Rapid Report

ibc Interdisciplinary Bio Central   Open Access, Open Review Journal

IntoPub: A Directory Server for Bioinformatics Tools and Databases

DongSoo Jung[1], Ji Han Kim[1], Sanghyuk Lee[1] and Byungwook Lee[1,*]

[1]Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806, Republic of Korea

Subject areas:
Biological computation/Database

Author contribution: D.J. designed the algorithm; J.H.K. carried out the programming and analyzed the results; S.L and B.L. directed the entire study; D.J., S.L. and B.L. wrote the manuscript; All authors read and approved the final manuscript.

***Correspondence** and requests for materials should be addressed to B.L. (bulee@kribb.re.kr).

Editor: Sun Shim Choi, Kangwon National University, Republic of Korea

Received July 29, 2011;
Revised September 08, 2011;
Accepted September 08, 2011;
Published September 20, 2011

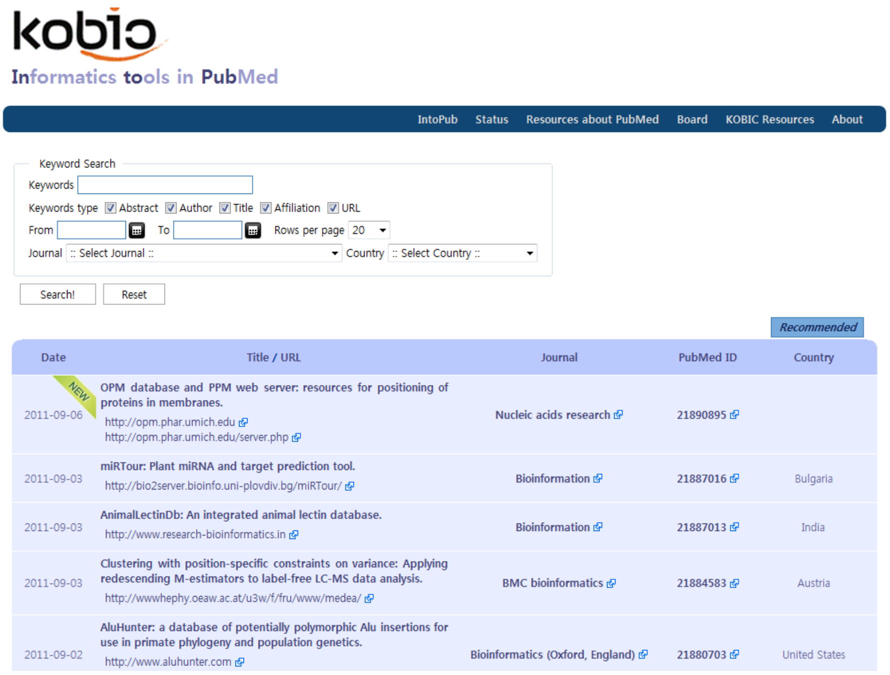
Citation: Jung, D., et al. IntoPub: A Directory Server for Bioinformatics Tools and Databases. IBC 2011, 3:12, 1-3.
doi: 10.4051/ibc.2011.3.3.0012

Funding: The work was supported by the KRIBB Research Initiative Program and the Korean Ministry of Science and Technology under grant number (2010-0029345).

Competing interest: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

Bioinformatics tools and databases are useful for understanding and processing various biological data. Numerous resources are being published each year. It is not a trivial task to find up-to-date relevant tools and databases. Moreover, no server is available to provide comprehensive coverage on bioinformatics resources in all biological fields. Here, we present a directory server called IntoPub that provides information on web resources. First, we downloaded XML-formatted abstracts containing web URLs from the NCBI PubMed database by using 'ESearch-EFetch' function in the NCBI E-utilities. The information is obtained from abstracts in the PubMed by extracting 'www' or 'http' prefixes. Then, we curate the downloaded abstracts both in automatic and manual fashion. As of July 2011, the IntoPub database has 12,118 abstracts containing web URLs from 174 journals. Our analysis shows that the number of abstracts containing web resources has increased significantly every year. The server has been tested by many biologists from several countries to get opinion on user satisfaction, usefulness, practicability, and ease of use since January 2010. In the IntoPub web server, users can easily find relevant bioinformatics resources, as compared to searching in PubMed. IntoPub will continue to update and incorporate new web resources from PubMed and other literature databases. IntoPub, available at http://into.kobic.re.kr/, is updated every day.



