PubDNA Finder in a Nutshell

Searching the Life Sciences Literature with Sequences of Nucleic Acids

Biomedical researchers and clinicians working with molecular technologies in routine clinical practice often need to review the available literature to gather information regarding specific sequences of nucleic acids. This includes, for instance, finding articles related to a concrete DNA sequence, or identifying empirically-validated primer/probe sequences to evaluate the presence of different micro-organisms.

Dr. Miguel García-Remesal, Universidad Politécnica de Madrid

Unfortunately, these hard and time-consuming tasks often need to be manually performed by researchers themselves since no publicly available biomedical literature search engine, e.g. PubMed, PubMed Central, etc., provides the required search functionalities. In this article, we describe PubDNA Finder, a web service that enables users to perform advanced searches on PubMed Central-indexed full text articles with sequences of nucleic acids.

Searching the Life Sciences Literature

PubDNA Finder [1] is a web service we developed linking more than 180,000 full text articles available at PubMed Central (PMC) at the time of writing, to the DNA/RNA sequences appearing in them. PubDNA Finder extends the functionality provided by the PMC search engine by enabling researchers to perform queries involving both keywords and DNA/RNA sequences. To our knowledge, PubDNA Finder is the first search engine providing such advanced search capabilities. PubDNA Finder can be accessed free of charge at http://servet.dia.fi.upm.es:8080/pubdnafinder

Search Functionalities provided by PubDNA Finder

Researchers using PubDNA Finder can perform three different types of queries: (1) sequence-based queries, (2) keyword-based queries and (3) combined queries. A detailed description of each type of query follows.

Sequence-based Queries

Sequence-based queries (SBQs) are targeted at retrieving all articles mentioning the DNA/RNA sequences specified by the user. Users can perform two different types of SBQs: simple and complex, depending on how the target sequences are specified.

Simple SBQs involve one or more DNA/RNA sequences linked by a single logical operator. Sequences are represented as strings composed of symbols belonging to the IUPAC standard

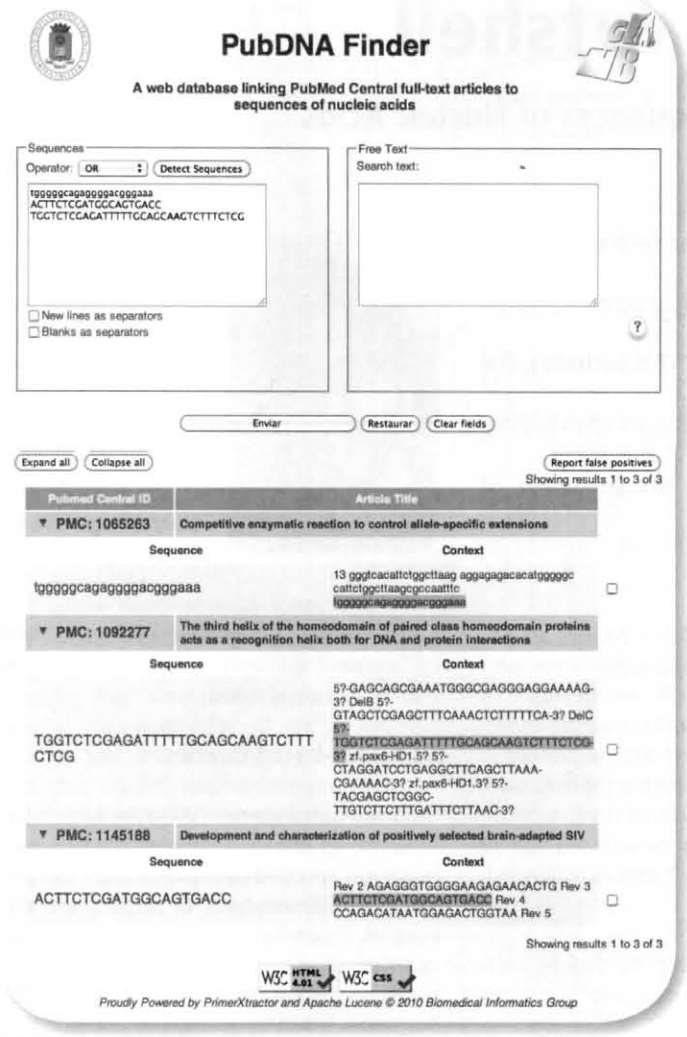

Fig. 1: A screenshot showing the results of the execution of a sample simple SBQ.

