MMsPred: a bioactivity and toxicology predictive system

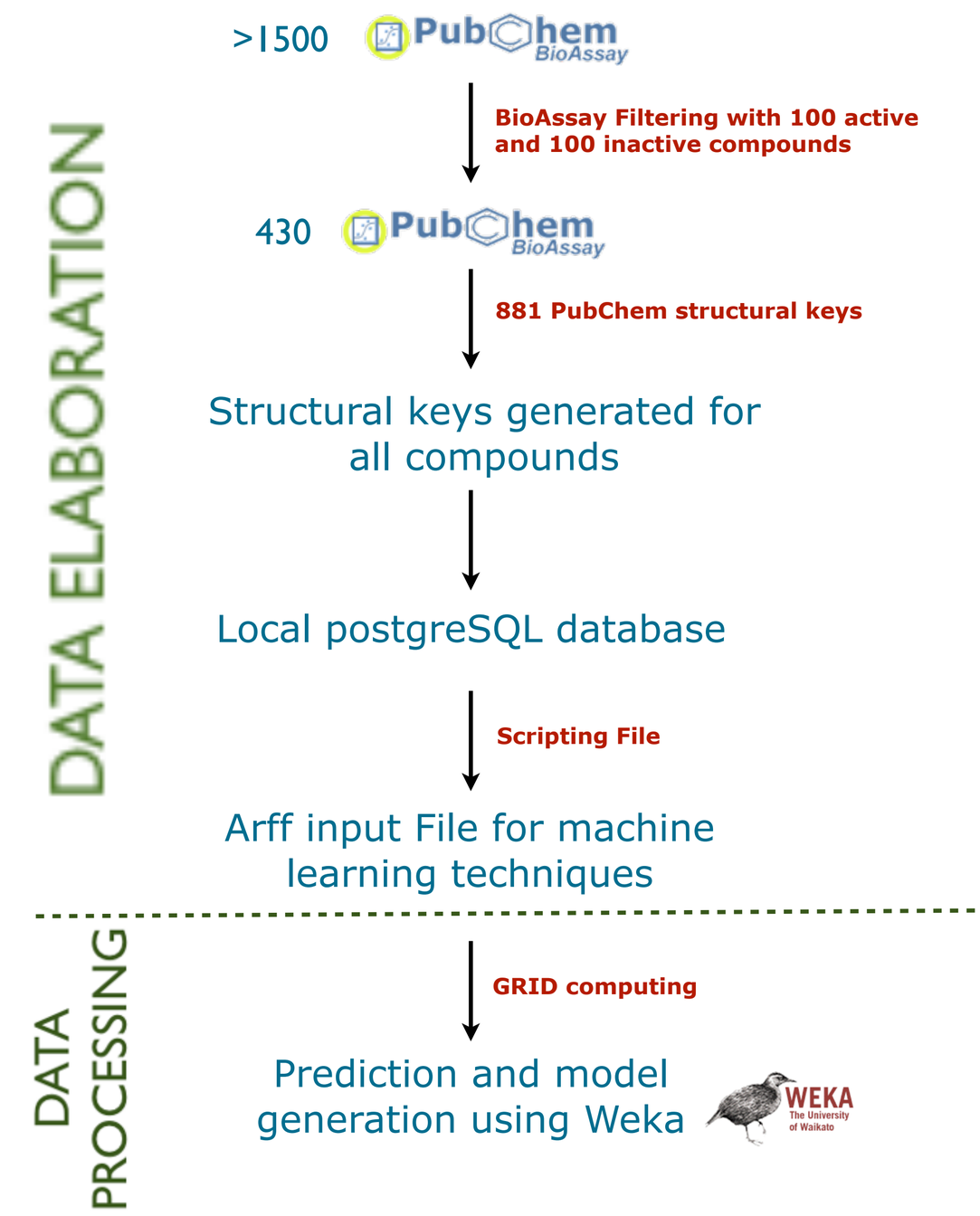
J. Masciocchi[1], L. Pireddu[1], R. Medda[1], P. Palla [1], M. Floris[1], L. Pani[2], P. Rodriguez-Tomé[1]

[1] CRS4, Parco Tecnologico della Sardegna, Ed1, 09010 Pula (CA) , Italy
[2] CNR, Istituto di Tecnologie Biomediche, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy

