Ontology-based expansion of virtual gene panels to improve diagnostic efficiency for rare genetic diseases

Jae-Moon Shin [1], Toyofumi Fujiwara [1] and Atsuko Yamaguchi [2]

[1] *Database Center for Life Science, Kashiwa, Chiba, Japan*
[2] *Tokyo City University, Setagaya, Tokyo, Japan*

Abstract

Recently, to diagnose rare genetic diseases, Virtual Gene Panels (VGP) comprising sets of disease-related causal genes have been used to interpret candidate genes detected through whole-genome sequencing and whole-exome sequencing. In a pilot study of the UK 100,000 Genome Project for diagnosing rare diseases, VGPs from PanelApp software were used to filter candidate genes instead of manual interpretation to improve diagnostic efficiency. However, in about 50% of cases, the PanelApp VGPs also filtered out disease-causing genes. Here, we propose several methods using the hierarchical structure of Mondo disease ontology to design VGPs that avoid filtering disease-causing genes with high probability. To determine the best VGP design method among those proposed, we evaluated our methods with various parameters via computational experiments with an evaluation data set composed of 74 patients. Our results show that our proposed method can contribute significantly to automatically filtering candidate genes as well as shortening the interpretation time for diagnosing rare diseases.

Keywords

Rare disease, Ontology, Genetic testing, Virtual gene panel

1. Introduction

Rare diseases are those with a very low prevalence rate, of which there are estimated to be about 10,000 in the world [1]. Worldwide, the total number of patients with these diseases is estimated to be more than 400 million, many of whom have been undiagnosed for years [2]. Approximately 80% of rare diseases are believed to have a genetic origin [1]. The advancement of next-generation sequencing (NGS) technology has decreased both the cost and time of decoding genetic sequences. As a result, using NGS for genetic testing is considered a powerful tool for diagnosing rare diseases [3]. However, diagnosing rare diseases using NGS involves a labor-intensive process of searching through literature to identify a single candidate disease-causing gene that can best explain a patient's symptoms. This manual interpretation process can take hours, even for trained experts [4].

Recently, to reduce the time needed for manual search, a Virtual Gene Panel (VGP) comprising a set of disease-related causal genes has been used to filter candidates. By taking the intersection of candidate genes and the set of genes of the VGP corresponding to an initial diagnosis, the number of potential candidates can be reduced significantly. As part of the UK 100,000 Genomes Project, a pilot study for rare disease diagnosis, the PanelApp software, including 332 VGPs, was developed and applied to the manual interpretation of whole-genome sequencing results [5]. However, they failed to effectively filter out candidates in 50% of cases [6].

There are two reasons why disease-causing genes were not effectively filtered in half of the pilot study cases. First, there were only 332 VGPs included in PanelApp. VGPs in PanelApp are designated through manual curation, which takes time to design. Therefore, there were not always appropriate

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences
EMAIL: shin@dbcls.rois.ac.jp (J. Shin.); fujiwara@dbcls.rois.ac.jp (T. Fujiwara.); atsuko@tcu.ac.jp (A. Yamaguchi.)
ORCID: 0000-0001-5859-601X (J. Shin.); 0000-0002-0170-9172 (T. Fujiwara.); 0000-0001-7538-5337 (A. Yamaguchi)

VGPs corresponding to initial diagnoses. Furthermore, if the initial diagnosis is even slightly wrong, the disease-causing gene may not be included in the VGP.

In this study, we propose a method to design VGPs automatically using a knowledge graph. Additionally, we propose several methods to expand VGPs using the hierarchical structure of Mondo disease ontology (Mondo) [7]. The basic premise of the expanded VGP design is adding genes associated with the superclass or sibling of the initially diagnosed disease in Mondo into an original set of genes. We evaluated the diagnostic efficiency of these methods using 74 cases of rare genetic diseases. Our results show that the best method succeeded in automatically filtering candidate genes while still containing the disease-causing genes with high probabilities.

