Novel computational tools based on bioinformatic and chemoinformatic data to complement zebrafish embryo teratogenicity test

Xabier Cendoya[1], Celia Quevedo[2] , Ángel Rubio[1], Maitane Ipiñazar[2], Arantza Muriana[2] and Francisco J. Planes[1]

[1] CEIT and Tecnun, University of Navarra, San Sebastian, Spain. fplanes@tecnun.es

[2] BIOBIDE; San Sebastian, Spain. quevedo@biobide.es

INTRODUCTION

Approximately 3% of newborns present congenital anomalies and around 5-10% of those are caused by exposure to teratogenic agents. For this reason, regulatory organisms and the industry demand for effective methods to test the developmental toxicity of drugs, industry chemicals or waste products. The use of the zebrafish embryotoxicity test is an attractive strategy to minimize *in vivo* assays and animal models. Overall, this assay has a good predictability; however, the outcome is based on morphologic evaluation, which is subjective and subtle effects might be neglected. With the increasing amount of molecular databases, the development of *in silico* tools that complement experimental assays is promising. In this work, we present an *in silico* platform that makes use of bioinformatics and chemoinformatics data, as well as machine learning methods, in order to predict the teratogenic potential of a particular compound. First, we show a combined systems biology and metabolomics study in order to identify metabolic biomarkers that improve the sensitivity of the zebrafish embryotoxicity test. Second, a learning algorithm using structural information is evaluated and compared using publicly available data, analyzing their complementarity with the zebrafish embryotoxicity test and metabolic biomarkers with newly generated proprietary data.

METHODS

This work consists of an *in silico* platform (TERATOOL) that integrates diverse sources of data to allow bioinformatics and chemoinformatics analysis (Figure 1). The information that was integrated included:

- ✓ A database of approximately 400 compounds with labels for their risk of teratogenicity. 290 of them were obtained from Enoch et al [1].
- ✓ Chemical structures and properties, together with biological target data, were obtained from ChEMBL [2], DrugBank [3] and HMDB [4].
- ✓ A number of publicly available transcriptomics experiments of the zebrafish embryotoxicity test (40 compounds) [5,6,7,8,9,10].
- ✓ A metabolic network reconstruction of the zebrafish was obtained from Bekaert et al [11].

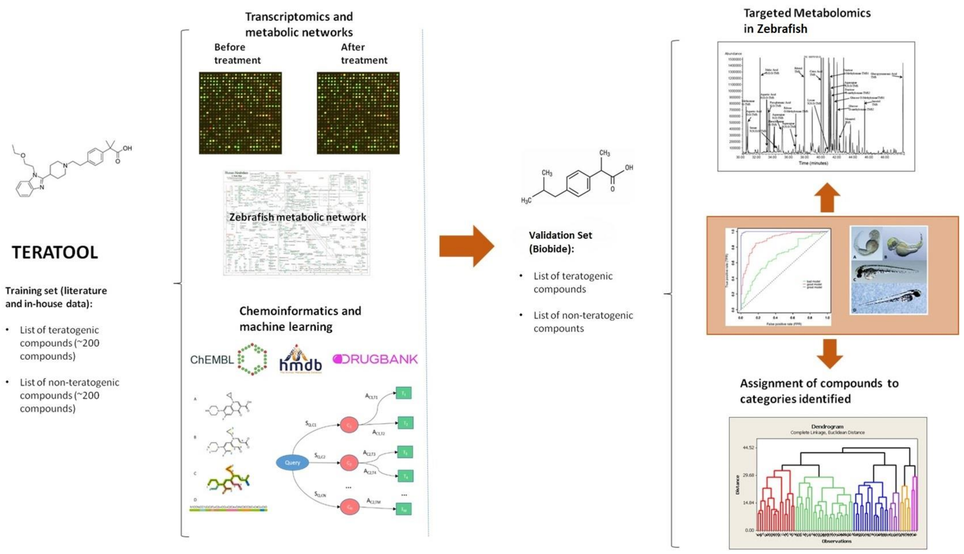

Figure 1. The TERATOOL platform integrates a list of approximately 400 molecules an includes chemoinformatics, transcriptomics and metabolic network data. With this, we can test several algorithms and identify potential biomarkers that can later be experimentally validated.

