Annotation driven optimized clustering for disease genes batch identification

Federico Francone Marianna De Santis Stefano Lucidi Paolo dell'Olmo
Lorenzo Farina Laura Palagi

February 28, 2019

Abstract

...

1 Introduction

The application of network science towards identifying, preventing, and treating diseases is known as Network Medicine, a field of research that has considerably grown over the last decade. Network medicine relies on the idea that using networks to represent complexity of gene regulation, metabolic reactions and protein-protein interactions may help to understand the causes and mechanisms of diseases [4]. Among the networks studied, the human interactome is the most important one, as it represents the interactions among cellular components, such as proteins and genes associated to them. One of the main goals of Network Medicine is to detect those genes whose mutations are involved in diseases, named as disease genes (or seed genes). It is still uncertain whether such genes have unique properties that distinguish them from non-disease genes. From a network perspective, this amounts to asking whether they are randomly localized across the human interactome or they show some kind of patterns in the topology, as for instance hubs or modules. In our work, we tried to identify such patterns of disease genes using a novel top-down approach.

2 Our method

By looking at a specific disease and at its known disease genes, we can notice that there are some biological functions that occur more frequently than others. In order to detect new disease genes, we use the idea that having those annotations rise up the probability for a gene to be seed. Furthermore, we also take into account topological characteristics of the interactome graph. In particular, we keep as a valid assumption the fact that new disease genes may be found by looking "close to" the known seeds in the interactome graph.

2.1 Selection of meaningful information

We considered two kinds of annotations: one using GO terms (Gene Ontology - Biological Process [2]) and the other one using KEGG pathways [3]. In our analysis we consider the interactome according to the Ghiassan et al. [5] dataset, made of 13458 genes.

Let $i = 1, \ldots, P$ be a generic seed gene related to a specific disease and $i = P+1, \ldots, 13458$ the remaining genes in the interactome. Let $j = 1, \ldots, \sharp_{ANN}$ be a generic annotation. For each database (i.e. GO terms and KEGG pathways), we define x_{ij} as

$$x_{ij} = \begin{cases} 1 & \text{if annotation } j \text{ occurs in gene } i \\ 0 & \text{otherwise} \end{cases}$$

Then, we define as N_j the number of times that annotation j occurs in the known disease genes:

$$N_j = \sum_{i=1}^{P} x_{ij}.$$

On the other hand, we produced M random samples of P genes collected from all the genes in the interactome that are not known as seed genes. Namely, we picked the random samples from the 13458 genes in the (known) interactome excluding the P seed genes.

Let $k = 1, \ldots, M$ be a generic sample, let $i = P + 1, \ldots, 13458$ be a generic gene and $j = 1, \ldots, \sharp_{ANN}$ be a generic annotation.

In order to measure how frequent an annotation occurs in a generic sample of genes, we define x_{ijk} as a boolean variable that is 1 if annotation j occurs in gene i belonging to sample k, 0 otherwise. Then, similarly as before, we count the number of times that annotation j occurs in the genes of the generic k-th sample:

$$N_{jk} = \sum_{i=1}^{P} x_{ijk}.$$

Finally, we consider the mean value of N_{jk} among the samples considered and we define $\bar{N}_j$ as a measure of the frequency that an annotation j occurs in the non-disease genes:

$$\bar{N}_j = \frac{1}{M} \sum_{k=1}^{M} N_{jk} = \frac{1}{M} \sum_{k=1}^{M} \sum_{i=1}^{P} x_{ijk}.$$

The rationale behind our approach is that annotations with highest N_j should characterize a disease gene. On the other hand, annotations with highest $\bar{N}_j$ should give us some information on how to distinguish a non-disease gene from a disease one.

In the following, we denote by J a set of $2N$ annotations: the first N elements of J correspond to the N annotations with highest N_j, the second N elements of J correspond to the N annotations with highest $\bar{N}_j$. The idea is that these $2N$ annotations are able to help us in identifying new disease genes as it will be explained in the next section.

2.2 Optimized clustering phase and selection of putative genes

For each gene i in the interactome, we built a vector A_i with $2N$ components defined as

$$A_{ij} = \begin{cases} 1 & \text{if annotation } j \in J \text{ occurs in gene } i \\ 0 & \text{otherwise} \end{cases}$$

Note that, the vector A_i related to a disease gene i should have several 1 in its first N components and few 1 in its second N components. On contrary, the vector A_i related to a

