CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity

Kamel Mansouri[1], Nicole Kleinstreuer[2], Eric Watt[1], Jason Harris[1] and Richard Judson[3]
[1]ORISE Fellow, NCCT, ORD/U.S. EPA, RTP, NC, USA
[2]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA
[3]NCCT, ORD/U.S. EPA, RTP, NC, USA

Kamel Mansouri | mansouri.kamel@epa.gov | ORCID: 0000-0002-6426-8036

www.epa.gov

Abstract

Background: In order to protect human health from chemicals that can mimic natural hormones, the U. S. Congress mandated the U.S. EPA to screen chemicals for their potential to be endocrine disruptors through the Endocrine Disruptor Screening Program (EDSP). However, the number of chemicals to which humans are exposed is too large (tens of thousands) to be accommodated by the EDSP Tier 1 battery

Objectives: To solve this, combinations of in vitro high-throughput screening (HTS) assays and computational models are being developed to help prioritize chemicals for more detailed testing.

Methods: Previously, CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) demonstrated the effectiveness of combining many QSAR models trained on HTS data to prioritize a large chemical list for estrogen receptor activity. The limitations of single models were overcome by combining all models built by the consortium into consensus predictions. CoMPARA is a larger scale collaboration between 35 international groups, following the steps of CERAPP to model androgen receptor activity using a common training set of 1746 compounds provided by U.S. EPA. Eleven HTS ToxCast/Tox21 in vitro assays were integrated into a computational network model to detect true AR activity. Bootstrap uncertainty quantification was used to remove potential false positives/negatives. Reference chemicals (158) from the literature were used to validate the model.

Results: The model combining ToxCast/Tox21 assays showed 95.2% and 97.5% balanced accuracies for AR agonists and antagonists respectively. The resulting data was used to build qualitative and quantitative models and a consensus combining the different structure-based and QSAR modeling approaches. Then, a library of ~80k chemical structures, including ~11k chemicals curated from PubChem literature data using ScrubChem tools are being integrated with CoMPARA's consensus predictions.

Conclusion: The results of this project will be used to prioritize a large set of more than 50k chemicals for further testing over the next phases of ToxCast/Tox21, among other projects.

Project planning

Steps	Tasks
1: Training and prioritization sets NCCT/ EPA	- ToxCast assays for training set data - AUC values and discrete classes for continuous/classification modeling - QSAR-ready training set and prioritization set
2: Experimental validation set NCCT/ EPA	- Collect and clean experimental data from the literature - Prepare validation sets for qualitative and quantitative models
3: Modeling & predictions All participants	- Train/refine the models based on the training set - Deliver predictions and applicability domains for evaluation
4: Model evaluation NCCT/ EPA	- Evaluate the predictions of each model separately - Assign a score for each model based on the evaluation step
5: Consensus modeling NCCT/ EPA	- Use the weighting scheme based on the scores to generate the consensus - Use the same validation set to evaluate consensus predictions
6: Manuscript writing All participants	- Descriptions of modeling approaches for each individual model - Input of the participants on the draft of the manuscript

U.S. Environmental Protection Agency
Office of Research and Development

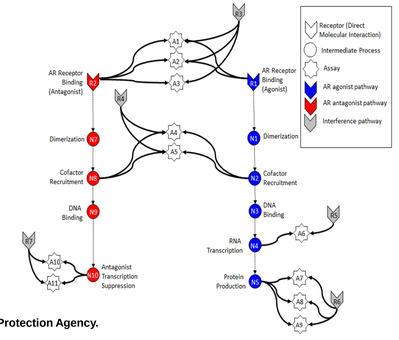
Disclaimer: The views expressed in this poster are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Participants

- **DTU/food:** Technical University of Denmark/ National Food Institute. **USA**
- **EPA/NCCT:** U.S. EPA/ National Center for Computational Toxicology. **USA**
- **TUM:** Technical University Munich. **Germany**
- **ILS&NIH/NIEHS:** ILS Inc & NIH/NIEHS. **USA**
- **IRCSS:** Istituto di Ricerche Farmacologiche "Mario Negri". **Italy**
- **MTI.** Molecules Theurapetiques in silico. **France**
- **LockheedMartin&EPA:** Lockheed Martin IS&GS/ High Performance Computing. **USA**
- **NIH/NCATS:** National Institutes of Health/ National Center for Advancing Translational Sciences. **USA**
- **IdeaConsult. Bulgaria**
- **NIH/NCI:** National Institutes of Health/ National Cancer Institute. **USA**
- **UFG.** Federal University of Golas. **Brazil**
- **UMEA/Chemistry:** University of UMEA/ Chemistry department. **Sweden**
- **UNC/MML:** University of North Carolina/ Laboratory for Molecular Modeling. **USA**
- **UniBA/Pharma:** University of Bari/ Department of Pharmacy. **Italy**
- **UNIMIB/Michem:** University of Milano-Bicocca/ Milano Chemometrics and QSAR Research Group. **Italy**
- **VCCLab.** Virtual Computational Chemistry Laboratory. **Germany**
- **NCSU.** NC State University, Bioinformatics Research Center. **USA**
- **EPA/NRMRL.** National Risk Management Research Laboratory. **USA**
- **FDA/NCTR/DBB:** U.S. FDA/ National Center for Toxicological Research/Division of Bioinformatics and Biostatistics. **USA**
- **INSUBRIA.** University of Insubria. Environmental Chemistry. **Italy**
- **Tartu. University of Tartu.** Institute of Chemistry. **Estonia**
- **NIH/NTP/NICEATM. USA**
- **Chemistry Institute.** Lab of Chemometrics. **Slovenia**
- **SWETOX.** Swedish toxicology research center. **Sweden**
- **LZU:** Lanzhou University. **China**
- **BDS.** Biodetection Systems. **Netherlands**
- **IBMC.** Institute of Biomedical Chemistry. **Russia**
- **UNIMORE.** University of Modena Reggio-Emilia. **Italy**
- **MSU.** Moscow State University. **Russia**
- **ZJU.** Zhejiang University. **China**
- **JKU.** Johannes Kepler University. **Austria**
- **CTIS.** Centre de Traitement de l'Information Scientifique. **France**
- **ECUST**. East China University of Science and Technology. **China**
- **UNISTRA/Infochim:** University of Strasbourg/ ChemoInformatique. **France**

