Interreg
ITALIA-SLOVENIJA
UNIONE EUROPEA
EVROPSKA UNIJA
TRAIN
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

R.1.WP3.1 REPORT ON IT SERVICES

Poročilo o IT storitvah za analizo podatkov, pridobljenih iz presejalnih testov genomov in spojin/ Report su servizi IT per l'analisi congiunta di dati da screening di composti e screening genomici

Responsible partner: Jožef Stefan Institute

Contributing partner: ICGEB

Authors of the report: Tomaž Stepišnik Perdih, Dragi Kocev, Sašo Džeroski

1. Introduction

In the first period of the project, we focused on understanding the bioanalytical workflows used to analyze screening data. Our aim was to familiarize ourselves with the typical screening methods and some representative data that come out of the screenings as well as data from auxiliary sources that can contribute to better understanding of the results. The workflows for analyzing data generally consist of (1) data understanding, (2) pre-processing, (3) machine learning task identification, (4) machine learning method selection and application, (5) visualization and understanding of the models and results and (6) deployment of the models and the results for practical use. The understanding of the typical workflows used for screening data analysis form the framework for design and development of IT services to support and facilitate knowledge discovery from screening data.

More specifically, we analyzed data coming from several screens investigating: (1) heart cell regeneration and (2) fibroblast and heart smooth muscle cell proliferation. From technological perspective, we analyzed data from compound screens with the aim to identify potentially effective drugs and data from miRNA screens with the aim to find a potential targets.

2. Machine learning tasks, datasets and software

We analyzed two sets of data provided by the ICGEB.

- The measurements of Alpha Smooth Muscle Actin (SMA) and Collagen intensities in mice hearts after the administration of different drugs. The goal was to discover new (previously untested) compounds that would lower the intensity of Alpha SMA.
- The measurements of fibroblast and heart smooth muscle cell proliferation after the administration of different miRNAs. We tried to identify which genes or biological pathways can be targeted to decrease the proliferation of smooth muscle cells, while not affecting the proliferation of fibroblast.

For our analysis, we used the following tools and online databases.

- CLUS system for predictive clustering (https://sourceforge.net/projects/clus/) was used to construct predictive clustering trees and ensembles thereof for classification, regression and multi-target regression tasks, and to perform feature ranking for those tasks.
- PubChem database (https://pubchem.ncbi.nlm.nih.gov/) was used to find protein and gene targets of compounds.
- ChEMBL database (https://www.ebi.ac.uk/chembl/) was used to get structural information for compounds and additional information for drugs.
- MiRTarBase database (http://mirtarbase.mbc.nctu.edu.tw/php/index.php) was used to find gene targets of miRNAs.
- KEGG database (https://www.genome.jp/kegg/) was used to find the pathways-gene associations for homo sapiens.

3. Analysis of Alpha SMA and Collagen screening data

Our plan for discovering novel compounds that lower the Alpha SMA intensity consisted of three parts. The first part was to use the screening data to learn a predictive model, that would take a compound's description as an input and predict its effect on the Alpha SMA intensity. Second, we needed to find a database of compounds to apply the predictive model on and search for interesting candidates. Third, due to the massive amount of different compounds that exist, we needed a way to filter/prioritize the candidates.