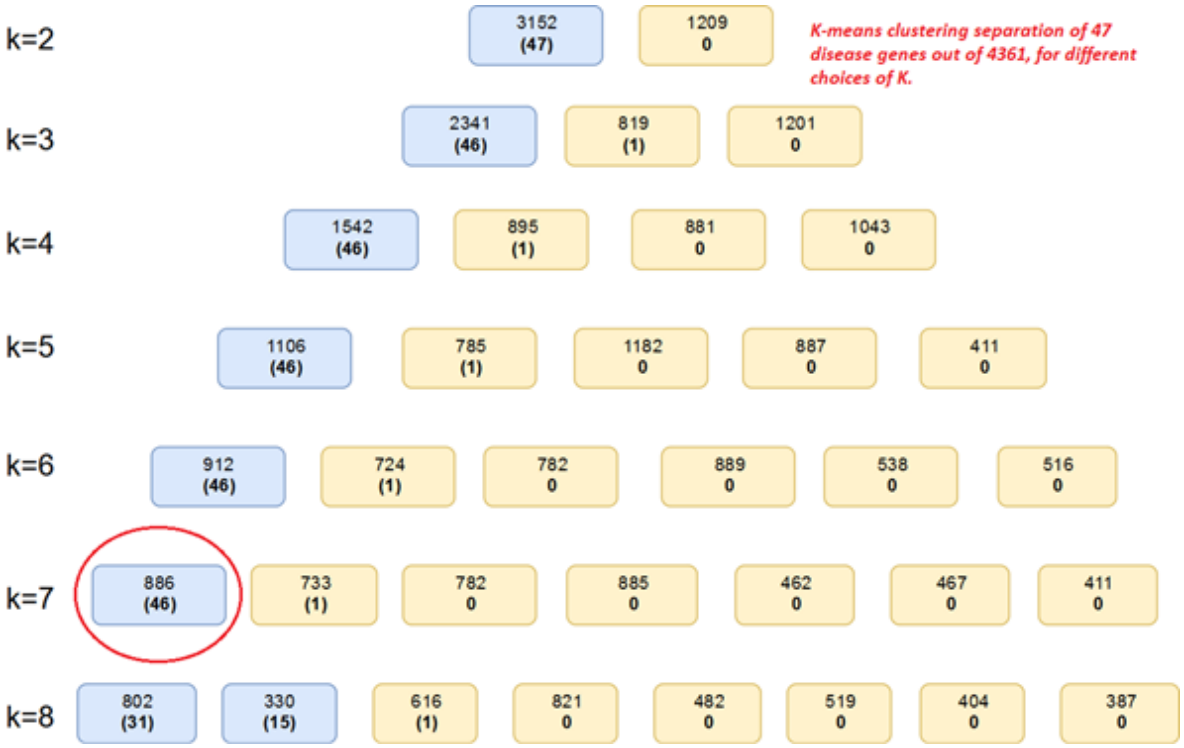

Figure 1: Clustering process applied to the "Carbohidrate metabolism inborn errors" - GO Database annotations [2]

non-disease gene i should have several 0 in its second N components and few 0 in its second N components. Our starting assumption is that it should be possible to distinguish between disease and non-disease genes on the basis of "similarities" among vectors A_i. Clearly, it is not a priori known how to recognize such "similarities", so that, in practice, we cannot use these features in order to decide if a generic gene is a disease gene or not. In order to overcome this issue, we rely on well known data mining techniques, such as clustering methods introduced in the literature in order to discover "hidden similarities" among set of vectors. In particular, there can be several "hidden similarities" and the main challenge in applying clustering methods is that of finding the "right ones". Our attempt has been that of using an adaptive strategy in order to guide clustering methods to group the majority of known disease genes in a specific cluster. The idea is that this specific cluster should contain unknown disease genes, besides containing known disease genes.

In particular our approach has been the following. First, we consider a k-means clustering algorithm [6]: let k be a fixed integer number, the k-means clustering algorithm separates the vectors A_i into k clusters. Starting from $k = 2$, we iteratively apply the k-means clustering algorithm to the set of vectors A_i, until we are able to identify a reasonable value for k. Note that, when applying the algorithm the first time, namely when $k = 2$, all the known disease genes belong to the same cluster, due to our particular choice of the set J. Then, we go on applying the algorithm increasing incrementally the number k of clusters by one.

We end up with the number of clusters k_{max}, that is the maximum number of clusters so that at least the 90% of disease genes belong to the same cluster.

In order to limit the possible introduction of false positives, we repeat the same procedure

onto two different datasets, namely GO terms (Gene Ontology - Biological Process [2]) and KEGG pathways [3]. We end up with two sets of genes C_i^*, $i = GO, KEGG$ (the first one obtained using GO terms and the other one obtained using KEGG pathways) and considering the intersection $C_{GO}^* \cap C_{KEGG}^*$, we obtain a batch of known and putative disease genes. In order to carry out the final predictions of new disease genes, we further use the topological structure of the interactome. More specifically, we selected the core set of putative new disease genes by considering those that are seed genes' first neighbors in the interactome network. This was motivated by the fact that disease genes of a specific disease tends to be grouped in a same region of the human interactome [4].

In our implementation, we used Python Sklearn library, provided by Scikit-learn website, in order to import the K-means algorithm implementation. Last but not least, the 'init' parameter has been set equal to "k-means ++", so that the initial cluster centers are selected in a smart way to speed up convergence.

2.3 Validation and numerical testing

The proposed procedure may strongly depend on the way we choose the disease genes considered as input. In order to reduce the sensitivity of our method, we repeated overal process 100 times, varying the choice of input disease genes. In particular, at every repetition, we randomly take 70% of the P known disease genes as input, so that the set J of $2N$ annotations used in the clustering phase changes. Therefore, at the end of each repetition, we obtain a different core of putative disease genes. We consider as final putative disease genes those that appear in at least the 40% of the 100 core sets obtained.