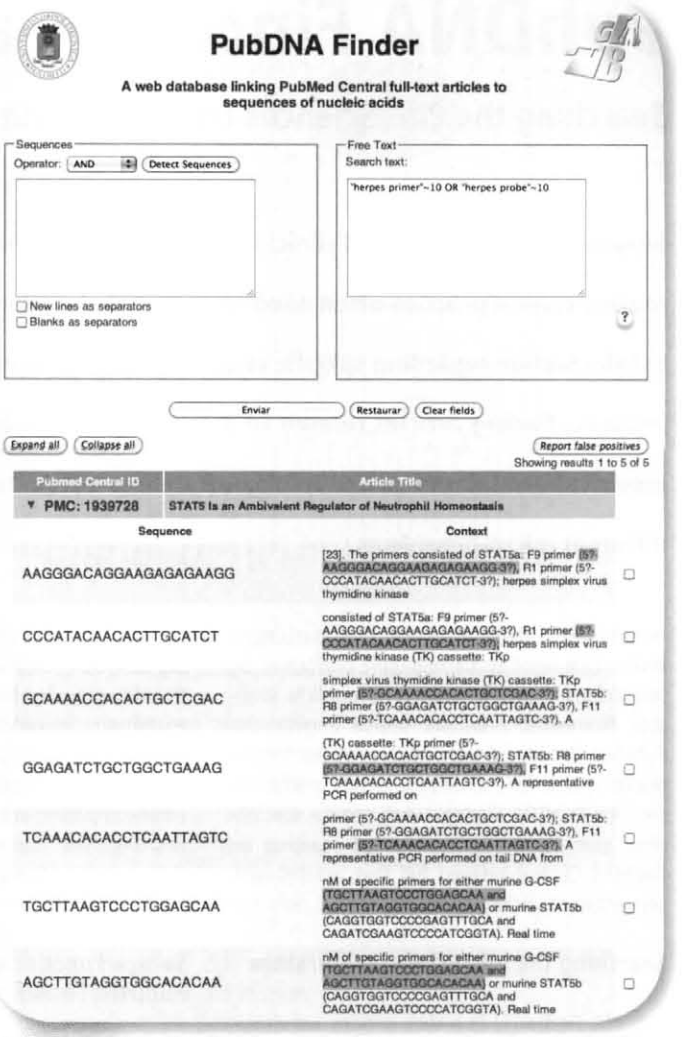

Fig. 2: A screenshot showing the results of the execution of a sample KBQ.

nucleotide codes. To execute a simple SBQ, we would have to type all the target sequences , one per line, in the text box labeled with "Sequences", select either the AND or OR operator in the "Operator" combo box, and click on the "Submit" button. For each hit in the results set, the user would be presented with the relevant information on the manuscript. This includes the PubMed Identifier (PMCID) associated with the article, the article's title, the genetic sequences – mentioned in the paper – that match the user query, the context in which each matched sequence occurs, and a link to the full text of the article. For instance, as shown in figure 1 if we launched the query "tgggggcagaggggacgggaaa OR actctcgatggcagtgacc OR tggtctcgagatttttgcagcaagtctttctcg", we would be presented with all papers in the database containing at least one of the three sequences specified in the query.

On the other hand, advanced SBQs involve complex sub-searches such as wildcard searches, fuzzy searches and proximity searches. We briefly describe each complex search type below.

Wildcard searches enable users to use the single and multiple character wildcard symbols, "?" and "*" respectively, to define patterns for matching the target sequences. For instance, the sample query "cga?ttg OR tta*" would retrieve papers containing sequences such as "cgacttg" or "ttatttcc".

By contrast, fuzzy searches are aimed at performing approximate matching by retrieving manuscripts containing sequences that are "similar" to these specified in the query. The similarity between two sequences is calculated using the Levenshtein Distance [2]. These searches can be performed by appending a tilde character at the end of the target sequence. It is also possible to optionally specify a similarity threshold. The latter is a value between 0 and 1. The greater the threshold, the more similar are the matched sequences to the target sequence. For instance, if we issued the query "cgattg~0.6", we would retrieve articles containing sequences such as "ctgatcg" or "tgcattg". Conversely, if we executed the query "cgattg~0.8", we would retrieve papers containing sequences such as "cggattg" or "cgacttg".

Proximity searches are aimed at retrieving articles that contain two specific sequences which are within a given distance, i.e. a number of words, away. Proximity searches can be performed by enclosing the target sequences between double quotes and appending the tilde

character plus the distance threshold after the last double quote character. For instance, the query „cac-ctttgaaaacgctacttcagacgct tcattcttgctgtttgtg"~3 would retrieve the article with PMID 2374257, which mentions both target sequences within a distance of two words – note that the original query requires both sequences being at a distance of at most three words.