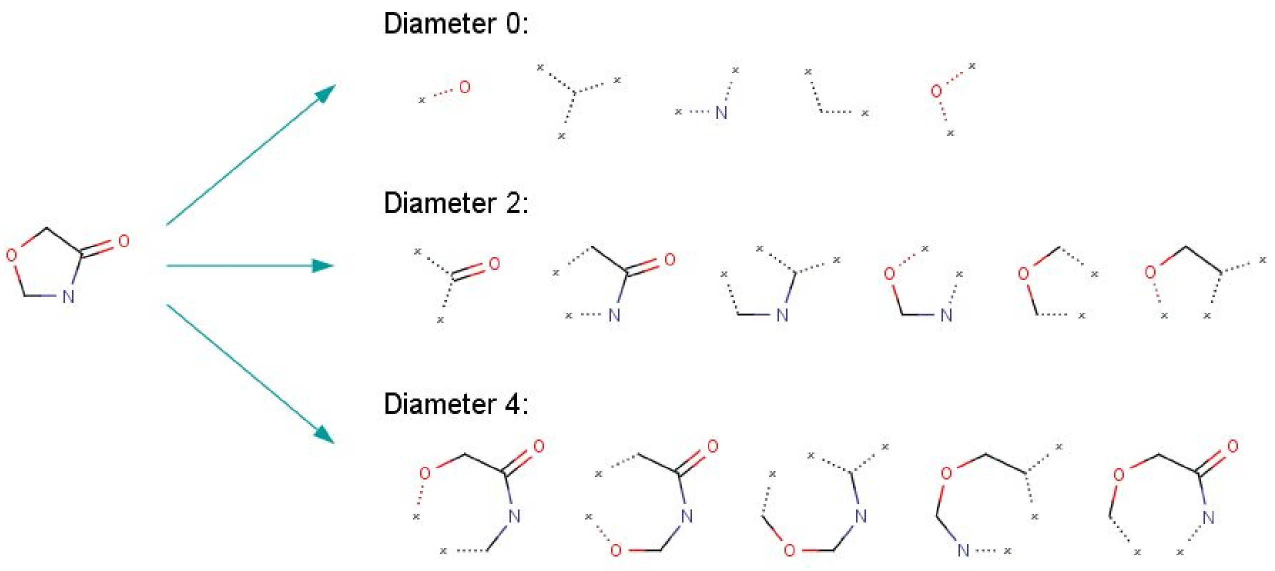

Fig 1: Illustration of the patterns that Extended Connectivity Fingerprint (ECFP) recognize in the molecules.

First, we had to decide how to describe the molecules in a way that machine learning algorithms understand. One option is to describe compounds with their bioactivity profiles. This means that each compound is described with an n-dimensional binary vector, where every component corresponds to one protein. If a compound is active on a specific protein, the corresponding component of the vector is 1, otherwise it is 0. To construct the bioactivity profiles of compounds one has to search a substantial amount of protein screening data to collect the necessary information. Online databases such as PubChem or ChEMBL are popular sources for this and we decided to use them as well. Another option to describe compounds are structural molecular descriptors. Molecules are typically represented as graphs, and over the years, multiple descriptors derived from their structure were developed. Among them, Extended Connectivity Fingerprints (ECFPs, Fig 1) are often used and typically prove well suited for predicting biological properties of molecules. With ECFPs, compounds are again described with binary vectors. Each vector component corresponds to a set of substructures. If that component has value 1, it means that at least one of those substructures is present in the molecule. We decided to try both bioactivity profiles and ECFPs to describe the molecules.

The screening data we received included the measurements of Alpha SMA and Collagen intensities for 640 compounds. The compounds were identified with CAS numbers. To search online databases for the SMILES strings detailing their structure and screens to construct their bioactivity profile, we had to match CAS numbers with other identifiers (ChEMBL and PubChem IDs). For this we used Chemical Translation Service (http://cts.fiehnlab.ucdavis.edu/) and some manual checking.