In order to validate our approach, we use a more recent dataset, namely the diseasome updated to 2018 [1]. We check wether the final putative disease genes found belong to this diseasome, with the aim of having a biological confirmation of our results.

In Table 1, we report the results obtained on the following diseases: "Breast neoplasm", "Alzheimer", "Ulcerative colitis", "Chron disease", "Multiple sclerosis", "Carbohydrate metabolism inborn errors". The first column (#p), reports the number of final putative disease genes, i.e. the number of genes that appear in at least the 40% of the 100 core sets obtained. In the second column (#ps), we report the number of putative genes that are known as seeds for other diseases. In the third column (#pvD), we have the number of putative genes that have been validated through the diseasome updated to 2018 [1]. The last columns (#pvDo), reports the number of putative genes seeds for other diseases that have been validated trhough the diseasome updated to 2018 [1] as disease genes for the specific disease in exam.

3 Biological interpretation

4 Conclusions

In this work, we devised a method to detect new seed genes involved into human diseases. Our procedure is based on an optimized clustering process that leads to the identification of a so called set of putative disease genes. Robustness of the method has been verified through validation. Applying our procedure to specific diseaseS we obtained predicted seed genes that could be further validated. Future work will be devoted to perform "in vitro" experiments.

Table 1: Results obtained on six different diseases

Disease name	#p	#ps	#pvD	#pvDo
Breast neoplasm	67	13	60	12
Alzheimer	17	2	14	2
Ulcerative colitis	74	25	31	8
Crohn disease	144	46	54	18
Multiple sclerosis	118	45	49	20
Carbohydrate metabolism inborn errors	12	5	5	2

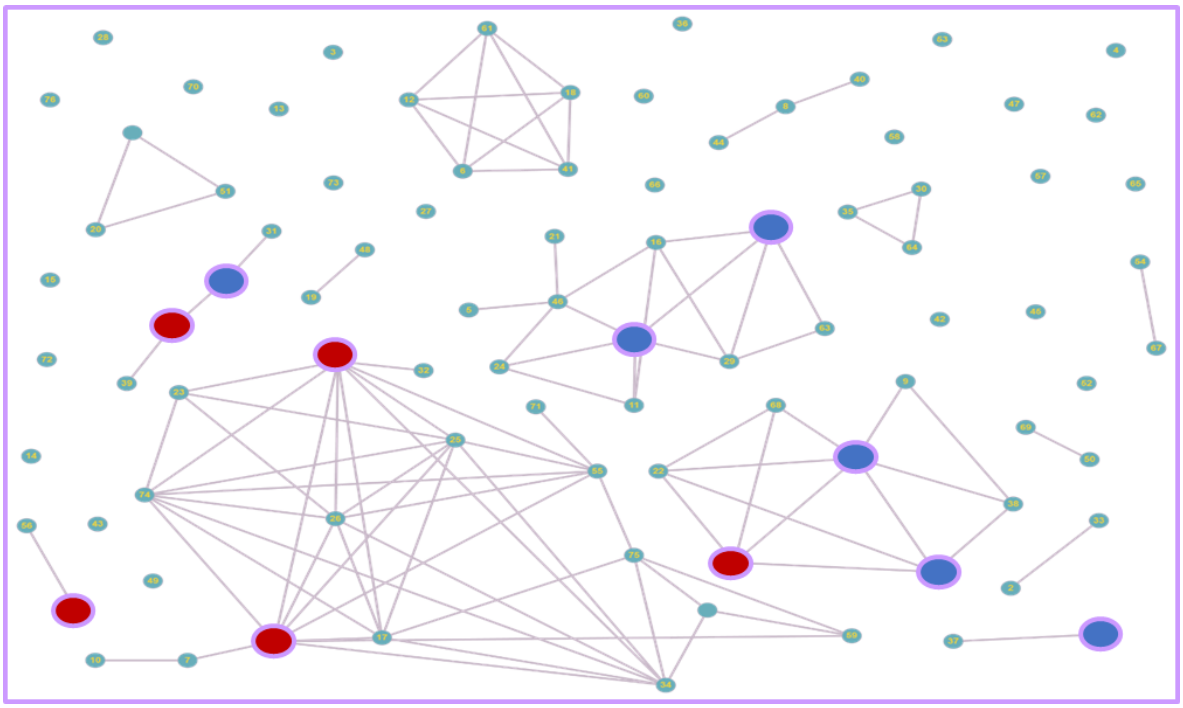

Figure 2: Network representing how the 11 putative genes are disposed and linked with the 77 disease genes for the "Carbohidrate metabolism inborn errors". Putative genes that are known as seeds for other diseases are highlighted in red.

References

[1] Disgenet, 2018.

[2] European go database, the official database of go tournaments in europe.

[3] Kegg pathways database.

[4] N. Gulbahce A.-L. Barabási and J. Loscalzo. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2011.

[5] A.-L. Barabási. S.D. Ghiassan, J. Menche. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Computation Biology*, 11(4), 2015.

[6] Shokri Z. Selim and Mohamed A. Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, (1):81–87, 1984.