Keyword Search
Keywords
Keywords type ☑ Abstract ☑ Author ☑ Title ☑ Affiliation ☑ URL
From   To   Rows per page 20
Journal :: Select Journal ::   Country :: Select Country ::

Date	Title / URL	Journal	PubMed ID	Country
2011-09-06	OPM database and PPM web server: resources for positioning of proteins in membranes. http://opm.phar.umich.edu http://opm.phar.umich.edu/server.php	Nucleic acids research	21890895	
2011-09-03	miRTour: Plant miRNA and target prediction tool. http://bio2server.bioinfo.uni-plovdiv.bg/miRTour/	Bioinformation	21887016	Bulgaria
2011-09-03	AnimalLectinDb: An integrated animal lectin database. http://www.research-bioinformatics.in	Bioinformation	21887013	India
2011-09-03	Clustering with position-specific constraints on variance: Applying redescending M-estimators to label-free LC-MS data analysis. http://wwwhephy.oeaw.ac.at/u3w/f/fru/www/medea/	BMC bioinformatics	21884583	Austria
2011-09-02	AluHunter: a database of potentially polymorphic Alu insertions for use in primate phylogeny and population genetics. http://www.aluhunter.com	Bioinformatics (Oxford, England)	21880703	United States

Keywords: directory server, bioinformatics databases, bioinformatics tools, web server, IntoPub, PubMed

Introduction

Numerous bioinformatics resources including biological databases and tools have been developed so far and many of them are publicly available[1]. These resources are tremendously useful for analyzing and managing a variety of biological data[2]. Efficient use of bioinformatics resources would be of great help in producing more comprehensive and accurate interpretation[3]. Moreover, it is more cost-effective and time-saving to use these resources compared to self development or manual data processing[4]. However, it is not an easy task to find proper relevant resources since the number of bioinformatics resources is increasing rapidly[5].

Bioinformatics resources are mainly published in scientific journals. Most of these published articles contain web URLs, where users can input a query and obtain results or download databases and tools for local use. Google search is powerful but it produces many irrelevant hits. Accordingly, researchers tend to prefer searching literature databases as proven resources. PubMed[6] is the most widely used database, comprised of approximately 20 million literatures in the biomedical field covering MEDLINE, life science journals, and online books. In spite of its usefulness and comprehensiveness, searching for appropriate resources in PubMed is still complicate since most hits would not contain the proper web links to bioinformatics resources.

There are several web sites dedicated to directory service for bioinformatics resources. The most popular one is the Nucleic Acids Research (NAR) Database and Web server collections[7,8]. They provide a good summary of bioinformatics tools and databases classified according to subject categories. Although the NAR collections are comprehensive in terms of subject coverage, they include resources published only in the NAR Database and Web server issues, which are published only once a year. Even though a few sites, such as NCBI[9], EBI[10], and ExPASy[11], publicly classify and provide bioinformatics resource information, they cover only part of diverse bioinformatics fields. To our knowledge, there is no directory server that provides comprehensive and up-to-date information on bioinformatics tools and databases in all biological fields.

Here we present a directory server, called IntoPub (Informatics tools in PubMed), to provide bioinformatics resource information useful for processing and analyzing biological data. To construct IntoPub, we obtained abstracts containing web resource information from the PubMed database, and annotated the abstracts automatically and manually. New web resources introduced in PubMed are updated in IntoPub on a daily basis.

