First results – Motivation

The first part of this project aimed to recreate the Di Bernardo's paper (https://www.nature.com/articles/s41540-017-0022-3 - Figure 2) plot of gene expression distance compared to the 3D chemical structure distance of the compounds that are included in the CMap O2 (https://portals.broadinstitute.org/cmap/index.jsp) and discover any correlation between them or another mapping method to get better results. In order for the results to be considered better, they should provide a better correlation coefficient than the initial ones as they are shown in Di Bernardo's research.

Gene-Level Distance

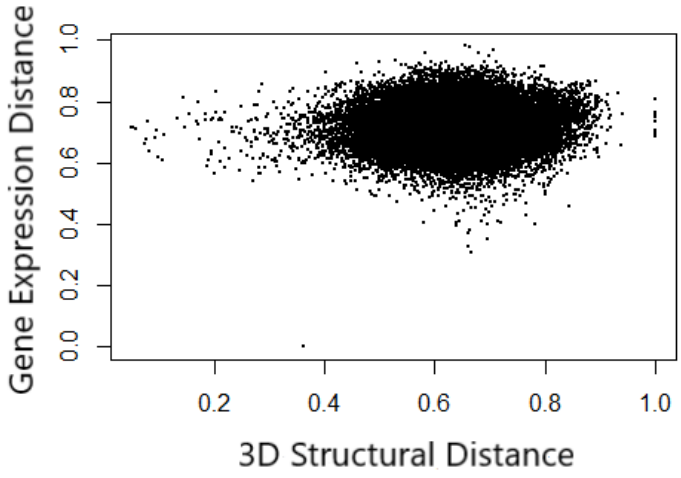
In order for the gene expression distance to be calculated, the rank matrix, which was available at CMap O2, was used. This matrix contained all the compound perturbations in ranked form, with the probes at the rows and the compounds at the columns. It was used as input at a GSEA score algorithm and its results were the gene level distances of the compounds. In this way, the gene level distance was defined for all the compounds. Compounds with the same name and cell line were firstly filtered according to their concentrations, keeping the most representative and then they were merged through a rank merging technique, to provide a unique perturbation column for each compound. In this way, a pairwise distance matrix was built, containing all the compounds of the dataset.

3D Structural Distance

The 3D distances of the compounds were provided by the Di Bernardo datasets.

Gene-Level Distance vs 3D Structural Distance

Using the above mentioned pairwise distances, the following plot was constructed, representing the pairwise gene-level distance of the compounds at the y-axis and the pairwise 3D chemical distance of the compounds at the x-axis. This plot is similar to Di Bernardo's result and it also has more pairs than Di Bernardo's, since there, only 784 compounds are present, because only these have at least one ATC annotation. With the available data, the plot constructed is shown below.

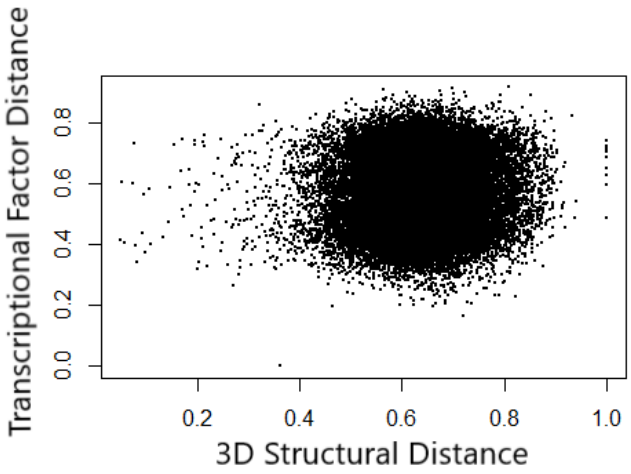


The plot does not seem to uncover any correlation between the two distances and this is also proved through the mathematical correlation computed between the two distance matrices. On this ground, the distance on the transcription factor level will be computed to investigate the possibility of a better correlation between this distance and the 3D structural distance of the instances.

Transcription Factor-Level Distance vs 3D Structural Distance

In order to calculate the distance between each pair of drugs in terms of transcription factor activity, the VIPER algorithm was used. The gene expressions of all compounds were calculated and bound to a total matrix, where the non-common probes of all the cell lines were removed (the number of probes removed is too small regarding to the whole set of probes, ~1%). Afterwards, this matrix was used as input to the VIPER algorithm and it provided a transcription factor activity matrix for the whole dataset. Every column of this matrix corresponds to a single compound and its rows are the transcription factors (279 in total).

Since the transcription factor activity matrix is constructed, its GSEA score is computed. The result of this score is a distance matrix, in which are all the pairwise distances between the compounds that were in the resulting matrix of VIPER. This distance is plotted against the 3D structural distances of the compounds and results to the following plot.



As it can be seen from the plot, there is hardly any correlation at the transcription factor level as well. The observation that can be made is that the data points are more scattered than the ones in the gene level distance. This fact may mean that the transcription factor distance is a better metric, since the differences between the points are more distinct. The 3D structural distances are the same with the first plot, so any difference between these two plots is coming from the transcription factor distances.

The better performance of the transcription factor as distance metric of the pairwise distances of the compounds might be result of the fact that the transcription factors are the direct target of a drug, and thus they provide a better insight into the biological differences of the drugs. Moreover, genes are the products of the transcription factors. So, studying the transcription factors brings the whole distance measurement to a higher biological hierarchy level. This level change could be proved less prone to noise inserted from the experimental measurements and provide better results.

