# 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) in order to reproduce them, and update the different datasets to check if the presented procedure, which intends to be a systematic analysis, still stands for current versions of both the interactome and the disease genes associations databases.

## 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 itself (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, the significantly lower separation indicator values for highly gene related disease pairs, {TODO: complete}.

The interactome used in Menche et al. (2015) contains 13460 genes and 141296 physical links. 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.

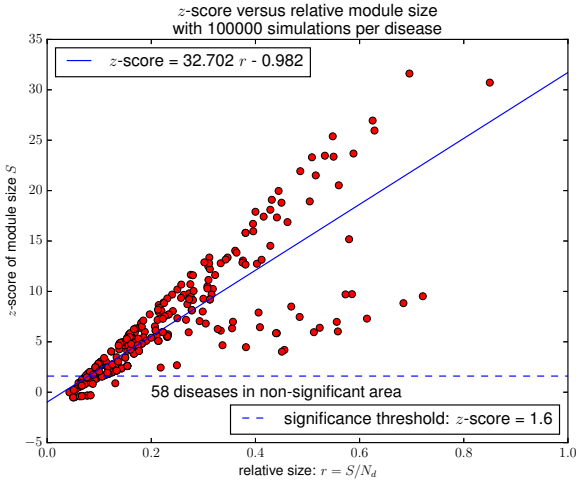


Figure 1: z-score of largest connected component size vs relative module size.

### 2.2 Separation distribution

Original paper's figure 3.K-L plots the separation distribution of the disease pairs according to their overlapping score ($C$-score and $J$-score defined respectively as $|A \cap B| / \min(|A|, |B|)$ and $|A \cap B| / |A \cup B|$ for $A$ and $B$ two diseases). Disease pairs $AB$ having a null $J$-score (and thus $C$-score) do not share any gene since $|A \cap B| = 0$. Disease pairs $AB$ having a $C$-score equal to 1 are either identical (if their $J$-score equals 1) or one is a strict subset of the other.

Figure 2 plots the distribution of $s_{AB}$ separation indicator. Distribution shown on (**A**) and (**B**) fit almost exactly the one presented in the article, the only difference being that for non-overlapping disease pairs, the amount of pairs having a

[1]Considering the distribution to be normal as in Barraez et al. (2000), a $z$-score $\geq 1.6$ represents a $p$-value $\leq 0.05$ which corresponds to considered *significant* results.