Results and Discussion

Statistics

We obtained bioinformatics resources from the PubMed database (see Material and Methods). As of July 2011, the IntoPub database has 12,118 abstracts containing web URLs from 174 journals. Our analysis shows that the number of abstracts containing web resources has increased significantly every year (Figure 1). Notably, the cumulative number has increased exponentially since 2003. Such an increase indicates that computational methods have become increasingly more important in biological fields. More than 90% of the resources have been introduced through the application notes in the Bioinformatics journal, Web server and Database issues in NAR, and BMC journals (see IntoPub website for more statistics).

Searching

IntoPub can be accessed through a web interface for querying. The query interface allows the user to search against web resources and abstracts. Web resources can be searched by abstract, author, title, affiliation, and URL. Users can also filter by date, journals, and countries. The search results contain the title, URL, journal, PubMed ID, and country. Users can access an abstract or article by clicking on ‘PubMed ID’.

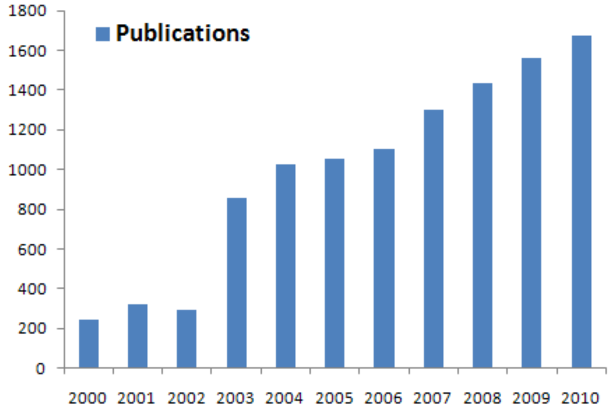


Figure 1. Yearly distribution of publications containing web resources.

In the IntoPub web server, users can easily find relevant bioinformatics resources, as compared to searching in PubMed. For example, given a simple query of ‘RNA-Seq’, PubMed results produced more than 374 results, for which most are research results without web resource information. However, IntoPub results generated 55 abstracts containing web information. Thus, users can find the proper web resources for their research efficiently.

Web interface of IntoPub

The IntoPub server consists of a web interface, a database management system (DBMS), and the core programs. The web interface is implemented with static HTML and PHP scripts and MySQL DBMS is used to store the IntoPub database.

The web interface of the IntoPub server has several menus, including Searching, Status, Resources about PubMed, and Board. In the Status menus, users can see several statistics. In the Resources about PubMed menu, text mining tools and databases are introduced.

Discussion

Not only being an information service for general users, IntoPub can serve as a good news provider for software development in the field of bioinformatics. Currently, it is not easy to keep track of all published software even for experts in bioinformatics. IntoPub, being updated daily, would be a good site to find the most recent programs, algorithms, and databases available on the web.

We obtained bioinformatics resources from PubMed by using two keywords: ‘http’ and ‘www’. Although this method is effective in collecting bioinformatics resources from PubMed, there are some caveats to our method and its implementation. First, if a web URL does not have one of the two words, it would not be included in our database. For instance, some URLs represented by words and periods with no space (e.g. ‘kobic.re.kr’) would be missing in IntoPub. Secondly, the web resource information comes only from the abstract. Thus, URLs within the body of an article will be missed. We did not expand the search scope to the full text deliberately since it would include many false positives such as citations for other resources used in the work.

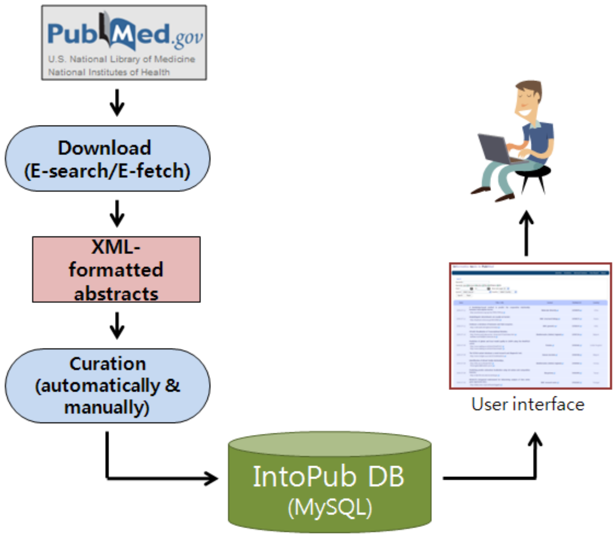


Figure 2. The workflow of constructing the IntoPub database.

