Journal of Biomedical Informatics 44 (2011) 919–926

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

MeSHy: Mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms

T. Theodosiou [a,c], I.S. Vizirianakis [b], L. Angelis [c], A. Tsaftaris [a,d], N. Darzentas [a,*]

[a] Institute of Agrobiotechnology, Centre for Research and Technology – Hellas (CERTH), P.O. Box 361, 6km Charilaou-Thermis, GR-57001, Thessaloniki, Greece
[b] Laboratory of Pharmacology, Department of Pharmaceutical Sciences, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece
[c] Department of Informatics, School of Natural Sciences, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece
[d] Department of Genetics and Plant Breeding, Aristotle University of Thessaloniki, GR-54006, Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 3 November 2010
Accepted 31 May 2011
Available online 13 June 2011

Keywords:
PubMed
MeSH terms
Ontology
Imatinib

ABSTRACT

Motivation: PubMed is the most widely used database of biomedical literature. To the detriment of the user though, the ranking of the documents retrieved for a query is not content-based, and important semantic information in the form of assigned Medical Subject Headings (MeSH) terms is not readily presented or productively utilized. The motivation behind this work was the discovery of unanticipated information through the appropriate ranking of MeSH term pairs and, indirectly, documents. Such information can be useful in guiding novel research and following promising trends.

Methods: A web-based tool, called MeSHy, was developed implementing a mainly statistical algorithm. The algorithm takes into account the frequencies of occurrences, concurrences, and the semantic similarities of MeSH terms in retrieved PubMed documents to create MeSH term pairs. These are then scored and ranked, focusing on their unexpectedly frequent or infrequent occurrences.

Results: MeSHy presents results through an online interactive interface facilitating further manipulation through filtering and sorting. The results themselves include the MeSH term pairs, along with MeSH categories, the score, and document IDs, all of which are hyperlinked for convenience. To highlight the applicability of the tool, we report the findings of an expert in the pharmacology field on querying the molecularly-targeted drug imatinib and nutrition-related flavonoids. To the best of our knowledge, MeSHy is the first publicly available tool able to directly provide such a different perspective on the complex nature of published work.

Implementation and availability: Implemented in Perl and served by Apache2 at http://bat.ina.certh.gr/tools/meshy/ with all major browsers supported.

1. Introduction

The increasing rate of publications entering biomedical literature databases, in conjunction with their differential source of production (clinical, pharmacological and genetic or genomic), make the process of relevant knowledge extraction difficult and time-consuming even for specialists. Furthermore, fast, accurate, reliable and user-friendly retrieval of biomedical information is in increasing demand for most disciplines in health-related areas (e.g. drug discovery and delivery, diagnosis and therapy in clinical practice, clinical translation of genomics data). Thus, it has become imperative to develop automated or semi-automated text-mining tools that are efficient and effective in extracting relevant information applicable in medicine, pharmacy and biology [1]. The same tools are critical in highlighting important associations between disease pathogenesis, drugs and their pharmacodynamic or pharmacokinetic parameters of action, (including drug interactions or adverse drug reactions), and their molecular targets (e.g. genes, genetic polymorphisms, proteins, receptors, drug metabolizing enzymes and transporters) [2–5]. However, upon searching the biomedical literature, critical connections within and across the various sources of information are often lost.

PubMed is by far the most popular gateway for access to biomedical literature. Despite several algorithms used by the PubMed retrieval system (like [6,7]), its effectiveness and impact on research is heavily dependent on the query itself. Study [8] describes methods to facilitate the formulation of precise queries using more relevant terms, showing that queries are usually up to three terms long, while study [9] indicates that the use of four to five terms is most likely to result in reading the abstracts of the retrieved titles. Similarly, many questions go unanswered due to lack of skills in formulating the necessary queries [10].

Another characteristic of the PubMed document retrieval system is that the semantic information contained in the Medical Subject Headings (MeSH) terms [11] is not readily presented to the

* Corresponding author. Fax: +30 2310 498 270.
E-mail address: ndarz@certh.gr (N. Darzentas).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.05.009

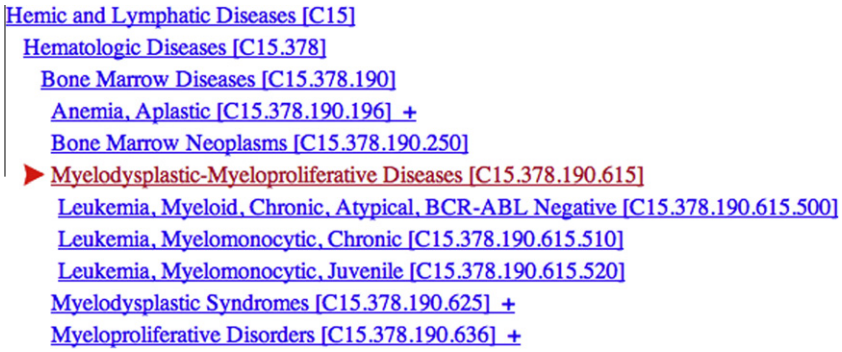

Fig. 1. An example of MeSH hierarchy. Inside the brackets is the number that indicates the location of each MeSH term in the hierarchy.

