IQC Validator

NCATS Informatics

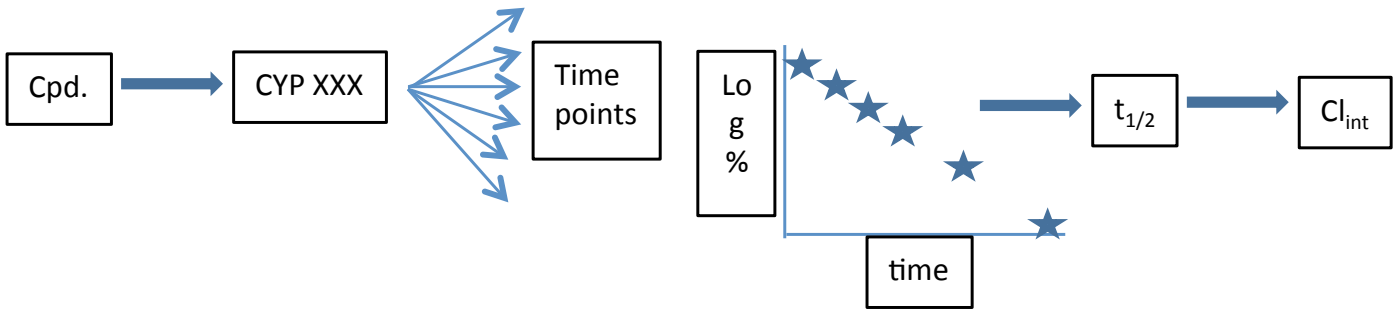
December 9, 2013

Overview

- Desktop client
- Proposed scoring function to evaluate data fitting

Desktop client
Features

- A simple desktop client to help validate the CYP isozyme data generation workflow

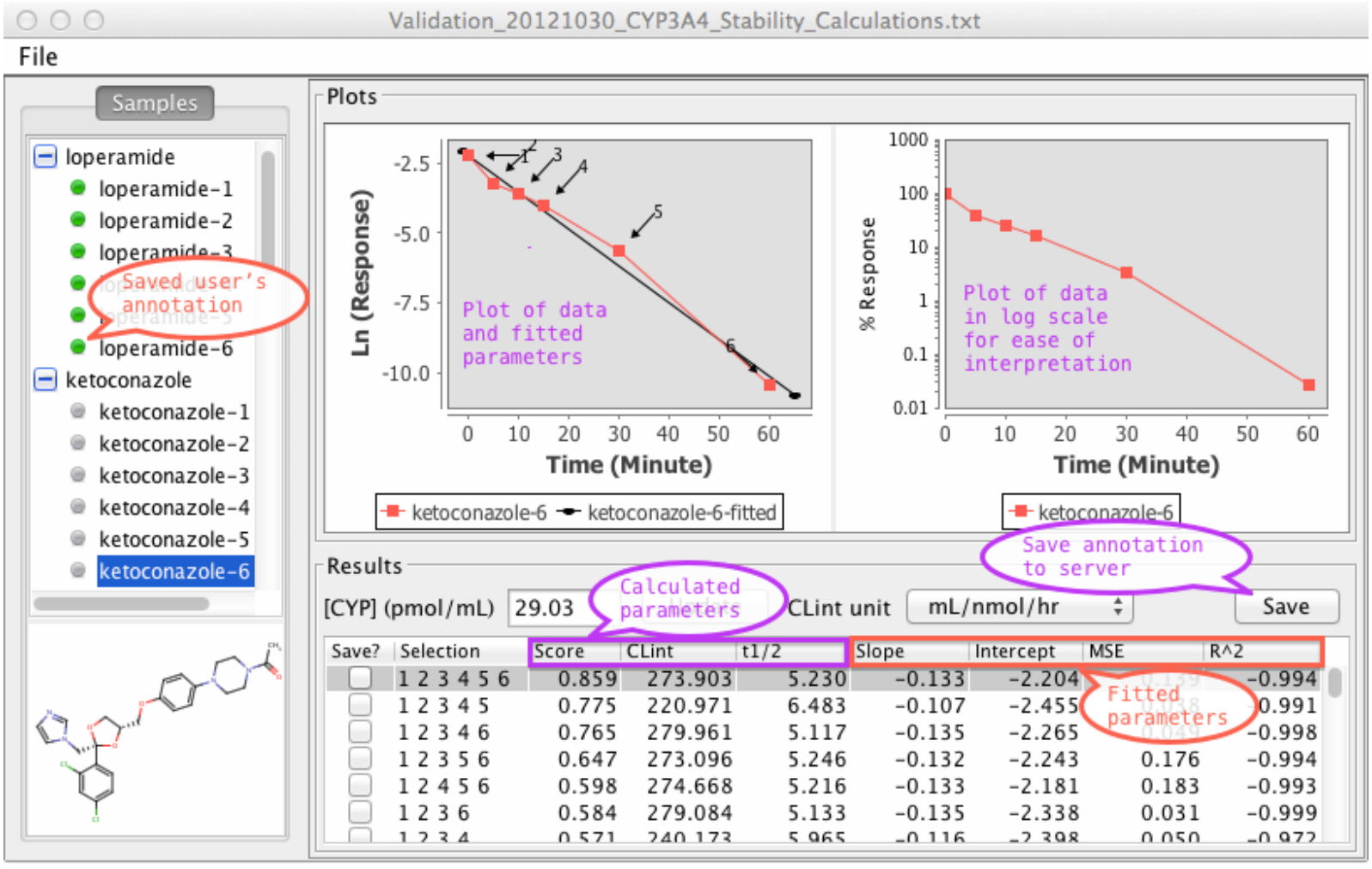


- Minimal user interaction
  - Precalculate results for all possible combinations
  - Rank results based on scoring function
- A web-based protocol for data management (i.e., data upload, download, and annotation)
- Available at

  `http://tripod.nih.gov/ws/iqc/iqc.jnlp`

Desktop client (cont'd)

Screenshot



Proposed scoring function

- Developed based on discussion with Scott Obach and Ed Kerns
- Let the raw score be defined as follows

$$\text{Score}_{\text{raw}} = NR^2 \exp(-\sigma_e) \sum_i 2^{-i}, \tag{1}$$

where N is the number of data points used to build the linear model, $\sigma_e \geq 0$ and $R \in [-1, 1]$ are the estimated mean square error and Pearson's correlation of the model, respectively.

- Weighted contribution of T_0, T_5, T_{10}, T_{15}, T_{30}, T_{60} as $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$, respectively.
- The best possible raw score (i.e., $\sigma_e = 0$, $|R| = 1$, and $N = 6$) is

$$\text{Score}_{\text{best}} = 6 \sum_{i=0}^{5} 2^{-i} = 11.8125$$

with i corresponds to the index of $\{T_0, T_5, T_{10}, T_{15}, T_{30}, T_{60}\}$.

Scoring function (cont'd)

- Normalized score is the ratio of raw and best

$$\text{Score} = \frac{\text{Score}_{\text{raw}}}{\text{Score}_{\text{best}}}$$

  where Score $\in [0, 1]$.
- Initial observations
  - Good fit when score ≥ 0.9
  - In good agreement with with manual evaluation (based on limited annotations from Ed Kerns)
- Require additional validation
  - Collect additional manual annotations
  - Evaluation metric (e.g., Spearman's rank correlation)