PROJECT IN DATA MINING

"EXPLORING BIOLOGICAL TARGET RELATIONSHIPS USING JACCARD-BASED CLUSTERING AND ASSOCIATION RULE MINING"

Cláudia Afonso (nº 36273, 30 hours), Rita Rodrigues (nº 54859, 30 hours)
Group 2

Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
10th of June 2023

Abstract: Drug discovery is a lengthy, expensive and complex multi-step process that can last several years and cost hundreds of millions of dollars. To reduce costs while expediting this process, *in-silico* approaches have been systematically employed in the pharmaceutical industry. In this context, data-driven approaches that combine the mining of chemical and biological information with computer and statistical methods can play a pivotal role in advancing the field of biomedical research and accelerate the drug discovery process. In this project, we aim to use two distinct approaches to explore relationships between biological targets that are known to be bound by a set of FDA approved drugs. The first approach consists of a Jaccard-based clustering method that groups targets into distinct clusters according to the molecular fingerprint data associated with the compounds to which these biological targets are known to bind. The second approach is based on association rule mining techniques to explore the existing relationships between targets without considering the molecular fingerprint data of the active compounds.

Keywords: Molecular fingerprints, Jaccard similarity coefficient, hierarchical agglomerative clustering

---

1. INTRODUCTION

1.1 The relevance of data mining in the pharmaceutical industry

Conventional approaches to drug discovery are lengthy and expensive, lasting up to 10 years and costing more than $600 million (Mohs & Greig, 2017). To reduce these costs while increasing the efficiency and predictability of the drug discovery process, computer aided-drug design approaches have been systematically employed in the pharmaceutical industry. Enabled by the emergence of large, multi-labelled datasets resulting from traditional approaches, these *in-silico* strategies have led to discovery of several clinically approved drugs (Zhang et al., 2022). More recently, the vast amounts of biomedical information resulting from 'omics' studies has also facilitated the drug discovery process by contributing to the identification of relevant biological targets (Yang et al., 2009). Therefore, a data-driven approach that combines the mining of chemical and biological information with computer and statistical methods has the potential to greatly advance the field of biomedical research and accelerate the drug discovery process.

A central aspect in the drug discovery process consists in understanding how active compounds bind to their biological targets. Such information can be beneficial in two distinct ways. First, it enables the identification of chemical substructures within compounds that are essential to elicit a specific biological response, effectively establishing structure-activity relationships (SARs). In this context, compounds are typically represented as molecular fingerprints consisting of binary vectors where each bit encodes the presence or absence of a particular substructural fragment (Capecchi et al., 2020). These fingerprints can subsequently be used to search

for structural similarity between compounds in the sparse and high-dimensional binary chemical space (Probst & Reymond, 2018). Since compounds with a high degree of structural similarity will bind to similar biological targets, the information resulting from molecular similarity studies forms the basis of virtual screening (VS). The latter can be defined as a set of computational methods that explore large datasets of compounds to identify potential candidates corresponding to molecules most likely to interact with a desired target (Sabe et al., 2021). Secondly, drugs are rarely specific for a single target, instead producing multiple and diverse effects beyond their intended primary mechanism of action. Therefore, the pleiotropic effects of drugs resulting from their binding to various biological targets can be analysed to construct drug-target and target-target networks (Yildirim et al., 2007). Dissecting these relationships can have important implications in elucidating drug targets in the context of both cellular and disease networks, such as pointing to a common structure, function or biological pathway between targets, thus facilitating future drug discovery efforts.

1.2 Aims of the Project

In this project, two main goals were pursued. First, the structural similarity between a set of FDA approved drugs was computed according to their molecular fingerprint data. Briefly, the Jaccard coefficient between each pair of drugs, *A* and *B*, represented by binary vectors was calculated as:

$$J(A,B) = \frac{c}{c+a+b} = 1 - d(A,B)$$

where c is the number of common molecular fingerprints (bit value equal to 1) for both drugs, a is the number of molecular fingerprints (bit value equal to 1) that are exclusive to A and b is the number of molecular fingerprints (bit value equal to 1) that are exclusive to B (Han, 2012). The Jaccard score was used as the similarity metric since it is particularly well-suited to handle binary feature vectors as is the case of molecular fingerprint data. The distance matrix, obtained by subtracting the Jaccard scores for all pairs of drugs from 1, was used to group similar drugs together by performing Hierarchical Agglomerative Clustering (HAC).

