A method for inferring candidate Disease-Disease Associations

Pietro Cinaglia$^{1,3[0000-0003-2237-6984]}$ and Marianna Milano$^{2,3[0000-0003-1561-725X]}$

[1] Department of Health Sciences, Magna Graecia University, 88100, Catanzaro, Italy
[2] Department of Experimental and Clinical Medicine, Magna Graecia University, 88100, Catanzaro, Italy
[3] Data Analytics Research Center, Magna Graecia University, 88100 Catanzaro, Italy.
cinaglia@unicz.it, m.milano@unicz.it

Abstract. The analysis of Disease-Disease Associations (DDA) and Gene-Disease Associations (GDA) is a relevant task in bioinformatics. These are analysed to investigate the interactions between sets of diseases and genes as well as their similarity, e.g., to improve the phases of diagnosis, prognosis and treatment in medicine. Generally, the extraction of information of interest from large-scale data, usually heterogeneous and unstructured, is performed via time-consuming processes. Therefore, several computational approaches have been focused on their prediction through data integration and machine learning techniques.
This paper presents a solution for Inferring DDA (*IDDA*) by integrating curated biomedical ontologies and medical dictionaries. It is able to extract a set of DDA using an in-house score based on the GDA. A preliminary step based on data enrichment retrieves the information about gene and disease, and it integrates these with a set of curated biological data ontologies and dictionaries. Specifically, *IDDA* extracts DDAs based on an in-house score, which uses GDAs for its evaluations. In a preliminary step, it performs data enrichment to retrieve concepts both for diseases and genes, by integrating several curated biomedical ontologies and medical dictionaries.

Keywords: bioinformatics · gene-disease · disease-disease · ontologies · data integration

1 Background

In recent years, a large amount of genomic and biological data is analysed in clinical research trials to evaluate novel treatments, to correlate human diseases with genomics data, as well as for knowledge extraction [3,5,7,8]. For instance, the genes involved in a disease are analysed for knowledge extraction, to understand its key factors (e.g., molecular basis and biological mechanisms), as well as to evaluate treatments and diagnosis. Furthermore, disease profiling also uses -omics data (e.g., genomics, transcriptomics, metabolomics) for evaluating susceptibility, progression and manifestation.

DOI: 10.1007/978-3-031-63772-8_8

Data integration allows cataloguing and analysing heterogeneous and unstructured information from different types and models, such as transcription factor binding sites, protein interactions, Gene-Disease Associations (GDAs), drug-target associations, medical ontologies and dictionaries, as well as literature repositories [4]. To give a non-exhaustive, genes associated with similar disorders show a higher likelihood of interaction, and diseases with common genes could share similar origins or mechanisms, by extension [13]. Similarly, Disease-Disease Associations (DDAs) represents relationships among diseases, and are useful to investigate diagnosis, prognosis, and treatments.

Experimental methods for GDA are expensive and time-consuming [9], therefore, several computational methods were developed to infer GDA. Generally, these identify concepts from medical literature, as well as by integrating protein interactions, functional annotation of signalling pathways, gene expression, medical vocabulary, disease concepts, and other biomedical data source.

In this scenario, ontologies play a crucial role in obtaining an interdisciplinary view from large and heterogeneous sources [6].

An ontology consists of a formal representation of relationships and properties existing among a set of concepts [2].

Usually, network-based scoring methods are applied to infer DDAs, establishing relationships between two or more diseases, based on biological assumptions; a non-exhaustive example may be: if two known disease gene sets are associated with related diseases, they should be close to each other in the protein or gene network.

DOSE [20] is a well-known tool for scoring similarities between diseases. It uses Disease Ontology (DO) [15] to associate each disease with an identifier, in order to compute semantic similarity between correlated concepts; genetic information and diseases not mapped by DO are disregarded. *DOSE* can apply both Jiang [14] and Wang [19] scores for inference.

In this paper, we present *IDDA*, a solution for Inferring Disease-Disease Associations (IDDA) by integrating curated biomedical ontologies and medical dictionaries.

2 Materials and Methods

In this section, we describe the methodology applied by our solution for integrating and processing the following sets of data: ClinVar [16], MedGen [11], DO, Gene Ontology (GO) [1], and DisGeNet [17].

IDDA integrates the mentioned datasets, to produce its own dataset which enables the enrichment of information related to genes and diseases.

2.1 Datasets

In this section, we propose a description for each dataset of curated biomedical ontologies and dictionaries used by *IDDA*.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_8

ClinVar provides an archive for human medically relevant variants and phenotypes. The phenotypic descriptions available in ClinVar are based on the information maintained by MedGen.

