BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 21 2010, pages 2801–2802
doi:10.1093/bioinformatics/btq520

Databases and ontologies

Advance Access publication September 9, 2010

PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids

Miguel García-Remesal[1,2,*], Alejandro Cuevas[2], David Pérez-Rey[1,2], Luis Martín[2], Alberto Anguita[2], Diana de la Iglesia[2], Guillermo de la Calle[2], José Crespo[2,3] and Víctor Maojo[1,2]

[1]Departamento de Inteligencia Artificial, Facultad de Informática, [2]Biomedical Informatics Group, Facultad de Informática and [3]Departamento de Lenguajes, Sistemas Informáticos e Ingeniería del Software, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegacedo S/N, 28660 Boadilla del Monte, Madrid, Spain

Associate Editor: John Quackenbush

ABSTRACT

Summary: PubDNA Finder is an online repository that we have created to link PubMed Central manuscripts to the sequences of nucleic acids appearing in them. It extends the search capabilities provided by PubMed Central by enabling researchers to perform advanced searches involving sequences of nucleic acids. This includes, among other features (i) searching for papers mentioning one or more specific sequences of nucleic acids and (ii) retrieving the genetic sequences appearing in different articles. These additional query capabilities are provided by a searchable index that we created by using the full text of the 176 672 papers available at PubMed Central at the time of writing and the sequences of nucleic acids appearing in them. To automatically extract the genetic sequences occurring in each paper, we used an original method we have developed. The database is updated monthly by automatically connecting to the PubMed Central FTP site to retrieve and index new manuscripts. Users can query the database via the web interface provided.

Availability: PubDNA Finder can be freely accessed at http://servet.dia.fi.upm.es:8080/pubdnafinder

Contact: mgarcia@infomed.dia.fi.upm.es

Received on July 16, 2010; revised on August 27, 2010; accepted on September 4, 2010

1 INTRODUCTION

The biological literature is the main source of information reporting empirically validated genetic sequences, such as for instance PCR primers and probes. As result, researchers usually need to review the available literature to search for sequence data, which can be a hard and time-consuming task. PubMed Central is currently the main source of open-access full-text papers reporting genetic sequence data. However, the search engine provided by PubMed Central does not support researchers to retrieve papers containing the genetic sequences specified by the user, and to automatically identify and extract the sequences of nucleic acids mentioned in the retrieved articles.

PubDNA Finder is an online repository linking PubMed Central manuscripts to the different genetic sequences appearing in them. It extends the search capabilities provided by PubMed Central by allowing researchers to (i) retrieve all articles containing the genetic sequences specified by the user—featuring both exact and approximate matching; (ii) retrieve all the sequences appearing in the manuscripts matching a keyword-based query; and (iii) retrieve all articles matching a keyword-based query and containing the sequences specified by the user. PubDNA Finder currently contains the 176 672 papers available from PubMed Central at the time of writing. The database is automatically updated on a monthly basis to retrieve and index new manuscripts.

2 METHODS

To create the index, we downloaded all the 176 672 XML-formatted manuscripts available from the PubMed Central FTP site[1] at the time of writing. We used Apache Lucene[2] 3.0.1 to index the different documents based on the full text of the manuscripts and the genetic sequences appearing in each manuscript. The latter were automatically identified and extracted—together with the context in which they appeared—using a method created by the authors and reported elsewhere (García-Remesal *et al.*, 2010). The adopted method resorts to a rule-based expert system to automatically identify and extract the sequences of nucleic acids. To enable users to interactively query the developed index, we created a web interface.

3 FEATURES

Users can perform three different types of queries using PubDNA Finder, as described below.

3.1 Sequence-based queries

Sequence-based queries (SBQs) are aimed at retrieving all manuscripts containing one or more genetic sequences specified by the user. There are two different types of SBQs: simple and advanced. Simple SBQs are composed of one or more complete sequences linked by a single logical operator, such as 'retrieve all manuscripts containing either the sequence TATGGAAMAGATC-GGCGG or the sequence ATTGGCGAAGTCGGTAGG'.To launch this query, we would type the target sequences—one per line—in the text-box labeled with 'Sequences' (Fig. 1) and then we would select the OR logical operator in the 'Operator' combo box.

*To whom correspondence should be addressed.

