OpenTox Tutorials

Drug Discovery Predictive Toxicology Application II: Building a Model to Predict Kinase Inhibitor Activity

Grant Agreement	Health-F5-2008-200787
Acronym	OpenTox
Name	An Open Source Predictive Toxicology Framework
Coordinator	Douglas Connect

Contract No.	Health-F5-2008-200787	
Document Type:	Tutorial	
Name	Drug Discovery Predictive Toxicology Application II Tutorial	
Document ID:	OpenTox DrugDiscoveryApplicationII Tutorial	
Date:	Mar 08, 2011	
Status:	Final Version	
Organisation:	Douglas Connect	
Contributors	Roman Affentranger (RA, Author)	Douglas Connect (DC)
	Nina Jeliazkova (NJ)	Ideaconsult Ltd. (IDEA)

Distribution:	Final version

Purpose of Document:	Downloadable PDF version of the OpenTox tutorial "Drug Discovery Predictive Toxicology Application II: Building a Model to Predict Kinase Inhibitor Activity".

Document History:	1 – RA (DC) and NJ (IDEA) authored

Table of Contents

Table of Figures

Acknowledgements

Research Funding

OpenTox - An Open Source Predictive Toxicology Framework, www.opentox.org, is funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).

Project Partners

Douglas Connect (DC), In Silico Toxicology (IST), Ideaconsult (IDEA), Istituto Superiore di Sanita' (ISS), Technical University of Munich (TUM), Albert Ludwigs University Freiburg (ALU), National Technical University of Athens (NTUA), David Gallagher (DG), Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences (IBMC), Seascape Learning (SL), Jawaharlal Nehru University (JNU), Fraunhofer Institute for Toxicology & Experimental Medicine (ITEM).

Advisory Board

European Centre for the Validation of Alternative Methods, European Joint Research Centre, U.S Environmental Protection Agency, U.S. Food & Drug Administration, Nestlé, Roche, AstraZeneca, Lhasa, Leadscope, University of North Carolina, Pharmatrope, Bioclipse, EC Environment Directorate General, Organisation for Economic Co-operation & Development, CADASTER, Bayer Healthcare.

Correspondence

Dr. Barry Hardy, OpenTox Project Coordinator and Director, Community of Practice & Research Activities, Douglas Connect, Baermeggenweg 14, 4314 Zeiningen, Switzerland

Email: barry.hardy -(at)- douglasconnect.com

Summary

This document represents the second part of a tutorial on the application of OpenTox facilities in a drug discovery workflow. The tutorial example of a predictive toxicology application in drug discovery is provided using the data on anti-malarial compounds made available at the ChEMBL Neglected Tropical Disease (NTD) archive (www.ebi.ac.uk/chemblntd/). Using this data, a model is built to predict protein kinase inhibitor activity of chemicals using the ToxCreate (www.toxcreate.org) application.

All tutorials and their updates are made available online under www.opentox.org/tutorials.

This tutorial is available online under http://opentox.org/tutorials/drug-discovery

Drug Discovery Predictive Toxicology Application II: Building a Model to Predict Kinase Inhibitor Activity

1 Introduction

Using the data on antimalarial compounds made available at the ChEMBL Neglected Tropical Disease (NTD) archive (http://www.ebi.ac.uk/chemblntd/), in this exercise subsets of the antimalarials are extracted to be used in a model building exercise via the OpenTox prototype application ToxCreate. 857 of the 13'519 compounds contained in the TCAMS dataset are annotated with a protein (class) target. Of these 857 compounds, 233 are annotated as Ser/Thr kinase inhibitors. In this exercise we'll use this information to create a dataset that can be used to build a model that predicts whether or not a given compound is likely to be a kinase inhibitor. The dataset for the model building therefore needs to consist of two columns: the SMILES string of the compound and its classification (Ser/Thr kinase inhibitor = 1, otherwise 0).

