2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Integrating categorical and structural proximity in Disease Ontologies*

Lorenzo Madeddu[1], Giorgio Grani[2] and Paola Velardi[3]

Abstract— **The purpose of the study described in this paper is to shed more light on disease similarities by analyzing the relationship between categorical proximity of diseases in human-curated ontologies and structural proximity of the related disease module (DM) in the interactome. We propose a methodology (and related algorithms) to automatically induce a hierarchical structure from proximity relations between DMs, and to compare this structure with a human-curated disease taxonomy.**

Clinical relevance— **Disease ontologies are extensively used for diagnostic evaluation and clinical decision support but still reflect the clinical reductionist perspective. We demonstrate that the proposed network-based methodology allows us to analyze commonalities and differences among structural and categorical similarity of human diseases, help refine human disease classification systems, and identify promising network areas where new disease-gene interactions can be discovered.**

I. Introduction

Disease taxonomies play a key role in defining the mechanisms of human diseases, potentially impacting both diagnosis and treatment. However, as remarked in [1], contemporary approaches to the classification of human diseases are mainly based on anatomical and pathological data, and clinical knowledge. Modern molecular diagnostic tools have shown the shortcomings of this methodology, reflecting both a lack of sensitivity in identifying pre-clinical diseases and a lack of specificity in defining diseases unequivocally. As a response to the limits of contemporary disease taxonomies, Zhou et al. [2] proposed a New Classification of Diseases (NCD) to capture the molecular diversity of diseases and define clearer boundaries in terms of both phenotypical similarity and molecular associations. Their study is based on the so-called "disease module hypothesis": proteins involved in the same disease show a high propensity to interact with each other [3]. Furthermore, if we identify a few disease components, the other disease-related components are likely to be located in their network-based "neighbourhood".

On the other hand, inducing disease relationships solely from DMs in the interactome is hindered by incomplete knowledge of disease-related genes. In this study we propose a methodology to integrate categorical relationships automatically induced from proximity of DMs in the human interactome network, with manually crafted categories in human-curated ontologies. Detected commonalities and differences may suggest latent and unknown molecular properties of diseases, help refine and extend disease classification systems, and facilitate precise clinical diagnosis consistent with molecular network properties.

II. Aims and Methods

DMs have been successfully used, for example, to prioritize diagnostic markers or therapeutic candidate genes, and in drug repurposing [4], [5]. However, according to Barabási et al. [5], these results have marginally influenced the disease taxonomies and, conversely, to the best of our knowledge, disease taxonomies have not been used to analyze DMs. In this study we aim for the first time to integrate taxonomic and network-based disease categorization principles, with the following innovative contributions:

1) to automatically induce a full-fledged hierarchical structure from proximity relations between DMs in the human interactome;
2) to compare this structure with a human-defined disease taxonomy (such as the Disease Ontology[1]));
3) to systematically identify categorical analogies and discrepancies between molecular and human-defined taxonomies.

Our research hypothesis is that a study of the relationships between molecular-based and human-curated disease taxonomies could help refine our knowledge on human diseases and identify limitations and perspectives of current module-based computational approaches to the study of diseases.
The main phases of the proposed approach are the following:

A. Induction of a Taxonomy of Disease Modules:

First, we automatically induce a taxonomic structure of diseases, hereafter referred to as the Interactome Taxonomy (I-T). We induce the I-T by applying hierarchical agglomerative clustering (with Average cluster-merge) to the human interactome network, exploiting proximity relations of DMs, as shown in Figure 1.

Given the interactome graph G, a set of diseases D_{it} and their DMs DM_{it} in G, hierarchical clustering is performed using a distance matrix of DMs, based on the following distance measure[2]:

$$dist(A,B) = \frac{\sum_{a\in A} min_{b\in B} SP(a,b) + \sum_{b\in B} min_{a\in A} SP(a,b)}{|A| + |B|} \quad (1)$$

*This research has been supported by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University and by the "Sapienza information-based Technology InnovaTion Center for Health - STITCH". This research was conducted without experimental procedures involving human subjects or animals.