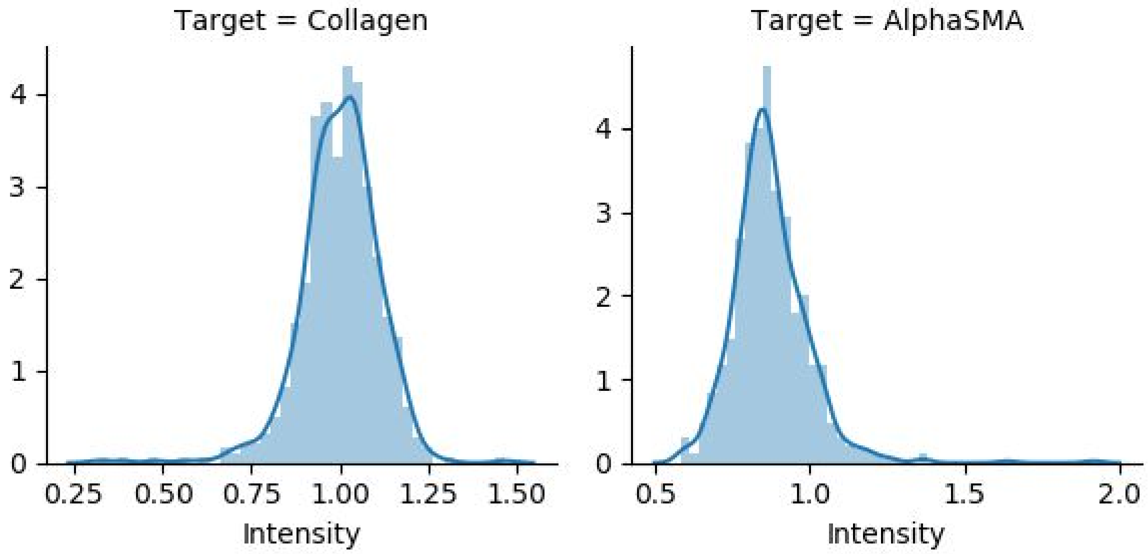

Fig 2: The distributions of Collagen and Alpha SMA intensity measurements.

For predictive models, we decided to use bagging ensembles (Fig 3) of predictive clustering trees (PCTs). Predictive clustering trees are constructed with a greedy top-down induction algorithm that recursively splits the data according to the attribute values. The algorithm selects the split that maximizes the variance reduction of the target variables. PCTs are a generalization of standard decision trees that supports structured output prediction and semi-supervised learning. In our case, the attributes are binary vectors (bioactivity profiles or ECFPs) and the target variables are Alpha SMA and collagen intensities. When no suitable test is found (too few examples or too low variance reduction), a leaf is created where the mean value is stored and used for prediction. Ensembles of trees improve the predictive performance over single trees: multiple trees are built on bootstrapped samples of the learning set. Predictions of individual trees are then aggregated (averaged) into the final prediction.

We decided to use tree-based models because they have several nice properties.

- Individual trees are interpretable: they can be inspected so that it is clear how a prediction was made.
- When used in an ensemble they offer state of the art performance.
- When used in an ensemble they can also be used to rank the attributes according to their importance for predicting the target.