user and therefore is often lost. MeSH terms form an ontology used for indexing and annotating PubMed documents. According to [12] an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The MeSH ontology is a structured vocabulary of semantically related terms, built for facilitating computational tasks. The more general MeSH terms appear at the top of the hierarchy whereas more specific terms appear at the bottom (Fig. 1). In PubMed, MeSH terms are manually assigned to each document by indexers (biomedical subject specialists) based on the context of the whole document and not only the abstract and/or the title. Thus, they contain high-density information from the whole document which may not be inferred from the title or the abstract that PubMed returns.

The importance of MeSH terms in knowledge discovery has been widely recognized and several tools have been based on their utilization. GoPubMed [13] is a system that addresses the issue of clearly presenting to the end-user the semantic information from MeSH terms through extensive use of them (along with Gene Ontology terms) for the presentation and analysis of the results of a PubMed query. In MiSearch [14] MeSH terms are used to build statistical models while in MScanner [15] and in Suomela and Andrade [16] they are used to create the training sets for classifiers. Also, Zhu et al. [17] and Struble and Dharmanolla [18] describe a method to enhance document clustering by using MeSH term semantic similarity. Furthermore, in our previous work [19] we have shown that the information contained in MeSH terms is highly correlated with the information contained in the title and abstract of each document and can therefore facilitate document classification and clustering.

MeSH term concurrence has been used as early as 1989 [20] for helping physicians to explore biomedical literature relevant to hepatological patient records and in [21] for obtaining medical knowledge via automated analysis of literature and using it to build medical knowledge bases. The MeSHMap tool [22] is a tool similar to GoPubMed which uses only MeSH terms and allows the user to manually explore term concurrences. An important semi-automated tool is also Arrowsmith [23] which uses MeSH term concurrence in order to associate two different sets of PubMed articles. Another approach used specifically for PubMed documents is sentence-level concurrence of query keywords used in Relemed [24]. FACTA [25] is a text search engine specific for MEDLINE abstracts that uses concurrence statistics and MeSH terms, along with other concepts from UniProt [26], KEGG [27], etc. in order to rank retrieved information. Finally, a recent addition is Epiphanet [28], an interactive knowledge discovery system that allows end-users to explore relations between concepts extracted from biomedical literature.

Significant efforts are being made to personalize information retrieval and ranking processes to assist the users in getting the information they need [29]. To accomplish this, methods either have to know the user's needs or provide customization for as many users as possible, both tasks being hard. Methods outlined above do indeed offer solutions towards both – e.g. Epiphanet not only adapts to the user by offering related concepts of the initial query, but also interactively allows for exploration of the resulting information.

However, there currently does not appear to be a system that directly promotes (scores higher than others) statistically unanticipated information through pairing of MeSH terms. This last point has been the motivation behind the development of MeSHy. That is, how to discover unusual or unanticipated information contained in MeSH term pairs and their relevant PubMed documents that usually remain out of reach of the average user due to the low date-based ranking of PubMed. Since the MeSH terms are essentially carriers of documents' information, the key idea for addressing the research question was to locate and reveal documents that contain unusually frequent occurrences of MeSH term pairs. MeSHy is a system for statistically scoring and ranking MeSH term pairs from a simple user query to PubMed. The tool is freely available as a web application.

In the following sections, we first describe the different stages of the methodology in detail, we then include an application report from an expert group in pharmacology, and finally we discuss implications, possible applications, limitations, and future plans.

2. Methodology

In summary, MeSHy is an implementation of an algorithm that extracts the MeSH terms from the retrieved documents, filters them excluding the trivial ones and then probabilistically scores and ranks pairs of MeSH terms derived from each document. Indirectly, the scored MeSH term pairs are used to rank the documents themselves. The filtering is performed in two stages and its purpose is to keep the most informative and descriptive MeSH terms of the query.

The main stages of the methodology are (also depicted in Fig. 2):

2.1. Stage *i*

In Stage *i* a query is submitted to the PubMed retrieval system and the relevant documents are retrieved.

2.2. Stage *ii*

Stage *ii* involves the extraction of the MeSH terms, and optionally of the chemical terms (hereafter implied under MeSH terms), from the documents and the construction of all possible MeSH term pairs of each document. Chemical terms are part of the MeSH