[1] L. M. is with the Translational and Precision Medicine Department, Sapienza University of Rome, Rome, 00185 Italy lorenzo.madeddu@uniroma1.it

[2] G. G. is with the Translational and Precision Medicine Department, Sapienza University of Rome, Rome, 00185 Italy giorgio.grani@uniroma1.it

[3] P. V. is with the Computer Science Department, Sapienza University of Rome, Rome, 00185 Italy velardi@di.uniroma1.it

[1] https://disease-ontology.org

[2] used, e.g., in [3]

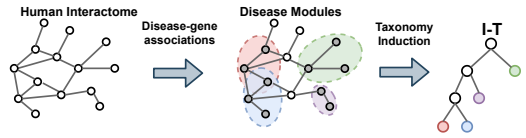

Fig. 1: A visual example of taxonomy induction.

where A, B are the set of nodes in modules DM_A and DM_B associated to diseases $d_A, d_B \in D_{it}$ and SP is the shortest path length between two given nodes in G.

B. Taxonomy alignment

The I-T taxonomy is not immediately comparable with a human-curated *reference ontology* (hereafter R-T). Whatever the choice of the R-T, the R-T and the I-T are expected to be defined on different sets of diseases nomenclatures, D_{rt} and D_{it}. Furthermore, they are also expected to be structurally diverse. For example, R-T has usually a polyhierarchical structure, while I-T is by construction a binary tree.

To compare I-T and R-T, we first need to create a mapping M from D_{it} to D_{rt} nomenclatures, and next, to prune the hierarchies so that they include the same set of leaf disease nodes, a process that we call *taxonomy alignment*. Let M be an available mapping of disease nomenclatures (see Section III for details). Our taxonomy alignment procedure consists of three algorithms: *merge*, *split*, and *prune*.

Merge and Split: In the I-T disease nodes are by construction leaf nodes, while this is often not the case for the R-T (see Figure 2 (left)). As visually shown in Figure 2 (right), the purpose of the *merge* and *split* phases is to move all disease nodes of the R-T on its leafs, without altering the direction of hyperonymy relations.

Prune: Next, the *prune* algorithm prunes both the R-T and the I-T, by recursively removing leaves not linked by any mapping relation in M and chains of inner nodes. Examples of removed nodes are highlighted with a double circle in Figure 3.

As a final result, the R-T and the I-T have as leaf nodes the same set of diseases, denoted as $D_{\cap}$.

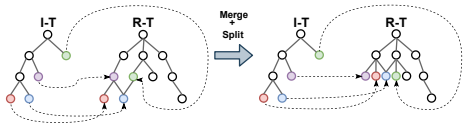

Fig. 2: A visual example of merge and split.

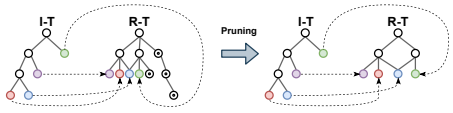

Fig. 3: A visual example of pruning.

C. Semantic Labeling of the Interactome Hierarchy (I-T)

After alignment, the two hierarchies are structurally comparable, however the inner nodes of the I-T do not have category labels, by construction. To further facilitate a comparative analysis of I-T and R-T, we defined an algorithm to label each inner node in the I-T with the most similar category label in the R-T. In order to find the most similar R-T category node, we exploit the notion of *cluster* C_c associated with a category node c in a taxonomy, defined as the set of all its descendant disease nodes that are also in $D_{\cap}$. For example, the red circles in Figure 4 show the clusters associated with the inner nodes A and A' belonging to the I-T and R-T respectively.

The *labeling* algorithm labels every I-T disease category c with the name of the R-T category c' with highest similarity score $sim(C_c, C_{c'})$ between the clusters of c and c'. To compute the similarity between two clusters, we use the Jaccard coefficient, a popular measure of set similarity. Always with reference to Figure 4, the label of node A' of the R-T will be associated to node A of the I-T. Note that only inner nodes with a similarity score higher than an experimentally defined threshold receive a label.

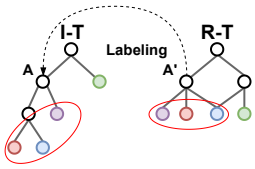

Fig. 4: A visual example of labeling.

III. Experimental set up

To conduct a DM analysis, we considered the most recent release of the human protein-protein interaction network published by Barabási et al. [4], which is an extension of a highly cited and popular interactome used by Menche et al. [3]. The network has $|V| = 16\,677$ proteins and $|E| = 243\,603$ physical undirected protein interactions.

