Improving Acronym Searches on PubMed

Kevin Williams and Garrett M. Dancik (Faculty Mentor)
Department of Computer Science,
E-mail: williamskev@my.easternct.edu and dancikg@easternct.edu
Eastern Connecticut State University, Willimantic, Connecticut, USA

Introduction

PubMed is a widely used online repository containing >30 million biomedical literature citations. Currently, many PubMed queries are mapped to Medical Subject Headings (MeSH), a controlled vocabulary derived by biomedical experts to categorize literature. Unfortunately, this coverage does not extend to many acronyms, which can yield incomplete and potentially undesired results. Here we present the PubMed Acronym Detector (PAD), a Google web extension that allows users to include acronyms in their searches. PAD currently maps 4,494 acronyms to 5,097 MeSH IDs and 6,158 MeSH terms.

PAD checks PubMed search queries to detect whether a MeSH-related acronym has been used. If so, the user can select a phrase, click 'search', and re-query PubMed using the specified MeSH term. For example, a researcher may search for 'next-generation sequencing' using 'NGS', yielding ~16,000 results based solely on the presence of this string in the text. This search will return any article using the acronym 'NGS' but will not include results for 'next-generation sequencing' that do not also include 'NGS'. Using PAD, results for the modified query will include all articles tagged as being related to 'next-generation sequencing", and will return more than twice as many articles (~37,000 results). Additional words in the search query can be specified to further narrow down the results. For example, adding 'bladder cancer' to a query containing 'NGS' returns 60 articles, but yields 115 results if PAD is used.

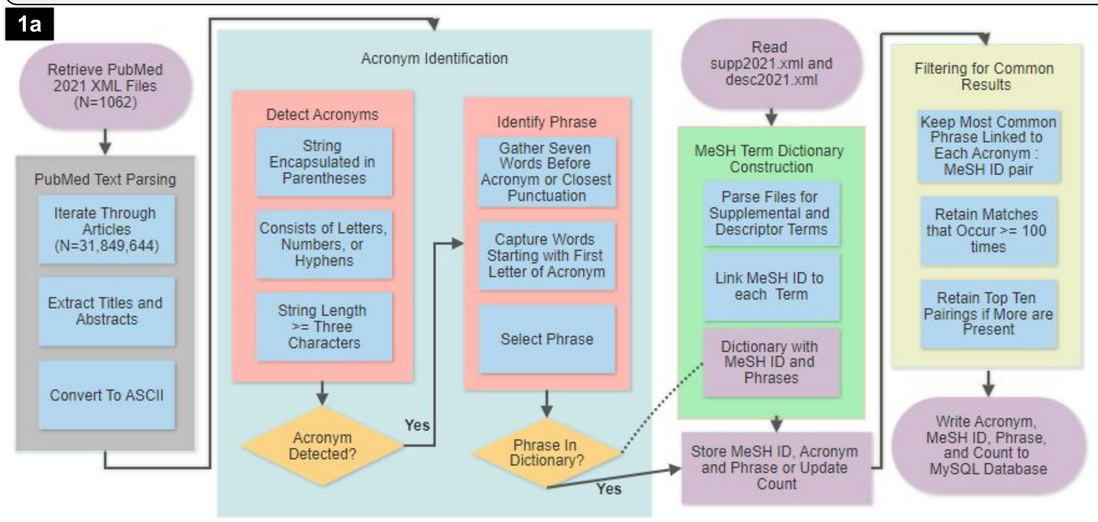
Availability:

- Processing: https://github.com/kewilliams86/acronym_detector (**Fig. 1**)
- Extension: https://github.com/kewilliams86/pubmed_extension (**Fig. 2**)

Methods

- PubMed title/abstracts (N=31,849,644) were downloaded from https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
- MeSH terms (desc2021.xml and supp2021.xml) were downloaded from https://www.nlm.nih.gov/databases/download/mesh.html
- Python was used to download PubMed baseline XML files (using *ftputil* module), extract PubMed text (*lxml* module), parse MeSH data (*xml* module), gather acronym/phrase data (*re* module), and write to the database (*mysql* module).
- Python code for PubMed retrieval and extracting PubMed text is modified from: https://github.com/kewilliams86/SummerBio
- Acronym, phrase, count, and corresponding MeSH IDs were stored in a single table in a MySQL database, with an index on the acronym.
- The web extension is written in JavaScript, CSS, and HMTL

Results



1b

Potential variants of the genes associated with CMT were screened by next-generation sequencing (NGS) of the members of the pedigree → next-generation sequencing (NGS)

Figure 1. Overview of MeSH-related Acronym Detection

(a) PAD parses PubMed XML files, identifies acronyms and related phrases, attempts to match them to a dictionary containing phrases and MeSH IDs, retains the acronym, MeSH ID, count, and phrase, filters results, and stores them in a database.

