A high throughput approach to keep alive a web-based database system for multiple search among published bioinformatics tools and databases

Vassilis Atlamazoglou, Trias Thireou, Anastasia Alexandridou and George Spyrou*

Abstract—**Since modern Biology has been transformed to a quantitative science dealing with tremendous amounts of data, the applications of information science and technology on Biology, i.e. Bioinformatics, are so many that they comprise a separate dataset needing organization and indexing for optimum search and information retrieval. Metabasis, a web-based database system for organizing and maintaining information relevant to published bioinformatics tools and databases has been developed by our group. However, such an effort requires rapid and massive information update, an issue quite difficult since the bioinformatics tools production rate is very high and there is no standard protocol in the way they are presented in the literature. We present here a high throughput automatic retrieval procedure to keep alive a database of this kind, using certain methodology for automatic downloading, parsing and filtering data from the related literature.**

I. Introduction

THE tremendous progress in modern Biology and the need for biological results quantification, data archiving, analysis, interpretation and information retrieval through the web led to a corresponding explosion in the development of bioinformatics tools and databases. Nevertheless, it is very dangerous especially for the newcomers in the field rather to sink than swim towards the tidal wave of available tools. Furthermore, the selection of the most appropriate tool for a particular application is neither simple nor obvious, especially if one faces the choice of a novel tool over one that is already accepted and widely used. Due to the increased user's inertia in trusting new tools, the new and often improved applications are very slowly tested and adopted.

There are a large number of available websites containing categorized lists of bioinformatics tools and databases. The most useful of these sites are typically those created by highly qualified scientists that propose the tools they use themselves [1-4]. For obvious reasons, such lists are neither always comprehensive nor targeted to a wide range of users. Furthermore, the majority of these lists fail to be kept updated because they are not designed as dynamic repositories (databases) of information. A few databases of biocomputing tools are also available [5-9]. The crucial point for them is the period they are kept updated and the quality of updating protocol they adopt. Among them there are two important efforts: BiowareDB and MetaDB. BiowareDB [5] is an attempt at collecting an exhaustive hyperlinked database links to the vast amounts of freely available and commercial bioinformatics and biocomputing software. MetaDB [6] is a sorted, searchable collection of biological databases. Most entries in the metadatabase include a relevant peer-reviewed abstract or excerpt along with a link to the abstract or full text article.

The MetaBasis project [9] aims to provide a one-stop source for bioinformatics tools and databases, especially helpful to newcomers to the field of bioinformatics. It intends to be an approach primarily used for basic research. However, it may also give the opportunity to experts working in the interfacing field (from the laboratory bench to the clinical bed) to find bioinformatics tools suitable to their needs, such as microarray analysis tools.

We have built an automatic high throughput updating procedure to support Metabasis and keep it alive in terms of reliable up to date information. This procedure includes use of Medline-Entrez Programming Utilities, scripts for XML parsing and filtering as well as rule-based classification scripts in order to massively collect the new bioinformatics tools and databases. Using this procedure, Metabasis contains up to now more than 3000 entries published bioinformatics tools and databases.