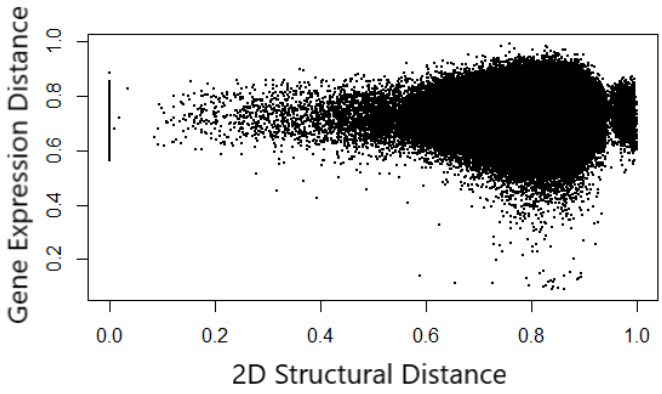
In conclusion, using the transcription factor activity processed with the GSEA algorithm as distance metric seems to result in a better biological metric, but still uncorrelated with the pairwise 3D structural distance of the compounds.

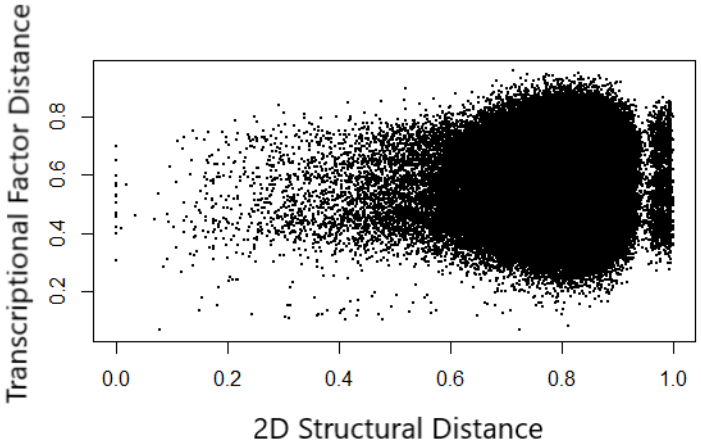
2D Structural Distance

Since none of the above plots and correlations was satisfying, an attempt is carried out in order to construct a better metric regarding the structural distance of the compounds, which was the only unaltered characteristic of both of the plots. This attempt is to find a 2D chemical structure similarity metric. This will be achieved through the SMILES of the compounds present in the dataset. Most of the compounds' SMILES can be obtained with R code. From them that are not found, the SMILES are searched through the Internet, and more specific the DrugBank database. Finally, some compounds that were absent from the database, were excluded from the research.

The pairs of compounds and their respective SMILES encoding were inserted to an algorithm in Python and its result is the whole similarity matrix of the compounds, using their 2D chemical structure as metric thanks to their SMILES.

The pairwise 2D chemical distance is then plotted against pairwise gene expression distance and also against the transcription factor activity distance. The results are the ones shown below.

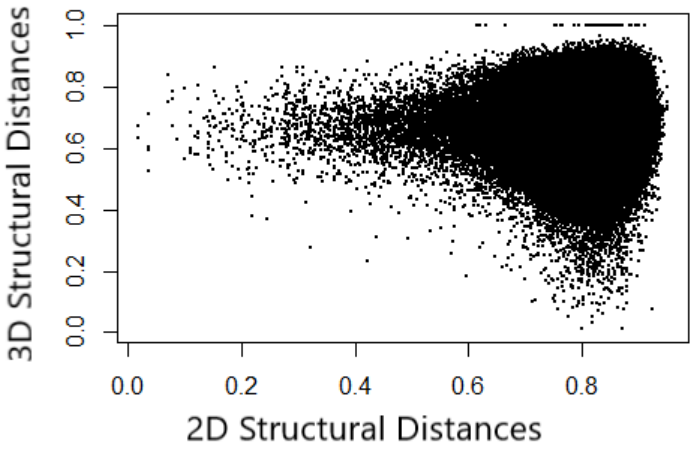




As it can be easily observed from both of the diagrams, the correlation between the gene expression distance and the 2D structural distance and also between the transcriptional factor activity distance and the 2D structural distance is near to zero.

Again, the mapping of the distances at a transcription factor level seems to provide more scattered and distinct points in the plot. So it can be supposed once more that the distance in terms of transcriptional factor activity is more reliable than the gene expression one. Yet, in both of the cases, there is no correlation between these biological-effect based distances and the 2D chemical structure distances of the compounds.

At this point, the calculated 3D chemical structure distances are plotted against the 2D chemical structure distances in an attempt to uncover any correlation between these two. The expected outcome of this plot would exhibit a correlation, since these are two chemical distance metrics based on the compounds' structure. But, in contrast to what is expected, the plot regarding these two structural distance metrics does not show any correlation between them.



The fact that there is no correlation between the two chemical structure metrics points out that they are probably not reliable to be used for any investigation on the compounds similarity. The algorithms used to obtain the chemical structure distances in 2D and 3D level are, as proved, not suitable to extract any safe results.

Outcome

Based on the provided results, it is obvious that the definition of a correlation between biological and chemical structure characteristics of the compounds is not possible. The main reason for this is that the chemical structure similarity algorithms that exist underperform in both 2D and 3D structure. So, the structural part will have to be taken out of the similarity metrics.

That's why the construction of a learning model is considered. This model will take as input two compounds and predict whether they are similar or not. The model's data will consist of four different biological metrics, which will be aggregated to a final function. They will be the gene expression levels, the transcription factors activity, the pathways and the networks that are linked to a compound. Through the dataset, these data are provided for three distinct cell lines (MCF7, HL60 and PC3). The model will consider each cell line as a separate dataset and thus will use all of the four metrics from

each cell line, aggregated, to decide whether the two input compounds are similar or not on biological level.