Keyword-based Queries

Keyword-based queries (KBQs) are aimed at retrieving all DNA/RNA sequences mentioned in papers matching the search terms, a functionality that is also missing in the PMC search engine. KBQs are composed of either keywords or phrases – i.e. sequences of keywords enclosed between double quotes – linked by explicitly using the AND and OR logical operators. For instance, the KBQ 'probe OR probe AND "E. coli"' would retrieve all the sequences mentioned in articles that contain the phrase "E. coli" and either the word "primer" or "probe" – or both. It is also possible to use wildcard, fuzzy and proximity modifiers in KBQs if required. For instance, to search for primer/probe sequences for the Herpes virus, we could execute the following KBQ '"herpes primer"~10 OR "herpes probe"~10'. As shown in figure 2, the system would return all the sequences mentioned in articles in which the word "herpes" co-occurs either with "primer" or "probe" within distance 10.

Combined Queries

Combiner queries (CQs) combine the results of a SBQ and a KBQ by means of an AND operation, thus retrieving the records of all articles matching both queries. For each hit in the results set – i.e. papers matching the KBQ and containing any sequence specified in the SBQ – , the system presents the user with the article's PMCID, its title, a link to its full text and a list of the sequences mentioned in the article that match the SBQ, together with the context in which they occur. Figure 3 shows the result of executing a combined query that is aimed at determining whether there are any sequences beginning with either "CTTCTAAC" or "ATAGTTC" that are somehow connected to the H1N1 virus or the swine flu disease.

Additional Features

PubDNA finder provides other additional features, such as automatically identifying and extracting all sequences mentioned in a plain-text document provided by the user, or retrieving all sequences mentioned in a concrete article identified by its PMCID.

References

[1] García-Remesal M. et al.: Bioinformatics 26(21), 2801–2802 (2010)

[2] Levenshtein V.I.: Soviet Physics Doklady 10, 707–10 (1966)

[3] García-Remesal M. et al.: BMC Bioinformatics 11, 410 (2010)

Contact

Dr. Miguel García-Remesal

Departamento de Inteligencia Artificial

Facultad de Informática

Universidad Politécnica de Madrid

Campus de Montegancedo S/N

Madrid, Spain

mgarcia@infomed.dia.fi.upm.es

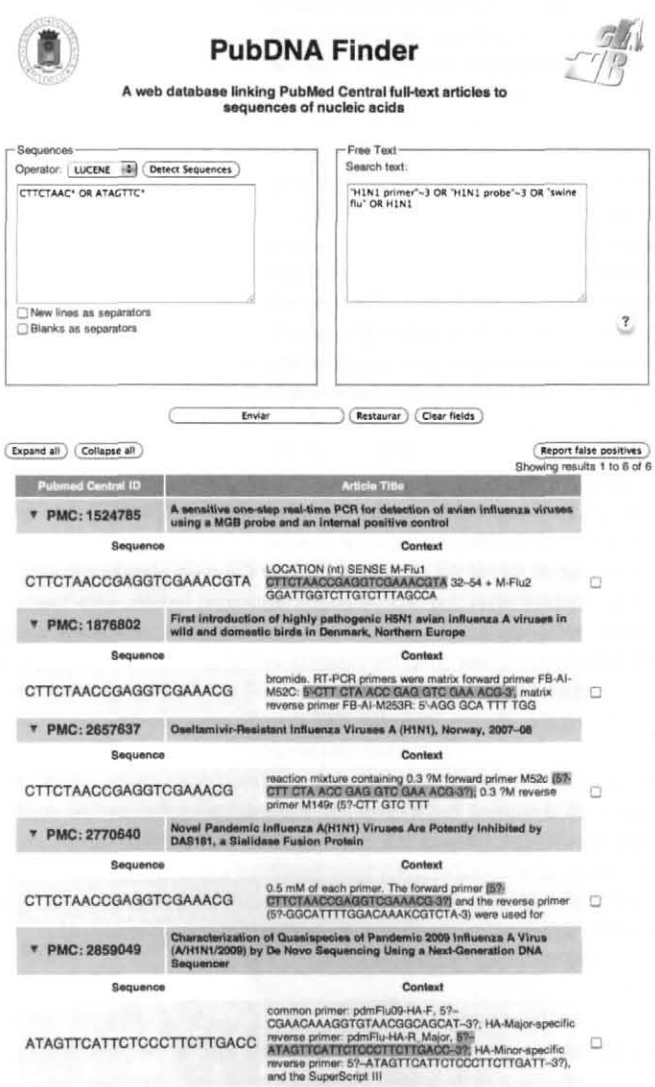

Fig. 3: A screenshot showing the results of the execution of a sample CQ.