SAPIENZA
UNIVERSITÀ DI ROMA

Digital Epidemiology and Precision Medicine

Final Project - 2nd Module

Drug Repurposing to improve the therapies of Rheumatoid Arthritis, Juvenile Arthritis, Gouty Arthritis and Infectious Arthritis

Jeremy Sapienza 1960498

Master Degree: Data Science
a.y. 2021/2022

12/23/2021

Abstract

Rheumatoid arthritis (RA) is a long-term autoimmune disorder that primarily affects joints. It typically results in warm, swollen, and painful joints. Pain and stiffness often worsen following rest. In this paper, I want to analyze also other kinds of Rheumatoid arthritis like: the juvenile arthritis, the infectious arthritis and the gouty arthritis. The aim of this analysis is to find through the network analysis the drugs repurposable considered as 'old drugs' that could be reused for new therapies. To do this task, I considered mainly SAveRUNNER tool. This is an incredible use of the network analysis that allows to be a valid, faster and cheaper alternative for different studies that require a lot of money and time to find a drug that could be optimal for a kind of disease. To finish this report is reported the real medical indication considering each drug taken in the analysis bringing some interesting correlation with the diseases studied. At the end is inserted an extra tool (GSEA) to filter more the drugs considered.

1 Introduction

Rheumatoid arthritis (RA) affects about 1% of the population and prefers women 3 to 1 over men. In women, its incidence varies over the years and increases from menarche to just before menopause. However, the disease is rarely observed under 45 years of age. Being young and male implies a low risk factor for sporadic rheumatoid arthritis, but exists the juvenile arthritis. It is a disease caused by inflammation of the joints. The peak of onset is between 1 and 4 years of age (females are more affected than males with a 3:2 ratio), this could be considered in pair with the known Rheumatoid arthritis. Most commonly, the wrist and hands are involved, with the same joints typically involved on both sides of the body. The disease may also affect other parts of the body, including skin, eyes, lungs, heart, nerves and blood. Two diseases are involved in this problem: the infectious and the gouty arthritis. The Infectious arthritis is an infection of the fluid and tissues of a joint usually caused by bacteria, but sometimes by viruses that can spread to a joint via the bloodstream or from a nearby infection, causing an infection. Instead, the Gouty arthritis is a type of arthritis due to the formation of small uric acid crystals in and around the joints, which causes sudden attacks of severe pain and swelling. Although, these diseases are similar today some of these diseases don't have an effective aware drug. So, our aim is to find through SAveRUNNER the repurposable drugs that could be recommended to doctors. The steps on which the tool is pre-processed and takes the analysis are described in the following sections.

2 Materials and Methods

In this section are explained which data are treated and which steps are considered for our analyses. In particular the steps are:

1. Computation of network proximity and p-values
2. Computation of network similarity
3. Selection of proximal drug-disease associations
4. Cluster detection
5. Adjustment of network similarity
6. Normalization of network similarity

SAveRUNNER is the network-medicine-based algorithm used for drug repurposing. This tool creates a bipartite drug-disease network quantifying the interplay between the drug targets and the disease-specific proteins in the human interactome via a novel network-based similarity measure that prioritizes associations between drugs and disease locating in the same neighborhoods.

2.1 Data

For doing these analyses with the tool, I provided these inputs (mandatories):

1. Human Interactome by Cheng et al.
2. Disease Genes by Phenopedia → Checking for 4 diseases in analysis
3. Drug targets by Drug Bank → handling 1858 drugs

SAveRUNNER takes the input list of diseases-genes and the list of drugs-targets. The representation of these lists are considered as graph networks. The intuition of the algorithm is based on the **network similarity measure** of the networks in the human interactome:

$$f(p) = \frac{1}{1 + e^{-c\left[\frac{(1+QC)(m-p)}{m} - d\right]}} \quad (1)$$

For which:

- p is the network proximity $\rightarrow$ $p(T,S) = \frac{1}{||T||}\sum_{t \in T} min\ d(t,s)\ \forall s \in S$
- QC is the quality cluster score
- m is max(p)
- c is the steepness of f(p) and d is the point such as f($\frac{(1+QC)(m-p)}{m} = d$) is equal to 0.5

To get statistical significance processes, I fixed a p-value ≤ 0.5 and leaved as default the original settings of SAveRUNNER. Now, let's see the main functions performed by SAveRUNNER.

2.2 Computation of network proximity and p-values

As the first step, It wants to find the modules of diseases and drugs that are near in the human interactome leveraging the network-based proximity mentioned before, if the proximity is equal to 0 means that the modules are closer and vice versa. This measure represents the average shortest path length between the drug targets in the drug module and the closest disease genes in the disease module. It's important to consider a statistical significance of the observed network proximity between the two modules, this process is repeated 1000 times and the observed values are z-score normalized.

