Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

Assessment of curated phenotype mining in neuropsychiatric disorder literature

Jean-Fred Fontaine [a,b], Josef Priller [c,d], Eike Spruth [c], Carol Perez-Iratxeta [e,f,*], Miguel A. Andrade-Navarro [a,b,*]

[a] Max Delbrück Center for Molecular Medicine, Germany
[b] Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany
[c] Charité – Universitätsmedizin Berlin, Department of Neuropsychiatry and Laboratory of Molecular Psychiatry, Berlin, Germany
[d] Cluster of Excellence "NeuroCure" and Berlin Institute of Health (BIH), Berlin, Germany
[e] The Sprott Center for Stem Cell Research, Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, ON K1H 8L6, Canada
[f] Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Article history:
Received 25 July 2014
Received in revised form 25 November 2014
Accepted 27 November 2014
Available online xxxx

Keywords:
Data mining
Text mining
Neuropsychiatric disorders
Clinical diagnostics
Data curation
Drug therapy

ABSTRACT

Clinical evaluation of patients and diagnosis of disorder is crucial to make decisions on appropriate therapies. In addition, in the case of genetic disorders resulting from gene abnormalities, phenotypic effects may guide basic research on the mechanisms of a disorder to find the mutated gene and therefore to propose novel targets for drug therapy. However, this approach is complicated by two facts. First, the relationship between genes and disorders is not simple: one gene may be related to multiple disorders and a disorder may be caused by mutations in different genes. Second, recognizing relevant phenotypes might be difficult for clinicians working with patients of closely related complex disorders. Neuropsychiatric disorders best illustrate these difficulties since phenotypes range from metabolic to behavioral aspects, the latter extremely complex.

Based on our clinical expertise on five neurodegenerative disorders, and from the wealth of bibliographical data on neuropsychiatric disorders, we have built a resource to infer associations between genes, chemicals, phenotypes for a total of 31 disorders. An initial step of automated text mining of the literature related to 31 disorders returned thousands of enriched terms. Fewer relevant phenotypic terms were manually selected by clinicians as relevant to the five neural disorders of their expertise and used to analyze the complete set of disorders. Analysis of the data indicates general relationships between neuropsychiatric disorders, which can be used to classify and characterize them. Correlation analyses allowed us to propose novel associations of genes and drugs with disorders. More generally, the results led us to uncovering mechanisms of disease that span multiple neuropsychiatric disorders, for example that genes related to synaptic transmission and receptor functions tend to be involved in many disorders, whereas genes related to sensory perception and channel transport functions are associated with fewer disorders. Our study shows that starting from expertise covering a limited set of neurological disorders and using text and data mining methods, meaningful and novel associations regarding genes, chemicals and phenotypes can be derived for an expanded set of neuropsychiatric disorders. Our results are intended for clinicians to help them evaluate patients, and for basic scientists to propose new gene targets for drug therapies. This strategy can be extended to virtually all diseases and takes advantage of the ever increasing amount of biomedical literature.

1. Introduction

Text and data mining of phenotype information from the biomedical literature have been used to characterize human diseases [1,2], and to find associations of diseases with genes and chemicals (e.g., drugs) [3–6]. However, the complexity of human phenotypes makes it difficult to obtain relevant results using this information;

* Corresponding authors at: Ottawa Hospital Research Institute, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada. Fax: +1 (613) 739 6294 (C. Perez-Iratxeta). Johannes-Gutenberg University of Mainz, Faculty of Biology, Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany. Fax: +49 6131 39 21589 (M.A. Andrade-Navarro).

E-mail addresses: cperez-iratxeta@ohri.ca (C. Perez-Iratxeta), Andrade@uni-mainz.de (M.A. Andrade-Navarro).

http://dx.doi.org/10.1016/j.ymeth.2014.11.022
1046-2023/

this is particularly problematic for neuropsychiatric disorders, which often include a variety of complex alterations of behavior.

A widely used source for methods that predict gene function based on human phenotype information is the Online Mendelian Inheritance in Man® database (OMIM), which also contains information of known genes and genomic regions linked to phenotypes. Studies using OMIM have shown for example that clusters of similar phenotypes are related to genes that are more likely to be functionally similar [7] or to physically interact between each other [8]. Yet, the database is focused on genetic disorders and diseases, and thus it is not appropriate to study human phenotypes accurately. Human phenotypes have been defined more specifically, for example in the MeSH database [9] and the Human Phenotype Ontology [10]. These resources can be used to annotate and study disorders (e.g., [11–13]). Yet, they may not be comprehensive enough to study alterations of behavior, and, while the sets of terms themselves (e.g., MeSH) may be comprehensive, available consensus phenotypic annotations of disorders with these terms may be unbalanced between disorders.

