Uncovering disease-disease relationships through the incomplete interactome

Robin Petit[1] and Tom Leenaerts[1]

[1]Université Libre de Bruxelles

Abstract

This paper intends to work on results exposed in Menche et al. [2015].

1 Introduction

In Menche et al. [2015], authors applied disease genes databases (in particular OMIM and GWAS) on the human interactome in order to determine the properties of their distribution in the graph. Major results were that: firstly diseases tend to *cluster* in denser subgraphs than the interactome (shown by bigger largest connected component than expected in random interactome subgraphs), secondly that phenotypically close diseases tend to overlap on a significant amount of genes.

NOTE: references expressed as Sx refer to the original paper's supplementary materials. Any other reference is to this very paper, unless explicitly mentioned.

2 Reproducing results

The first part of this paper focuses on the reproduction of exposed results in Menche et al. [2015], namely the disease modules propensity to cluster into highly connected components, {TODO: COMPLETE}.

The interactome used in Menche et al. [2015] contains 13460 genes and 141296 physical genes. OMIM and GWAS databases allowed the authors to work on 299 diseases.

2.1 Clustering of disease modules

Figure S4.b plots the relative size of each disease module versus its relative size (defined as the quotient of the largest connected component size by the number of genes related to the disease).

When plotting the same data making 10^5 random simulations per disease and setting the significance threshold to be 1.6[1], the obtained result is shown on Figure 1, which fits the one presented in the original paper.

[1]Considering the distribution to be normal as in Barraez et al. [2000], a z-score ≥ 1.6 represents a p-value ≤ 0.05 which corresponds to considered *significant* results since the test is right-tailed.

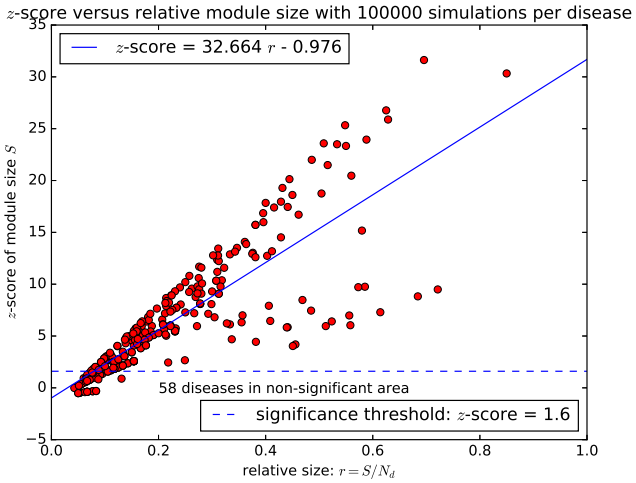

Figure 1: z-score of largest connected component size vs relative module size

2.2 Separation distribution

Original paper's figure 3.K-L plots the separation distribution of the disease pairs according to their overlapping score (J-score and C-score defined respectively as $|A \cap B| / \min(|A|, |B|)$ and $|A \cap B| / |A \cup B|$ for A and B two diseases).

3 Databases update

4 Updated results

5 Interpretation

6 Improvements

6.1 Subgraph largest connected component distribution

The z-score plotted in Figure 1 requires a null hypothesis, being the random one. Those are computed as follows: if S_D is the disease module associated with a given disease D, then its z-score is given by:

$$z\text{-score} = \frac{|S_D| - \mu(S^{\text{rand}})}{\sigma(S^{\text{rand}})}, \tag{1}$$

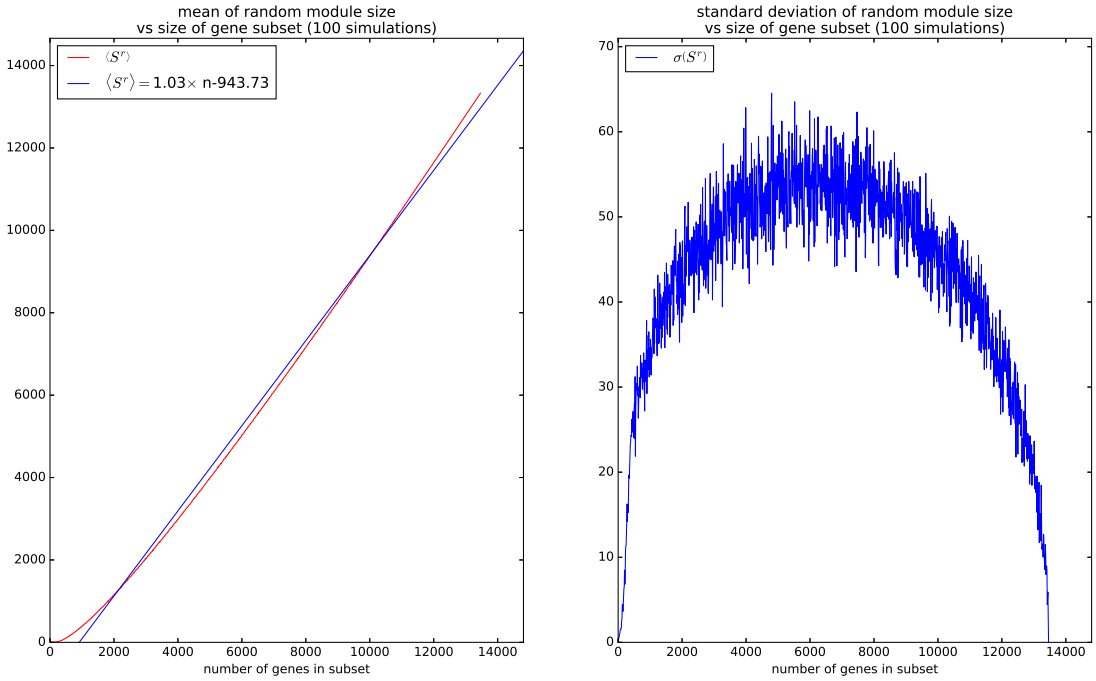

Figure 2: S^{rand} mean and standard deviation distribution

with $\mu(S^{\text{rand}})$ and $\sigma(S^{\text{rand}})$ being respectively the mean and the standard deviation of the largest connected component size of a random subgraph of size $|D|$ in the interactome.

These values are obtained by simulations: taking subgraphs at random of given size in the interactome. With 10^2 simulations per subgraph size, Figure 2 plots simulated mean and standard deviations of largest connected component size versus subgraph size.

7 Conclusion

References

Barraez, D., Boucheron, S., and Fernandez De LaVega, W. (2000). On the fluctuations of the giant component. *Comb. Probab. Comput.*, 9(4):287–304.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224).