2.3 Computation of network similarity

As the second step, It translates the proximity measure between a range more strict between 0 and 1. To have normalized values. The corresponding normalization is the one reported below:

$$similarity = \frac{max(p) - p}{max(p)} \quad (2)$$

p is the network proximity.

2.4 Selection of proximal drug-disease associations

As the third step, it needs to pick the statistical significant drug-disease associations, so it wants to leave the default fixed value of p-value ≤ 0.05. If we have a drug-disease associations lower than this p-value means that probably the off-label drugs (the predicted one by the tool) could be considered as repurposed for the corresponding disease.

2.5 Cluster detection

As the fourth step, it wants to do a cluster analysis to highlight the groups of drugs and diseases that are similar between each other. The steps on doing this cluster analysis are based on calculating the network modularities to detect the clusters (also called as communities). At the end we have a quality cluster score (QC) used to quantify the essence of each cluster:

$$QC = \frac{W_{in}}{W_{in} + W_{out} + P} \quad (3)$$

For which:

- W_{in} reports the total weight of edges within the cluster
- W_{out} reports the total weight of edges for the cluster connected to the rest of the network
- P is a penalty term

2.6 Adjustment of network similarity

As the fifth step, SAveRUNNER does an adjustment for the similarity measure treated as:

$$similarity = (1 + QC) \cdot similarity \quad (4)$$

if two nodes are in the same cluster the similarity is increased by a factor proportional to the QC score of the cluster which they belong, if not their similarity doesn't change the value of adj. similarity. This value could be ≥ 1.

2.7 Normalization of network similarity

As the sixth step, it applies a normalization of the values from 0 to 1 applying a sigmoid function:

$$f(x) = \frac{1}{1 + e^{-c(x-d)}} \tag{5}$$

3 Results and Discussion

At the end of these steps, I got a list of predicted associations between drugs and diseases considered as **weighted bipartite drug-disease network**. The link is associated if the modules are closer in the human interactome and the p-value ≤ 0.05, the weights of the association are adjusted and normalized. In this section are presented the results given by SAveRUNNER tool plus an extra filtering final network with suitable discussions.

3.1 Drug-Disease network with p-value ≤ 0.05

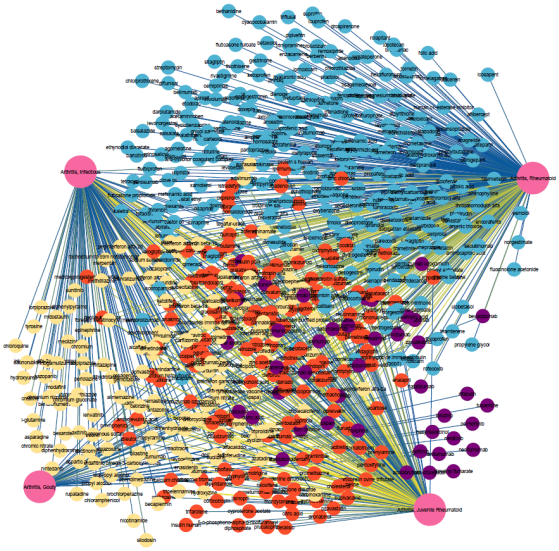

Figure 1: Drug-disease network

As we can see in Figure 1, the main important information extracted by SAveRUNNER is the drug-disease network of the most repurposable drugs that have p-value ≤ 0.05. As we can see, we represented 604 nodes and 908 edges with adj. similarity values as edge values colored from 0 to 1. The pink diseases are the main diseases treated in the analysis. The drugs are clustered by four groups of these main diseases (clusters).

3.2 Drugs repurposable for the diseases

It's possible to see the heatmap where is applied the dendrogram on it, here it's handled the Rheumatoid Arthritis and other 3 diseases of the drug-disease network. The Disease-Drug network is composed by 608 genes and 904 edges (4 diseases and 1858 associated drugs). The network is clustered according by diseases (as rows) and drug (as columns) by a complete linkage of hierarchical clustering algorithm and by using the Euclidean distance as distance metric (default setting). The colors of the adjusted similarity value shows in the heatmap the repurposability of each drug respect to the disease, going from the blue (0 value) to the yellow color (1 value) it's highlighted the fact that the drug is repurposable, Figure 2:

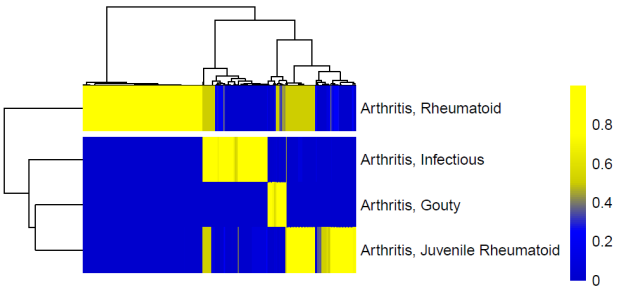

Figure 2: Repurposable drugs for each disease

3.3 Common repurposable drugs in each disease

SAveRUNNER let's the possibility to have a heatmap to see how much common/predicted drugs are repurposable for each pair of diseases:

