A Grid Algorithm for High Throughput Fitting of Dose-Response Curve Data

Yuhong Wang*, Ajit Jadhav, Noel Southal, Ruili Huang and Dac-Trung Nguyen

National Institutes of Health, NIH Chemical Genomics Center, 9800 Medical Center Drive, Rockville, MD 20850, USA

Abstract: We describe a novel algorithm, Grid algorithm, and the corresponding computer program for high throughput fitting of dose-response curves that are described by the four-parameter symmetric logistic dose-response model. The Grid algorithm searches through all points in a grid of four dimensions (parameters) and finds the optimum one that corresponds to the best fit. Using simulated dose-response curves, we examined the Grid program's performance in reproducing the actual values that were used to generate the simulated data and compared it with the DRC package for the language and environment R and the XLfit add-in for Microsoft Excel. The Grid program was robust and consistently recovered the actual values for both complete and partial curves with or without noise. Both DRC and XLfit performed well on data without noise, but they were sensitive to and their performance degraded rapidly with increasing noise. The Grid program is automated and scalable to millions of dose-response curves, and it is able to process 100,000 dose-response curves from high throughput screening experiment per CPU hour. The Grid program has the potential of greatly increasing the productivity of large-scale dose-response data analysis and early drug discovery processes, and it is also applicable to many other curve fitting problems in chemical, biological, and medical sciences.

Keywords: Curve fitting, Hill equation, Grid algorithm, Dose-response curve, High throughput screening.

1. INTRODUCTION

The technological advances in high throughput screening (HTS) [1] shifted the rate-limiting step in drug discovery from data generation to data analysis. In today's drug discovery process, thousands of dose-response curves are routinely generated in a typical secondary screening project. At the National Institutes of Health Chemical Genomics Center (NCGC), we have developed a new quantitative high-throughput screening (qHTS) paradigm [2] to profile every compound in large collections of chemicals and search for chemical compounds of great potential in probing the chemical, genomic and biological universes and thus the biological pathways. In this new paradigm, compound titration is performed in the primary screening, and hundreds of thousands of dose-response curves are routinely generated.

High throughput curve fitting and outlier detection of such amount of dose-response data remain a great challenge [3,4]. First, the current nonlinear curve fitting algorithms, such as Levenberg–Marquardt (LM) algorithm [5], are based upon derivatives, their solutions correspond to local optimum, and the quality of the solutions to a large degree depends upon data quality and starting point. To find a good curve fit and solution, outlier data points, a common phenomenon for HTS data, need to be manually detected and masked, and different starting points need to be tried. As a consequence, these algorithms are difficult to automate and are generally not scalable.

A number of computer programs are currently available for fitting dose-response curves. The most widely used programs are the package DRC [6], an add-on for the language and environment R [7], and XLfit(R), a Microsoft(R) Excel add-in (www.idbs.com). The XLfit add-in is very popular in pharmaceutical and biotech industries, and the DRC package is used more in academic and government research institutions.

In this study, we proposed a novel algorithm, the Grid algorithm, for high throughput curve fitting. We will describe the algorithm and its performance first and then compare it with DRC and XLfit on simulated dose-response data.

2. HILL EQUATION AND GRID ALGORITHM

2.1. Hill Equation

The dose-response curve is modeled by the four-parameter symmetric logistic model or Hill equation [8]:

$$y = y_{min} + \frac{y_{max} - y_{min}}{1 + 10^{(EC50 - x) \times slope}}$$

Where y is the biological response of a chemical compound or biological agent, x the dose or concentration of an agent in log unit, y_{min} the biological activity without the compound, y_{max} the maximum saturated activity at high concentration, *EC50* the inflection point in log scale at which y is at the middle of y_{min} and y_{max}, and *slope* is Hill slope.

The goal of a curve fitting algorithm is to determine a statistically optimized model that best fits the data set. One commonly used quantitative measure for the fitness of a model is R^2, which is defined as

$$R^2 = 1.0 - \frac{SS_h}{SS_c}$$

*Address correspondence to this author at the National Institutes of Health, NIH Chemical Genomics Center, 9800 Medical Center Drive, Rockville, MD 20850, USA; Tel: 215-358-0186; Fax: 301-217-5728; E-mail: wangyuh@mail.nih.gov

1875-3973/10

Where SS_h and SS_c are the mean sum-of-square of the vertical distances of the points from the fitted curve (Hill fit) or the line (constant fit), respectively. Typical non-linear optimization algorithms start from initial values of the four variable parameters, evaluate first and/or second order derivative of R^2 with respect to each variable, and adjust these variables to optimize target function either SS_h or R^2.