2. Methods

2.1. Knowledge graph

We constructed a knowledge graph based on the Resource Description Framework to design VGPs automatically (https://integbio.jp/rdf/dataset/pubcasefinder), using existing classes for interoperability with other knowledge graphs. All diseases and genes were defined as instances of "med2rdf:Disease" and "med2rdf:Gene" classes, respectively, by Med2RDF-ontology (http://med2rdf.org/) [8]. Gene-disease associations were defined as instances of the "sio:SIO_000983" class defined by Semantics Science Integrated Ontology [9].

We collected gene-disease associations from the following three data sources.

1. MIM2Gene (https://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene_medgen).
   Associations between MIM Numbers with type "phenotype" and NCBI Gene IDs were extracted.
2. OrphaData (http://www.orphadata.org/data/xml/en_product6.xml).
   Associations between OrphaCodes and Gene symbols were extracted.
3. GenCC (https://search.thegencc.org/download) [12].
   Associations between disease IDs including MIM Numbers, OrphaCodes, and Mondo IDs and HGNC IDs were extracted.

We connected to MIM numbers in Online Mendelian Inheritance in Man [10] and OrphaCodes in Orphanet [11] to Mondo IDs using the "equivalentTo" classes in Mondo. In our constructed knowledge graph, genes are connected for a disease ID in Mondo through med2rdf:Disease and sio:SIO_000983. For a disease ID d in Mondo, we denote a set of genes by $G_0(d)$ connected in the knowledge graph. Then, we define $G(d) = G_0(d) \cup (\cup_{d' \text{ is a subclass of } d} G(d'))$.

2.2. Development of VGPs

Initially, when a disease, *d*, is diagnosed, *G*(*d*) can be regarded as a VGP for *d*. **Original set (OR):** For a given disease *d*, output a set *G*(*d*) of genes. Using OR, we can obtain as many VGPs as the number of Mondo IDs. However, like PanelApp, OR also has the problem of filtering out the disease-causing gene if the initial diagnosis is slightly different. To overcome this problem, we propose an expansion of the VGP designed by OR.

Mondo forms a directed acyclic graph, in which nodes represent diseases and edges represent "rdfs:subClassOf" relationships between diseases. Using this structure, we expanded the VGPs designed by OR by adding genes of superclasses in Mondo.

Here we describe four simple rolling-up methods to expand VGP using rdfs:subClassOf. **One class up for all paths (1UAP):** For disease d, compute a set of genes $G = \{g \mid g \in G(d'), d \text{ rdfs:subClassOf } d'\}$. **Two classes up for all paths (2UAP):** For disease d, compute a set of genes $G = \{g \mid g \in G(d''), d \text{ rdfs:subClassOf } d' \text{ and } d' \text{ rdfs:subClassOf } d''\}$. **One class up for minimum paths (1UMP):** For disease d, compute a set of genes $G = argmin_{|G(d')|} \{g \mid g \in G(d'), d \text{ rdfs:subClassOf } d'\}$. **Two classes up for minimum paths (2UMP):** For disease d, compute a set of genes $G = argmin_{|G(d'')|} \{g \mid g \in G(d''), d \text{ rdfs:subClassOf } d' \text{ and } d' \text{ rdfs:subClassOf } d''\}$.

Additionally, we describe the following two methods to expand VGP by rolling-up until the number of genes exceeds the threshold *k*.

- All paths up with threshold *k* (TH *k* AP)
- Minimum path up with threshold *k* (TH *k* MP)

For an initially diagnosed disease *d*, TH *k* AP and TH *k* MP compute a set *D* of Mondo IDs, such that each d' in *D* is *d* or ancestor of *d* and the size of $G(d')$ is equal to or greater than the threshold *k*. TH *k* AP outputs the union of $G(d')$ for all d' in *D*, and TH *k* MP outputs $G(d')$ for d' with minimum size of $G(d')$. The algorithms of TH *k* AP and TH *k* MP are shown in Algorithms 1 and 2.