Conclusion and Prospects

In this article, we have described development of the IntoPub directory server and its unique features. Major merit is that it provides manually-curated, thus highly reliable information on bioinformatics resources. After the construction of IntoPub in January 2010, the server has been tested by many biologists from several countries to get opinion on user satisfaction, usefulness, practicability, and ease of use. Users' comments were actively accommodated in the IntoPub service. IntoPub will continue to update and incorporate new web resources from PubMed and other literature databases.

Materials and Methods

Data source

We downloaded XML-formatted abstract containing web URLs from the NCBI PubMed database by using 'ESearch-EFetch' function in the NCBI E-utilities. Abstracts containing bioinformatics resources were identified by searching the web site address (URL) with 'www' or 'http'. Then, the abstracts were filtered by the publication date. For example, abstracts published during 2009 can be queried with '(http[Title/Abstract]) AND www[Title/Abstract]) AND "2010/01/01"[Publication Date] : "2010/12/31"[Publication Date])' as the PubMed search input. The PubMed search results were saved in the XML format.

Data parsing

We developed an annotation pipeline to obtain abstracts with valid web URLs from the downloaded data (Figure 2). The pipeline can be divided into three steps. First, we parsed XML-formatted abstracts. Then, we filtered out the duplicate abstracts by comparing PubMed IDs in the IntoPub database. Invalid abstracts were automatically screened out by comparing with a list of invalid abstracts and URLs. Second, we manually checked the URLs of web servers to determine whether abstracts contain invalid web information. For example, all abstracts published in the British Journal of Cancer contain the journal URL, 'www.bjcancer.com'. Such erroneous URLs were added to the invalid URL list, which is used in the first step of screening. We checked web URLs are valid. We also manually removed space and comma (,) from web URLs, and corrected misspelled web addresses. If new journals containing valid URLs were identified, they were added into a journal list containing web server. Lastly, we saved new abstracts with web resource information in the IntoPub database. The database is being updated on a daily basis since most of the task is automated. We also extract web URL from the abstract and the country of origin from the author's affiliation using regular expressions to provide more detailed information.

Acknowledgements

This work was supported by the KRIBB Research Initiative Program and the Korean Ministry of Education, Science and Technology (MEST) under grant number (2010-0029345). DJ thanks Jemin Kim for correcting and updating XML-parsing modules.

References

1. Eaton, A.D. (2006). HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res* 34, W745-747.
2. Gilbert, D. (2004). Bioinformatics software resources. *Brief Bioinform* 5, 300-304.
3. Brooksbank, C., Cameron, G., and Thornton, J. (2010). The European Bioinformatics Institute's data resources. *Nucleic Acids Res* 38, D17-25.
4. Tsai, R.T., Dai, H.J., Lai, P.T., and Huang, C.H. (2009). PubMed-EX: a web browser extension to enhance PubMed search with text mining features. *Bioinformatics* 25, 3031-3032.
5. Dai, H.J., Huang, C.H., Lin, R.T., Tsai, R.T., and Hsu, W.L. (2008). BIOSMILE web search: a web application for annotating biomedical entities and relations. *Nucleic Acids Res* 36, W390-398.
6. Beebe, D.C. (2006). Public access success at PubMed. *Science* 313, 1571-1572.
7. Benson, G. (2010). Editorial. Nucleic Acids Research annual Web Server Issue in 2010. *Nucleic Acids Res* 38, W1-2.
8. Cochrane, G.R., and Galperin, M.Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res* 38, D1-4.
9. Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S.H. (2010). The NCBI BioSystems database. *Nucleic Acids Res* 38, D492-496.
10. Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* 38, W695-699.
11. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31, 3784-3788.