*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Application Note

|  |
| --- |
| Data and text miningDARTpaths, an *in silico* platform to investigate molecular mechanisms of compoundsDiksha Bhalla1\*, Marvin N. Steijaert2\*, Eefje S. Poppelaars3\*, Marc Teunis4\*, Monique van der Voet3, Marie Corradi4, Elisabeth Dévière2, Luke Noothout5, Wilco Tomassen5, Martijn Rooseboom6, Richard A. Currie7, Cyrille Krul4, Raymond Pieters4,8, Vera van Noort1,9^, and Marjolein Wildwater3^1 KU Leuven, Centre of Microbial and Plant Genetics, Faculty of Bioscience Engineering, 3001, Leuven, Belgium2 Open Analytics, 2600, Antwerp, Belgium3 Vivaltes, 1704 NA, Heerhugowaard, the Netherlands4 Utrecht University of Applied Sciences, Innovative testing in Life Sciences & Chemistry, 3584 CH, Utrecht, the Netherlands5 CLEVER°FRANKE, 3512 CC, Utrecht, the Netherlands6 Shell Global Solutions International B.V., 2596 HR, The Hague, the Netherlands7 Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, UK 8 Utrecht University, Institute for Risk Assessment Sciences, 3584 CM, Utrecht, the Netherlands9 Leiden University, Institute of Biology Leiden, 2333 BE, Leiden, the Netherlands\*Equal contribution ^Corresponding authorsAssociate Editor: XXXXXXXReceived on XXXXX; revised on XXXXX; accepted on XXXXX Abstract**Summary:** Xpaths is a collection of algorithms that allow for the prediction of compound-induced molecular mechanisms of action by integrating phenotypic endpoints of different species; and proposes follow-up tests for model organisms to validate these pathway predictions. The Xpaths algorithms are applied to predict developmental and reproductive toxicity (DART) and implemented into an *in silico* platform, called DARTpaths.**Availability and implementation:** for reviewers an installed version of the application is available: dartpaths.openanalytics.eu with reviewer login. All code is available on GitHub <https://github.com/Xpaths/dartpaths-app> under Apache license 2.0, detailed overview with demo is available at <https://www.vivaltes.com/dartpaths/> **Contact:** vera.vannoort@kuleuven.be**,**  m.wildwater@vivaltes.com **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

Regulatory compliance to address the human health risks of chemicals commonly still requires both rodents and non-rodents as test species. In addition to these species not always being reliable due to species-specific differences, they are not ethically sustainable in the current century. For this reason, a broad portfolio of New Approach Methods (NAMs) have been established over the last years such as a wide diversity of cell- and organoid-based protocols, assays that use small organisms like *Caenorhabditis elegans (C. elegans-nematodes), Danio rerio (zebrafish)* or *Dictyostelium discoideum (D. discoideum-slime mold)*, as well as several *in silico* platforms.

The largest challenge remains how to integrate and combine the study data of all these different and available screening platforms into a single framework. Such a framework connects and enhances the predictivity of the scientific and regulatory framework for hazard and risk assessment. With Xpaths we aimed to create a data model and algorithms for interspecies integration of test outcomes. This cross-species profiles of compound mechanistic activity, can support regulatory acceptance and ultimately, approval of NAMs. To predict effects in higher organisms like humans, compound-affected evolutionary conserved molecular pathways, are influenced by compounds *and* predicted in multiple test methodologies. Those pathways seem to form a key to identify the most relevant and conserved toxicity endpoint effects. The ontology-linked approach of Xpaths allows translation of toxicity (phenotypic) endpoints of different studies in different species, into a - mechanistical knowledge-based prediction output, using statistical pathway ranking. Data from a broad set of cell-based assays as well as six major model organisms: *Dictyostelium discoideum* (slime mold/social amoebe)*, Caenorhabditis elegans* (nematode)*, Drosophila melanogaster* (fruit fly)*,* *Danio rerio* (zebrafish), *Mus musculus* (mouse),and *Homo sapiens* (human) form the foundation of the Xpaths model. Moreover, the model can also be used to predict which assays in *C. elegans* or *D. rerio* could be used for further validation of predictions of possible (adverse outcome) mechanisms. With Xpaths, we aimed to create a single data model for interspecies connection of test outcomes, with the overall goal to improve the reliability of compound effect predictions.