MedGen is a catalogue of human disorders and phenotypes with a genetic component, released by the National Center for Biotechnology Information.

Human disorders are also catalogued in DO, which consists of a set of terms linked hierarchically by using interrelated subtypes.

GO describes the fundamental characteristics of genes and their products in a species-independent manner.

A set of curated GDAs is available in DisGeNET using the UMLS Concept Unique Identifier (UMLS-CUI) and Entrez gene unique integers (GeneID) to identify the disease and the gene, respectively.

2.2 Gene-Disease Associations

Formally, let D be a set of diseases and G be a set of genes, such that $D = [d_1, d_2, ..., d_n]$ and $G = [g_1, g_2, ..., g_m]$, with n and m respectively the size of D and G.

IDDA performs the cross-referencing as the Cartesian product to build a domain for GDA:

$$\forall g \in G \; \exists d \in D : f(d, g) \rightarrow GDA$$

2.3 Disease-Disease Associations

Let $D = [d_1, d_2, ..., d_n]$ be a set of unique diseases, and $G = [g_1, g_2, ..., g_m]$ be a set of unique genes, respectively with a size of n and m.

Assuming $GDAs$ as the complete set of GDAs extracted by *IDDA*, and each GDA as pair (d_x, g_y) with $d_x \in D$, $g_y \in G$, $1 <= x <= n$ and $1 <= y <= m$.

Let denote $Gd_1d_2 = [cg_1, cg_2, ..., cg_k]$ the subset of k common genes (cg) identified for a specific DDA.

A DDA (d_1, d_2) is formally identified in accordance with the following conditions:

$$\forall cg_i \in Gd_1d_2 \; \exists (d_1, cg_i) \in GDAs, (d_2, cg_i) \in GDAs : \; Gd_1d_2 \; \subseteq \; (GDAs \cap d1, d2)$$

with $1 \leq i \leq k$, $d_1 \in D$, and $d_2 \in D$.

2.4 Score evaluation

IDDA calculates an own score useful to provide a weight for each association.

Similarly to *DOSE*, Jiang is used by *IDDA* as default method to calculate the semantic similarity based on MF between genes.

DOI: 10.1007/978-3-031-63772-8_8

Formally, Jiang is an Information Content (IC)-based score that can be defined as follows:

$$sim(d_1, d_2) = 1 - \min(1, IC(d_1) + IC(d_2) - 2 \cdot IC(MICA))$$

IDDA calculates its own score for a DDA, to evaluate an associative rank. Let DDA be the association between two diseases $D1$ and $D2$. Let $G1 = g_{11}, \ldots, g_{1n}$ be a set of genes with $n = |G1|$ related to $D1$, and $G2 = g_{21}, \ldots, g_{2m}$ be a set of genes with $m = |S|$ for $D2$ networks with $m = |G2|$. A DDA is evaluated when there exist one or more common genes between $D1$ and $D2$, formally if $G1 \cap G2 \neq \{\}$.

IDDA measures the pairwise gene similarity by using Jaccard index [12]. The latter computes the proportion of shared genes between $G1$ and $G2$ relative to the total number of genes of $D1$ and $D2$, normalizing the number of common genes in each DDA ($|G1 \cap G2|$).

Formally, the Jaccard index (J) between $G1$ and $G2$ is defined as:

$$J(G1, G2) = \frac{|G1 \cap G2|}{|G1| + |G2| - (|G1 \cap G2|)}$$

with $0 \leq J(G1, G2) \leq 1$. More generally, $J(G1, G2) = 1$ when $G1 = G2$, otherwise, $J(G1, G2) = 0$ when $G1 \cap G2 = \{\}$.

IDDA evaluates two main concepts that we denoted as internal and external similarity: IS and ES, respectively.

The former concerns the average semantic similarity among the common genes between $D1$ and $D2$, assuming $DDA(D1, D2)$, related to $DDA(D1, D2)$.

The latter concerns the average semantic similarity between the other genes belonging to $D1$ and $D2$. Both of these are normalized applying the Jaccard index (reported as J).

Formally, IS and ES are defined below, as well as the semantic similarity cross-function (f). The latter is based on the Jiang method; it is denoted below with $jiang(A)$ or $jiang(A, B)$, with A and B two generic sets of genes without duplicates. Note that $jiang(A)$ performs a score for each of the combinations of A, while $jiang(A, B)$ performs a score between all pairs of genes (A_i, B_j) with $1 \leq i \leq |A|$ and $1 \leq j \leq |B|$.