2.2. Grid Program

The Grid algorithm is implemented as first of the three main components in the Grid program; the second is for outlier detection, and the third for statistical evaluation. The Grid algorithm is conceptually simple: it goes through all points in a grid of four parameters or dimensions and finds the point that has the optimum SS_h or R^2. To make it efficient, the Grid program searches a coarse grid first followed by a fine one; it consists of four major steps:

Step 1: Define the coarse grid. The upper boundary for *EC50* is fixed at -2 log(M) (10 mM). The lower boundary defaults at -10 log(M) (0.1 nM) or 1.2 times the log value of the lowest dose for whichever is smaller. The most commonly tested concentrations in HTS are within these boundaries. The interval defaults at 0.5.

To determine the boundaries for y_{min} and y_{max}, we find the minimum and maximum responses, minY and maxY first, and then perform linear regression on the dose-response data in order to determine the direction of the data points: increasing or decreasing. Increasing data corresponds to biological activator, and decreasing inhibitor. For increasing curves, if minY>0.0, the boundaries for y_{min} default between 0.0 and 1.2*minY, otherwise between 1.2*minY and 0.0. If maxY >0.0, the boundaries for y_{max} default between 0.0 and 1.2*maxY, otherwise between 1.2*maxY and 0.0. The boundaries for y_{min} and y_{max} are similarly decided for decreasing curves. Here, 1.2 is the default scaling coefficient. The intervals for both y_{min} and y_{max} default at 20^{th} of their ranges or 2.0 for whichever is greater, and they are denoted as dy_{min} and dy_{max}, respectively.

The boundaries for *slope* are dynamically calculated. For each sampled values of *EC50*, y_{min} and y_{max} using the above determined boundaries and intervals, we estimated *slope* by linear regression of the following rewritten Hill equation:

$$log(\frac{y_{max}-y_{min}}{y_i-y_{min}}-1)=(EC50-x_i)\times slope$$

Where x_i and y_i are the *i*th concentration and response of a dose-response curve. The *slope* from linear regression is denoted as s_0. $s_0/2$ and $2s_0$ are used as the default lower and upper boundaries of *slope*; the interval defaults at 0.1.

All default values of both boundaries and intervals can be changed in the Grid program's preference.

Step 2: Find the coarse solution. R^2 for each point in the grid of four dimensions is calculated, and the one with the greatest R^2 is the coarse solution. Let us denote the solution as $\hat{}EC50$, $\hat{}slope$, $\hat{}y_{min}$, and $\hat{}y_{max}$.

Step 3: Define the fine grid. For *EC50*, the boundaries are ($\hat{}EC50$-1.0, $\hat{}EC50$+1.0), and the interval is 0.05 by default. For y_{min}, the boundaries are ($\hat{}y_{min}$-2*dy_{min}, $\hat{}y_{min}$+2*dy_{min}), and the interval is 0.1*dy_{min} or 0.5 for whichever is greater. For y_{max}, the boundaries are ($\hat{}y_{max}$-2*dy_{max}, $\hat{}y_{max}$+2*dy_{max}), and the interval is 0.1*dy_{max} or 0.5 for whichever is greater. The boundaries *slope* is calculated according to the same procedure as described in Step 1.

Step 4: Find the fine solution. The procedure is the same as Step 2.

2.3. Outlier Detection and Data Masking

Outlier detection and curve fitting are two inseparable processes. Current available computer programs generally require visual inspection and manual intervention for outlier detection. This is apparently not feasible for high throughput fitting of a large number of dose-response curves. At the same time, manual process tends to be subjective and error-prone.

Our outlier detection algorithm consists of mainly two steps. First, we used the deviation of a data point from the extrapolated one on the line connecting the previous two data points, next two data points, or previous and next data points. A data point is masked if the deviation is > 70% (default) of maxY-minY. As tested in a large number of qHTS data, 70% seems to be a good default value. After a curve is fitted with the outlier data points masked by their deviations from the extrapolated lines, in the second step, we recalculate each data point's deviation from the fitted curve. If a data point has a deviation < 30% (default) of maxY-minY, it is unmasked. The curve will be refitted if there is any change in data masking in the second step.