Algorithms 1 TH *k* AP

Input: a Mondo ID d, the Mondo Ontology
Output: a set G of genes
$D = \emptyset$
D'.push(d) $//D'$ is a stack
while$(D'$ is not empty$)$
  $d' = D'$.pop()
  if$(|G(d')| \geq k)$
    $D = D \cup \{d'\}$
  else
    D'.push(d'') for d'' such that d' is a subclass of d''
$G = \cup \{d \in D\}\, G(d)$
return G

Algorithms 2 TH *k* MP

Input: a Mondo ID d, the Mondo Ontology
Output: a set G of genes
$D = \emptyset$
D'.push(d) $//D'$ is a stack
while$(D'$ is not empty$)$
  $d' = D'$.pop()
  if$(|G(d')| \geq k)$
    $D = D \cup \{d'\}$
  else
    D'.push(d'') for d'' such that d' is a subclass of d''
$d'' = argmin_{|G(d)|}\, \{d|\, d \in D\}$
$G = G(d'')$
return G

3. Results

To evaluate the diagnostic efficiency of VGPs developed according to the proposed methods, a data set from 74 patients obtained via whole-exome sequencing (WES) was used. The data set includes candidate genes from WES, initially diagnosed diseases, and disease-causing genes. The median of the number of candidate genes in the 74 patients was 384. For each patient, the number of disease-causing genes, which would be included in *G*, was one. The number of initially diagnosed diseases, which is the input of the methods, takes from one to three, with a median of one.

Using the knowledge graph, VGPs for about 25,000 diseases contained in Mondo can be automatically designed by the OR method. To evaluate the performance of the seven expansion methods, we computed Coverage, Median, and Expectation. Table 1 shows the experimental results of OR, 1UAP, 1UMP, 2UAP, and 2UMP. Table 2 and Table 3 show the results for TH *k* AP and TH *k* MP when increasing the threshold from 50 to 300 at increments of 50. Coverage refers to the ratio of VGPs that include disease-causing genes. Median is the median of the sizes of intersections of the patient's candidate genes and VGPs. Expectation is the expected number of genes to analyze for determining the

final disease-causing gene. If VGPs include disease-causing genes, only genes in intersections of the patient's candidate genes and VGPs should be analyzed. However, if VGPs do not include disease-causing genes, all candidate genes should be analyzed. Therefore, the expectation can be computed using the following formula.

$$\text{Expectation} = (\Sigma_{g(p) \in \text{VGP}(p)} \mid C(p) \cap \text{VGP}(p) \mid + \Sigma_{g(p) \notin \text{VGP}(p)} \mid C(p) \mid) / 74$$

where, for patient p, a set of genes in VGP, the disease-causing gene, and a set of candidate genes are denoted by VGP(p), $g(p)$, and $C(p)$, respectively.

Table 1
The results of the experiments using OR, 1UAP, 1UMP, 2UAP and 2UMP

	OR	1UAP	1UMP	2UAP	2UMP
Coverage	0.4864	0.9324	0.8378	0.9729	0.8918
Median	1	23.5	10	60	18
Expectation	197.22	52.77	74.67	67.14	64.87

Table 2
The results of the experiments using TH *k* AP

	TH 50 AP	TH 100 AP	TH 150 AP	TH 200 AP	TH 250 AP	TH 300 AP
Coverage	0.9054	0.9189	0.9324	0.9459	0.9459	0.9459
Median	19.5	22.5	23	24	24	26.5
Expectation	55.14	54.39	51.24	48.09	48.54	49.35

Table 3
The results of the experiments using TH *k* MP

	TH 50 MP	TH 100 MP	TH 150 MP	TH 200 MP	TH 250 MP	TH 300 MP
Coverage	0.8513	0.8648	0.9054	0.9324	0.9324	0.9324
Median	10.5	12	13	14.5	15	15
Expectation	70.59	66.5	51.93	41.97	42.29	42.56

2UAP performed best in terms of Coverage, although this had the worst Median. Conversely, OR had the best Median but worst Coverage. Regarding Expectation, TH 200 MP performed best.

4. Discussion

Among the seven methods, the best performance observed was for TH *k* MP, especially when *k* = 200. As *k* increases, both the coverage of TH *k* MP and the median of the number of genes to be analyzed become larger. Therefore, there is a trade-off between coverage and the median. By selecting suitable values of *k*, users can design a VGP with preferred coverage and median.

