Locate multiple **transcription factor** binding sites for multiple **transcription factors** simultaneously in unaligned DNA sequences that may be heterogeneous in DNA composition. ...more info

4. DBTBS -- Bacillus transcriptional regulation database [new window] [preview]
Search for information on Bacillus **transcription factors** and conduct comparative genomics study on TFs. ...more info

5. JASPAR -- database for eukaryotic transcription factor binding profiles [new window] [preview]
Search for **transcription factor** binding site profiles for multicellular eukaryotes. ...more info

6. Target Explorer -- an automated tool for the identification of new target genes for a specified set of transcription factors [new window] [preview]
Predict clusters of binding sites for **transcription factors** in the context information taken from genome annotations. ...more info

7. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes [new window] [preview]
Search for comprehensive information on eukaryotic **transcription factors**. ...more info

8. MATCH -- a tool for searching transcription factor binding sites in DNA sequences [new window] [preview]
Search potential binding sites for **transcription factors** (TF binding sites) nucleotide sequences. ...more info

9. TRRD -- Transcription Regulatory Regions Database [new window] [preview]
Search for information on structural and functional organization of **transcription** regulatory regions of eukaryotic genes. ...more info

10. The MAPPER database -- A multi-genome catalog of putative transcription factor binding sites [new window] [preview]
Search for putative **transcription factor** binding sites in multiple genomes (human, mouse and D. melanogaster). ...more info

Topics
Clustered Results

- transcription factor or factors (118)
  - Cis-Elements Prediction Tool, Regulatory Element (34)
    - Tool, transcription starting site prediction tool (12)
      - Tss Prediction Tool, Transcription Initiation Site (4)
      - Responsive element prediction tool (2)
      - Tool, DNA Sequence Analysis Tool, Genomic (2)
      - Phylogenetic Footprinting Tool (2)
      - Other Topics (2)
    - DNA -binding sites prediction tool, cis (10)
    - Alignment tool, multiple sequence alignment tool (6)
    - Matrices Based (2)
    - Binding sites, cis-elements, cis (3)
    - Sequence sets (2)
    - Other Topics (1)
  - Gene expression (29)
  - Eukaryotic Transcription Factors (4)
  - Elements prediction tool, gene regulatory element (7)
  - Protein interactions (7)
  - SNPs, Single nucleotide polymorphisms (7)
  - Annotation, The Prediction (3)
  - Escherichia coli (4)
  - Transcriptional regulation, transcription (3)
  - Binding sites , DNA - binding sites, promoter (3)
- More

Figure 2. (a) The screenshot of the first page results for the testing query 'transcription factor or factors' from searching the OBRC using the Zope®-based search engine coupled with the Vivísimo Clustering Engine®. (b) The expanded view of the major clusters of the search results.

Analysis Tools' (325), 'Protein Sequence Databases and Analysis Tools' (306), 'Genomic Databases and Analysis Tools' (270), 'Structure Databases and Analysis Tools' (244) and 'RNA Databases and Tools' (130). The top five specific topics are 'Protein structures' (214), 'Regulatory sites and transcription factors' (112), 'Protein sequence motifs, active or functional sites, and functional annotations' (77), 'Human mutations and diseases' (76) and 'General protein sequence databases, sequence similarity search, analysis, and alignment tools' (68). Some resources were listed in multiple categories.

DISCUSSION

Studies have shown that the clustered results display is more efficient and user friendly than the traditional sequential search results display (20,21). Applying the Vivísimo

Clustering Engine® to the search results offers the users not only a quick overview of all the search results requiring little scrolling, but also shows how the search results are related to each other, as represented by the themes (Figure 2). This advantage becomes compelling in cases where a large number of search results are returned, as the clustered results display drastically reduce the effort needed to navigate through the results set in order to locate the most relevant ones. The sequential display, as employed by popular Web search engines, requires users to scroll down page by page in order to find the results specific to their needs. Another benefit brought by the Vivísimo Clustering Engine® is that users can use relatively broad query terms and may still able to find specific results quickly. This could be particularly helpful to users during their searches as it may reduce the efforts on query reformulation. Furthermore, with Vivísimo's document clustering, there is little need for the expensive and laborious tasks of creating a controlled vocabulary and/or to extensively indexing or pre-labeling the documents.

Our preliminary evaluation study suggests that OBRC search strategy performs much better than Web search engine based strategy, largely attributed to its centralized collection and curated keywords (data not shown). However, the recall and precision are still imperfect. A close examination of the search results indicates that the false negatives, which lower the recall, are primarily due to the synonym problems that have long plagued information retrieval in the biological literatures (22). Another main cause is the singular or plural form of terminologies. Such problems can be largely circumvented by implementing a special online thesaurus or synonym mapping protocol in OBRC. The false positives, which lower the precision, are mainly attributed to the fact that the Zope®-based search engine searches all the text fields of each OBRC entry, and sometimes words in some of the fields match with the queries despite their irrelevance to the major content/function of the corresponding database/software tool. Such false positives could be entirely eliminated if the Zope®-based search engine searched only the keyword field of each OBRC entry. A tradeoff of such strategy is that the keywords are generated to represent only the main concepts, contents and functions of an underlying database/software tool, thus restricting the search to only the keywords field may result in lower recall as the less relevant database/software tools are likely to be left out.