2.4. Statistical Evaluation

We performed two statistical evaluations on the curve fitting results. The first is an F test to compare a flat line fit model with the Hill model. In this test, the F ratio of the mean sum-of-square of differences (SS) of the flat model over that of the Hill model is calculated, and the corresponding significance or *p* value is obtained from the F distribution.

The second is confidence interval (CI) evaluation. The CI of a parameter is defined as the interval that has a certain probability (95% for example) of containing the true value of the parameter. CI values can be directly calculated from estimated standard errors or Monte Carlo simulation [3]. In the Grid program, we used reversed F test. We start from a *p* value, for example 0.05 for a confidence of 95%, and estimate the F ratio based upon the given *p* value and number of degrees of freedom. After that, we sample the space of four parameters and find the intervals of these parameters in which the corresponding mean sum-of-squares are less than the F ratio times the sum-of-squares of the best-fit model.

2.5. Generation and Fitting of Simulated Dose-Response Data

We used computer-simulated dose-response data with or without random noise of moderate amplitude to test the performance of the Grid program and compare it with DRC and XLfit. The main idea is that a good fitting program should be able to reproduce the actual values of the four parameters (*EC50*, *slope*, y_{min}, and y_{max}) that are used to generate the simulated dose-response data.

To generate the simulated data, y_{min} and y_{max} are randomly chosen between -25% and 25%, and 25% and 125%,

respectively. *EC50* is between -10.0 or 0.1nM and -4.0 or 10mM. The concentration starts at 0.1nM and is serially increased by two folds. The biological response is calculated according to the Hill equation. These ranges are typical for high throughput screening. s*lope* is between 0.5 and 5.0. We randomly generated two sets, each 100, of dose-response curves of 20 data points. 20 data points are quite standard in secondary compound screening on 384 well plates. The first set of 100 curves is directly used. Since the randomly sampled *EC50* values fall within the used concentration range, most curves in this set are complete: they start from a plateau of y_{min} at the low concentrations and end at the plateau of y_{max} at the high concentrations. From the second set of 100 curves, 10 consecutive data points are randomly selected to form an incomplete or partial curve. For high throughput screening, most of the generated dose-response curves are incomplete, and the ability to recover actual parameters from such partial curves is crucial. These 200 curves of 20 and 10 data points are without any random noise.

Four sets, each 100, of 20 data points curves are similar generated but with 2, 4, 6, and 8 data points randomly selected and perturbed by a value from -50% to 50%. Four sets, each 100, of 10 data points and incomplete curves are also generated with 1, 2, 3, and 4 data points randomly selected and perturbed by a value from -50% to 50%. Most of the noises observed in HTS assays are of similar nature. The alternative method is to apply random perturbation at all data points. This method is not used in this study. First, this kind of uniform noises is rarely observed in experimental HTS data. Second, as seen from our numerical experimentations, perturbation of even moderate amplitude (20-40% for instance) at all data points could easily disrupt the original dose-response curve and make it impossible to recover the four actual parameters.

Curve fitting using the Grid program and DRC was performed on Mac OS X 10.6. The Grid program is implemented using Java JDK1.5 and its source codes and binary are available at NCGC's web site (http://ncgc.nih.gov/resources/software.html) together with sample data. We used the R program for Mac OS X version 2.9 and DRC package version 1.7 (http://cran.r-project.org/web/packages/drc/index.html) and wrote a simple Unix shell script to perform the curve fitting. Fitting using XLfit add-in version 5.1 for Microsoft Excel was performed on Microsoft Windows XP.

2.6. Comparison of Grid, DRC and XLfit Programs

For each set of 100 dose-response curves and for each of parameters (*EC50*, *slope*, y_{min} and y_{max}), we calculated and used six numbers to compare the performance of Grid, DRC and XLfit programs: the median and maximum differences between the fitted and the actual values; the number of curves without fit; the number of curves with absolute difference > 2.0 for *EC50*, absolute difference > 200% for y_{min} and y_{max}, ratio >2.0 or < 0.5 for *slope*; Pearson correlation coefficient r and significance (*p* value).