To construct DMs, we collected disease-gene associations from DisGeNET [6] with a GDA[3] score greater or equal of 0.3. Finally, we selected as DMs the 948 diseases with a set of disease genes of size at least 10[4].

We selected the Disease Ontology (DO) as Reference Taxonomy (R-T) [7], since it exploits the molecular insights of disease phenotypes with the purpose of identifying "commonalities of diseases located in a common molecular location, originating from a particular cell type or resulting from a common genetic mechanism". For this reason, the DO is a good categorical framework for integrating network biology-based disease properties. By parsing the DO "obo" file[5], we generated a directed acyclic network hierarchy of 10012 diseases and disease categories, 10061 edges and 12 levels. To create a mapping M between the two different nomenclatures, we used partial mappings directly provided in DisGeNET and in the DO, that we further extended with the support of clinicians to cover all the 948 DMs.

IV. Results

In this Section, we summarize the major outcomes of a clinical analysis supported by the methodology presented in

[3]GDA is a "reliability" score, for details see `www.disgenet.org/dbinfo#section43`

[4]smaller modules imply a limited knowledge of the related disease-gene associations to date, and may lead us to unreliable results.

[5]`http://www.obofoundry.org/ontology/doid.html`

previous Sections. Our analysis is based both on the study of matching and unmatching pairs of R-T and I-T categories. Figure 5 is a visual representation of our findings[6]. The red (i.e., disease of cellular proliferation), light purple (i.e., genetic disease), and orange (i.e., disease of anatomical entity) network areas represent dense neighborhood identified as explained in Section IV-A. The three dashed circles represent a zoom-in of the interactome. They show the three "unexpected" disease category relationships, described in Section IV-B.

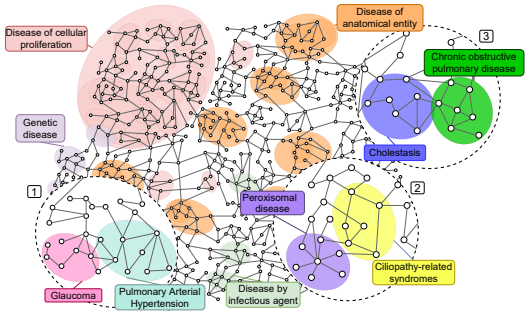

Fig. 5: A visual example of the labeled interactome.

A. Finding disease categories with a corresponding dense neighbourhood in the interactome.

First, we conducted an analysis to reveal in the human interactome large neighbourhoods of DMs associated with disease categories in R-T. Dense neighborhoods of diseases in the interactome network are useful to identify promising disease categories for disease gene prediction, drug repurposing and detection of comorbidities. To find these large neighbourhoods, we verified the existence of topmost disease categories of the DO (our selected R-T) with a high overlapping with some inner (categorical) nodes in the I-T. A DO disease category c' that is "well-represented" by an I-T category c implies a strong molecular proximity relationship among the diseases in cluster C_c. Symmetrically, this implies that there exists a molecular mechanism that strengthens the classification principle of the DO category.