# Methods

The Xpaths algorithms are written in R (R Core Team, 2021) and Python (3). Data are made interoperable using standard vocabularies and are integrated from databases that contain regulatory study information (QSAR toolbox), model organism-specific data sources (WormBase, DictyBase, FlyBase, Zebrafish Information Network, Mouse Genome Informatics), biology databases (Reactome, ENSEMBL). Xpaths’ pathways are based on Reactome (Jassal *et al.*, 2019) and orthology information is retrieved from ENSEMBL compara (Herrero *et al.*, 2016). The sources (Chisholm *et al.*, 2006; Harris *et al.*, 2019; Larkin *et al.*, 2020; Ruzicka *et al.*, 2018; Bult *et al.*, 2019; Köhler *et al.*, 2020) and the process of implementing the backend database for each of the eight organisms are described in the supplementary materials.

1. **Phenotype pathway enrichment analysis**

The phenotype pathway enrichment analysis (PPEA) module takes as input a list of phenotype ontology identifiers observed in model organisms after exposure to a chemical. The algorithm first retrieves genes associated to the phenotypes and uses an area under the curve algorithm to find the best matching pathways, similar to a gene set enrichment analysis (Subramanian et al., 2005). For ease of interpretation of the AUC score, *p*-values are estimated by Monte Carlo random sampling. Cross-species *p*-values are combined using the harmonic mean. Details of the PPEA are discussed in the supplementary materials.

1. **Phenotype prediction from evolutionary conserved pathway data**

Phenotype enrichment analysis identifies phenotypes that are enriched for gene sets comprising of genes in a human pathway and their orthologs in model species. Expected phenotypes in model organism can be predicted. Ortholog data were retrieved from ENSEMBL Compara (Herrero *et al.*, 2016). The *p*-values in the phenotype enrichment module are computed from hypergeometric distribution. *P*-values in enrichment score are corrected by FDR-correction (Benjamini & Hochberg, 1995). Details of the phenotype prediction algorithms are discussed in the supplementary materials.

# Results

****Xpaths is designed for cross-species connection of test outcomes to increased reliability in compound effect prediction. Compared to currently-available phenotype enrichment tools, Xpaths is designed to be easy to use and is based on FAIR principles (i.e., findable, accessible, interoperable, and reusable).

 **Fig. 1. Semantic data model of Xpaths.** Connection of data types in de DARTpaths application. 1. Gene-organism from ENSEMBL 2. Orthology from ENSEMBL Compara 3. Gene - Protein – pathway from Reactome 4. Genetic variants to phenotypes from organism specific databases 5. Compound target *in vitro* data from EPA 6. In vitro target names to genes from ENSEMBL 7. Computed fingerprint similarities 8. Regulatory reports from QSAR toolbox 9. Toxicity conclusion extracted with xpaths scripts 10. Text-mining of regulatory reports and literature (mammalian), compound-phenotype links from model organism databases. The paths between phenotypes/endpoint effects in mammals, pathways/mechanisms, phenotypes/endpoint effects in New Alternative Methods, and phenotypes/endpoint effects in humans are calculated by Xpaths as an alternative to the traditional paths between phenotypes/endpoint effects in mammals and those in humans.

The data model can be reused for other comparative pathways projects by providing all Extraction, Transformation and Loading scripts and enrichment algorithms via a GitHub repository. At the same time the user-friendly interface makes the tool widely applicable.Future work might be able to make use of the Xpaths algorithms as a general data model. Currently, the Xpaths algorithms have been implemented in the DARTpaths platform.

 **DARTpaths**

DARTpaths is designed as an *in silico* platform to predict developmental and reproductive toxicity (DART). The DARTpaths platform is created as a R Shiny app. All code is available on GitHub.

* + 1. **Substance exploration**

Users can start by searching for a single chemical or a complex substance (i.e. UVCB), either by name or by identifier (CAS, EC, SMILES). Endpoint data from different test methodologies (cell-based assay, rat, rabbit, *Dictyostelium discoideum, C. elegans, Drosophila melanogaster,* *Danio rerio*, *Mus musculus*)is collected and shows the predicted impacted pathways in humans that are potentially affected by the chemical. A summary score indicates the strength of evidence for that pathway. Results are ranked based on this score and can be filtered by type of evidence from five categories: based on non-mammalian phenotype evidence, based on mammalian phenotype evidence, based on *in-vitro* evidence, based on all evidence (all of the first three), pathways with any type of evidence (any of the first three). Furthermore, the user can select a pathway from this list, which will then show information on this selected pathway. This includes: summary score, number of genes in this pathway in humans, cross-species conservation of genes between humans and each model species, and phenotypes linked to this pathway categorized by evidence type (mammalian, non-mammalian, *in-vitro*). Detailed information is available for similar substances of this chemical, including additional information to assist in selecting related chemicals, such as: *in vitro* test data of this chemical (with assay type and link to the source database), phenotypes, and regulatory information of the chemical.