3. RESULTS AND DISCUSSION

3.1. Fitting of Simulated Data with Grid Program

For simulated data without noise, the fitted four parameters (*EC50*, *slope*, y_{min} and y_{max}) by the Grid program are significantly correlated with the actual values ($p < 2.2e-16$) (Table **1**). For *EC50*, actual values are recovered with maximum difference < 0.5 (in log scale). For curves of 10 data points, the median difference of y_{min} is 0.42%, but the maximum difference is 53.49%. We examined a few dose-response curves with difference of y_{min} > 10.0%, and the one with the maximum difference is plotted in Fig. (**1**). For this curve, the actual y_{min} is -15.9, but the fitted value is 37.5%. The median difference of y_{max} is 0.36%, and the maximum one is 38.75%. The dose-response curve with the maximum difference is plotted in Fig. (**2**). For this curve, the actual y_{max} is 90.75%, but the fitted value is 64.0%. For these a few partial curves, the Grid program apparently has difficulty in reproducing the actual y_{min} and y_{max} values. For curves of 20 data points, as expected, the differences between fitted and actual y_{min} and y_{max} values are smaller for two reasons. First, more data points add more constraints for and thus reduce the possible space of solutions. Second, most of the 20 data points curves are complete curves.

With increasing noise, the differences between fitted and actual values grow larger. Among the four parameters (*EC50*, *slope*, y_{min}, and y_{max}), *EC50* and y_{max} are biologically the most important; and here we will focus on these two parameters. For curves of 10 data points, as the number of randomly perturbed data points increases from 0 to 4, correlation coefficient for *EC50* decreases from 0.997 to 0.673, 0.612, 0.370, and 0.278; the number of curves with maximum difference of EC50 > 2.0 increases from 0 to 5, 10, 18, and 28; and there are no curves with maximum difference of y_{max} > 100%. For curves of 20 data points, the results are remarkably better. The correlation coefficients remain > 0.65, and the number of curves with EC50>2.0 only rises from 0 to 1, 6, 9, and 11. Overall, the Grid program is able to recover the actual parameters for most of the simulated data even with 4 of 10 data points randomly perturbed, and the fitted values remain significantly correlated with the actual ones ($p < 0.05$ with one exception of 0.055).

The Grid program failed to fit 49 curves of 10 data points out of 400 curves with noise. We examined these failed dose-response curves and plot a typical one in Fig. (**3**). This curve has 3 data points randomly perturbed. Apparently, the perturbation disrupted the original curve and made it insignificant.

For recovering the actual parameters of dose-response curves, 20 data points are substantially better than 10 data points. But the Grid program is still able to recover the actual parameters for most of the simulated and partial curves of 10 data points. For high throughput data screening, this is critical. First, 10 data points curves are more economical to generate than 20 data points ones. Second, for HTS, it is difficult to optimize the concentration range to generate a full dose-response curve for compounds of great difference in potency.

3.2. Comparison Between Grid and DRC Program

The fitting results by DRC are given in Table **2**. The starting parameters are automatically assigned by DRC. For data without noise, DRC performs better than the Grid program in recovering the actual parameters with only one exception: the maximum difference of y_{min} for 20 data points curves is 138.35%. In the Grid algorithm, each parameter is

Table 1 Differences Between Fitted and Actual Values of *EC50*, *Slope*, y_{min}, and y_{max} by Grid Program

