論文の内容の要旨

論文題目　A new computational method to assist interpreting genomic variants of rare diseases based on disease-phenotype associations extracted from case reports

（症例報告由来の疾患−症状関連情報を用いた希少疾患ゲノムの変異解釈支援手法の開発）

氏　名　藤原　豊史

Introduction

At present over 6,000 rare diseases have been identified, and ~80% of them are genetic in origin. Unfortunately, a quarter of rare disease patients waited 5 to 30 years for a diagnosis, and 40% of rare disease patients were misdiagnosed at first. Such patients will likely lose opportunities such as optimization of clinical management and early intervention. To tackle this situation, next-generation sequencing (NGS)-based analysis is being undertaken to identify candidate diseases for undiagnosed patients. Following analysis, clinicians rank candidate diseases through a differential diagnosis process based on symptoms and signs observed in the patient that are collectively called "phenotypes".

Even though the analysis and process improve diagnostic rates, the differential diagnosis process is time-consuming. Recently, to speed up the process, phenotype-driven differential diagnosis systems have been implemented. These systems provide a ranked list of diseases or genes based on the similarity score, and the top-listed diseases represent the most likely differential diagnosis. The performance of these systems is greatly influenced by the quantity and quality of underlying databases of disease–phenotype associations (DPAs). Note that these databases rely on manual curation and show a limited coverage. In the case of DPA database from Orphanet, more than half of the rare diseases (~60.5% of 6,268) are not associated with a phenotype. The development of DPA databases is based on the curation of papers by human experts, which is time-consuming and labor-intensive because of the large volume and rapid growth of life sciences papers.

To improve the performance of phenotype-driven differential diagnosis systems, the limited coverage problem of DPA databases needs to be overcome, which is an important problem for diagnosis of rare disease. In this study, to address the problem, we empirically explore one question in a large scale: Can automatically extracted DPAs from case reports contribute to improving the performance of phenotype-driven differential diagnosis systems for rare diseases? First, we extract DPAs from case reports in PubMed using a text mining approach and compare those with DPAs from Orphanet. Second, we develop a new phenotype-driven differential diagnosis system PubCaseFinder and demonstrate that automatically extracted DPAs without manual screening can contribute to improve the performance of automated differential diagnosis.

Material and Methods

We used PubMed E-utilities to obtain a large collection of case reports and used the following query to collect case reports and record titles and abstracts: “case reports” [Publication Type] OR “case reports” [ti] OR “case report” [ti]. We found that 1,895,021 PubMed entries were initially collected as case reports, among which only 1,083,283 had both titles and abstracts (as of July 20, 2017). We extracted DPAs from our collection of case reports using a text mining approach. At first, we annotated titles and abstracts of case reports with Human Phenotype Ontology (HPO; releases/2017-06-30) terms and Orphanet Rare Disease Ontology (ORDO; version 2.3) terms using ConceptMapper with HPO and ORDO. Using the processed annotations, we identified DPAs that are co-occurrences of an ORDO term and an HPO term within a sentence.

We developed PubCaseFinder based on a DPA database where phenotypes are associated with diseases defined in Orphanet. Some of the DPAs are from Orphanet, whereas some originate from text mining results. PubCaseFinder takes as input a set of HPO terms that describe the signs and symptoms of the patient. The set of HPO terms is then compared with diseases in the database. Note that each disease in the database is also represented by a set of HPO terms. Thus, the comparison is performed as a similarity computation between two sets of HPO terms. As a result, PubCaseFinder outputs a ranked list of candidate diseases according to the similarity score. PubCaseFinder uses the GeneYenta algorithm represents the similarity ranging from 0% for no phenotypic overlap to 100% for complete phenotypic overlap.

Results

We annotated titles and abstracts of 1,083,283 case reports with HPO terms and ORDO terms and identified DPAs that are co-occurrences of an ORDO term and an HPO term within a sentence. As a result, 810,705 case reports were annotated with 6,380 HPO terms and 316,674 case reports were annotated with 3,788 ORDO terms. Using these annotations, we identified 70,011 DPAs consisting of 3,881 HPO terms and 3,072 ORDO terms. We also obtained 51,590 DPAs consisting of 4,832 HPO terms and 2,478 ORDO terms from Orphanet. Figure 1 shows the overlap between the two sets of ORDO terms included in DPAs from case reports and from Orphanet. We found that 1,483 ORDO terms were common to the two data sources and 1,589 ORDO terms included in DPAs from case reports were not found in DPAs from Orphanet.