Using VGPs, the number of candidate genes to be analyzed can be reduced. If a VGP was not used, all candidate genes had to be analyzed (median of 388). However, by using VGPs designed by TH 200 MP, the median of the number of candidate genes was reduced to 14.5 if the VGPs included disease-causing genes. Therefore, VGPs designed by our proposed methods may be useful for gene ranking systems, such as PubCaseFinder [13], by filtering candidate genes using VGPs.

5. Conclusion

In this study, seven methods were presented to develop expanded VGPs for initial clinical diagnosis using knowledge graphs including Mondo. As a result, the TH 200 MP method, i.e., the minimum path navigation using a threshold of 200, achieved the best performance regarding Expectation. We found that this could contribute significantly to automatically reducing candidate genes as well as shortening

the interpretation time for diagnosing rare diseases. We expect our methods to be widely used by clinicians to diagnose rare diseases with NGS analysis technology.

Acknowledgments

We are very grateful to Prof. Hirotomo Saitsu for providing the evaluation dataset. This work was supported by the National Bioscience Database Center of the Japan Science and Technology Agency, by "Challenging Exploratory Research Projects for the Future" grant from Research Organization of Information and Systems and by JSPS KAKENHI grant number 21K12148.

References

[1] M. Haendel, N. Vasilevsky, D. Unni, C. Bologa, N. Harris, H. Rehm, et al, How many rare diseases are there?, Nat Rev Drug Discov 19 (2020) 77-78. doi:10.1038/d41573-019-00180-y.
[2] S. Marwaha, J. W. Knowles, E. A. Ashley, A guide for the diagnosis of rare and undiagnosed disease: beyond the exome, Genome Med 14 (2022) 23. doi:10.1186/s13073-022-01026-w.
[3] D. Bick, M. Jones, S. L. Taylor, R. J. Taft, J. Belmont, Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases, J Med Genet 56 (2019) 783-791. doi:10.1136/jmedgenet-2019-106111.
[4] A. L. Wise, T. A. Manolio, G. A. Mensah, J. F. Peterson, D. M. Roden, C. Tamburro, et al, Genomic medicine for undiagnosed diseases, Lancet 394 (2019) 533-540. doi:10.1016/S0140-6736(19)31274-7.
[5] A. R. Martin, E. Williams, R. E. Foulger, S. Leigh, L. C. Daugherty, O. Niblock, et al, PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels, Nat Genet 51 (2019) 1560-1565. doi:10.1038/s41588-019-0528-2.
[6] Investigators GPP, D. Smedley, K. R. Smith, A. Martin, E. A. Thomas, E. M. McDonagh, et al, 100,000 Genomes pilot on rare-disease diagnosis in health care - preliminary report, N Engl J Med 385 (2021) 1868-1880. doi:10.1056/NEJMoa2035790.
[7] K. A. Shefchek, N. L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, et al, The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species, Nucleic Acids Res 48 (2020) D704-D715. doi:10.1093/nar/gkz997.
[8] M. Kamada, T. Katayama, S. Kawashima, R. Kojima, M. Nakatsui, Y. Okuno, Med2RDF: Semantic biomedical knowledge-base and APIs for the clinical genome medicine, in: Proceedings of the 12th. International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. SWAT4HCLS 2019, Edinburgh, Scotland, 2019, pp. 161–162.
[9] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, et al, The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, J Biomed Semantics 5 (2014) 14. doi:10.1186/2041-1480-5-14.
[10] OMIM, An online catalog of human genes and genetic disorders, 2022. URL: https://www.omim.org/
[11] Orphanet, The portal for rare diseases and orphan drugs, 2022. URL: https://www.orpha.net/consor/cgi-bin/index.php
[12] GenCC, The gene curation coalition, 2022. URL: https://thegencc.org/
[13] T. Fujiwara, J. Shin, A. Yamaguchi, Advances in the development of PubCaseFinder, including the new application programming interface and matching algorithm, Hum Mutat 43 (2022) 734-742. doi:10.1002/humu.24341.