We considered the 8 disease categories in the first level of the DO as the most general disease categories (the categories in Table I). To evaluate the degree of similarity between these DO categories and their most similar correspondents in the I-T, we used the Jaccard similarity, i.e. the "label score" computed by the labeling algorithm of Section II-C. We also calculated the statistical significance of our results by computing the p-value over a random distribution. Table I provides an overview of the topmost DO disease categories and their similarity degree with correspondent I-T categories induced from DM molecular network-proximity. In particular, we found that the DO disease categories that show a higher localization in a network neighborhood are "disease of cellular proliferation" and "genetic disease". This means that tumors and genetic diseases are highly localized in two neighbourhoods of the human interactome (see upper-left clusters in Fig. 5). From a biological network perspective, close DMs of "disease of cellular proliferation" are motivated by the fact that cancer diseases have similar genetic causes in differentiation and proliferation control genes such as the well-known $P53$ [5]. The second best matching category is "disease of anatomical entity", i.e., disease grouped by human experts according to an anatomical localization principle. However, as shown in the Table I, the similarity value is high but not statistically significant. This is motivated by the fact that diseases belonging to this topmost category are grouped in diverse sub-categories scattered over the network rather than in a large "anatomical" neighbourhood (see orange clusters in Fig. 5). To confirm this hypothesis, we performed a systematic automated pair-wise comparison among sub-categories of "disease of anatomical entity". We found that *very rarely* category pairs belonging to different anatomical sub-systems have overlapping clusters in the I-T, with some obvious and well documented exception, like nervous and respiratory systems, gastrointestinal and integumentary systems, musculoskeletal and cardiovascular systems [8]–[10]. In other terms, our experiments show that the validity of the anatomical classification principle is not disproved by the DM localization hypothesis, at least, given our state-of-the-art knowledge of disease-gene associations. This observation leads us to consider one limitation of the study presented in this Section, which stems from the high incompleteness of the human interactome [5]. It follows that, while positive results (disease categories corresponding to highly overlapping DMs) are useful pieces of evidence to identify interesting areas of the interactome to discover new disease-gene associations, the absence of such evidence could be either motivated by the non existence of a similarity relation, or by a lack of knowledge on gene interactions in specific areas of the interactome.

TABLE I: Correspondence among topmost DO categories and the induced taxonomy.

R-T (Disease Ontology)	Induced I-T
Disease Category Name (size)	**Best Label Score** (P-value)
disease of cellular proliferation (255)	54.77% ($3.14 \cdot 10^{-20}$)
disease of anatomical entity (434)	50.05% (0.08)
genetic disease (12)	41.66% ($6.14 \cdot 10^{-10}$)
disease by infectious agent (10)	30% ($1.92 \cdot 10^{-7}$)
physical disorder (21)	26.09% ($1.51 \cdot 10^{-9}$)
disease of mental health (76)	21.51% ($1.06 \cdot 10^{-13}$)
syndrome (42)	21.27 % ($8.69 \cdot 10^{-11}$)
disease of metabolism (55)	16.36% ($4.66 \cdot 10^{-11}$)

B. Finding unexplored structural relations between disease categories.

A more interesting result would be to identify "unexpected" and unexplored neighborhoods in the I-T, e.g., disease categories that are not presently connected in human-curated taxonomies but whose strong molecular similarities suggest that one such connection should be exploited to en-

[6]Clearly, the Figure is only a simplified representation of the human interactome, and furthermore, it does not preserve the dimensional differences among highlighted areas

rich the R-T ontology. To help finding these relations we developed a visual tool to explore the I-T in a more systematic way. Supported by this tool, clinical experts have identified, among the others, the following interesting results: there exist strong unexpected molecular relationships between glaucoma and pulmonary arterial hypertension, cholestasis and chronic obstructive pulmonary diseases (COPD), peroxisomal diseases and ciliopathy-related syndromes (see Fig. 5). *We remark that we detected these unexpected categorical disease relationships thanks to the algorithm for labeling the I-T*. By delving into these relationships, we were able to find confirmations in very recent clinical studies. For example, Lewczuk et al. [11] shed light on common molecular mechanisms and manifestations between pulmonary hypertension and glaucoma through multiple case reports. Instead, Tsechkovski et al. [12] observed that cholestasis and COPD patho-mechanisms are mediated by common molecular components like the Alpha 1-antitrypsin protein. However, the relationship between Alpha 1-antitrypsin mutations and liver disease is debated and yet to be elucidated [13]. Finally, Zaki et al. [14] found biological mechanisms between peroxisomal diseases and ciliopathy related syndromes (e.g. Joubert syndrome, Bardet-Biedl syndrome, Jeune syndrome). In conclusion, recent clinical evidence confirms that these detected relationships could be used to extend the DO. Other unexplored strong relationships that we identified lack at the moment support from published studies[7], however the results reported above demonstrate the relevance and potentials of our proposed methodology.