CONCLUSIONS

We have created the OBRC, covering the most widely used and authoritative open source bioinformatics databases and software tools on the Web. The implementation of the Vivísimo Clustering Engine® in OBRC enhances the output of search results and may help users to navigate through large numbers of results with ease. The rich content in OBRC coupled with the advance search features represents a novel search solution for online bioinformatics resources that will benefit biomedical researchers at large. Its aggregated content may also be useful as part of an integrated biological information system.

A future direction will be to continue to expand OBRC to include databases and software tools published in other journals. We will also explore new methods, such as constructing an embedded synonym mapping protocol, implementing the Vivísimo domain-specific controlled vocabularies to further boost the recall and precision, as well as to enhance the results clustering process. Additionally, we will improve the usability of OBRC by studying user experiences and implementing other features, such as adding RSS feed and user/curator preferences/ratings of each resource. We welcome any comments and suggestions on further improvement of OBRC.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Ms Barbara Epstein for her support of this project, and Ms Jill Foust for Web maintenance of OBRC. This research was supported by the National Library of Medicine Biomedical Informatics Training Grant to Y.-B.C. under contract no. 5T15LM007059-18. Funding to pay the Open Access publication charges for this article was provided by the Health Sciences Library System, University of Pittsburgh.

REFERENCES

1. Liolios,A., Tavernarakis,N. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects world-wide. *Nucleic Acids Res.*, **34**, D332–D334.
2. Galperin,M.Y. (2006) The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res.*, **34**, D3–D5.
3. Kostoff,R. (2001) The extraction of useful information from the biomedical literature. *Acad. Med.*, **76**, 1265–1270.
4. Kostoff,R. (2003) Role of technical literature in science and technology development and exploitation. *J. Inform. Sci.*, **29**, 223–228.
5. Cannata,N., Merelli,E. and Altman,R.B. (2005) Time to organize the Bioinformatics Resourceome. *PLoS Comput. Biol.*, **1**, e76.
6. Grivell,L. (2002) Mining the bibliome: searching for a needle in a haystack? *EMBO Rep.*, **3**, 200–203.
7. Schilling,L.M., Wren,J.D. and Dellavalle,R.P. (2004) Letter to the editor: Bioinformatics leads charge by publishing more Internet addresses in abstracts than any other journal. *Bioinformatics*, **20**, 2903.
8. Wren,J.D. (2004) 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, **20**, 668–672.
9. Lu,D. (2006) Information needs of biologists for online bioinformatics resources: implications for health science information professionals. In *Proceedings, 105th Annual Meetings Medical Library Association, Inc.*, 94, E21 May 14–19, 2005, San Antonio, TX.
10. Teufel,A., Krupp,M., Weinmann,A. and Galle,P.R. (2006) Current bioinformatics tools in genomic biomedical research. *Int. J. Mol. Med.*, **17**, 967–973.
11. Fox,J.A., Butland,S.L., McMillan,S., Campbell,G. and Ouellette,B.F.F. (2005) The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Res.*, **33**, W3–W24.
12. Fox,J.A., Butland,S.L., McMillan,S. and Ouellette,B.F.F. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **34**, W3–W5.
13. Discala,C., Benigni,X., Barillot,E. and Vaysseix,G. (2000) DBcat: a catalog of 500 biological databases. *Nucleic Acids Res.*, **28**, 8–9.
14. Babu,P.D., Boddepalli,R., Lakshmi,V.V. and Rao,G.N. (2005) DoD: Database of Databases—updated molecular biology databases. *In Silico Biol.*, **5**, 605–610.
15. Bader,G.D., Cary,M.P. and Sander,C. (2006) Pathguide: a Pathway Resources List. *Nucleic Acids Res.*, **34**, D504–D506.

16. Salton,G. (1971) Cluster search strategies and the optimization of retrieval effectiveness. In Salton,G. (ed.), *The SMART Retrieval System.*. Prentice Hall, Englewood Cliffs, NJ, pp. 223–242.
17. Zamir,O. and Etzioni,O. (1998) Web document clustering: a feasibility demonstration. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'98),* August 24–28, Melbourne, Australia. ACM Press, NY, pp. 46–54.
18. Zeng,H., He,Q., Chen,Z., Ma,W. and Ma,J. (2004) Learning to cluster Web search results. In *Proceedings of the 27th annual international conference on research and development in information retrieval (SIGIR'04),* July 25–29, Sheffield, UK. ACM Press, NY, pp. 210–217.
19. Chattopadhyay,A., Tannery,N.H., Silverman,D.A.L., Bergen,P. and Epstein,B.A. (2006) Design and implementation of a library-based information service in molecular biology and genetics at the University of Pittsburgh. *J. Med. Libr. Assoc.*, **94**, 307–313.
20. Leuski,A. (2001) Evaluating document clustering for interactive information retrieval. In *Proceedings of 10th International Conference on Information and Knowledge Management (CIKM'01),* November 5–10, Atlanta, GA. ACM Press, NY, pp. 33–40.
21. Wu,M., Fuller,M. and Wilkinson,R. (2001) Using clustering and classification approaches in interactive retrieval. *Inf. Proc. Manage.*, **37**, 459–484.
22. Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.