* + 1. **Pathway exploration**

Users can also choose to start by selecting a pathway. Pathway exploration shows type of orthology relations (one-to-one, one-to-many, many-to-many) for each organism for all genes in this human pathway. This can inform which species may have a diversified response to a chemical, since chemical effects can largely differ between orthologs. Furthermore, it can be used to target a specific pathway in an experiment by knowing which species has better conservation for this pathway in advance. This is particularly useful for testing mixtures as reported by EPA (Krewski *et al.*). Finally, it can be used to predict phenotypic endpoints in model species *C*. *elegans, Drosophila melanogaster* (fruit fly)*,* and *Danio rerio* (zebrafish) to further validate pathway predictions.

* + 1. **Database**

# Currently, the DARTpaths database contains information on DART-related phenotypic endpoints for model organisms (Table 1) and *in vitro* assays. Compound-phenotype connections are obtained from model organism databases where they are described using structured vocabularies. For mammalian data, compound-phenotype links are only available in text format. To add data on phenotypic endpoints of mammals from scientific articles and study reports, we have developed text mining algorithms (to extract phenotypes, compounds, organisms, doses, exposure routes, parent vs. offspring, and in vitro vs. in vivo) using natural language processing with SpaCy (spacy.io; Neumann et al., 2019) and mammalian phenotype ontology (Smith & Eppig, 2009) linking with PhenoTagger (Luo et al., 2021). This approach was applied for one compound (Table 1). This text mining can be used on demand to further populate the database and thereby improve the pathway predictions.

|  |  |  |
| --- | --- | --- |
|  | Compound-phenotype connections | Unique compounds with phenotypes |
| Zebrafish(Zebrafish Phenotype Ontology) | 10,893 | 990 |
| Nematode (Wormbase Phenotype) | 3,359 | 281 |
| Mammals(Mammalian Phenotype) | 27 | 1 |

Table 1 Compound-phenotype connections in the DARTpaths application

Funding

This work was supported by The National Centre for the Replacement Refinement & Reduction of Animals in Research (NC3Rs) for the CRACK IT project DARTpaths (project number CRACKITDP-P1-2) and by EFRO 3R TOXFLOW (project number KVW-00181).

*Conflict of Interest:* none declared.

Author contributions

The study was designed by M.N.S., V.v.N., M.T., R.P., M.R., R.C., and M.W. The manuscript draft was written by D.B., V.v.N, E.S.P., M.N.S., and M.W. All authors corrected, amended, and complemented the manuscript.

**Acknowledgments**

We thank Nadha Sholiha, Miguel Cisneiros, Jan-Willem Lankhaar, Jens van Erp Machteld Varewyck, Marijke van Moerbeke for their contributions during development of the application and Catherine Vickers (NC3Rs) and Tobias Verbeke (Open Analytics) for discussions during the project.

References

Benjamini, Y., Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. Doi: 10.1111/j.2517-6161.1995.tb02031.x.

Bolker, J.A. (2019) Selection of Models: Evolution and the Choice of Species for Translational Research. *Brain Behav Evol*, 93, 82–91.

Bult, C.J. *et al.* (2019) Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*, 47, D801–D806.

Chisholm, R.L. *et al.* (2006) dictyBase, the model organism database for Dictyostelium discoideum. *Nucleic Acids Res*, 34, D423–D427.

Harris, T.W. *et al.* (2019) WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res*, 48, D762–D767.

Herrero, J. *et al.* (2016) Ensembl comparative genomics resources. *Database J Biological Databases Curation*, 2016, bav096.

Jassal, B. *et al.* (2019) The reactome pathway knowledgebase. *Nucleic Acids Res*, 48, D498–D503.

Krewski, D. *et al.* Toxicity Testing in the 21st Century: A Vision and a Strategy: Journal of Toxicology and Environmental Health, Part B: Vol 13, No 2-4. *Journal of Toxicology and Environmental Health, Part B*.

Larkin, A. *et al.* (2020) FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res*, 49, D899–D907.

Luo, L., Yan, S., Lai, P. T., Veltri, D., Oler, A., Xirasagar, S., ... & Lu, Z. (2021). PhenoTagger: a hybrid method for pheno-type concept recognition using human phenotype ontology. *Bioinformatics, 37*(13), 1884-1890.

Neumann, M., King, D., Beltagy, I., Ammar, W. (2019) ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proc. 18th BioNLP Workshop Shar. Task*, pp. 319–327. Doi: 10.18653/v1/W19-5034.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www>.R-project.org/.

Ruzicka, L. *et al.* (2018) The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res*, 47, gky1090-.

Smith, C.L., Eppig, J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med, 1*(3):390-399. Doi: 10.1002/wsbm.44.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545–15550. Doi: 10.1073/pnas.0506580102

Sullivan, J. *et al.* (2013) InterMOD: integrated data and tools for the unification of model organism research. *Sci Rep-uk*, 3, 1802.