After clustering drugs into similar groups, the second aim was pursued, which consisted of exploring relationships between biological targets using two distinct avenues. In the first, these relationships were explored by computing the Jaccard similarity between targets belonging to each drug cluster. The resulting distance matrices were then used to group similar targets belonging to the same drug cluster by performing HAC. It is expected targets that are bound by similar molecules share a degree of similarity themselves. By grouping targets in the same drug cluster through HAC, it is anticipated that the degree of similarity between the resulting target groups will be even greater. In the second avenue, the possible relationships between targets were explored through association rule mining.

2. DATASET

The dataset for this project includes two separate files. The first, named “fps.txt”, includes molecular fingerprint data (with 2048 attributes) for a set of 1101 FDA approved drugs. Here, the data was already encoded into binary form, where the presence or absence of each fingerprint in a molecule is encoded as 1 or 0, respectively. The second, titled “acts.txt”, contains the names of the known biological targets for the molecules. This file was processed similarly to a transactions type file, where each column represents a possible target. Here, the data was converted into binary form, where the presence or absence of an interaction between a target and a molecule is encoded as 1 or 0, respectively.

3. Results and Analysis

3.1 Aim 1 – Finding similarity between compounds based on their molecular fingerprints

Following a simple pre-processing of the data in the "fps.txt" file, the first aim of the project was pursued. It consisted in finding similarity between compounds based on their molecular fingerprint data. As previously mentioned, similarity between compounds was computed through HAC by using the distance matrix as the metric. Here, the Silhouette coefficient was employed to heuristically derive the optimal number of clusters and the most appropriate HAC linkage method for the data at hand. As a result, the Silhouette coefficient was computed for each linkage method at various discrete values for the number of clusters with a certain range. As observed from the resulting plot (Figure 1), the most suitable linkage method appears to be the 'ward' method, since the respective Silhouette score surpasses its initial value when the data is grouped in approximately 200 clusters. The resulting dendrogram following HAC through the 'ward' method (shown attached to this report due to technical constraints related to its size) was visualized to fine-tune the optimal number of clusters for the molecular fingerprint data. By cutting the dendrogram at a value of 2.5, the obtained number of clusters was 137. This value indicates that based on their similarity computed from molecular fingerprint data, the compounds can be grouped into 137 separate groups.

Another graph was plotted to understand whether the resulting division of molecules by each of the 137 clusters was balanced or not. As observed in Figure 2, this distribution is uneven with approximately a third of clusters containing more than 10 molecules and two-thirds containing less than 10 molecules. In addition, five clusters contain more than 20 molecules. This imbalance in the number of molecules per cluster suggests that a predominance in certain molecular fingerprints exists in the original dataset. Thus, the dataset of FDA approved drugs at hand likely favours some molecular substructures in detriment of others. This result is not necessarily out of the ordinary considering that some properties are desired and required for a molecule to have therapeutic effects and receive FDA approval, and this fact is subsequently reflected in the molecular composition of drugs.

3.2 Aim 2 – Exploring relationships between biological targets through clustering and association rule mining

Following a simple pre-processing of the data in the "acts.txt" file, the second aim of the project was pursued. It consisted in exploring relationships between targets using two distinct avenues. However, before proceeding to do this, the processed data from the "acts.txt" file was merged with the clustering results obtained from the first aim and a graph was then plotted to visualize the number of bound targets per drug cluster (Figure 3). As observed, this distribution is quite unbalanced, with an increased number of targets being bound in some drug clusters in comparison to others. Indeed, in some drug clusters compounds bind to relatively few biological targets (< 20), while in others the compounds are active against an extraordinarily high number of targets (> 150).