[1]ftp://ftp.ncbi.nlm.nih.gov/pub/pmc
[2]http://lucene.apache.org/

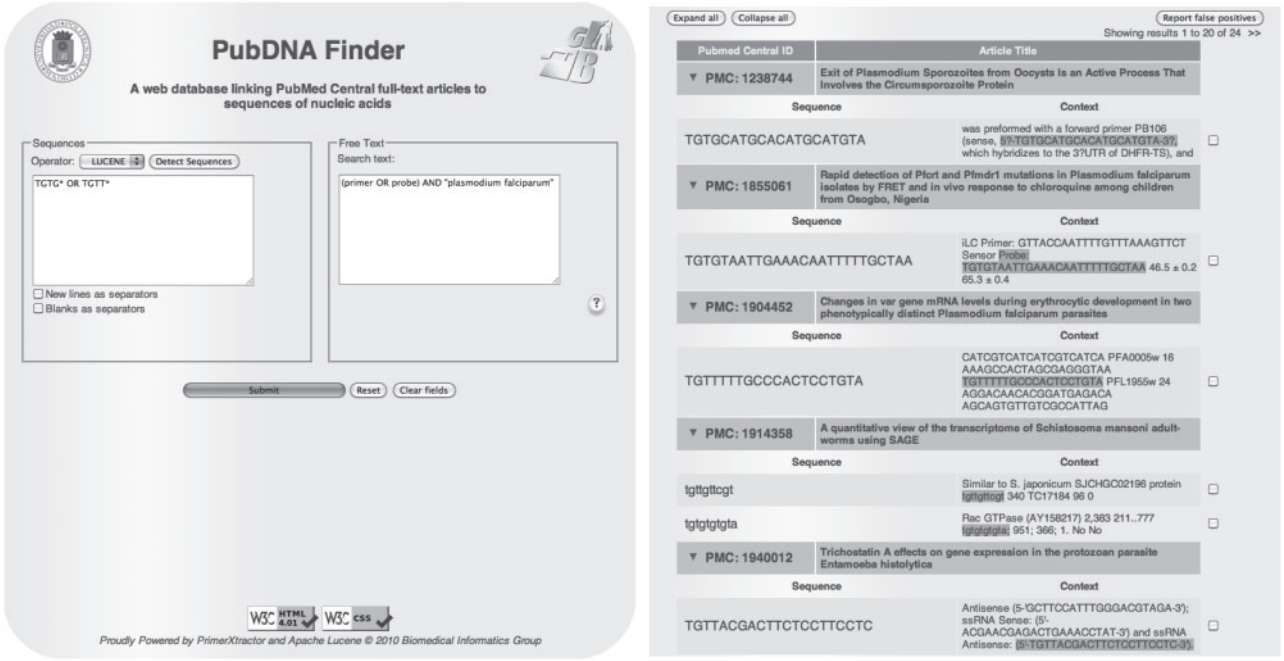

Fig. 1. Overview of the database query interface.

In contrast, advanced SBQs can involve complex expressions following the Lucene query syntax. This includes, among other features, the use of wildcards—e.g. '?' to match a single character or '*' to match any number of characters—range-based queries, fuzzy queries (to perform approximate matching) and complex logic combinations of different expressions. An example of advanced SBQ would be 'retrieve all manuscripts containing genetic sequences matching any of the expressions AT??TGAA*TA or AT??TGAA*GA'. To launch this query, we would type the string 'AT??TGAA*TA OR AT??TGAA*GA' in the textbox labeled as 'Sequences' and then we would select the 'LUCENE' operator in the 'Operator' combo box.

3.2 Keyword-based queries

Keyword-based queries (KBQs) are designed to retrieve the sequences appearing in manuscripts matching the search terms. To perform a KBQ, we would type a sequence of keywords in the textbox labeled as 'Free Text' (Fig. 1). After executing the query, we would be presented with the sequences appearing in all manuscripts containing any of the search terms. Additionally, we can create more complex queries by following the Lucene query syntax. For instance, the query ' "TaqMan probe" AND "Brucella Melitensis" ' would retrieve the sequences mentioned in all the manuscripts containing both the strings 'TaqMan probe' and 'Brucella Melitensis'.

3.3 Combined queries

Combined queries (CQs) are aimed at retrieving all articles matching a keyword-based query and containing the sequences specified by the user. Figure 1 shows an example of a CQ, where we are interested in retrieving all manuscripts matching the KBQ '(primer OR probe) AND "plasmodium falciparum" ' and containing sequences that match the expression 'TGTG* OR TGTT*'.

As shown in Figure 1, users are presented with the different sequences (and manuscripts) matching the user query, together with the context in which they appear. Users can also access the full text of the retrieved manuscripts at PubMed Central by following the provided links.

3.4 Additional features

PubDNA Finder can also retrieve all the sequences mentioned in any specific manuscript identified by its PubMed Central identifier (PMCID). For instance, to retrieve all sequences appearing in the article whose PMCID is 1253840, we would type the query string 'pmcid:1253840' in the text-box labeled with 'Free Text'.

Additionally, PubDNA Finder features a built-in recognizer and extractor of genetic sequences. Users can paste any text in the 'Sequences' text-box to automatically identify and extract all the sequences mentioned in the pasted text simply by clicking the button 'Detect Sequences'.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Casimir A. Kulikowski (Rutgers University) for his valuable comments and suggestions.

Funding: European Commission through the DICODE project; the Spanish Ministry of Science and Innovation through the ImGraSeg project, FIS/AES PS09/00069 and COMBIOMED-RETICS; Comunidad de Madrid, Spain.

Conflict of Interest: none declared.

REFERENCE

García-Remesal,M. *et al.* (2010) A knowledge engineering approach to recognizing and extracting sequences of nucleic acids from scientific literature. In *Proceedings of the 32nd Annual Conference of the IEEE Engineering in Medicine and Biology Society.* Buenos Aires, Argentina.