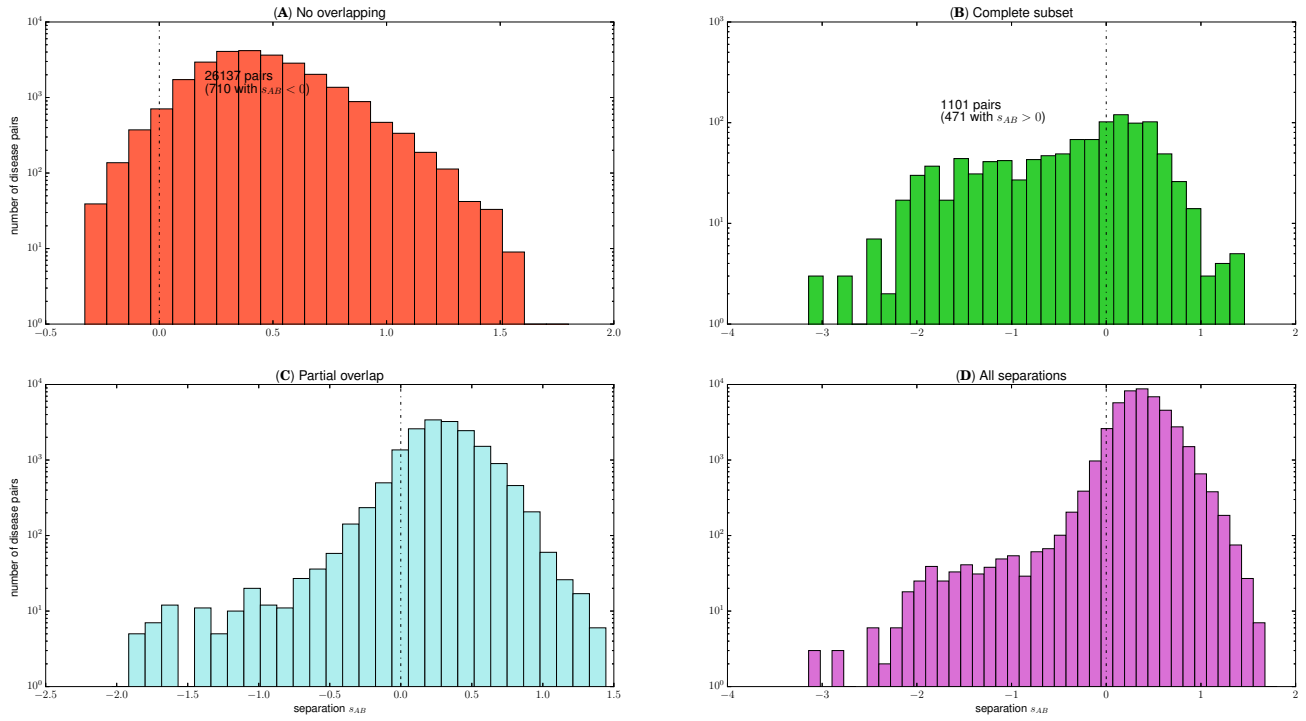


Figure 2: Disease pairs separation.

separation value below 0 is 710 versus 717 in the original paper, being totally non-significant since 7 pairs represent less than $3\%$ of all the non-overlapping pairs set.

# 3 Databases update

## 3.1 Interactome update

In order to update the interactome, the tool *inter-tool* presented in Catabia et al. (2017) has been used. Latest datasets from BioGRID, IntACT, and the Database of Interacting Proteins (DIP) (on July 25th) and of MINT (on July 27th) have been downloaded and merged via inter-tools. This newer version merged with the original interactome yield a new graph having 17786 nodes and 370326 edges (so a bit more than 1.3 times the initial amount of nodes, and more than 2.6 times the initial amount of edges).

## 3.2 Disease genes update

Disease genes considered in this update are the same 299 diseases studied in Menche et al. (2015). An update of these is yet to be done in further investigations.

## 3.3 Updated results

By defining the density of a graph $\Gamma = (V, E)$ as being $d(\Gamma) := |E| / \binom{|V|}{2}$, we find that the newer interactome is more than 1.5 times denser than the one used in Menche et al. (2015).

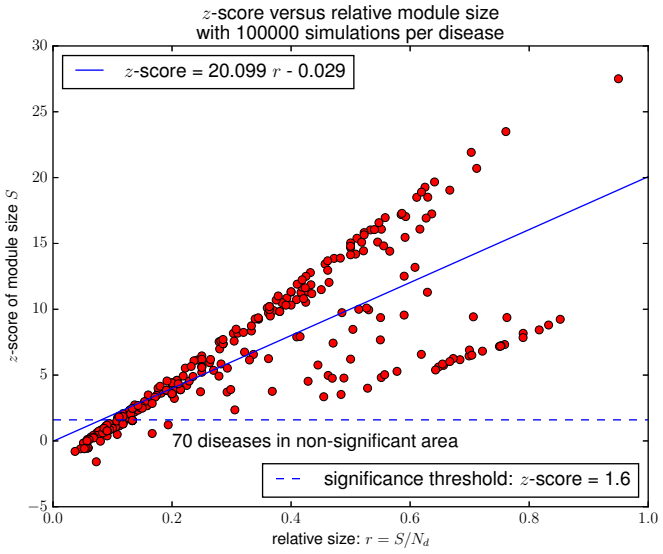


Figure 3: z-score of the largest connected component size vs relative module size of the newer interactome

Figure 3 is the adaptation of Figure 1 for the newer interactome. We observe that 12 more diseases have a $z$-score below the threshold of 1.6, which makes results less significant, whereas general relative size has shifted towards right as seen in Figure 4. This is explained by the increase in the gene number, leading to a higher coverage of the disease genes.