II. Materials And Methods

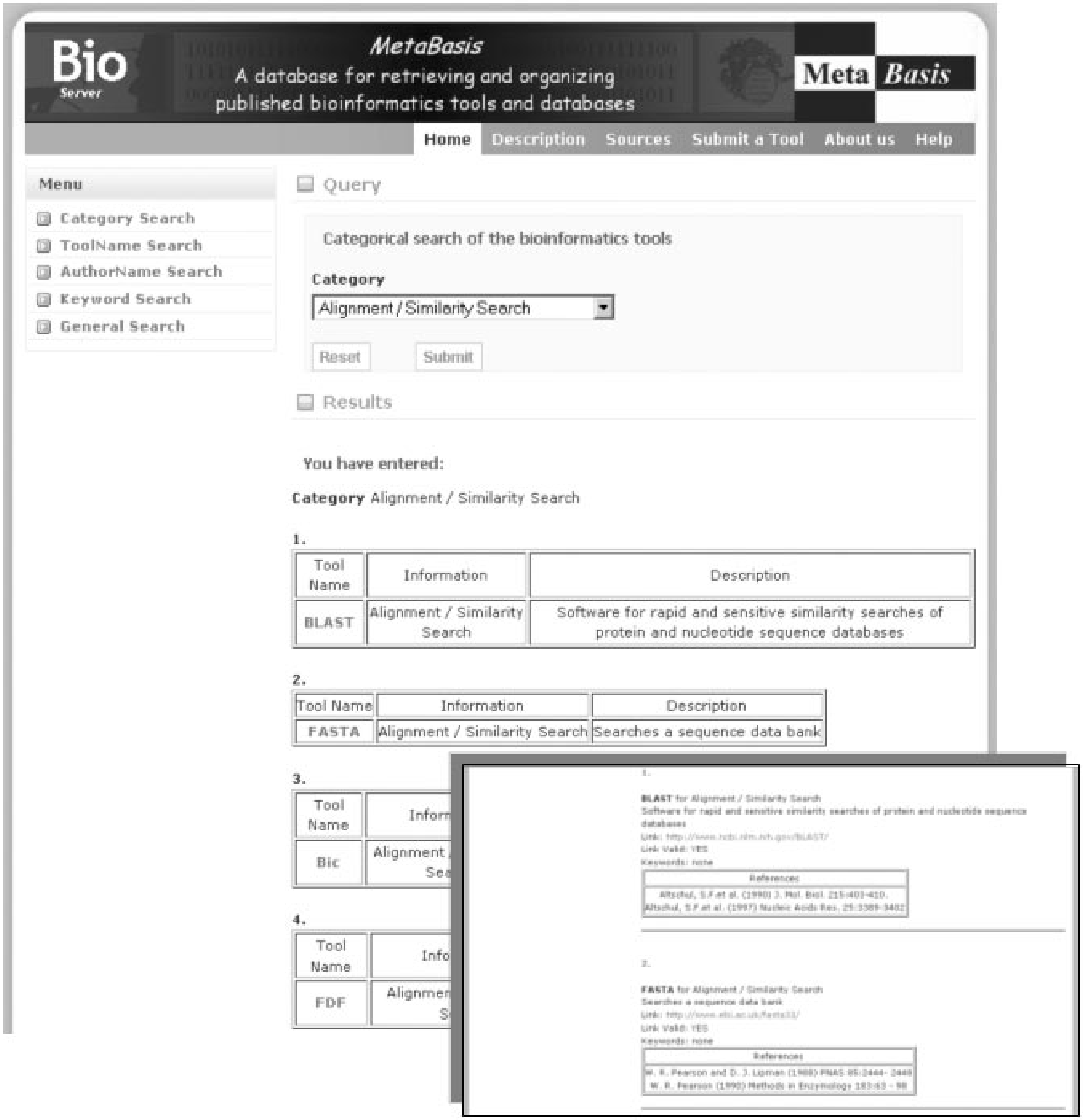
A. Description of the Database System

We have developed a web-based relational database system for organizing and maintaining information relevant to bioinformatics tools and databases, called MetaBasis (from Greek 'metabaino', "to go over"). Although MetaBasis has been presented in previous publication, its new interface and functionalities are presented here. MetaBasis is a multiple way searchable database. The user may search asking for tools (i) that belong to one among thirteen general categories, (ii) with a certain name or part of the name, (iii) published by a certain author, (iv) that have in their description certain non-predefined keywords or

Manuscript received July 5, 2008. V. Atlamazoglou and T. Thireou are with the School of Electrical and Computer Engineering, National Technical University of Athens, 9, Iroon Polytechniou str. Athens, GR 15780, Greece.

A. Alexandridou and G. Spyrou are with the Biomedical Informatics Unit, Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephessiou , 115 27 Athens, Greece

* Corresponding Author



Categorical search of the bioinformatics tools

Category

Alignment / Similarity Search

You have entered:

Category Alignment / Similarity Search

1.

Tool Name	Information	Description
BLAST	Alignment / Similarity Search	Software for rapid and sensitive similarity searches of protein and nucleotide sequence databases

2.

Tool Name	Information	Description
FASTA	Alignment / Similarity Search	Searches a sequence data bank

3.

Tool Name	Inforn
Bic	Alignment Sea

4.

Tool Name	Info
FDF	Alignmen S

1.

