ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-189

Discovering Monogenic Causes of Multi-Diseases by Mining Electronic Medical Records and Genetics Repositories

Adnan KULENOVIC, Azra LAGUMDZIJA-KULENOVIC,
A ! A – Absolute Information Age, Inc. Toronto, Canada.

Abstract. We present a method called *SMDG* (Single Multi-Disease Genes) for systematic discovery of monogenic causes of multi-diseases. Multi-disease conditions, quite common in older populations, are difficult to treat due to missing their precise medical guidelines and need for attention of multiple health care providers. Finding monogenic causes of these diseases would enable introducing new therapeutic approaches, focused on the remediation of mutations of single genes. *SMDG* is based on the hierarchical divisive clustering of electronic medical records *(EMR)* that include genetic data, and on the analysis of the public gene-to-disease and gene-to-gene repositories. The method was tested on the database of the Harvard Personal Genome Project (*PGP*), the gene-to-disease repository *DisGeNET* and the gene-to-gene interactions repository *BioGRID*. It identified possible new monogenic causes of selected multi-diseases, which were confirmed as valid hypotheses by examining related research papers.

Keywords. Multi-diseases, *EMR* Data Mining, Hierarchical Clustering, Gene-to-Disease Associations, Gene–to-Gene Interactions, Set-valued Categorical Data.

1. Introduction

We are proposing a method for discovering monogenic causes of multi-diseases. We assume that a multi-disease is a set of co-occurring diseases without a primary one, while co-morbidity is a disorder characterized by a primary disease and several secondary diseases.

This method, called *Single Multi-Disease Genes* (*SMDG*), is based on the clustering of electronic medical records (*EMR*) that include patient genetic data, and the analysis of the public gene-to-disease and gene-to-gene association repositories. Its output is a set of hypothetical monogenic causes of multi-diseases, which can be further investigated and confirmed in targeted research projects.

Why are we interested in this research? Patients with multi-diseases require attention of several health care providers, which complicates their diagnostic and treatment processes. In addition, medical guidelines for multi-diseases are limited or non-existent, which could cause adverse effects of overmedication. Also, the monogenic causes of multi-diseases are not systematically tracked in the public gene-to-disease repositories, nor are there effective methods for their finding.

The goal of this research is to introduce and elaborate a method that would aid the identification of monogenic causes of multi-diseases, which in turn could speed up finding their effective therapies focused on mutations of single genes.

The inputs, outputs, steps and formulas of *SMDG* are specified in Sections 2 and 3. *SMDG* was validated by using the *Harvard Personal Genome Project* database (*PGP*) [1], and genetics repositories *DisGeNET* [2] and *BioGRID* [3]. The results, discussed in section 4, identify a new hypothetical monogenic cause of a selected multi-disease, which was confirmed by examining available research papers. In section 5 we compare *SMDG* with similar methods and suggest its possible extensions.

2. SMDG inputs and outputs

SMDG inputs are: *(i)* an ***EMR*** **repository** *(EMR)*, *(ii)* a **Gene-to-Disease Association Repository** *(GDR)* and *(iii)* a **Gene-to-Gene Interaction Repository** *(GGR)*.

The current version of *SMDG* requires only the patient diseases and mutated genes from *EMR,* i.e. $EMR \subset P \times \mathbb{P}(DIS) \times \mathbb{P}(GEN)$ (*P* is a set of patient IDs, *DIS* is the set of all diseases, *GEN* is the set of all genes and $\mathbb{P}$ powerset symbol). *EMR* element, $r = (p, D, G) \in EMR$, is a triplet of patient ID, a set of patient's diseases *D* and a set of patent's mutated genes *G*.

SMDG needs only the gene-to-disease pairs from *GDR* ($GDR \subset GEN \times DIS$). Similarly, required content from *GGR* is only a set of gene pairs $(g_m, g_n) \in GGR \subset GEN \times GEN$. (g_m, g_n) is an ordered pair, meaning that gene g_m can change behavior of gene g_n (i.e. g_m is a "bait" and g_n a "hit"). Other components of *EMR, GDR* and *GGR,* such patient demographics, gene-to-disease association types and gene-to-gene interaction types, can be added to an extended version of *SMDG* (more in Section 5).

SMDG outputs are: *(i)* **Set of multi-diseases,** $MD \subset \mathbb{P}(DIS)$ *(*single multi-disease is denoted as $md \in MD$*),* and *(ii)* **Set of weighted hypothetical multi-disease genes** $HMDG \subset MD \times GEN \times W$. W=(0,100] is a set of weights of the hypothetical genes.