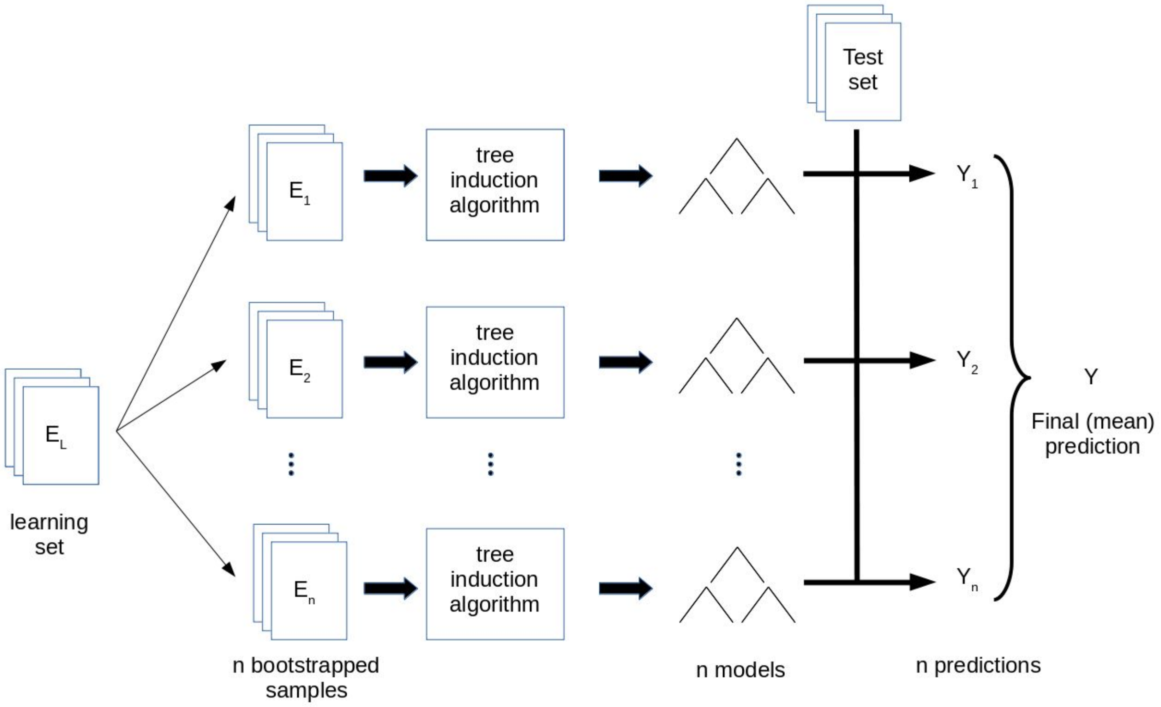

Fig 3: A schema representing ensemble learning. Multiple models are learned on bootstrapped samples of the learning set. Each model makes its predictions for the training set examples, before predictions of individual models are aggregated into the ensemble prediction.

Performance was measured with mean squared error (MSE): $\Sigma(y_i - y'_i)^2$, where y and y' are true and predicted values, respectively. To estimate the MSE we used 10-fold cross validation. The image above shows the distribution of the two target variables. The mean absolute values of bagging ensembles of regression trees were 0.0836 for Collagen and 0.0871 for Alpha SMA (Fig 4).

Since the goal was to identify interesting candidates for wet-lab experiments, we also wanted to investigate the reliability of individual predictions. For this we looked at the standard deviation of predictions of individual trees in the ensemble - if ensemble members agree with each other, we expect the prediction to be more accurate. We showed there was correlation between lower standard deviation of predictions of individual trees and lower prediction error of the ensemble prediction (Fig 5).