2 Selecting a subset to create a model with ToxCreate

To create the dataset required for model building go to http://pirin.uni-plovdiv.bg:8080/malaria/dataset

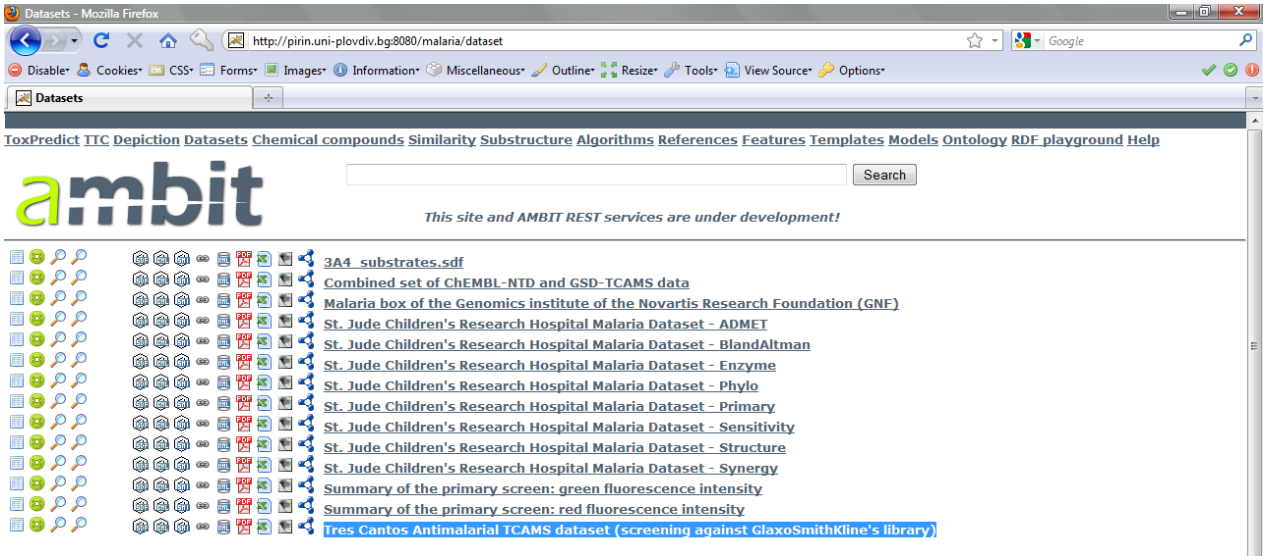

Figure 1 The list of antimalarial datasets on http://pirin.uni-plovdiv.bg:8080/malaria/dataset

Click on "Tres Cantos Antimalarial TCAMS dataset (screening against GlaxoSmithKline's library)"

Browse the dataset and find the column "Target hypothesis". You will note that most entries are empty (only ~6% of the compounds have a target hypothesis annotated). In the 100 compounds displayed by default when following the link to the TCAMS data, you will only find one entry with value "Adrenergic receptor antagonist". You could click on the link, which would filter out only compounds with this potential target.

For our purpose, we want the list of compounds annotated to be kinase inhibitors. You could try to increase the number of displayed compounds until you find one, or you could enter "Ser/Thr protein kinase" in the searching text box at the top of the page and click the "Search" button. The results will be displayed as below (see Figure 2).