RESULTS

Identifying metabolic biomarkers

The reporter metabolites algorithm [12] was used with the transcriptomics data and the zebrafish metabolic network to search for metabolites that indicated highly altered regions of the metabolic network, by establishing an integrative score based on differential expression analysis of neighbor genes (Figure 2). We found several metabolites potentially reporting teratogenic action with a substantially higher redundancy than gene biomarkers, and these are currently being analyzed experimentally.

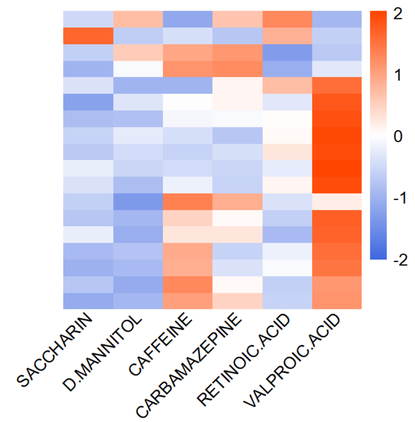

Figure 2. Gene expression heatmap of genes associated with a biomarker metabolite for a number of tested molecules.

Prediction of teratogenicity with machine learning

Molecular fingerprints were obtained and compared for similarity using the Tanimoto metric. Assuming that similar molecules tend to have similar properties, new targets were annotated for each molecule when they appeared recurrently as targets of similar molecules. The new relationships were evaluated computationally using Autodock [13] whenever protein structures where available, obtaining good affinities (< -6kcal/mol).

Using the molecular fingerprints and annotated genes, a machine learning algorithm was developed. A variable selection was carried out, leaving out properties that had low variance or were not good predictors of teratogenicity. We used a support vector machine with a polynomial kernel of degree 1, obtained the cost with tuning function, and obtained good predictive power (85% in training, 81% in validation) for the molecular characteristics of the molecules in the database (Figure 3).

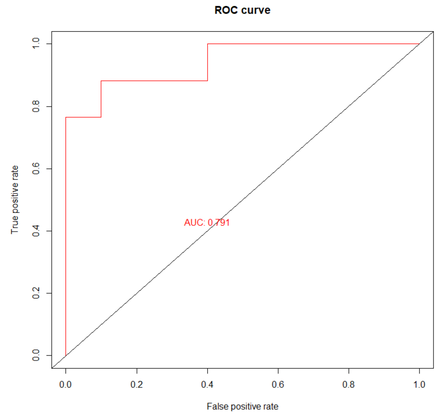

Figure 3. ROC curve of prediction of teratogenicity with the computational algorithm.

CONCLUSIONS

Bioinformatic and chemoinformatic methods seem to be capable of complementing current experimental methods in the testing of teratogenicity by proposing biomarkers and consistent, data-driven approach to the prediction of the teratogenic potential of a certain molecule.

REFERENCES

[1] Enoch SJ. QSAR & Combinatorial Science 2009; 28:696–768.
[2] Bento AP. Nucleic Acids Research 2014; 42 1083-1090.
[3] Law V. Nucleic Acids Research 2014; 42(1):D1091-7.
[4] Wishart DS. Nucleic Acids Research 2013; 41(D1):D801-7.
[5] Schiller V. Reproductive Toxicology 2013, 42, 210-223.
[6] Hermsen SA. Toxicology and applied pharmacology 2013, 272(1), 161-171.
[7] Tzima E. International Journal of Molecular Sciences 2017, 18(2), 365.
[8] Choi JS. PloS one 2016, 11(8), e0160763.
[9] Haggard DE. Toxicology and applied pharmacology 2016, 308, 32-45.
[10] GSE89780: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89780
[11] Bekaert M. PloS one 2012, 7(11), e49903.
[12] Patil KR. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(8), 2685-2689.
[13] Trott O. Journal of Computational Chemistry 2010, 455-461.

This work was supported by the Basque Government.