Example of a multi-disease found in *PGP EMR* is md_2 = *{Hyperlipidemia, Hypertension, Allergy}*, and one of its hypothetical genes is *TP53* with weight 100, i.e. $(md_2, TP53, 100) \in HMDG$. *SMDG* test results are discussed in Section 4.

3. SMDG Steps

Step 1: Prepare the EMR repository *(EMR)* in a relational database, which stores patients' diseases and genes as set-valued categorical data along with patient IDs. Align the disease names and gene symbols between the *EMR, GDR* and *GGR*.

Step 2: Apply a clustering algorithm to find multi-diseases *md* in *EMR*. A review of projects that used the clustering of *EMR* data can be found in paper [4]. We developed a version of the divisive hierarchical clustering algorithm in PL/SQL language. The algorithm assumes that the diseases are stored as set-valued categorical attributes. This type of data mining requires no supervision and can be further optimized for large *EMR* databases by using various database optimization techniques.

Distance function *dist* of this algorithm is based on the Jaccard factor: If D_1 and D_2 are disease sets of records $r_1, r_2 \in EMR$ then $dist(r_1, r_2) = 1 - |D_1 \cap D_2| / |D_1 \cup D_2|$, $dist \in [0, 1]$. The Jaccard distance function (also called dissimilarity function) is used in clustering of set-oriented categorical data. For example, in paper [5] it was applied within a k-means clustering algorithm. We used it to find sub-cluster centroids and to allocate *EMR* records to these clusters. The algorithm output is a set of clusters $C \subset$

$\mathbb{P}(EMR)$. The algorithm will stop dividing a cluster $C_i \in C$ if $|C_i| < cmin$ or $dist(r_1, r_2) \leq distmax$ for any records $r_1, r_2 \in C_i$. *cmin* and *distmax* are *SMDG* parameters.

The most common diseases found in the leaf cluster records r ($r \in C_i$) define multi-diseases md_i. The maximum number of diseases in each md_i is controlled by *SMDG* parameter *dmax*, i.e. $|md_i| \leq dmax$. For example, in our testing, discussed in Section 4, the most frequent diseases found in cluster C_2 were md_2= *{Hyperlipidemia, Allergy, Hypertension }*, for *dmax* = 3.

Step 3: Find candidate multi-disease genes of each *md* as set $NMDG \subset MD \times \mathbb{P}(GEN)$. $NMDG = \{(md, NG): md \in MD, NG = \cup\{G: \exists(p, D, G) \in EMR, md \subseteq D\}\}$. For example, candidate genes of multi-disease md_2 are *{APOB, TP53, …}*.The complete list can be found in Section 4.

Step 4: Use *GDR* to **find confirmed multi-disease genes** for each *md* as set $CMDG \subset MD \times \mathbb{P}(GEN)$. A confirmed gene can cause each disease of a multi-disease, i.e. $CMDG = \{(md, CG): md \in MD, CG = \{g: (\forall d \in md)(\exists(g, d) \in GDR)\}\}$. For example, for the above md_2, *GDR DisGeNET* returns $CG = \{SELE, AHR, PLAT, ALB\}$.

Step 5: Use *GGR* to **find precursor genes of the confirmed genes,** denoted as set $PMDG \subset MD \times \mathbb{P}(GEN)$. If *PG* is the set of precursor genes of *md* then $(md, PG) \in PMDG$. The precursor genes are either direct predecessors of confirmed genes or there is a path of interacting genes between them. The maximum count of interacting genes is an *SMDG* parameter called $maxpath \in \mathbf{Z}$ (set of non-negative integers). For example, if $g_p, g_{p1}, g_{p2} \in PG$, g_c is a confirmed gene, and $(g_{p2}, g_c), (g_{p1}, g_{p2}), (g_p, g_{p1}) \in GGR$ then valid paths are $(g_{p2}, g_c), (g_{p1}, g_{p2}, g_c)$ *and* $(g_p, g_{p1}, g_{p2}, g_c)$ for $maxpath = 2$.