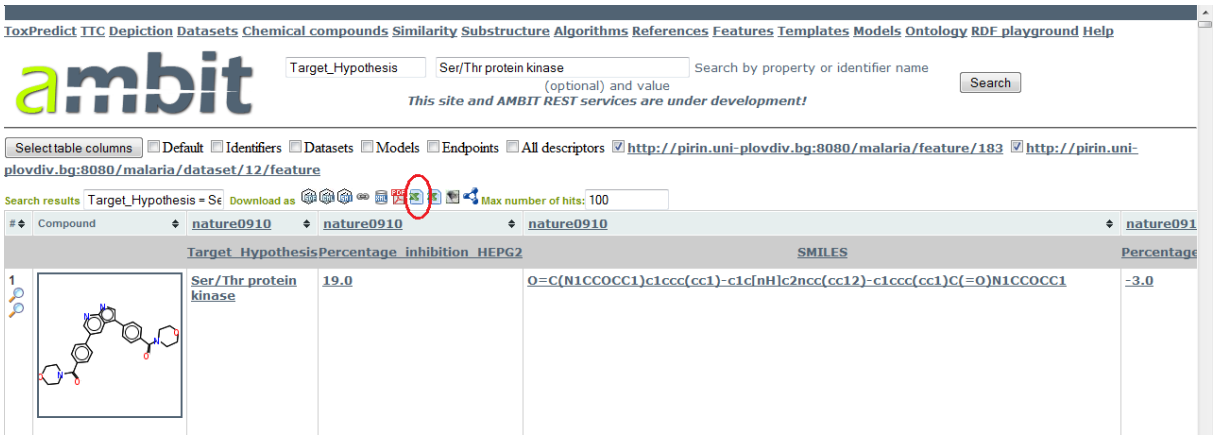

Figure 2 Search results for "Ser/Thr protein kinase" on the TCAMS antimalarial dataset

To build a model, it is not enough to have a list of Ser/Thr kinase inhibitors. We also need some "negatives". Although strictly speaking we don't have any true negatives, we will use the compounds that do have a target hypothesis annotation – but one that is not "Ser/Thr kinase" – as negatives. So, we extract the whole list of compounds with non-empty target hypothesis, and replace "Ser/Thr kinase" with a "1", and all the other target hypotheses with "0".

To extract the list of compounds with non-empty target hypotheses, use the following URL:

http://pirin.uni-plovdiv.bg:8080/malaria/compound?type=smiles&property=Target_Hypothesis**&search=+**&feature_uris[]=http://pirin.uni-plovdiv.bg:8080/malaria/feature/183&feature_uris[]=http://pirin.uni-plovdiv.bg:8080/malaria/dataset/12/feature&max=1000**&condition=!%3D**

This operation is not (yet) possible via the "Search" text field (it does not allow negation, e.g. something like Target_Hypothesis !=""), but only via the URL: briefly, the search for non-empty Target Hypothesis is done in the above URL, first with **&search=+** (the "+" stands for empty) – thus searching for all the empties – and then negating the search by **&condition=!%3D** (%3D stands for the "=" sign, thus !%3D stands for !=, or "not equal").

When following the above URL you'll get a table with compounds that have a non-empty Target_Hypothesis. The next step will be to export data. Click on the left one of the two little Excel icons (when moving the mouse pointer on top of it, a small text box "text/csv" should appear) to save the selected data as CSV.

For the model building, we will use the OpenTox application ToxCreate (www.toxcreate.org). Thus, first we need to format the data as explained at www.toxcreate.org/help. That is, we leave only the SMILES column and the Target_Hypothesis column. Now you should have the Target_Hypothesis in column 1 (or A), and the SMILES in column 2 (or B). If you are using Excel, go to the cell C2. Type

=IF(A2="Ser/Thr protein kinase"; 1; 0)

and hit "Enter". Again click on cell C2 to activate it. Now double-click on the little black square at the bottom-right corner of the cell's border to fill the column with this formula.

Now, copy the whole column C, and paste it (at the same place) using Excel's "Paste Special" function, pasting only the values. Once that's done, delete column A (holding the text entries for the Target_Hypothesis). Delete as well row 1 and save the resulting table as text CSV file to TCAMS-kinase_full.csv.

In your web browser, navigate to www.toxcreate.org. Read the instructions, and try to create a model using your dataset. As ToxCreate is currently a prototype, there are still some limitations. You might get an error in the model building, in which case you could try to reduce the number of compounds used to build the model to about 600. Just delete some rows until that table contains 600 rows or less. Save the resulting table to TCAMS-kinase-subset.csv.