Parameter	No1	No2	No3	No4	Median Dif	Max Dif	pvalue	r
EC50	10	0	0	0	0.01	0.43	<2.2e-16	0.997
	10	1	17	5	0.05	6.6	3.07E-12	0.673
	10	2	19	10	0.32	6.67	1.24E-09	0.612
	10	3	10	18	0.3	8.93	0.0003379	0.37
	10	4	3	28	0.81	6.02	0.005845	0.278
	20	0	0	0	0.01	0.32	<2.2e-16	0.999
	20	2	5	1	0.03	3.86	<2.2e-16	0.971
	20	4	2	6	0.09	7.91	4.06E-14	0.671
	20	6	0	9	0.11	8.62	1.07E-13	0.658
	20	8	0	11	0.22	6.55	4.35E-14	0.665
slope	10	0	0	2	0.07	2.7	<2.2e-16	0.91
	10	1	17	15	0.38	3.85	2.20E-11	0.653
	10	2	19	24	0.92	3.23	3.12E-07	0.532
	10	3	10	41	1.17	4.63	6.13E-4	0.354
	10	4	3	50	1.51	4.52	0.07407	0.182
	20	0	0	0	0.05	0.74	<2.2e-16	0.99
	20	2	5	7	0.29	3.33	<2.2e-16	0.726
	20	4	2	30	0.83	4.37	6.97E-06	0.437
	20	6	0	42	1.02	4.52	5.47E-09	0.543
	20	8	0	41	1.27	4.46	6.89E-05	0.387
y_{min}	10	0	0	0	0.42	53.49	<2.2e-16	0.795
	10	1	17	0	1.26	51.62	2.67E-11	0.651
	10	2	19	0	3.7	51.59	2.68E-08	0.571
	10	3	10	0	5.7	123.68	0.005533	0.29
	10	4	3	0	8.24	152.71	0.055	0.195
	20	0	0	0	0.19	21.53	<2.2e-16	0.954
	20	2	5	0	0.47	28.13	<2.2e-16	0.918
	20	4	2	0	2.52	59.25	<2.2e-16	0.804
	20	6	0	0	3.8	65.35	6.66E-12	0.619
	20	8	0	0	4.88	61.99	1.34E-11	0.612
y_{max}	10	0	0	0	0.36	38.75	<2.2e-16	0.937
	10	1	17	0	2.1	159.53	7.89E-12	0.664
	10	2	19	0	8.74	159.53	0.005633	0.305
	10	3	10	0	11.14	148.41	0.005119	0.293
	10	4	3	0	21.17	159.53	0.001426	0.319
	20	0	0	0	0.19	24.51	<2.2e-16	0.99
	20	2	5	0	0.79	65.92	<2.2e-16	0.876
	20	4	2	0	1.83	73.2	<2.2e-16	0.789
	20	6	0	0	3.49	74.54	1.47E-14	0.674
	20	8	0	0	3.07	85.75	7.55E-15	0.68

No1: number of data points; No2: number of randomly perturbed data points; No3: number of curves without fit; No4: number of curves with absolute differences in *EC50*, y_{min} and y_{max} > 2.0, 200.0% and 200.0%, respectively and maximum difference in ratio of *slope* > 2.0 or < 0.5; Median Dif: median difference; Max Dif: maximum difference; pvalue: significance of Pearson correlations; and r: Pearson correlation coefficient.

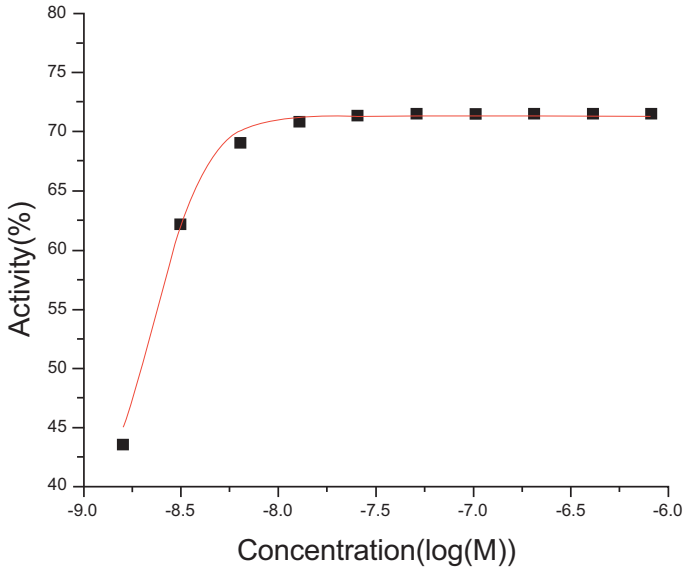

Fig. (1). Simulated 10 data points dose-response curve with actual y_{min} of -15.98%. y_{min} from Grid program is 37.55%. The difficulty in reproducing the actual y_{min} by the Grid program is apparently due to the fact that y_{min} is only supported by one data point.