Step 6: Calculate the weights of the precursor genes, which indicate possibility that a gene can cause a multi-disease. Let introduce functions *gdc* and *w,* as follows. *gdc* returns from *GDR* the count of known associations between a gene and all individual *md* diseases. $gdc = MD \times GEN \rightarrow \mathbf{Z}$. $gdc(md, g) = |\{d: d \in md, \exists(g, d) \in GDR\}|$. $gdc = |md|$ for confirmed genes.

w is the gene weight function defined as $w = MD \times GEN \rightarrow (0,100]$. If g_p is an *md* precursor or a confirmed *md* gene, then weight of g_p is calculated as:

$$w(md, g_p) = max\left(gdc(md, g_p) / |md| * 100 + glw * \sum_{s=1}^{Np} w(md, g_{ps}),\ 100\right).$$

Component $gdc(md, g_p) / |md| * 100$ returns the count of individual *md* diseases caused by gene g_p, which is mapped to a value in *[0,100]*. It has value *100* for confirmed genes. Component $glw * \sum_{s=1}^{Np} w(md, g_{ps})$ returns a modified sum of weights of all direct successor genes g_{ps} of g_p from paths connecting g_p with the confirmed genes. In a special case, g_{ps} is a confirmed gene with weight *100*. N_p is the number of these paths, and *glw* is the relative weight of a gene-to-gene interaction when compared to the weight of a direct gene-to-disease association. *glw* is an *SMDG* parameter and its recommended values are between *0.01* and *0.1*. Function *max* limits value of *w* to *100*.

Function *w* is recursive, and its value is calculated first for direct precursors of confirmed genes then for their precursors at the next level and so on.

Step 7: Find hypothetical multi-disease genes, as set $HMDG \subset MD \times GEN \times W$, $W = (0,100]$. A hypothetical gene of *md* is an element from the intersection of the candidate *md* genes *NG* and the precursor *md* genes *PG*, which is then extended with its weight. $HMDG = \{(md, g, wght): md \in MD, (\exists(md, NG) \in NMDG, \exists(md, PG) \in PMDG, g \in NG \cap PG), wght = w(md, g)\}$.

Step 8: Test the hypothetical multi-disease genes in targeted research projects.

4. SMDG testing and validation

We used the following inputs to develop, test and validate the *SMDG* method: *(i)* Database of *Harvard Personal Genome Project - PGP* [1], which includes about 2000 medical records with the demographic, disease, genetic, medication, lab tests and other *EMR* data, *(ii) DisGeNET* [2], a curated gene-to-disease repository and *(iii) BioGRID* [3], a gene-to-gene repository that records the human genetic and protein interactions.

In step 3.1 we extracted the disease and genetic data from *PGP* and aligned the disease names and gene symbols between *EMR*, *GDR* and *GGR*. Then we applied the *SMDG* hierarchical clustering algorithm (step 3.2) with the minimal cluster size *cmin* = 60, maximum distance between cluster elements *dismax* = 0.6 and the maximum number of *md* diseases *dmax* = 3. Seven multi-diseases were found in leaf clusters, for example, md_1= *{Depression, Anxiety, Allergy}*, md_2= *{Hyperlipidemia, Hypertension, Allergy}* and md_3= *{Hypertension, Hyperlipidemia, Depression}*.

We selected md_2 for further analysis, and step 3.3 returned the following candidate genes: *{APOB, PCSK9, WFS1, SP110, SLC45A2, AGTR1, COL4A1, FUT2, GUCY2D, KCNJ11, MTRR, PKP2, PMS2, PTCH1, RP1, TP53, BRCA2, ELAC2, H6PD, IL7R, MLH1, TAS2R38, TYR}*.

Next step (3.4) produced 4 confirmed md_2 genes from *DisGeNET: {SELE, AHR, PLAT, ALB}*. Then we found their precursor genes (Step 3.5) by taking *maxpath* = 2 and calculated their weights (step 3.6). These steps resulted in 13,103 precursor genes and 495,208 of their dependency paths to confirmed genes *SELE, AHR, PLAT* and *ALB*.

We reduced the set of these precursor genes to the md_2 candidate genes from *PGP*. Finally, in Step 3.7 we selected the following hypothetical genes for further analysis: *(APOB, 80.44), (TP53, 100), (MLH1, 71.98), (BRCA2, 60.7)*.

Figure 1 illustrates the above steps.

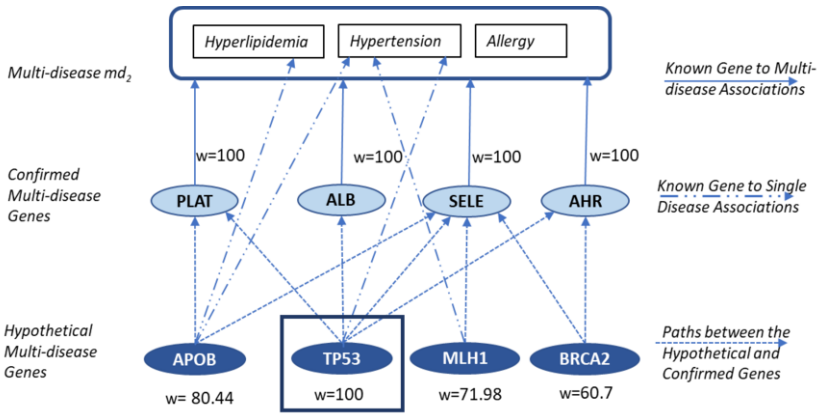

Figure 1. Example of hypothetical monogenic causes of multi-disease md_2.