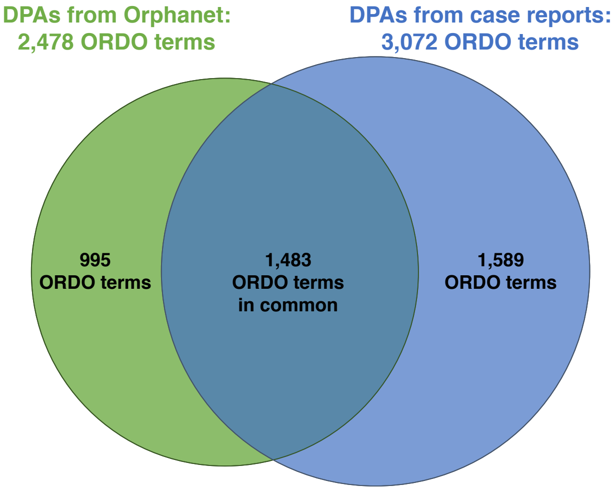

Figure 1. Overlap between two sets of ORDO terms found in disease–phenotype associations (DPAs) from Orphanet and from case reports

Within the overlapping 1,483 ORDO terms, we compared 40,512 DPAs from case reports with 35,172 DPAs from Orphanet. We regarded ORDO terms as the same if their related HPO terms were located in the same, superordinate, or subordinate part of the ontology hierarchy. As a result, 11,593 DPAs were in common, and 28,919 new DPAs were added to 1,483 rare diseases included in DPAs from Orphanet. We also identified 29,499 DPAs for 1,589 rare diseases that are not associated with a phenotype in Orphanet. In total, our text mining approach could identify 58,418 new DPAs and increase the coverage of DPAs in Orphanet by 113.2%.

To evaluate the performance of PubCaseFinder as a phenotype-driven differential diagnosis system, we collected 1,584 clinical cases from PhenomeCentral, which were registered by the Care4Rare Canada Consortium. It turned out only 243 cases out of them had both phenotypes and diagnoses, the former represented by HPO terms and the latter represented by MIM IDs. We used them as the test cases of our evaluation. All MIM IDs of the cases were converted to Orpha numbers using connections between MIM IDs and Orpha numbers in ORDO. To evaluate the effect of DPAs form case reports, we compared the performance of PubCaseFinder in three different settings: one with DPAs only from Orphanet (PubCaseFinder-O), one with DPAs only from case reports (PubCaseFinder-CR), and one with DPAs from both (PubCaseFinder-O/CR). For a reference purpose, we included Orphamizer (a customized system of Phenomizer for Orphanet) in our comparison because it was the most comparative system among phenotype-driven differential diagnosis systems, using DPAs from Orphanet and targeting the diseases defined in ORDO. For the evaluation, we compiled two exclusive sets of diseases as targets of differential diagnosis; one consisted of 2,323 diseases that were associated with phenotypes in Orphanet and consequently could be potentially solved by both PubCaseFinder and Orphamizer (Target-A), the other consisted of 1,589 diseases that were not associated with a phenotype in Orphanet (Target-B).

First, we evaluated the performance of PubCaseFinder (in the three different settings) and Orphamizer, when targeting target-A. The 135 cases out of the 243 PhenomeCentral cases were used for this evaluation, as they had diagnoses which belonged to Target-A. The result of each run was obtained as a ranked list of diseases. They were represented in terms of “recall at ranks” (i.e., the fraction of cases where the correct diagnosis appeared in the top-listed diseases). Figure 2 shows the recall rates by PubCaseFinder in the three settings and by Orphamizer. The top 10 recall rate of PubCaseFinder-O/CR is 57% (Fig. 2), which means that there is a correct diagnosis in the top 10 of a ranked list of 2,323 diseases for about one in every two cases. All recall rates of PubCaseFinder-O/CR are higher than those of PubCaseFinder-O, PubCaseFinder-CR, and Orphamizer (Fig. 2). The top 50 recall rate of PubCaseFinder-O is lower than the top 20 recall rate of PubCaseFinder-O/CR, which means that even if a user checks the top 50 diseases of PubCaseFinder-O, the diagnostic rate is lower than when checking the top 20 diseases of PubCaseFinder-O/CR. We also evaluated the