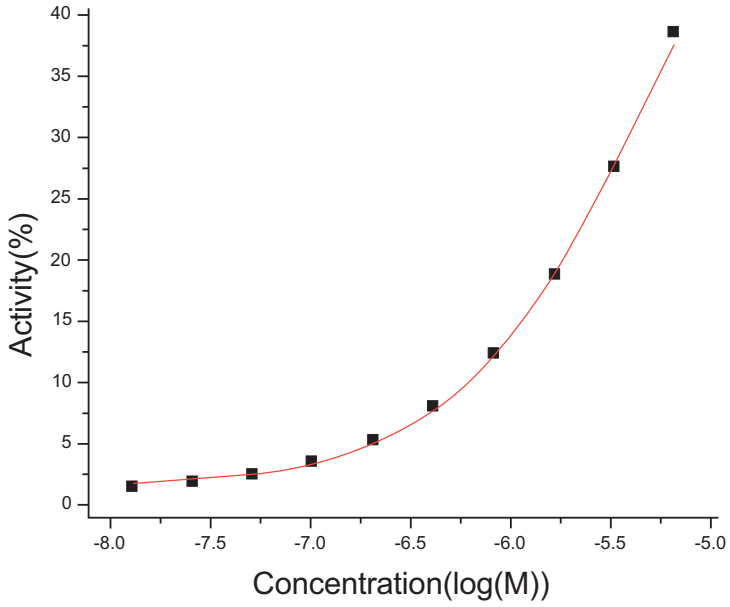

Fig. (2). Simulated 10 data points dose-response curve with actual y_{max} of 90.74%. y_{max} from Grid program is 64.00%. The difficulty in reproducing the actual y_{max} by the Grid program is due to the default constraint that the boundaries for y_{max} is between 0.0 and 1.2*maxY. In this special case, if we remove the constraint, the solution can be much closer to the actual values. For experimental data, however, this default constraint produces more reasonable solutions.

sampled at a defined interval. For practical application, the accuracy in the fitted values by Grid algorithm is sufficient. For example, the default sampling interval for *EC50* is 0.05, and this corresponds to a difference in concentration by $10^{0.05}$ or 1.12 times. For drug screening process, two times difference is generally acceptable. From a purely mathematical and numeric point of view, however, the solutions from DRC have better accuracy than the Grid algorithm.

For 10 data points curves with noise, however, DRC's performance degrade substantially. Out of 16 correlations (four parameters (*EC50*, *slope*, y_{min} and y_{max}) by four data sets), only three have significant correlation ($p<0.05$). Some correlation coefficients are even negative. The largest median difference in *EC50* is 1.51 or 31 ($10^{1.51}$) times different, and the largest maximum difference is 43.53 or $3x10^{43}$ times different. The largest median differences in y_{min} and y_{max} are 25.9% and 23.65%, and the largest maximum ones are 1,870% and 1,520%, respectively. As the number of randomly perturbed data points increases from 0 to 4, the correlation coefficient for *EC50* decreases from 1.0 to 0.136, -0.042, 0.150, and 0.072; the number of curves with maximum difference of *EC50* > 2.0 increases from 0 to 10, 26, 28, and 33; and the number of curves with maximum difference of y_{max} > 100% increases from 0 to 2, 6, 6, and 5. For

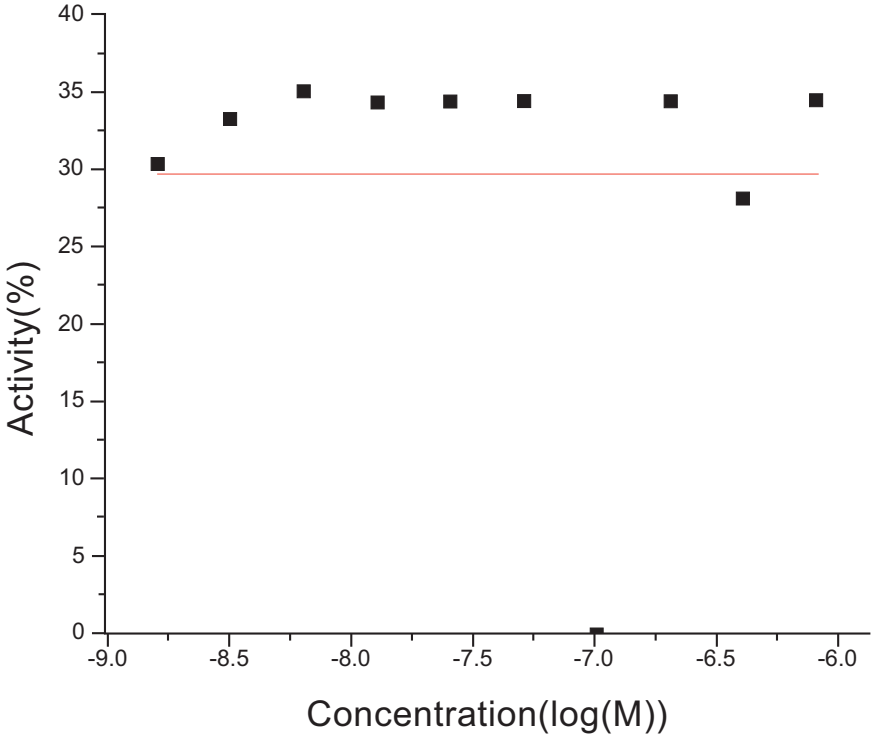

Fig. (3). Simulated 10 data points dose-response curve with three data points randomly perturbed by a random number between -50% and 50%. Grid program failed to fit this curve. Note the data point at bottom is automatically identified and masked.