V. Discussion and Concluding Remarks

We believe that the biomedical understanding of diseases is on the edge of a radical change. The disease module hypothesis, with its relevant applications to disease-gene discovery and drug repurposing, is leading the revolution of bio-medical research of the future. For these reasons, we deem it fundamental to discover the degree of correspondence between disease similarity relations induced from the proximity of their related DMs, and categorical similarity in human-curated disease taxonomies. We developed a methodology to analyze relationships between diseases by leveraging, in a novel way, both taxonomic and molecular aspects. The proposed methodology supported a systematic analysis of human-crafted disease categories and their relationships with the DM molecular network-proximity. In particular, we found that some disease in "disease of cellular proliferation" and "genetic disease" form promising large disease network-neighbourhoods that could be exploited by network analysis methods for disease-gene detection. Next, we evaluated the consistency of the "disease anatomical entities" at the molecular level and found that there is no strong evidence of a network-neighbourhood of anatomical entities but, contrarily, disease neighbourhoods related to anatomical systems are scattered. Finally, we used our methodology to find unexplored strong molecular relationships between "specific" disease categories, such as glaucoma and pulmonary hypertension, diseases that are distant in human-crafted taxonomies but appear to be related by comorbidities and pathogenesis at the molecular level.

One limitation of our study arises from the highly incomplete state of the art knowledge on disease-related genes. This resulted in a limited mapping between human-crafted taxonomies and our induced hierarchy of DMs (about 12% of DO diseases), and furthermore prevented the interpretation of some evidence concerning unobserved molecular relationships, which could be either motivated by the non existence of such relations, or by the lack of knowledge on gene interactions in specific areas of the interactome.

[7]A clinical confirmation of our findings is clearly outside the scope of this research, although it represents a study hypothesis for further research by clinicians in the field.

References

[1] J. Loscalzo, I. Kohane, and A.-L. Barabasi, "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology," *Molecular systems biology*, vol. 3, no. 1, p. 124, 2007.

[2] X. Zhou, L. Lei, J. Liu, A. Halu, Y. Zhang, B. Li, Z. Guo, G. Liu, C. Sun, J. Loscalzo, *et al.*, "A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks," *EBioMedicine*, vol. 31, pp. 79–91, 2018.

[3] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, 2015.

[4] F. Cheng, I. A. Kovács, and A.-L. Barabási, "Network-based prediction of drug combinations," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[5] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nature reviews. Genetics*, vol. 12, pp. 56–68, 01 2011.

[6] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D845–D855, 2020.

[7] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: a backbone for disease semantic integration," *Nucleic acids research*, vol. 40, no. D1, pp. D940–D946, 2012.

[8] S. Chhabra and S. De, "Cardiovascular autonomic neuropathy in chronic obstructive pulmonary disease," *Respiratory medicine*, vol. 99, no. 1, pp. 126–133, 2005.

[9] B. L. Huang, S. Chandra, and D. Q. Shih, "Skin manifestations of inflammatory bowel disease," *Frontiers in physiology*, vol. 3, p. 13, 2012.

[10] B. J. Maron and M. S. Maron, "Hypertrophic cardiomyopathy," *The Lancet*, vol. 381, no. 9862, pp. 242–255, 2013.

[11] N. Lewczuk, A. Zdebik, J. Bogusławska, A. Turno-Krecicka, and M. Misiuk-Hojło, "Ocular manifestations of pulmonary hypertension," *Survey of Ophthalmology*, vol. 64, no. 5, pp. 694–699, 2019.

[12] M. Tsechkovski, V. Boulyjenkov, and C. Heuck, "A1-antitrypsin deficiency: Memorandum from a who meeting¿ l," *Bull World Health Organ*, vol. 75, no. 5, pp. 397–415, 1997.

[13] C. V. Schneider, K. Hamesch, A. Gross, M. Mandorfer, L. S. Moeller, V. Pereira, M. Pons, P. Kuca, M. C. Reichert, F. Benini, *et al.*, "Liver phenotypes of european adults heterozygous or homozygous for pi* z variant of aat (pi* mz vs pi* zz genotype) and non-carriers," *Gastroenterology*, 2020.

[14] M. S. Zaki, R. Heller, M. Thoenes, G. Nürnberg, G. Stern-Schneider, P. Nürnberg, S. Karnati, D. Swan, E. Fateen, K. Nagel-Wolfrum, *et al.*, "Pex6 is expressed in photoreceptor cilia and mutated in deafblindness with enamel dysplasia and microcephaly," *Human mutation*, vol. 37, no. 2, pp. 170–174, 2016.