Genome Medicine

RESEARCH HIGHLIGHT

Mining the literature: new methods to exploit keyword profiles

Miguel A Andrade-Navarro*

See related research articles: http://www.biomedcentral.com/1471-2105/13/249/abstract
http://genomemedicine.com/content/4/9/75/abstract

Abstract

Bibliographic records in the PubMed database of biomedical literature are annotated with Medical Subject Headings (MeSH) by curators, which summarize the content of the articles. Two recent publications explain how to generate profiles of MeSH terms for a set of bibliographic records and to use them to define any given concept by its associated literature. These concepts can then be related by their keyword profiles, and this can be used, for example, to detect new associations between genes and inherited diseases.

Keywords Data mining, databases, genes, disease, drugs

Research highlight

Biomedical concepts such as genes, diseases and drugs are described in multiple interrelated databases, which include links to relevant literature and keyword annotations. Data mining can be applied to these databases to infer novel associations between such concepts; for example, to assess whether mutations in a gene could cause a certain phenotype, or whether a drug could interact with a protein. Specifically, it is possible to create keyword profiles for biological concepts and then to use those profiles to relate concepts by the similarity of their profiles. This is the basis of methods such as ENDEAVOR [1], coPub [2], or the method developed by Kumar [3], among others.

In two companion articles recently published in *BMC Bioinformatics* [4] and in *Genome Medicine* [5], Canada-based computational biologists Warren Cheung, Francis Ouellette and Wyeth Wasserman describe a way to generate keyword profiles based on headings from PubMed called Medical Subject Heading (MeSH) terms, and to use them to infer associations between genes and inherited diseases, respectively. Pre-computed profiles for human genes, diseases and chemicals, as well as pre-computed associations between human genes and diseases are available on the Medical Subject Heading Over-representation Profiles (MeSHOPs) website [6]. The new method can also be used to generate a profile for any set of PubMed records depending on the MeSH terms attached to those records, and thereafter, comparisons are possible between any concepts that can be related to a set of bibliographic records via profile comparisons (Figure 1).

The value of curators

These new publications and many others that use data and text mining to try to infer biological knowledge by computational means rely on databases maintained at the United States National Library of Medicine (NLM) [7]; in this case, PubMed and Entrez Gene. Structured database records, annotations, and cross-linking between databases not only facilitate manual querying of the data but also allow such computational analyses.

In particular, the annotation of each PubMed record with MeSH terms that summarize its content requires a monumental amount of human effort, as it necessitates understanding the main points of each new PubMed entry; and currently, new PubMed records are being created at a rate of 3,000 per day.

Moreover, MeSH terms are a controlled vocabulary arranged hierarchically in 16 categories; they are complex and numerous (at present, there are more than 26,000). To make matters worse, new MeSH terms are sometimes added or old ones are modified, so that old records in PubMed may need re-annotation.

However, it should be noted that although using MeSH terms for querying PubMed records results in more specific retrieval than using words found in abstracts [8], querying with MeSH terms is an option seldom chosen by PubMed users [9]. So ultimately, it is possible that the

*Correspondence: Miguel.andrade@mdc-berlin.de
Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125, Berlin, Germany

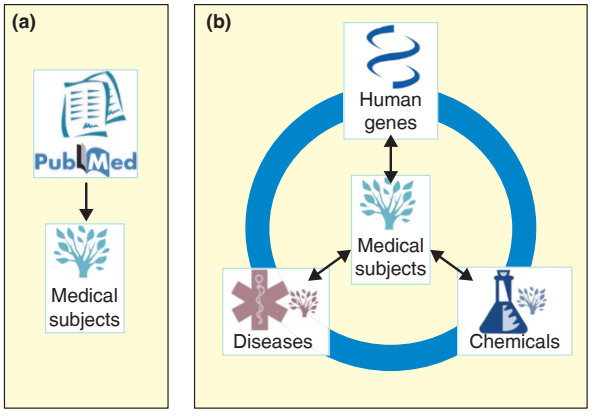

Figure 1. Comparing biological concepts using keyword profiles. (a) A set of references from PubMed can be associated with a profile of keywords according to the Medical Subject Heading (MeSH) terms assigned to them. By extension, this can be done for any concept or database record with an associated set of PubMed references. **(b)** Entrez (human) genes, diseases (defined as MeSH terms) or chemicals (defined as MeSH terms) can be associated with bibliographic records in PubMed, and therefore can be defined by profiles based on any MeSH terms (middle). Conversely, it is possible to find the genes, chemicals and diseases associated with particular MeSH terms. Associations between concepts (genes and diseases, but also genes and chemicals, or diseases and chemicals) can be defined by comparing such profiles. This figure is partly taken, with permission, from the MeSHOPs website [6].

huge value of the work that the NLM curators do in creating and using MeSH terms lies in enabling computational mining methods such as the ones highlighted here.