Table 2. Differences Between Fitted and Actual Values of *EC50*, *Slope*, y_{min}, and y_{max} by DRC

Parameter	No1	No2	No3	No4	Median Dif	Max Dif	pvalue	r
EC50	10	0	0	0	0	0.02	<2.2e-16	1
	10	1	6	10	0.1	4.35E+01	0.1912	0.136
	10	2	6	26	0.82	1.95E+01	0.6864	-0.042
	10	3	4	28	1.17	14.02	0.1434	0.15
	10	4	2	33	1.51	13.85	0.4833	0.072
	20	0	1	0	0	0.48	<2.2e-16	1
	20	2	7	4	0.05	4.85	<2.2e-16	0.896
	20	4	4	9	0.11	8.82E+00	2.67E-15	0.698
	20	6	6	8	0.16	2.01E+01	1.67E-05	0.428
	20	8	3	13	0.22	1.06E+01	1.34E-07	0.505
slope	10	0	0	0	0	0.17	<2.2e-16	1
	10	1	6	40	0.92	2.06E+01	0.8273	0.023
	10	2	6	65	2.81	2.43E+01	0.1721	0.142
	10	3	4	72	2.56	31.32	0.6223	-0.051
	10	4	2	78	4.17	32.96	0.006632	-0.273
	20	0	1	0	0	1.16	<2.2e-16	0.996
	20	2	7	25	0.58	12.17	0.0003296	0.364
	20	4	4	38	1.17	1.59E+01	0.04568	0.204
	20	6	6	49	1.58	1.29E+01	0.2353	0.124
	20	8	3	59	1.9	2.18E+01	0.3444	0.097

(Table 2). Contd…..

Parameter	No1	No2	No3	No4	Median Dif	Max Dif	pvalue	r
y_{min}	10	0	0	0	0	6.95	<2.2e-16	0.999
	10	1	6	6	5.86	8.94E+02	0.5452	0.063
	10	2	6	5	11.12	1.12E+03	0.3011	0.108
	10	3	4	8	13.38	1.87E+03	0.3933	-0.088
	10	4	2	12	25.93	638.85	0.4196	0.082
	20	0	1	0	0	138.35	4.44E-15	0.686
	20	2	7	1	2.38	255.71	0.003301	0.302
	20	4	4	0	2.92	191.66	0.001566	0.318
	20	6	6	1	4.25	2.53E+02	0.001778	0.318
	20	8	3	3	6.46	6.40E+02	0.8283	0.022
y_{max}	10	0	0	0	0	0.52	<2.2e-16	1
	10	1	6	2	6.12	4.24E+02	0.0003332	0.362
	10	2	6	6	16.09	400.79	0.1238	0.16
	10	3	4	6	17.66	857.84	0.02081	0.236
	10	4	2	5	23.65	1.52E+03	0.1223	0.157
	20	0	1	0	0	0.15	<2.2e-16	1
	20	2	7	4	2.18	2.71E+03	0.4401	0.081
	20	4	4	3	4.58	667.11	0.05901	0.193
	20	6	6	3	5.5	3.34E+03	0.3959	0.089
	20	8	3	3	6.89	4.85E+02	0.04059	0.208

No1: number of data points; No2: number of randomly perturbed data points; No3: number of curves without fit; No4: number of curves with absolute differences in *EC50*, y_{min} and y_{max} > 2.0, 200.0% and 200.0%, respectively and maximum difference in ratio of *slope* > 2.0 or < 0.5; Median Dif: median difference; Max Dif: maximum difference; pvalue: significance of Pearson correlations; and r: Pearson correlation coefficient.

curves of 20 data points, the results are better. Out of 16 correlations, most have a significant correlation ($p<0.05$).

In a word, DRC is very sensitive to noises, it tends to produce biologically unreasonable solutions, and the Grid program is much more robust. One main reason for the Grid program's robustness is that it effectively has multiple starting points; in fact, it tries all possible starting points in a grid. Using multiple starting points is expected to improve the performance of DRC.