The characterization of neuropsychiatric disorders, which can have multiple symptoms of varied severity, with some of them present only in a fraction of patients, requires sub-classifying the variants of each disease. However, the current resources annotating neuropsychiatric disorders with keywords representing their associated phenotypes are incomplete, which complicates this classification task. Automatic text mining algorithms based on these resources are not optimal for such analyses and this prevents the study of such less frequent phenotypes or disorders.

Here we have used the assistance of clinician experts to help with the selection of relevant phenotypic terms from biomedical documents in the field of neuropsychiatric disorders. Drawbacks of this approach are the time that clinicians can spend in the task, and their limited breadth of experience. To overcome these limitations, we hypothesized that terms (concepts) found manually to be good descriptors for a concise set of disorders can be advantageous for the automated annotation of a larger set.

In a first step, we used a semi-automated method to extract informative phenotype terms associated with five neurodegenerative disorders (Alzheimer's Disease, Parkinson's Disease, Spinocerebellar Ataxia, Amyotrophic Lateral Sclerosis, Huntington's Disease) for which a large amount of bibliography exists, including genetic information and therapeutic drugs, and for which we could manually evaluate their relevance based on our clinical expertise.

The selected set of informative phenotypic terms was then applied to 26 further neuropsychiatric disorders with enough associated literature in the PubMed database [9]. We then associated genes and drugs with the 31 disorders by data mining the corresponding PubMed abstracts. Finally, we further associated other genes and drugs with these 31 disorders by their correlation with phenotypes associated with these disorders.

Our approach illustrates how expert curation of phenotypic terms might be of great help to increase the value of text and data mining outputs in order to retrieve or predict relevant associations between phenotypes and diseases.

2. Methods

2.1. Literature data

The complete XML version of PubMed [9] was stored locally. English abstracts were extracted and stored in a MySQL database after part-of-speech processing used to select only nouns (TreeTagger, Helmut Schmid, University of Stuttgart). If existing, MeSH annotations were also extracted and stored. A stop word list was used to remove common and non-meaningful terms, such as "the" and "a".

2.2. Predictions of word-, gene- and chemical-disorder associations

In order to predict the association of a word with a disorder, we compared word usage in a set of abstracts related to the disorder (positive set) to the rest of PubMed (background set). The positive sets were constructed by selecting all the abstracts annotated with the MeSH term of the corresponding disorder (Table 1). As there was no MeSH term corresponding to corticobasal degeneration (CD), we used 930 abstracts returned by a PubMed keyword search. For each word, we computed the number of its occurrences in the positive set and in the background set. We did not take into account multiple occurrences of a word within an abstract, and we discarded low occurrence words (less than 10 occurrences in each set).

In order to predict the association of a gene with a disorder, we used the Génie text-mining tool [14], which is able to prioritize all human genes (defined as records in the Entrez database [9]) for a disorder given a positive set of abstracts possibly retrieved from a MeSH term or a PubMed query. For each disorder, the positive set was retrieved using corresponding MeSH terms or a PubMed query for CD, the background set was all PubMed, and all human protein-coding genes were ranked.

In order to predict the association of a chemical with a disorder, we used the Alkemio's web service [6], which is able to prioritize chemicals for a disorder given a positive set of abstracts as retrieved from association with a MeSH term or a PubMed query. For each disorder, the PubMed positive set was formed by the abstracts associated with the corresponding disease MeSH term or by those resulting from a PubMed query in the case of CD. The background set was the complete PubMed. All PubMed abstracts published during the last 3 years were used to find chemicals and to rank them.

Fisher's exact tests were performed to compute *P*-values for word-, gene- and chemical-disorder associations. Association scoring values were defined as $-\log(P\text{-value})$.

2.3. Inferring further associations of chemicals and genes to disorders

We applied further data mining to the association data created in the literature mining steps with Genie and Alkemio (see the above paragraphs). Given a disorder X, we first calculated the average of all non-zero phenotype association scoring values in order to consider only phenotypes with association scores above average. This filter was applied to use only the most strongly associated phenotypes for this analysis. Next, we calculated the correlation across the other 30 disorders between the phenotypes associated with disorder X and chemicals (not associated with disorder X) considering only phenotype-chemical pairs associated with three or more disorders; *P*-values were calculated according to the asymptotic *t*-distribution and corrected (Benjamini & Yekutieli). There is no further global correction for the whole set of disorders. At a cut-off of 0.05 for the *P*-values we obtained 61 extra associations of chemicals with disorders. Thus, capturing the strong relation between an individual phenotype (symptom) and a drug may suggest the novel use of the drug to treat a disease for which the phenotype is prevalent.

A similar approach was used to infer novel associations of genes with diseases. This resulted in 239 novel associations.

2.4. Deriving communities of phenotypes

We started with a graph constructed from the association data, where vertices are the 292 phenotypes. If two phenotypes are very