What I expected in Figure 3 is having higher number of common drugs on similar diseases like the Juvenile and the Rheumatoid Arthritis, also for Gouty Arthritis and Infectious Arthritis. In clinical context seems (in my opinion) to have similar problems and formations when they happen in patients. Is interesting to see a higher number of common drug repurposable in the case of the first comparison, but this doesn't happen for the second comparison. This is a surprise result got by the tool and this could be pass to doctors.

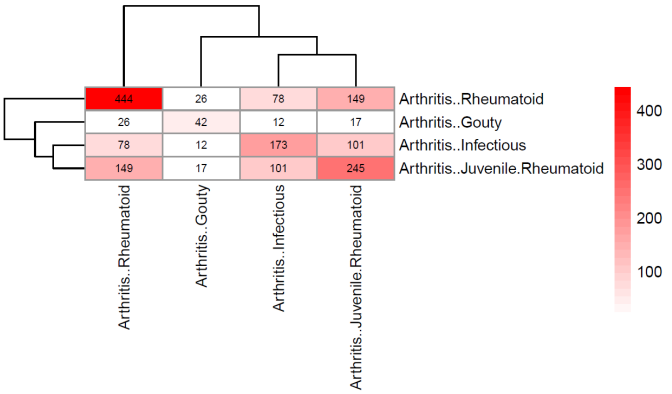

Figure 3: Heatmap clustered and represented by a dendogram of drug repurposable for common drugs for diseases

3.4 Correlation plots with adjusted similarity

SAveRUNNER gives the correlation plots that allows to split the large amount of predicted drugs repurposable between 0 and 1 by adjusted similarity value (adjusted by the size and color of the adj. similarity) for each disease treated in the analysis, splitted by 20 for each group. In fact, as column there are twenty drugs for each group generated and by rows the diseases analyze. In this case, I have 31 groups, let's have a look a piece of these groups in Figure 4

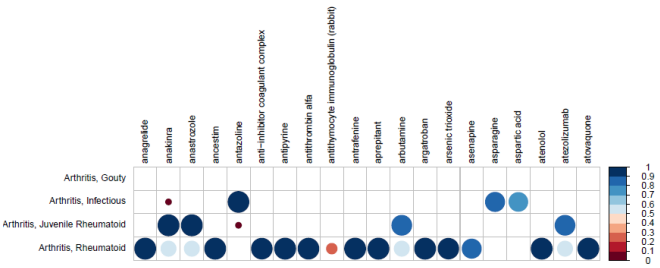

Figure 4: A correlation plot group out of the 31 groups given by SAveRUNNER

3.5 The most relevant disease

As we see, the main disease that gives interesting results is the Arthritis Rheumatoid, for which the subnetwork is reported in Figure 5

In the next sub-section, is possible to find some medical indication of this disease, inserting different other information and the corresponding real drugs used.

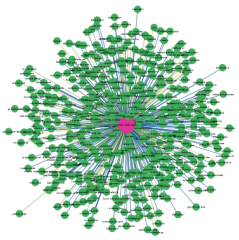

Figure 5: Arthritis Rheumatoid subnetwork

3.5.1 Medical Indication

The main data sets to handle in the medical indications are:

- Drug_Disease_network_pval0.05.txt → for the subnetwork of the rheumatoid athritis (RA)
- TTD_association.txt → for having the actual drugs used for some diseases (real case)

In this case using the main configuration file provided by an extra script outside the main tool discussed before, I get the onLabeled drugs that effectively are used, comparing each drug to the original medical indication for each disease. I consider each off-Label target drugs and then I merge these results with the original indications, and these are the results:

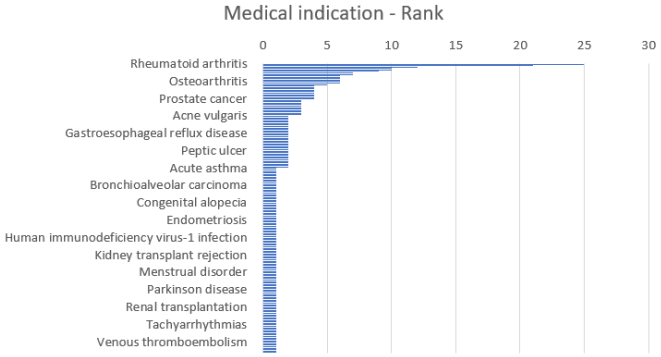

Figure 6: Medical Indication rank on our off-labels

In this case, I found 25 drugs used for the Rheumatoid Arthritis.