Formally, the semantic similarity function implemented in *IDDA* is defined as follows:

$$f(G) \longleftarrow \frac{\sum_{i=1}^{|G|-1} \sum_{j=i+1}^{|G|} G_{ij}}{n}$$

with $G = jiang((G1 \cup G2) \otimes (G1 \cup G2))$ (duplicates are discarded).

Internal similarity (IS):

$$IS = f(G1 \cap G2) \cdot J$$

External similarity (ES):

$$ES = jiang(G1 - (G1 \cup G2), G2 - (G1 \cap G2)) \cdot (1 - J)$$

***IDDA*'s score**:

$$IDDA\_score(D1, D2) = 1 - (IS + ES)$$

The value related to "IDDA_score" is expressed within the range $[0, 1]$, where 0 represents a condition of no similarity while 1 represents perfect similarity (the latter can be obtained by comparing a disease with itself, or by comparing two concepts related to the same disease).

3 Results and Discussion

This section reports the results performed to evaluate the efficiency and the validity of *IDDA*.

In preprocessing, a list of $318,001$ diseases was acquired using the ClinVar, that was integrated with DO for extracting $102,851$ disease's identifiers. Based on preprocessed data, a set of $461,633$ GDAs are extracted from DisGeNET.

IDDA identified a preliminary set of $19,957,259$ DDAs with a high redundancy, that was processed producing $10,283,680$ DDAs. The latter are reported as associations (d_1, d_2, cg), with cg the number of common genes between d_1 and d_2. Additionally, these are linked to DO terms to allow a comparison with other methods that supports only DO as source for the information. Therefore, the resulting dataset consists of $5,705$ DDA associated to a unique term in DO.

Homogeneous subgroup are isolated to identify similarity within the samples in *IDDA*'s dataset; this task was performed by using K-Means [18] as clustering algorithm. Furthermore, the elbow method [10] is applied, to determinate the optimal no. of clusters (k). Briefly, the elbow method selects the number of clusters to be such that adding a cluster does not significantly reduce the within-group sum of squares.

In our experimentation, *IDDA* was compared with *DOSE* (see Section 1), to evaluate its performance.

Note that *DOSE* applies the Jiang score on the DO's graph, thus the result is not related to the no. of genes (or other genomic information), contrarily to *IDDA*. Furthermore, *DOSE* was used to map the *IDDA* results on DO graph, by applying the Wang method on pathways related to each pair of diseases and the Jiang method for evaluating gene similarities.

We performed One-Way ANOVA tests as statistical analysis, to check the following hypothesis:

- differences among clusters in testing dataset are statistical significant both for *IDDA* and *DOSE* based on Jiang score.
- for each disease exists a correlation between its *IDDA*and the related pathway in DO.

Table 1 shows results for the first hypothesis. The One-Way ANOVA test between *IDDA* and *DOSE* is statistically significant. This confirms that the

DOI: 10.1007/978-3-031-63772-8_8

clustering (for $k = 3$) produced relevant groups that are effectively able to identify subgroups for the testing dataset.

Furthermore, this test verifies that both *IDDA*'s score and *DOSE* are able to identify the degree for a DDA. This hypothesis suggests that (i) there is also a correlation between the two dependent variables, and (ii) *IDDA*'s score is able to evaluate a DDA in according to *DOSE* method. Note that the two methods use different approaches, respectively the first evaluates the gene similarity, while the second evaluates the similarity by using DO information which do not contain genomic data.

		Sum of Squares	**Df**	**Mean Square**	**F**	**sig.**
IDDA	**Between Groups**	0.269	2	0.135	41.680	< 0.01
	Within Groups	0.446	138	0.003		
	Total	0.715	140			
DOSE	**Between Groups**	1.223	2	0.612	41.736	< 0.01
	Within Groups	2.022	138	0.015		
	Total	3.245	140			

Table 1. First hypothesis. The One-Way ANOVA test was performed between *IDDA* and *DOSE*. It confirms that the clustering (with $k = 3$) has produced relevant groups that are effectively able to identify subgroups for the testing dataset. The result is statistically significant. *Note: Df is the degrees of freedom.*

The second hypothesis is evaluated by performing a Bivariate (Pearson, two-tailed) Correlation between *IDDA* and *DOSE*, as shown in Table 2. The result is statistically significant.

		IDDA	**DOSE**
IDDA	**Pearson Correlation**	1	0.483**
	Sig. (2-tailed)		< 0.01
	N	141	141
DOSE	**Pearson Correlation**	0.483**	1
	Sig. (2-tailed)	< 0.01	
	N	141	141

Table 2. Second hypothesis. Bivariate correlation (Person, two-tailed) between *IDDA* and *DOSE* scores.

