Estimating the probability that two phenotypes are the same

The parameter sensitivity calculations and clustering approach provide an estimate of solution sensitivity and redundancy. This is essential for estimating the reliability of a solution in the context of a particular experimental data set. The parameter sensitivity analysis provides an estimate of the effect of any one parameter in a solution on the quality of the fit, while the clustering attempts to agglomerate the information from numerous solutions to determine the non-overlapping maximum-likelihood parameter sets that exist within a population of solutions.

In addition, it is important to estimate if two particular phenotypes are biologically different (e.g. that a mutant phenotype differs from the wildtype). Since a phenotype is essentially a model fit for a particular log-fluorescence time course, determining if two solutions are different can be estimated in the context of one or more datasets. Furthermore, if multiple possible solutions are found after clustering, it may be important to estimate which solution cluster is the likeliest to describe the experimental data. Therefore, a statistical approach was used to identify the representative solution for each phenotyped experimental condition.

The algorithm for estimating the conservative probability that best-fit parameter set X^1 is equivalent to best-fit parameter X^2 in the context of experimental datasets, F^1 and F^2.

ALGORITHM: *ProbabilityPhenotypesEquivalent*

GIVEN model fitting solutions, X^1 and X^2

GIVEN experimental log-fluorescence time courses, F^1 and F^2 (F^1 need not differ from F^2).

DEFINE Sets D_{11}, D_{12}, D_{21}, D_{22}

$D_{11} = D_{12} = D_{21} = D_{22} = \{\}$

$< Upper^1, Lower^1 > \leftarrow CalcParamSensitivities(X^1)$

$< Upper^2, Lower^2 > \leftarrow CalcParamSensitivities(X^2)$

REPEAT N times:

FOREACH fitted parameter $x_i \in X_1$

$x_i \leftarrow rand \cdot (Upper_i^1 - Lower_i^1) + Lower_i^1$, where rand is uniform on [0,1]

ENDFOREACH

$d_{11} \leftarrow CalcScore(X^1, F^1)$

$d_{12} = CalcScore(X^1, F^2)$

$D_{11} = D_{11} \cup d_{11}$

$D_{12} = D_{12} \cup d_{12}$

FOREACH fitted parameter $x_i \in X_2$

$x_i \leftarrow rand \cdot (Upper_i^2 - Lower_i^2) + Lower_i^2$, where rand is uniform on [0,1]

ENDFOREACH

$d_{21} \leftarrow CalcScore(X^2, F^1)$

$d_{22} = CalcScore(X^2, F^2)$

$D_{21} = D_{21} \cup d_{21}$

$D_{22} = D_{22} \cup d_{22}$

END REPEAT

$r_1 = \Pr(U_1 \mid \eta(\mu = \mid D_{11} \mid \cdot \mid D_{12} \mid \cdot 0.5, \sigma = \sqrt{\frac{\mid D_{11} \mid \cdot \mid D_{12} \mid +1}{12}})$ where

$$U_1 = \sum_{d_i \in D_{11}} \sum_{d_j \in D_{12}} \delta\left(d_i, d_j\right), \quad \delta(d_i, d_j) = \begin{cases} 1, d_i < d_j \\ 0, d_i \geq d_j \end{cases}.$$

$r_2 = \Pr(U_2 \mid \eta(\mu = \mid D_{21} \mid \cdot \mid D_{22} \mid \cdot 0.5, \sigma = \sqrt{\frac{\mid D_{21} \mid \cdot \mid D_{22} \mid +1}{12}})$,where $U_2 = \sum_{d_i \in D_{21}} \sum_{d_j \in D_{22}} \delta\left(d_i, d_j\right)$.

$r_3 = \Pr(U_3 \mid \eta(\mu = \mid D_{11} \mid \cdot \mid D_{21} \mid \cdot 0.5, \sigma = \sqrt{\frac{\mid D_{11} \mid \cdot \mid D_{21} \mid +1}{12}})$, where $U_3 = \sum_{d_i \in D_{11}} \sum_{d_j \in D_{21}} \delta\left(d_i, d_j\right)$.

$r_4 = \Pr(U_4 \mid \eta(\mu = \mid D_{12} \mid \cdot \mid D_{22} \mid \cdot 0.5, \sigma = \sqrt{\frac{\mid D_{12} \mid \cdot \mid D_{22} \mid +1}{12}})$, where $U_4 = \sum_{d_i \in D_{12}} \sum_{d_j \in D_{22}} \delta\left(d_i, d_j\right)$.

Then estimate the conservative probability that the two phenotypes are different given the calculated probabilities as:

$$\mathrm{p_{diff}} \leq \max\left(min\left\{r_3, 1 - r_3\right\}, min\left\{r_4, 1 - r_4\right\}\right)$$

RETURN $1 - p_{diff}$

END

In brief, the known null hypothesis distribution for U, the Mann-Whitney U statistic [10] is used to calculate the probabilities r_1, r_2, r_3, and r_4, that two difference distributions are statistically identical given their respective sum of difference ranks. Then the conservative probability that

the two phenotypes are different is the maximum probability that each phenotype does a better job of describing any one of the data sets.

To estimate the probabilities that one of the models best describes a particular data set given the other model and data set we can use the calculated probabilities, r_1, r_2, r_3, r_4 from the above algorithm to calculate the joint probability that both the solution in question fits the specific dataset better than the other dataset, and that the other solution does a worse job (negative control). Thus the optimistic probability that phenotype $X^{1,2} \rightarrow$ (describes) $F^{1,2}$ is calculated:

$$\Pr\left(X_1 \rightarrow F_1 | D_{11}, D_{12}, D_{21}, D_{22}\right) = r_1 \cdot r_3$$

$$\Pr\left(X_1 \rightarrow F_2 | D_{11}, D_{12}, D_{21}, D_{22}\right) = (1 - r_1) \cdot r_4$$

$$\Pr\left(X_2 \rightarrow F_1 | D_{11}, D_{12}, D_{21}, D_{22}\right) = r_2 \cdot (1 - r_3)$$

$$\Pr\left(X_2 \rightarrow F_2 | D_{11}, D_{12}, D_{21}, D_{22}\right) = (1 - r_2) \cdot (1 - r_4) \, .$$