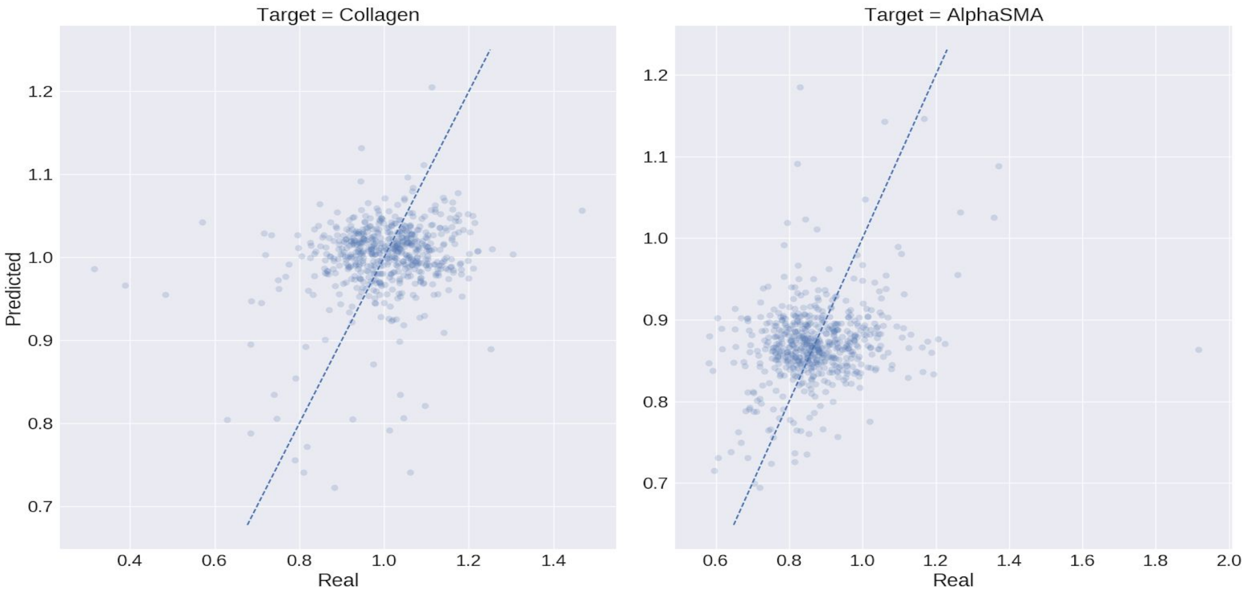

Fig 4: Plots of predicted and true values of Collagen and Alpha SMA intensities.

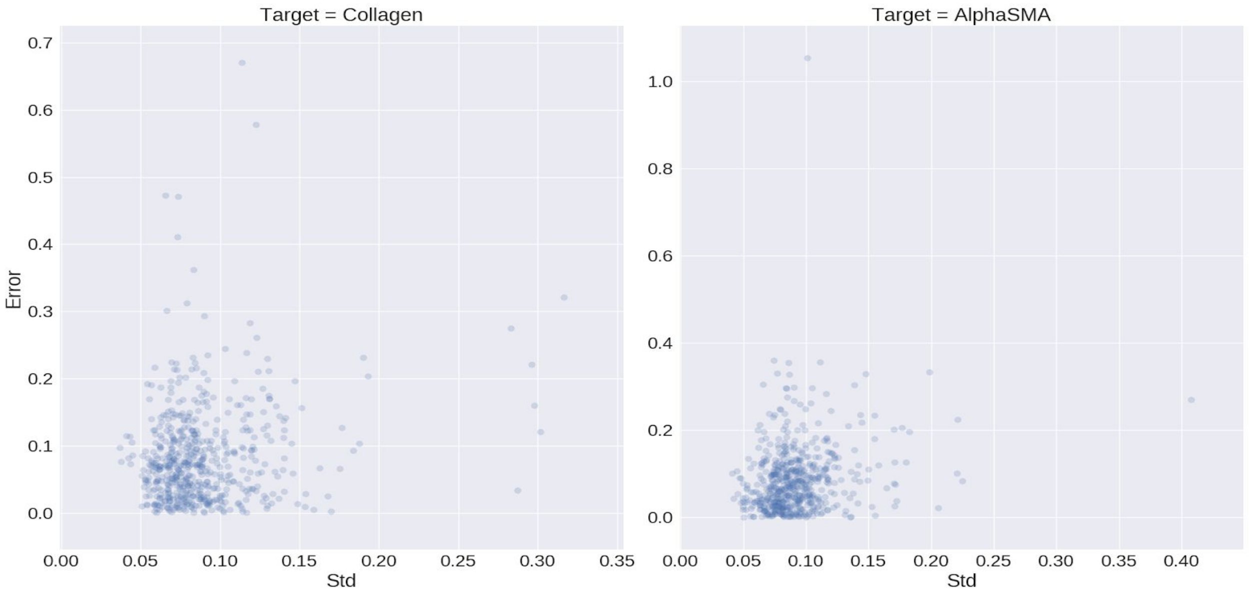

Fig 5: Plots of prediction errors in relation to the standard deviations of individual ensemble member predictions.

We then used the model constructed on all 640 learning examples to predict the Alpha SMA and Collagen intensities of all other compounds in the ChEMBL database (over 1.7 million in total). Experts at ICGEB decided that Alpha SMA intensity is a more reliable measurement than collagen, so we used it as a main criterion for sorting the candidates. Based on the distribution of Alpha SMA intensity measurements, we discarded all the compounds with predicted intensity above 0.75, leaving us with 622 interesting candidates. To make the inspection of candidates easier, we also clustered them. We repeatedly selected the candidate with the lowest predicted intensity as the prototype, and put all other candidates similar enough to the prototype to its cluster. Among the remaining candidates the next prototype was selected, and so on. For similarity we used the Tanimoto index, calculated from the ECFPs. With the similarity threshold set to 0.5, this resulted in 34 clusters of candidates. Where available, we also included the maximum phase achieved in clinical trials of the candidates, as found in ChEMBL.