Statistical analysis confirms that *IDDA* is able to identify a set of DDA that can be checked by other relevant methods applied to DO.

4 Conclusion

In this paper, we proposed *IDDA*, a solution to infer DDAs by integrating ontologies, gene set enrichment analysis, and semantic similarity among GO terms.

IDDA extracts DDAs based on an in-house score, which uses GDAs for its evaluations. In a preliminary step, it performs data enrichment to retrieve concepts both for diseases and genes, by integrating several curated biomedical ontologies and medical dictionaries: ClinVar, MedGen, DO, DisGenet and GO.

Our experimentation has been conducted to evaluate *IDDA*'s score validity, by comparing results with other relevant methods, as well as by mapping each DDA to DO.

Acknowledgements

This work was funded by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of 'Innovation Ecosystems', building 'Territorial R&D Leaders' (Directorial Decree n. 2021/3277) - project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009. This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**(1), 25–29 (May 2000)
2. Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, W., Abbasi, H.M.: A survey of ontology learning techniques and applications. Database (Oxford) **2018** (Jan 2018)
3. Cinaglia, P., Cannataro, M.: Network alignment and motif discovery in dynamic networks. Network Modeling Analysis in Health Informatics and Bioinformatics **11** (10 2022). https://doi.org/10.1007/s13721-022-00383-1
4. Cinaglia, P., Cannataro, M.: Identifying candidate gene-disease associations via graph neural networks. Entropy (Basel) **25**(6) (Jun 2023)
5. Cinaglia, P., Cannataro, M.: A method based on temporal embedding for the pairwise alignment of dynamic networks. Entropy **25**(4) (2023). https://doi.org/10.3390/e25040665
6. Cinaglia, P., Guzzi, P.H., Veltri, P.: Integro: an algorithm for data-integration and disease-gene association. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2076–2081 (2018). https://doi.org/10.1109/BIBM.2018.8621193
7. Cinaglia, P., Milano, M., Cannataro, M.: Multilayer network alignment based on topological assessment via embeddings. BMC Bioinformatics **24**(1) (Nov 2023). https://doi.org/10.1186/s12859-023-05508-5, http://dx.doi.org/10.1186/s12859-023-05508-5
8. Cinaglia, P., Tradigo, G., Cascini, G.L., Zumpano, E., Veltri, P.: A framework for the decomposition and features extraction from lung dicom images. In: Proceedings of the 22nd International Database Engineering & Applications Symposium. p.

31–36. IDEAS '18, Association for Computing Machinery, New York, NY, USA (2018)

9. Cinaglia, P., Vázquez-Poletti, J.L., Cannataro, M.: Massive parallel alignment of rna-seq reads in serverless computing. Big Data and Cognitive Computing **7**(2) (2023). https://doi.org/10.3390/bdcc7020098, https://www.mdpi.com/2504-2289/7/2/98
10. Fukuoka, Y., Zhou, M., Vittinghoff, E., Haskell, W., Goldberg, K., Aswani, A.: Objectively Measured Baseline Physical Activity Patterns in Women in the mPED Trial: Cluster Analysis. JMIR Public Health Surveill **4**(1), e10 (Feb 2018)
11. Fung, K.W., Bodenreider, O.: Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc pp. 266–270 (2005)
12. Fuxman Bass, J.I., Diallo, A., Nelson, J., Soto, J.M., Myers, C.L., Walhout, A.J.: Using networks to measure similarity between genes: association index selection. Nat. Methods **10**(12), 1169–1176 (Dec 2013)
13. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabasi, A.L.: The human disease network. Proc. Natl. Acad. Sci. U.S.A. **104**(21), 8685–8690 (May 2007)
14. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR **cmp-lg/9709008** (1997)
15. Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H., Schriml, L.M.: Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. **43**(Database issue), D1071–1078 (Jan 2015)
16. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J.B., Kattman, B.L., Maglott, D.R.: ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. **46**(D1), D1062–D1067 (Jan 2018)
17. Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., Sanz, F., Furlong, L.I.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. **45**(D1), D833–D839 (01 2017)
18. Steinley, D., Brusco, M.J.: Initializing k-means batch clustering: A critical evaluation of several techniques. J. Classification **24**(1), 99–121 (2007)
19. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F.: A new method to measure the semantic similarity of GO terms. Bioinformatics **23**(10), 1274–1281 (May 2007)
20. Yu, G., Wang, L.G., Yan, G.R., He, Q.Y.: DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics **31**(4), 608–609 (Feb 2015)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_8