The previous result might be indicative of several possible scenarios. First, since it was observed earlier that an imbalance in the number of molecules per cluster exists, then it should not be surprising that some clusters bind to more targets than others. However, when looking attentively to the plots in Figure 2 and Figure 3, it is possible to conclude that some of the drug clusters containing more compounds (for instance, clusters 61 and 6) actually bind to a comparatively low number of targets. Thus, the fact that some drug clusters exhibit a higher number of bound targets is not necessarily related to the number of molecules within that cluster. Thus, another possible explanation for the obtained results might be related to the fact that molecules in certain clusters, which share a particular set of fingerprints, exhibit pleiotropic effects and are able to bind to a broader set of biological targets. This pleiotropy might indicate the existence of relationships between targets, which can point to a common structure, function or pathway involving these targets in biological networks. A third possible scenario might be the existence of a bias in the dataset of FDA approved drugs towards particular biological targets and associated diseases. This is also not particularly surprising, since some targets are more "druggable" and relevant to disease-associated processes than others and thus the effort expended to develop and approve drugs in such cases is considerably greater.

To explore the second scenario, meaning the existence of possible relationships between targets, two distinct avenues were pursued. The first followed a similar rationale to the one described previously for the molecular fingerprint data. Briefly, the Jaccard coefficients between targets belonging to each drug cluster were determined, with the resulting distance matrices serving as the basis to perform HAC and group together similar targets belonging to the same drug cluster. To exemplify the results from this approach, the drug cluster 61 was selected at random. As observed in Figure 4, the distribution of targets per resulting cluster in the case of drug cluster 61 was also uneven, with most belonging to two separate sub-clusters (sub-cluster 8 and 11). The resulting target names in each resulting sub-cluster were identified. As observed in the attached Jupyter notebook, there appears to be a high degree of similarity between targets in each sub-cluster, as evidenced by their names. For instance, in case of subcluster 61_8, several common targets exist, such as 'ABCB11', 'ABCB1B', 'ABCC1', 'ABCC2' and 'ABCG2', which are all members of ATP Binding Cassette subfamilies (*GeneCards - Human Genes | Gene Database | Gene Search*). Other targets present this subcluster, such as 'ALK', 'ALOX5', 'APEX1', 'ASPH', 'ATP1A1' for instance, are not evidently related to the first ones. However, it is possible that some known or unknown relationship in the literature exists between them.

The second avenue pursued in this project to explore the existence of possible relationships between targets consisted in using association rule mining techniques. The latter can enable patterns of co-occurrence or dependency between targets to be uncovered based on the molecules to which they bind. To this end, the FP-growth algorithm was employed due to its demonstrated efficiency in a previous data mining project. Based on this method, the support threshold to compute frequent targetsets was defined. From this analysis it was possible to generate 273 relevant rules satisfying minimum support (≥ 0.035) and minimum confidence (> 0.95) values, as well as additional interestingness measures such as lift (> 10), leverage (> 0.035), conviction (> 20.0) and zhang's metric (> 0.95).

Interestingly, the rules appear to be connected since they display quite similar target names but in reverse order. For instance, rules 533 and 534 have the same target name sets ("DRD3", "ADRA1A", "HTR2A" and "HTR2C"). While "HTR2A" appears as the antecedent in the same set as "DRD3" and "ADRA1A" for rule 533 with "HTR2C" showing as the consequent, for rule 534 "HTR2C" appears together with "DRD3" and "ADRA1A" and "HTR2A" is now the consequent. These types of target exchanges between antecedent and consequent sets occur throughout the generated association rules and provide additional information to the insights retrieved from the Jaccard-based clustering approach. However, it should be noted that the generated rules, while displaying high values for the interestingness measures, revolved mostly around a few targets such as "HTR2A", "HTR2B", "HTR2C", "DRD3", "ADRA1A", "ADRA1B", "ADRA2A" and "HRH1", for instance.

4. CONCLUSIONS

Using the two proposed approaches, it was possible to extract meaningful and complementary information concerning existing relationships between biological targets in the data. The first approach, which consisted of a Jaccard-based clustering method to group targets into distinct clusters according to the molecular fingerprint data associated with the compounds to which these biological targets are known to bind, provided several relationships between targets. Some of these were obvious, since the names of the involved targets were mostly the same, indicating these belong to the same or related families. However, other relationships uncovered here did not elicit the same obvious sentiment and exploring these further in the literature would be interesting. The second approach, which consisted in the application of association rule mining techniques, also provided several relationships between targets, which were distinct and thus complementary to the ones given by the Jaccard-based clustering method. Some of the resulting relationships were obvious as well based on the target names, while others were not. However, considering the size of the dataset, most generated relationships centered on a few selected targets.