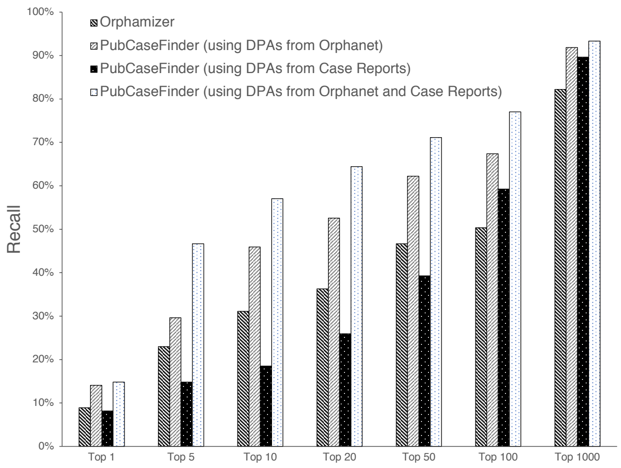

Figure 2. Performance comparison of three different settings of PubCaseFinder and Orphamizer

statistical significance of a correct diagnosis appearing in the top 10 with a binominal test and found that the *p*-value of PubCaseFinder-O/CR was 4.01×10^{-144}, whereas those of PubCaseFinder-O, PubCaseFinder-CR, and Orphamizer were 2.83×10^{-108}, 4.89×10^{-33}, and 4.73×10^{-65}, respectively. Those results clearly show the potential of DPAs from case reports to improve the performance of phenotype-driven differential diagnosis systems.

Second, we evaluated the performance of PubCaseFinder-CR when targeting target-B. We narrowed down 243 cases of PhenomeCentral to 59 cases whose diagnoses were part of the target-B. For the 59 cases, we obtained ranked lists of target-B using PubCaseFinder-CR and then calculated recalls on the basis of the results. PubCaseFinder-CR showed the recall number (rate), 2(3.4%)@1, 3(5.1%)@5, 5(8.5%)@10, 6(10.2%)@20, 13(22.0%)@50, 24(40.7%)@100, and 56(94.9%)@1000. We evaluated statistical significance of a correct diagnosis appearing in the top 10 by using a binominal test and found a *p*-value of 3.72×10^{-5}. Although the recall rates of PubCaseFinder-CR are low, the *p*-value shows the potential of PubCaseFinder for differential diagnosis of rare diseases that were not associated with a phenotype in Orphanet. Note that the recall rates of PubCaseFinder-CR for target-B were lower than those of PubCaseFinder-CR for target-A even though they both exploited DPAs from case reports. From examining the number of associated DPAs from case reports for target-A and target-B, on average, each disease in target-A had 27.3 DPAs, while each disease in target-B had 18.6 DPAs. It is interpreted that the difference of recall rates is caused by the difference of the number of associated DPAs.

Conclusion

In this study, we developed a text mining approach to extend the coverage of DPAs in manually curated databases like Orphanet. By applying the approach to a million case reports from PubMed, we could increase the coverage of DPAs from Orphanet more than 2 times. Based on the extended DPA database, we also developed PubCaseFinder, a new phenotype-driven differential diagnosis system. A series of experiments which was conducted using clinical cases from PhenomeCentral showed that the performance of phenotype-driven differential diagnosis could substantially be improved thanks to the extension of the DPA database. Note that automatic text mining techniques are often regarded as assistive tools to help manual curation of databases, due to its potentially high chance of noisy results. However, our experimental results suggest that even before a manual screening process, text mining approach can be substantially helpful. Our proposed approach is deemed practically useful, because manual curation will take enormous time and cost, particularly considering the large volume and rapid growth of case reports.

Previous studies reported that case reports were an important tool for extracting valuable information for rare diseases in spite of low certainty evidence due to its small samples. We therefore targeted the one million case reports included in PubMed, and this is, to our knowledge, the first demonstration that such a large collection of case reports was useful for tackling rare disease issues by using a text mining method. We will extend the collection of case reports to those of European in Europe PMC and those of Japanese in J-STAGE and will believe that further case reports will contribute more for tackling rare disease issues.