

**RS-PREDICTOR — CREATION OF CYTOCHROME
P450 REGIOSELECTIVITY MODELS**

By

Jed Mikhail Zaretski

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: CHEMISTRY AND CHEMICAL BIOLOGY

Approved by the
Examining Committee:

Dr. Curt M. Breneman, Thesis Adviser

Dr. Kristin Bennett, Member

Dr. Steve Cramer, Member

Dr. Dominic Ryan, Member

Dr. Mark Wentland, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2011
(For Graduation May 2011)

ABSTRACT

This thesis describes RS-Predictor, a new *in silico* method for generating predictive models of P450-mediated metabolism for drug-like compounds. Within this method, potential sites of metabolism (SOMs) are represented as "metabolophores": A concept that describes the hierarchical combination of topological and quantum chemical descriptors needed to represent the reactivity of potential metabolic reaction sites. RS-Predictor modeling involves the use of metabolophore descriptors together with multiple-instance ranking (MIRank) to generate an optimized descriptor weight vector that encodes regioselectivity trends across all cases in a training set. The resulting pathway-independent (ex. O-dealkylation versus Csp³ Hydroxylation), isozyme-specific regioselectivity model may be used to predict potential metabolic liabilities. In one of the first applications of rank aggregation within the chemoinformatics community, independently-generated regioselectivity rankings for a given compound are merged into single optimized consensus predictions. A new Lift metric for assessing prediction quality is introduced, where each substrate is assigned a lift weight that expresses the statistical likelihood of randomly picking the CYP-oxidized SOM(s) out of all putative SOMs on the substrate. The prediction quality of each model is also assessed through the number of correct and incorrect predictions made on a pathway-by-pathway basis.

The broad applicability of RS-Predictor is demonstrated through the creation of regioselectivity models for substrate sets of the following CYPs: 1A2(271), 2A6(105), 2B6(151), 2C19(218), 2C8(142), 2C9(226), 2D6(270), 2E1(145) and 3A4(475), as well as a Merged set of all 680 curated substrates. A comprehensive investigation into the relative signal content of descriptors from different classes for each isozyme is made through the generation of seven separate RS-Predictor models for each substrate set. Two of these models involve the incorporation of high quality DFT derived reactivity information from SMARTCyp, a technology developed separately by another research group. Optimal combinations of RS-Predictor and SMARTCyp are shown to have stronger performances than either method alone, cor-

rectly identifying a large proportion of the metabolites of each dataset in the top two rank-positions: 1A2(83.0%), 2A6(85.7%), 2B6(82.1%), 2C19(86.2%), 2C8(83.8%), 2C9(84.5%), 2D6(85.9%), 2E1(82.8%), 3A4(82.3%), Merged(86.0%). A cross-isozyme (CI) study is made through the application of the regioselectivity QSAR of one isozyme to predict the CYP-mediated metabolism of the substrates of a different isozyme. Comparing CI model performances against the original results of cross-validated (CV) models for the same set of substrates lets future users of RSPredictor gauge the relative benefits of creating isozyme-specific models.