5. ANNEXES

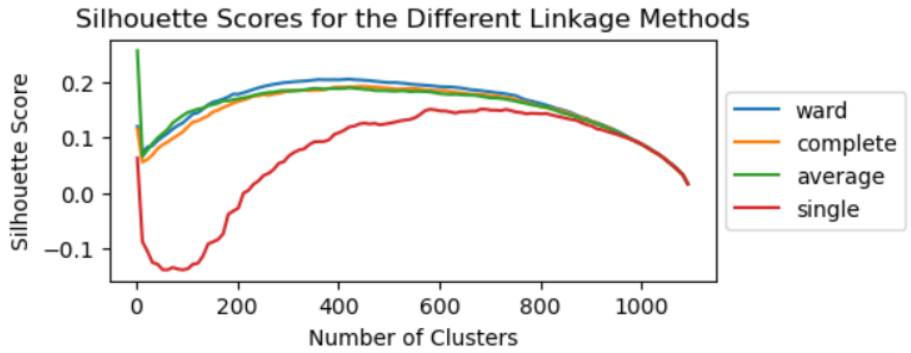

Figure 1. The Silhouette score was used to determine the most appropriate linkage method and the optimal number of clusters for the molecular fingerprint data.

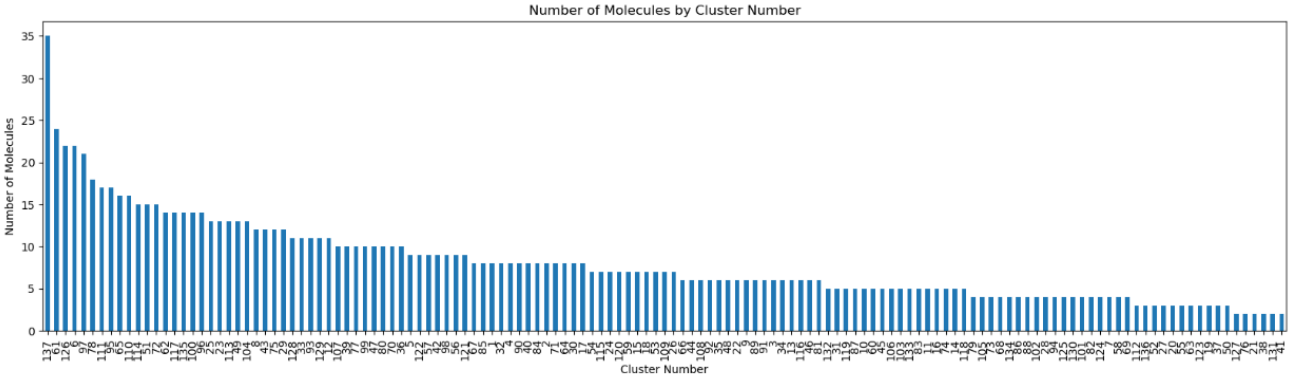

Figure 2. Distribution of the number of molecules by each of the resulting clusters.