In the last decade, the development and use of new methods in combinatorial chemistry and high-throughput screening has dramatically increased the number of known biologically active compounds. Paradoxically, the number of drugs reaching the market has not followed the same trend, often because many of the candidate drugs present poor qualities in absorption, distribution, metabolism, excretion, and toxicological properties (ADME-Tox). The ability to recognize and discard bad candidates early in the drug discovery steps would save lost investments in time and money. Machine learning techniques could provide solutions to this problem.
The goal of my research is to develop classifiers that accurately discriminate between active and inactive molecules for a specific target. To this end, I am comparing the effectiveness of the application of different machine learning techniques to this problem. As a source of data we have selected a set of PubChem's public BioAssays[1]. In addition, with the objective of realizing a real-time query service with our predictors, we aim to keep the features describing the chemical compounds relatively simple.
At the end of this process, we should better understand how to build statistical models that are able to recognize molecules active in a specific bioassay, including how to select the most appropriate classification technique, and how to describe compounds in such a way that is not excessively resource-consuming to generate, yet contains sufficient information for the classification. We see immediate applications of such technology to recognize compounds with high-risk of toxicity, and also to suggest likely metabolic pathways that would process it.

Approach that we have used to develop the software

This software must be able to predict biological or toxicological activity of molecules for any kind of BioAssay. In order to respect this concept, we have developed a tool which use structural keys to do prediction and not molecular descriptors which depend more on the BioAssay.



Results

For confidentiality reasons, the poster displays only the results fort just one PubChem BioAssay, the AID:852 or factor XIIa 1536 HTS Dose Response Confirmation.

PubChem BioAssay AID: 852
Compounds All: 641
Active: 146 Inactive: 495

	Accuracy	Precision	Recall
SVM	92,4 %	0.92	0,81
Neural Networks	92,2 %	0.92	0,78
Random Forest	92,2 %	0.91	0,74
Naive Bayes	84 %	0,65	0,75

Software Improvements

In order to have an approach more efficient in term of classification, our software should be able to identify specific patterns responsible for the activity.
We have developed a tool[2] that crop the compound into fragments which consist in single rings, Murko fragments, conjugated structures and linkers.
This approach will be used for features selection and to determine if the presence of specific patterns or combination of them can modify the activity of the compound since each fragment is related to the activity score of the compound. This second concept is similar to the structure-activity relationship process (SAR).

Fragmentation of all compounds of PubChem BioAssay 852

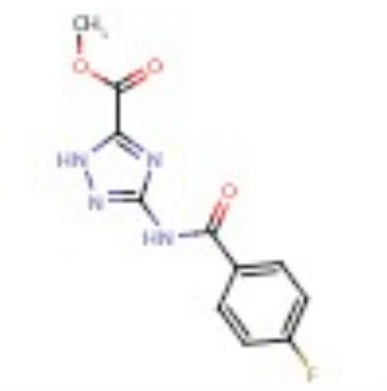
We get 5315 unique fragments on which we apply a filter, We removed all the fragments which are present in:

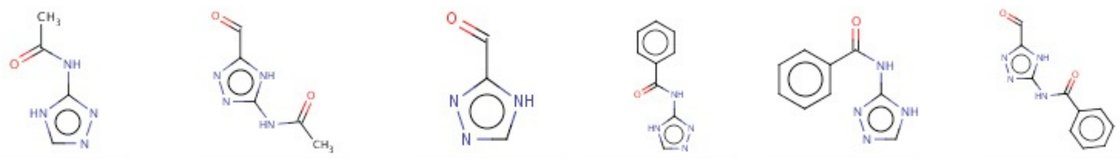
- **1 active and 0 inactive**
- **0 active and 1 inactive**
- **x active and x inactive (where x is any number)**

Finally we retrieve 694 fragments:

- **397 fragments present only in inactive molecules**
- **173 fragment present only in active molecules**

for example the compound with the accession number CID:2317506 who has a high activity score for this BioAssay will be cropped like this.





The fragmentation tool can also crop the compound in single rings and linkers but there are not present in the example above.

All the fragment are then used to generate a network in order to display if the combination of them have an impact on biological activity.

Future

Test the software with experimental data synthesized by the CNR laboratory.

Try to improve the prediction on BioAssays which present an unbalanced data distribution.

Add a function of the fragmentation tool to keep the side chain of the compound which contains specific atoms like halogens or sulfur in order to generate a better compound network

References

1. Eric W. Sayers*, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko and Jian Ye. **Database resources of the National Center for Biotechnology Information.** Nucleic Acids Research, 2009, Vol. 37, Database issue D5-D15
2. Joel Masciocchi, Gianfranco Frau, Marco Fanton, Mattia Sturlese, Matteo Floris, Luca Pireddu, Piergiorgio Palla, Fabian Cedrati, Patricia Rodriguez-Tomé, and Stefano Moro. **MMsINC: a large-scale chemoinformatics database**. Nucleic Acids Research Advance Access published on October 17, 2008.

Nature Precedings : doi:10.1038/npre.2010.4188.1 : Posted 28 Jan 2010