Training data

Tox21/ToxCast AR Pathway Model



This model was used to combine the following ToxCast assays and generate AUC scores that were used to train CoMPARA models after removing potential false positives/negatives.

Assay Name	Biological Process
NVS_NR_hAR	receptor binding
NVS_NR_cAR	receptor binding
NVS_NR_rAR	receptor binding
OT_AR_ARSRC1_0480	cofactor recruitment
OT_AR_ARSRC1_0960	cofactor recruitment
ATG_AR_TRANS	mRNA induction
OT_AR_ARELUC_AG_1440	gene expression
Tox21_AR_BLA_Agonist_ratio	gene expression
Tox21_AR_LUC_MDAKB2_Agonist	gene expression
Tox21_AR_BLA_Antagonist_ratio	gene expression
Tox21_AR_LUC_MDAKB2_Antagonist	gene expression
Tox21_AR_LUC_MDAKB2_Antagonist*	gene expression

1720 unique structures
- Agonist: ~50 actives
- Antagonist: ~160 actives
- Binding: ~170 actives

2 types of data:
- Continuous AUC scores
- Discrete hit calls

Prediction set

The list of chemicals to be predicted and prioritized for AR activity:

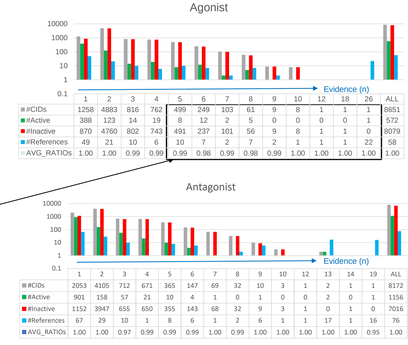
- **CERAPP list**: 32,464 unique QSAR-ready structures (standardized, organic, no mixtures…)
  - EDSP Universe (10K)
  - Chemicals with known use (40K) (CPCat & ACToR)
  - Canadian Domestic Substances List (DSL) (23K)
  - EPA DSSTox (version 1)– structures of EPA/FDA interest (15K)
  - ToxCast and Tox21 (In vitro ER data) (8K)
- **EINECS**: European INventory of Existing Commercial chemical Substances
  - ~60k structures
  - ~55k QSAR-ready structures
  - ~38k non overlapping with the CERAPP list
  - ~18k overlap with DSSTox

⇨ **29,904 + 17,984 = 47,888 unique standardized QSAR-ready structures**

Evaluation set

The tool **ScrubChem**, a curated version of PubChem Bioassay, was used to build datasets from public data. Starting from **~80K bioactivities & 11K chemicals**, results were grouped by target, chemical, modality, and outcome in order to derive hit calls summarizing results from different experiments. Hit calls were further filtered for quality by the number of pieces of evidence (n) and ratio of agreement between those values. The effect of a quality filter on the size of a dataset is shown in the example to the right.
This list will then be filtered and used as an evaluation set for the built models.

e.g., There are 932 chemicals in the selection grid (28 active and 924 inactive) with a hit call derived from 5 or more pieces of evidence (n). Each bin contains its own average for the ratio of agreement on hit calls in that bin. For example, the first bin (n=5) has 499 chemicals which on average have 98.8% agreement in the 5 data points used for their hit calls.



	1	2	3	4	5	6	7	8	9	10	12	18	26	ALL
#CIDs	1258	4883	816	762	499	249	103	61	9	8	1	1	1	8651
#Active	388	123	14	19	8	12	2	5	0	0	0	0	1	572
#Inactive	870	4760	802	743	491	237	101	56	9	8	1	1	0	8079
#References	49	21	10	6	10	7	2	7	2	1	1	1	22	58
AVG_RATIOs	1.00	1.00	0.99	0.99	0.99	0.98	0.99	0.98	0.99	1.00	1.00	1.00	1.00	1.00

	1	2	3	4	5	6	7	8	9	10	12	13	14	19	ALL
#CIDs	2053	4105	712	671	365	147	69	32	10	3	1	2	1	1	8172
#Active	901	158	57	21	10	4	1	0	1	0	0	2	0	1	1156
#Inactive	1152	3947	655	650	355	143	68	32	9	3	1	0	1	0	7016
#References	67	29	10	1	8	6	1	2	6	1	1	17	1	16	76
AVG_RATIOs	1.00	1.00	0.97	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	0.95	1.00	

Conclusions

- This project is prioritizing ~50K chemicals in a fast, accurate, and economic way.
- Generated high quality data and models that can be reused.
- Free & open-source code and workflows shared with the community.
- Data and predictions will be available for visualization on the EDSP dashboard: http://actor.epa.gov/edsp21/ and on the CompTox dashboard: https://comptox.epa.gov/dashboard/
- The prioritized lists will help in the selection of chemicals that will be tested in the next phases of ToxCast/Tox21.
- A joint paper with all participants and multiple satellite papers will be published in peer reviewed journals.