4. Analysis of fibroblast and heart smooth muscle cell proliferation data

We received measurements of effects of miRNAs on fibroblast and heart muscle cell proliferations. For heart muscle cell proliferation, two rounds of screening were performed. In total, the data included 2046 unique miRNAs, 2042 of them were screened on muscle cells and 856 on fibroblast. Among them, 852 were part of all 3 screens, whereas the rest are partially labeled.

Table 1: Illustration of the miRNA data we received.

miRNA ID	fibroblast	muscle R1	muscle R2
hsa-miR-1180-3p	0.00	0.51	0.03
hsa-miR-99b-5p	0.49	0.48	0.23
hsa-miR-98-3p	?	0.10	0.19
hsa-let-7c-3p	0.48	?	?
...			

There were three main avenues of analysis of this data that we explored.

1. We described each miRNA with the genes it targets and tried to determine which genes are important for fibroblast and muscle cell proliferation with interpretable predictive models and feature ranking.
2. We used a model learned on the miRNA screening data to try and find drugs that would reduce the cell muscle proliferation while not affecting the fibroblast

proliferation. This is possible because both miRNAs and drugs can be described in the same space (with gene targets).

3. We tried an alternative description of miRNAs using the biological pathways they affect, and repeated the analysis we performed with genes.

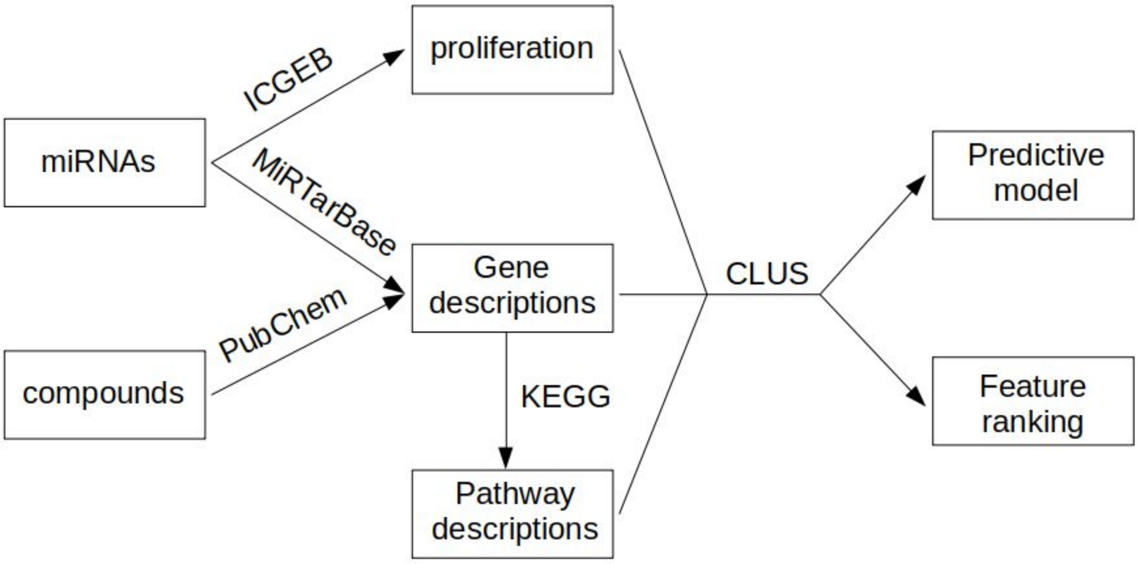

Fig 6: A schema of the experimental pipeline.