BLAST for Alignment / Similarity Search
Software for rapid and sensitive similarity searches of protein and nucleotide sequence databases
Link: http://www.ncbi.nlm.nih.gov/BLAST/
Link Valid: YES
Keywords: none

References
Altschul, S.F.et al. (1990) J. Mol. Biol. 215:403-410. Altschul, S.F.et al. (1997) Nucleic Acids Res. 25:3389-3402

2.

FASTA for Alignment / Similarity Search
Searches a sequence data bank
Link: http://www.ebi.ac.uk/fasta33/
Link Valid: YES
Keywords: none

References
W. R. Pearson and D. J. Lipman (1988) PNAS 85:2444- 2448 W. R. Pearson (1990) Methods in Enzymology 183:63 - 98

Fig. 1. Characteristic snapshots from Metabasis

combination of keywords using Boolean Logic (Figure 1). Lately, MetaBasis database has been incorporated in a new sever gathering tools, databases and web services developed in the Biomedical Research Foundation, Academy of Athens, called "BioServer". For this purpose the application has been transferred in MySQL platform and the dynamic web interface is implemented in PHP language.

The basic issue that we face in this paper is the high throughput curation approach we have built in order to have updated the database. According to this approach, we have established a certain retrieval and curation protocol in order to have the application updated and functional in a permanent way. This protocol includes (i) content update, described below and (ii) link validity check for all the deposited records (with proper script running against the corresponding URL links list), in a periodical basis.

B. High throughput data retrieving and database curation

MEDLINE/PubMed is one of the most significant information resources for bioinformatics text mining. In order to automate the procedure of retrieving relative articles from MEDLINE and to extract a set of rules to decide whether an article is describing a bioinformatics tool/database or not, we chose to focus our study, in the beginning, on the Journals Bioinformatics and Nucleic Acids Research. Bioinformatics is the longest running publication (launched as CABIOS — Computer Applications in the Bioscience) for original papers in this field. Likewise, in Nucleic Acids Research, the first issue of each year is devoted to biological databases and the July issue to articles describing web-based bioinformatics and computational biology software resources.

For this purpose, in order to search for and retrieve the requested data from MEDLINE, the Entrez Programming Utilities (eUtils) were used. Having specified the journal name and a time period to query, the PMIDs (PubMed Unique Identifier) of all the contained articles were retrieved. Using these PMIDs we downloaded, in XML format, the abstracts of all articles containing a URL. Each XML file contained one citation unit and it was properly parsed to extract selected XML text fields.

For the downloaded abstract xml files, a set of rules was examined in order to decide whether an abstract refers to a bioinformatics tool or not. Using regular expressions and a list of words (and their synonyms) that covered concepts of interest (e.g availability, online, download, source code, software etc) a representation of a text pattern was created to facilitate abstract automatic curation. Another pattern representation concerns abstracts in which the toolname appears both in the title and abstract text (in close proximity with the URL). Special care has been taken to exclude cases where the given URL refers only to supplementary material. Additionally a named entity recognition procedure was followed to extract information automatically (key items extracted: toolname and URL). If an abstract is selected as a relevant article (describes a software tool or database) but does not contain a tool name or the automatic procedure fails to recognize one, the abstract title is used as the toolname.

The final XML file of an abstract of interest includes the following data: PMID, abstract title, abstract text, authornames, journal reference, MeSH terms and the extracted toolname and URL.

Furthermore, the high throughput procedure is empowered with automatic keyword extraction using the KEA algorithm [10]. KEA is an open source software implemented in Java for extracting keyphrases from text documents. It can be either used for free indexing or for indexing with a controlled vocabulary. For a number of selected abstracts we have created the corresponding keyword files containing the related keywords extracted manually from each abstract. These files comprise the training model that KEA uses to produce the keywords for any other abstract examined. The flowchart of the automatic updating procedure is demonstrated in Figure 2.