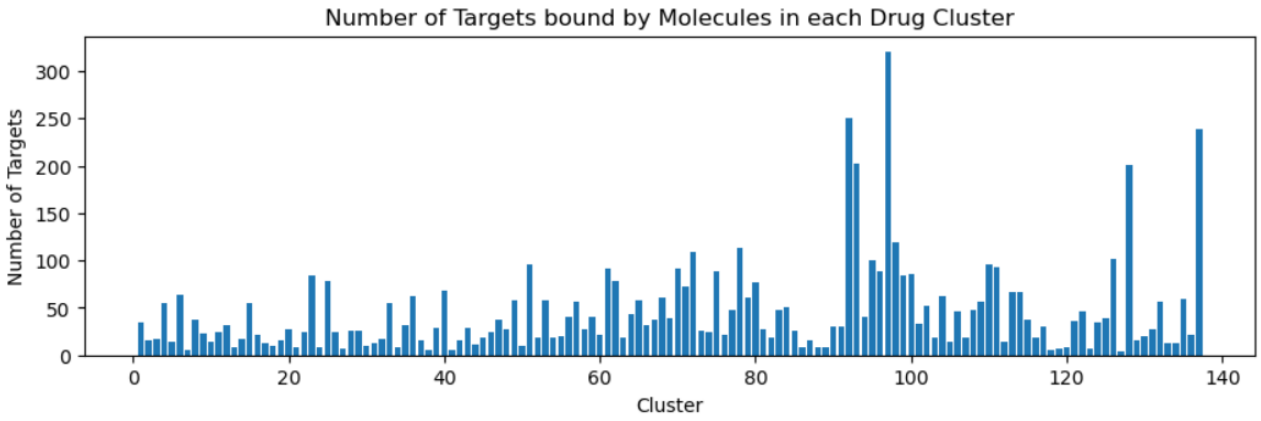

Figure 3. Distribution of targets bound by molecules in each drug cluster.

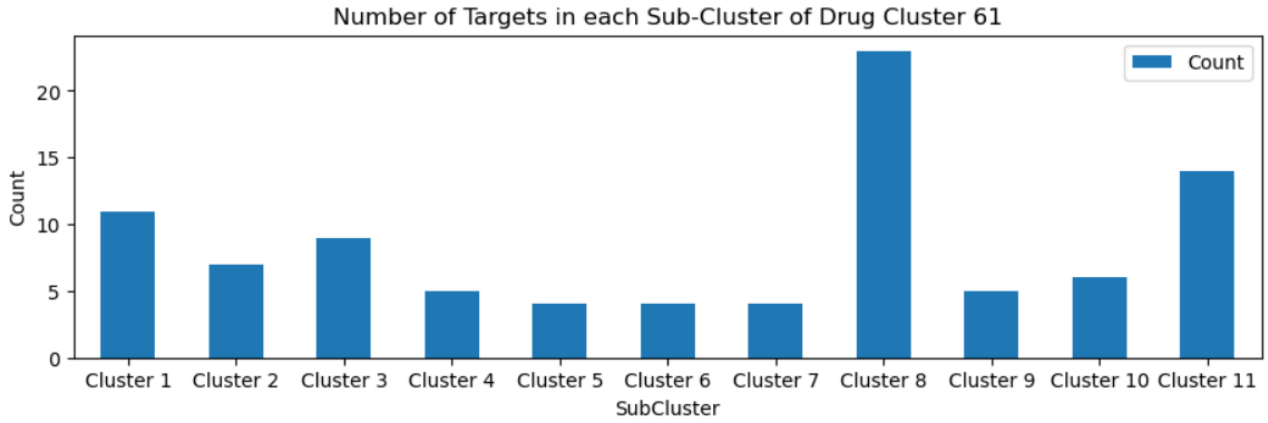

Figure 4. Distribution of targets belonging to drug cluster 61 in each resulting sub-cluster.

6. REFERENCES

Capecchi, A., Probst, D., & Reymond, J. L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminform*, *12*(1), 43. https://doi.org/10.1186/s13321-020-00445-4

GeneCards - Human Genes | Gene Database | Gene Search. Weizmann Institute of Science. Retrieved 10th of June, 2023 from https://www.genecards.org

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (M. K. Publishers, Ed. 3rd edition ed.).

Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimers Dement (N Y)*, *3*(4), 651-657. https://doi.org/10.1016/j.trci.2017.10.005

Probst, D., & Reymond, J. L. (2018). A probabilistic molecular fingerprint for big data settings. *J Cheminform*, *10*(1), 66. https://doi.org/10.1186/s13321-018-0321-8

Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur J Med Chem*, *224*, 113705. https://doi.org/10.1016/j.ejmech.2021.113705

Yang, Y., Adelstein, S. J., & Kassis, A. I. (2009). Target discovery from data mining approaches. *Drug Discov Today*, *14*(3-4), 147-154. https://doi.org/10.1016/j.drudis.2008.12.005

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., & Vidal, M. (2007). Drug-target network. *Nat Biotechnol*, *25*(10), 1119-1126. https://doi.org/10.1038/nbt1338

Zhang, Y., Luo, M., Wu, P., Wu, S., Lee, T. Y., & Bai, C. (2022). Application of Computational Biology and Artificial Intelligence in Drug Design. *Int J Mol Sci*, *23*(21). https://doi.org/10.3390/ijms232113568