4.1 Gene target analysis

We used the miRTarBase database to find gene targets of the miRNAs. It contains 2599 miRNAs in total, 586 of them were not part of the screening. The miRNAs included in the screening target 2535 different genes with strong evidence, and 15053 different genes with less strong evidence. There are 1363 miRNAs that do not target any genes with strong evidence, but only 33 of them do not target any genes with at least less strong evidence. For this reason, we decided to include less strong evidence in the miRNA gene target profiles. This means that each miRNA was described with 15053 binary features, each feature denoting the targeting of a particular gene.

We defined a classification problem: a miRNA is interesting, if it significantly reduces the heart muscle cell proliferation and does not affect the fibroblast proliferation. The second round of screening on heart muscle cells was better for determining how much a miRNA reduces the proliferation, so we disregarded the results of round 1. Specifically, miRNAs were interesting if they resulted in muscle cell proliferation below 0.5 (in round 2) and fibroblast proliferation between 0.8 and 1.2 (1 fold proliferation means no effect). This labeling was only possible for the 852 miRNAs that were screened for both fibroblast and muscle cell proliferation. The rest were unlabeled.

On this data, we constructed a PCT for predicting whether a miRNA is interesting, as well as a PCT that predicts the fibroblast and muscle cell proliferation directly (multi-target regression task). The trees were then inspected to see which genes affect the predictions and in what way. We also constructed ensembles of PCTs for both classification and multi-target regression tasks and used them to perform feature ranking. The importance of a feature (gene) is determined by the number of times it is used for tests in the trees, and how high in the trees it appears (the lower it is, the fewer examples it influences, lowering its importance).

4.2 From miRNAs to compounds

Like miRNAs, compounds can also be described by the genes they target. To do this, we searched the PubChem database to find all compounds that target at least one of the 15053 genes used to describe the miRNAs. This gave us 532149 compounds in total, however no compounds were found that target 12846 of the genes.

We used the multi-target regression ensemble of PCTs learned on the miRNA screening data to predict the fibroblast and muscle cell proliferation effect of these compounds. This gave us a long list of candidates for inspection that we again filtered and clustered as described previously for the Alpha SMA analysis.

Additionally, we tried to determine what structures in the compounds affect the proliferation effect. To do this, we described the compounds with MACCs key fingerprints. They are 166 dimensional binary vectors where every component corresponds to a specific structural pattern in the molecule (1 if present, 0 if not). While ECFPs typically give better predictive performance, MACCs keys are much easier to interpret, because of the bijection between the fingerprint bits and molecular structures. We performed feature ranking of the fingerprints using our predicted proliferation effects as the targets, to see which molecular substructures are the most important.

4.3 From genes to pathways

We also tried an alternative approach to describe the miRNAs. In the previous analysis, each miRNA was described with 15000-dimensional binary vectors, which was a very sparse representation. However, multiple different genes are a part of the same biological pathway, meaning they can affect the proliferation in the same way. Instead of describing the miRNAs with the genes they target, we can describe them with the pathways that the targeted genes are a part of. This way we compress the representation and make it less sparse, hopefully without losing much of the relevant information for making the predictions.

To get the pathway representation of miRNAs, we used the KEGG database to find which pathways are associated with which genes. There are 330 pathways in the KEGG database for homo sapiens. Using these descriptions, we repeated the analysis performed on gene descriptions earlier.

5. Summary

We used the measurements of Alpha Smooth Muscle Actin (SMA) and Collagen intensities to build a predictive model. With it, we predicted the intensities for all the compounds in the ChEMBL database. In collaboration with experts from ICGEB, these compounds were filtered and clustered so that the most appropriate drug candidate was found. Initial experiments confirmed the candidate's potential, further validation is in progress.

We used the measurements of fibroblast and heart smooth muscle cell proliferation to try to discover which genes and biological pathways are responsible for lower heart smooth muscle cell proliferation, without affecting the fibroblast proliferation. With this information we can then go on to find drugs that inhibit those genes/pathways. That would result in better heart muscle regeneration, without toxic effects (affecting fibroblast).