To end up with these medical indications, I want to plot the whole network showed in Figure 6, the major number of drugs found are used for Arthritis Rheumatoid (RA).

A more consideration is given by the subnetwork of Rheumatoid Arthritis where the adjusted similarities are ≈ 1, as we see in Figure 8, also in Figure 7 is possible to review the RA sub-network.

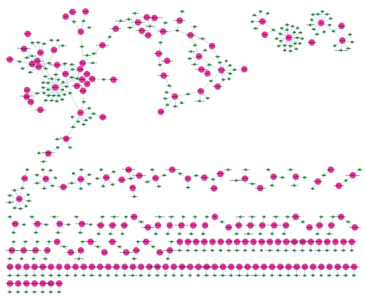

Figure 7: Medical Indications network

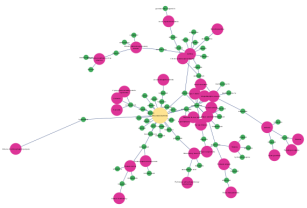

Figure 8: Medical Indications of RA subnetwork

3.6 Drugs (predicted) used by Rheumatoid Arthritis

Here are reported 3 type of drugs used for fighting the Rheumatoid Arthritis:

3.6.1 Adalimumab

Adalimumab is used, alone or in combination with other drugs, to relieve the symptoms of certain autoimmune diseases, for example rheumatoid arthritis, juvenile idiopathic arthritis, Crohn's disease, ulcerative colitis, ankylosing spondylitis, psoriatic arthritis and chronic plaque psoriasis. More details are attached in the references. This drug got these results:

p_{val}	adj_{sim}
9.08E-19	≈ 0.99

3.6.2 Infliximab

Infliximab is a medication used to treat a number of autoimmune diseases in which there is also the rheumatoid arthritis. It is given by slow injection into a vein, typically at six- to eight-week intervals. This was originally developed in mice as a mouse antibody. Because humans have immune reactions to mouse proteins, the mouse common domains were replaced with similar human antibody domains. They are monoclonal antibodies and have identical structures and affinities to the target. They are a combination of mouse and human antibody amino acid sequences, called as "chimeric monoclonal antibody". This drug got these results:

p_{val}	adj_{sim}
4.10E-5	≈ 0.99

3.6.3 Tolmetin

Tolmetin is a nonsteroidal anti-inflammatory drug. It is used primarily to reduce hormones that cause pain, swelling, tenderness, and stiffness in conditions such as the rheumatoid arthritis, including juvenile rheumatoid arthritis. In the United States it is marketed as Tolectin and comes as a tablet or capsule. This drug got these results:

p_{val}	adj_{sim}
1.14E-4	≈ 0.99

3.7 Extra Analysis (GSEA)

Here, I want to test whether the drugs predicted as repurposable drugs for rheumatoid arthritis could counteract the gene expression perturbations caused by the disease of interest. I'm interested to collected only the GSEAs < 0. The inputs in this case are the differentially expressed genes in particular the up-regulated and down-regulated of RA (could be catched by Gene Expression Omnibus) and then in CMAP, I need to put the differentially expressed genes of candidate drugs. CMAP calculates a score of correlation between drugs and disease, I want to find drug candidates that have potential treatment effect against the disease of interest, negative correlation are considered (CMAP score < 0). This is important, because we can find the drugs which their functions are useful to fight the disease. I only consider one data set provided by CMAP and the result is this plotted as multi-edges graph (by each gene corresponded to a drug) in Figure 9 with 216 nodes and 2987 edges.

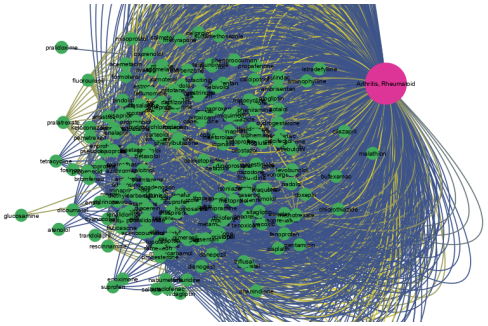

Figure 9: GSEA Rheumatoid Arthritis results

References

[1] What's the Rheumatoid Arthritis? - https://en.wikipedia.org/wiki/Rheumatoid_arthritis

[2] SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19 - https://journals.plos.org/ploscompbiol

[3] Humanitas - Adalimumab - https://www.humanitas.it/enciclopedia/principi-attivi/antinfiammatori/adalimumab/

[4] Wikipedia - Infliximab - https://en.wikipedia.org/wiki/Infliximab

[5] Cytoscape Functionalities - https://cytoscape.org/