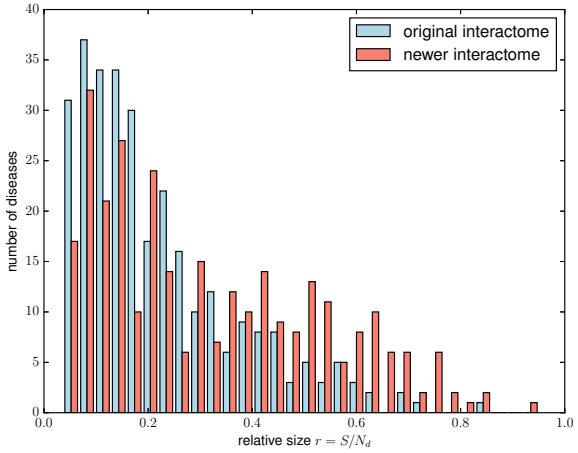


Figure 4: Comparison of relative size distribution between original and newer interactomes.

Due to the density increase, the degree distribution has changed as well. Figure 5 shows the degree distribution of both the original interactome and the new one. The most visible change is that the newer interactome contains more connected nodes which implies a mean twice as big.

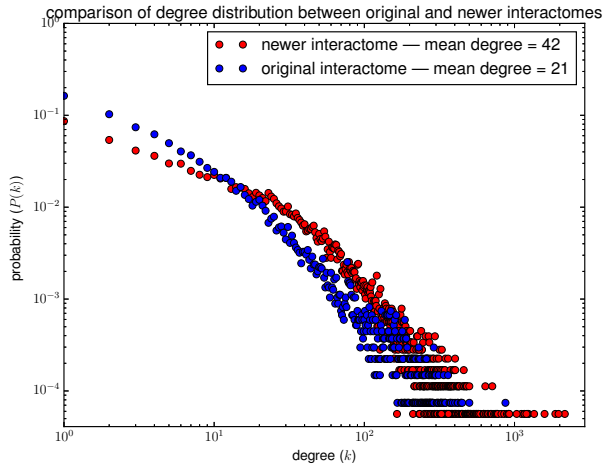


Figure 5: Degree distribution comparison.

The 12 diseases which have a decrease of $z$-score below 1.6 is due to the increase of the interactome density implying that a subgraph taken at random tends to have a wider LCC at equal size.

# 4 Interpretation

# 5 Improvements

## 5.1 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:

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

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 $10^3$ simulations per subgraph size, Figure 6 plots simulated mean and standard deviations of largest connected component size versus subgraph size.

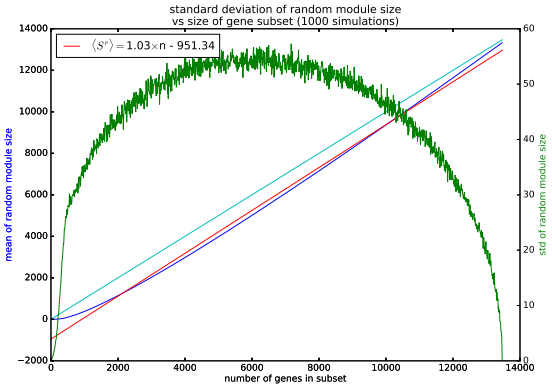


Figure 6: $S^{\text{rand}}$ mean and standard deviation distribution.

# 6 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.

Catabia, H., Smith, C., and Ordovás, J. (2017). Inter-tools: a toolkit for interactome research.

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).