Generating and comparing profiles

In the first publication [4], Cheung *et al*. describe how to build a profile out of a collection of PubMed references. Since these references are all annotated with MeSH terms, it is possible to identify the terms that are over-represented in the collection compared with their background prevalence in all publications in PubMed. Specifically, they measure over-representation by using *P*-values from a Fisher's exact test. The resulting profiles show how a particular collection of literature records defines something that is specific.

In their second publication [5], Cheung *et al*. apply the profiles generated to establish comparisons between genes and inherited diseases. The approach is familiar: similar genes can be expected to be associated with similar diseases [10]. Therefore, if a profile is established for a particular disease based on its associated bibliography, particular features would be expected to be highlighted that also appear in the profiles of genes related to the disease. Given a chromosomal region linked to an inherited disease by analysis of the genomes of patients and healthy controls, a novel candidate gene in that region might be identified based on previous associations.

One interesting point in the work of Cheung *et al.* is that the metrics they use to compare profiles do not take into account the similarity between whole profiles of enrichment values, but only the similarity between the enrichment values of the overlapping terms. This sets their method apart from other methods of keyword profile comparison: even a small, but significant, number of overlapping terms can be enough to highlight an association. It is easy then to go back to the original source papers (of which there may only have been three or four) that were annotated with the enriched MeSH terms for evidence of the association.

What next?

As shown in Figure 1, the way these profiles can be compared is by no means limited to genes and diseases. The web tool described in the highlighted manuscripts already supports the generation of profiles for three concept types (human genes, diseases and chemicals), but potentially allows the annotation of any concept (by their associated list of PubMed records). Similarly, any concept that is given one of these profiles can be compared to any other concept receiving another profile. So, there is no reason to stop at the gene-to-disease comparison: one could potentially examine associations between concepts of the same type (for example, genes with genes), or between other concepts (for example, protein sequence features and functions).

However, these calculations are computationally costly. This imposes limitations, precluding the authors from offering such general functionality. Rather, Cheung *et al.* focused on arguably the most pressing need; namely, to identify genes linked to disease. But they have clearly set out their method, and we can expect that this and other groups will use these ideas to create other metrics for the generation and comparison of keyword profiles, which may be even more efficient. Cheung *et al.* already give some ideas of how to benchmark such methods. These new developments are surely an encouragement for the support of continued efforts in high-quality annotation of large public databases.

Abbreviations

MeSH, Medical Subject Headings; NLM, National Library of Medicine.

Competing interests

The author declares that he has no competing interests.

Published: 30 October 2012

References

1. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic Acids Res* 2008, **36**:W377-384.
2. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W: **Literature**

mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010, **6:** e1000943.
3. Kumar V: **Omics and literature mining.** *Methods Mol Biol* 2011, **719:**457-477.
4. Cheung WA, Ouellette BF, Wasserman WW: **Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs).** *BMC Bioinformatics* 2012, **13:**249.
5. Cheung WA, Ouellette BF, Wasserman WW: **Inferring novel gene-disease associations using medical subject heading over-representation profiles.** *Genome Med* 2012, **4:**75.
6. **Medical Subject Heading Over–representation Profiles** [http://meshop.oicr.on.ca]
7. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40:**D13-25.
8. Jenuwine ES, Floyd JA: **Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals.** *J Med Libr Assoc* 2004, **92:**349-353.
9. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV: **A day in the life of PubMed: analysis of a typical day's query log.** *J Am Med Inform Assoc* 2007, **14:**212-220.
10. Tiffin N, Andrade-Navarro MA, Perez-Iratxeta C: **Linking genes to diseases: it's all in the data.** *Genome Med* 2009, **1:**77.

doi:10.1186/gm382
Cite this article as: Andrade-Navarro MA: **Mining the literature: new methods to exploit keyword profiles.** *Genome Medicine* 2012, **4:**81.