According to *DisGeNET,* mutations of *APOB* can cause *Hyperlipidemia* and *Hypertension*, *TP53* and *MLH1 Hypertension* and *BRCA2* none of the md_2 diseases.

Gene *TP53* was taken for further analysis due to its high weight resulting from its numerous dependency paths to the confirmed md_2 genes (Step 3.8). *DisGeNET* did not indicate that this gene can cause *Hyperlipidemia* and *Allergy*, however our further investigation showed that it could indeed cause these diseases [10, 11].

5. Discussion

To the best of our knowledge, methods for finding hypothetical single multi-disease genes have not been addressed yet, despite their significance discussed in the introduction. Several projects examined disease co-morbidities, which correspond to the *SDMG* step 3.2. For example, projects [6] and [9] aim at finding frequent disease co-morbidities by analyzing the *EMR* and genetics repositories. Projects [7] and [8] build clusters of co-occurring diseases by investigating the genetics repositories only.

Method *SMDG,* presented in this paper, is an innovative approach to finding hypothetical single multi-disease genes, which is based on the analysis of both *EMR* repositories and genetics repositories. We find that data mining of *EMR* is pivotal, as it helps finding not only frequent multi-diseases but also their candidate monogenic causes. Without identifying candidate genes in *EMR*, the number of hypothetical multi-disease genes obtained only from the genetics repositories could grow rather high making it difficult to select the hypothetical genes for further investigation. For example, our tests returned from *BioGRID* about 13k precursor genes of a sample multi-disease discussed in Section 4 for paths with maximum two internal genes.

In the future research, the *SMDG* method can be extended in different ways, which could improve its effectiveness and efficiency in finding frequent multi-diseases and their monogenic causes. Examples of these extensions are: *(i)* inclusion of the environmental, demographic and other data into *EMR* and the clustering distance function, *(ii)* extension of that function with the disease frequencies, *(iii)* comparison of various *EMR* clustering algorithms for finding multi-diseases and their candidate genes in large *EMR* repositories, *(iv)* weighting the gene-to-disease associations according to their type (i.e. *biomarker, therapeutic..*) and *(v)* weighting the gene-to-gene interactions according to their types (i.e. as *genetic, physical..*).

References

[1] Madeleine P Ball et al., Harvard Personal Genome Project: Lessons from participatory public research, *Genome Medicine* **6 (**2014).

[2] Janet Piñero, et al., DisGeNET: a comprehensive platform integrating information on human disease associated genes and variants, *Nucleic Acids Research* **45** (2017), D833–D839.

[3] Chatr-aryamontri, Andrew et al., The BioGRID Interaction Database: 2017 Update. *Nucleic Acids Research* **45** (2018), D369–D379.

[4] Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon, Mining Electronic Health Records: A Survey, *ACM Computing Surveys* (CSUR) 50 (2018).

[5] Fuyuan Cao, et al., An Algorithm for Clustering Categorical Data with Set-Valued Features. *IEEE Trans. Neural Netw. Learning Syst.* **29** (2018), 4593-4606

[6] Darcy A, Devis, Nitesh V. Chavla, Exploring and Exploiting Disease Interactions from Multirelational Gene and Phenotype Networks, *PLoS ONE* **6** (2011).

[7] Carlota Rubio-Perez. et al., Genetic and functional characterization of disease associations explains comorbidity, *Scientific Reports* **7** (2017).

[8] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, Roded Sharan, Associating Genes and Protein Complexes with Disease via Network Propagation, *PLoS Comput Biol.* **6** (2010).

[9] Francisco S. Roque, et. al., Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts, *PLoS Comput Biol.* **7** (2011), Issue 8.

[10] Yael Aylon & Moshe Oren, The Hippo pathway, p53 and cholesterol, *Cell Cycle*, **15** (2016), 2248-2255.

[11] Saccucci, P., et. al., p53 Codon 72 Genetic Polymorphism in Asthmatic Children: Evidence of Interaction with Acid Phosphatase Locus 1, *Allergy, Asthma & Immunology Research*, **6** (2014), 252–256.