(b) Sample sentence with NGS as acronym and "next-generation sequencing" as the phrase. A regular expression was used to identify acronyms and the seven previous words, to locate words that start with the first letter of the acronym, identify the appropriate starting word, and gather the phrase.

(c) Sample MeSH data. All terms are linked to the Unique ID (MeSH ID).

Example: The phrase identified in **(b)** is an entry term in **(c)**, and is relatively common. Therefore, the acronym NGS, the phrase next-generation sequencing, and MeSH ID D059014 are stored in the database. A user searching PubMed for the acronym will be able to find all articles associated with the MeSH term (see **Fig. 2**)

1c

High-Throughput Nucleotide Sequencing MeSH Descriptor Data 2021

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading	High-Throughput Nucleotide Sequencing
Tree Number(s)	E05.393.760.319
Unique ID	D059014
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D059014
Annotation	coordinate with SEQUENCE ANALYSIS, DNA or SEQUENCE ANALYSIS, RNA
Scope Note	Techniques of nucleotide sequence analysis that increase the range, complexity, sensitivity, and accuracy of results by greatly increasing the scale of operations and thus the number of nucleotides, and the number of copies of each nucleotide sequenced. The sequencing may be done by analysis of the synthesis or ligation products, hybridization to preexisting sequences, etc.
Entry Term(s)	Deep Sequencing High-Throughput DNA Sequencing High-Throughput RNA Sequencing High-Throughput Sequencing Illumina Sequencing Ion Proton Sequencing Ion Torrent Sequencing Massively-Parallel Sequencing Next-Generation Sequencing Pyrosequencing
Previous Indexing	Sequence Analysis or specifics (1998-2010)
Public MeSH Note	2011
History Note	2011
Date Established	2011/01/01
Date of Entry	2010/06/25
Revision Date	2020/01/16

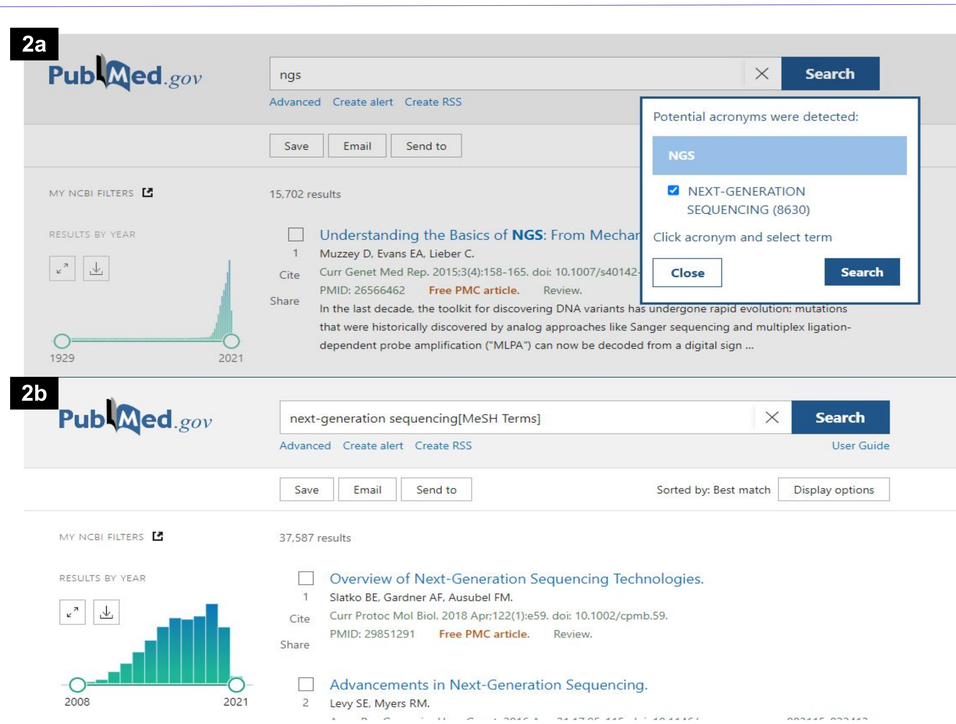


Figure 2. Screenshots of PubMed Acronym Detector.

(a) When searching PubMed using PAD, a dialog box will appear if a biomedically relevant (MeSH-related) acronym is detected. Upon clicking the accordion button for the desired acronym(s), the potential phrase(s) are revealed which have an associated checkbox. The user can choose to either close the dialog box without altering the search, or they can click the search button to re-query the server using the appropriate MeSH term.

(b) The updated query and results after clicking the search button in **(a)**.

Note that using PAD finds >37,000 articles associated with NGS while a standard search finds only ~15,000 articles that contain the acronym.