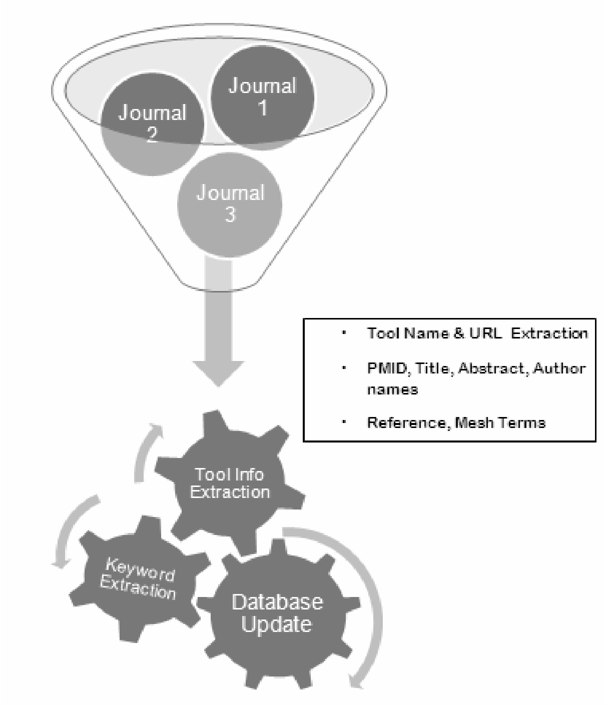


Fig. 2. Flowchart of the automatic updating procedure

III. Results

To assess the performance of the automatic retrieval procedure, we used a set of 600 abstract texts (300 from Bioinformatics and 300 from Nucleic Acids Research Database and Webserver issues) that was both automatically and manually curated, to decide whether they described bioinformatics tools or databases. The sensitivity and specificity measures are shown in Table 1. Based on the abstracts of the selected articles of interest from the above performance measurement procedure, we checked the

TABLE I
Sensitivity and specificity of the automatic selection of publications of interest

Journal	Sensitivity (%)	Specificity (%)
Nucleic Acid Research*	100	100
Bioinformatucs	97.01	91.55

* Databases & Web Server special issues.

TABLE II
Accuracy of the automatic extraction procedure of the tool's name and its URL (for the selected publications of interest)

Journal	Total name accuracy (%)	URL accuracy (%)
Nucleic Acid Research*	84.14	96.12
Bioinformatucs	83.08	99.23

* Databases & Web Server special issues.

accuracy with which the tool's name and its URL were automatically extracted. The corresponding measures are shown in table 2. As far as the URL extraction is concerned, there are abstracts where more than one URL is mentioned, related to other tools (used to or compared to) and supporting information or supplementary material of the publication.

IV. Discussion

From the demonstrated results, we can see that the presented procedure performs quite well in terms of automatic selection of publications of interest. In addition, it has very good performance in terms of URL extraction accuracy. The most difficult task is the accurate extraction of the Tool name. Our procedure performs well in that field (accuracy > 83%) but there is space for further optimization. On the other hand, journals should adopt a common protocol in order to present bioinformatics tools and databases. If the manuscripts follow a standard way of presentation, then it is expected for our automatic updating procedure as well as for others like this to reach to almost 100% accuracy.

References

[1] ExPASy life science directory [online]. Available from URL: http://www.expasy.org/links.html

[2] ExPASy proteomics tools [online]. Available from URL: http://www.expasy.org/tools/

[3] Fox JA, Butland SL, McMillan S, et al. The bioinformatics links directory: a compilation of molecular biology web servers. Nucleic Acids Res 2005 Jul; 33: W3-24

[4] Koonin EV, Galperin MY. Principles and methods of sequence analysis. In: Koonin EV, Galperin MY, editors. Sequence-evolution-function: computational approaches in comparative genomics [online]. Available from URL: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?.rid = sef.chapter.166

[5] Matthiessen MW. BioWareDB: the biomedical software and database search engine. Bioinformatics 2003 Nov; 19 (17): 2319-20

[6] Neurotransmitter.net. MetaDB: a metadatabase for the biological sciences [online]. Available from URL: http://www.neurotransmitter.net/metadb/metadb.php

[7] Institute Pasteur. The bio netbook [online]. Available from URL: http://www.pasteur.fr/recherche/BNB/bnb-en.html

[8] ExPASy & health on the net. BioHunt [online]. Available from URL: http://ca.expasy.org/BioHunt/

[9] Atlamazoglou V, Thireou T, Hamodrakas Y, Spyrou G., MetaBasis: a web-based database containing metadata on software tools and databases in the field of bioinformatics. Appl Bioinformatics. 2006;5(3):187-92, [online at URL: http://bioserver-1.bioacademy.gr/Metabasis

[10] KEA [online]. Available from URL: http://www.nzdl.org/Kea/