3.3. Comparison Between XLfit Programs with and Without Prefit

XLfit program has the option of prefitting four parameters. The fitting results by XLfit with and without prefit are given in Tables **3** and **4**, respectively. The results from prefit are generally worse than those without prefit. For example, for *EC50*, seven out of the eight correlations (four data sets of 20 data points with noise and four data sets of 10 data points with noise) are significant ($p<0.05$) without prefit, but only two are significant with prefit. The maximum differences between the fitted and actual values tend to be larger for results with prefit than those without prefit.

3.4. Comparison Between Grid, DRC and XLfit Programs

Like DRC, for data without noise, XLfit performs better in recovering the actual parameters than the Grid program. For data with noise, however, XLfit generally performs worse than DRC. First, the maximum differences between the fitted and actual parameters by XLfit are much larger than those by DRC. Second, even for curves of 20 data points, most of the 16 correlations become insignificant ($p>0.05$). We tried various options provided by XLfit such as locking a parameter at a prefitted value, but it did not improve.

Comparing with DRC and XLfit programs, the Grid program has two major advantages. The most important one is its robustness and accuracy. The Grid program is able to consistently reproduce the actual values for both complete and partial curves without or with noise. Both DRC and XLfit perform well for data without noise, but they are very sensitive to noise and their performance degrades rapidly with increasing noise. As discussed above, being able to recover actual values of four Hill parameters for partial curves with noise is very important for HTS.

Second, the Grid program is automated and scalable: It does not need to try various starting points in order to achieve good fitting, and it automatically identifies and masks outlier data points. At NCGC, the Grid program is routinely used to fit hundreds of thousands of curves with minimum human interaction. For high quality and high throughput fitting of large number of dose-response curve data, the Grid program has been proven indispensable.

The Grid program's robustness can be attributed to the very nature of the Grid algorithm. The Grid algorithm

Table 3. Differences Between Fitted and Actual Values of *EC50*, *Slope*, y_{min}, and y_{max} by XLfit with Prefit

Parameter	No1	No2	No3	No4	Median Dif	Max Dif	pvalue	r
EC50	10	0	0	0	0	5.00E-04	<2.2e-16	1.000
	10	1	0	22	0.13	37.17	0.2758	0.110
	10	2	0	38	1.27	60.67	0.2034	-0.128
	10	3	0	38	1.28	58.79	0.3286	0.099
	10	4	0	45	1.69	45.08	0.3929	0.086
	20	0	0	0	0	1.41	<2.2e-16	0.996
	20	2	1	11	0.05	226.3	0.1308	-0.153
	20	4	0	14	0.13	66.41	0.0007174	0.333
	20	6	0	18	0.19	26.76	0.003044	0.293
	20	8	0	17	0.25	53.92	0.7481	-0.033
slope	10	0	0	0	0	4.96E-04	<2.2e-16	1.000
	10	1	0	46	1.13	6.84E+15	0.0933	-0.169
	10	2	0	69	3.54	4.47E+29	0.7961	-0.026
	10	3	0	75	4.36	1.83E+66	0.3176	0.101
	10	4	0	78	8.79	7.84E+14	0.2788	-0.109
	20	0	0	0	0	1.16	<2.2e-16	0.994
	20	2	1	27	0.63	2.66E+22	0.5388	0.063
	20	4	0	43	1.21	1.85E+35	0.3934	-0.086
	20	6	0	55	1.75	1.08E+17	0.1767	0.136
	20	8	0	61	2.05	3.52E+21	0.1286	-0.153
y_{min}	10	0	0	0	0	4.96E-04	<2.2e-16	1.000
	10	1	0	13	5.73	2.75E+10	0.3146	0.102
	10	2	0	13	12.1	3.05E+04	0.759	-0.031
	10	3	0	13	10.24	3.09E+07	0.1349	-0.151
	10	4	0	14	14.36	1.74E+10	0.6675	0.043
	20	0	0	1	0	9.97E+04	0.3379	-0.097
	20	2	1	4	2.21	1.21E+09	0.5864	-0.055
	20	4	0	4	3.45	1.06E+08	0.5062	0.067
	20	6	0	5	4.7	2.00E+03	0.7532	-0.032
	20	8	0	5	5.49	2.76E+03	0.9608	0.005
y_{max}	10	0	0	0	0	4.97E-04	<2.2e-16	1.000
	10	1	0	1	5.49	1.86E+03	0.01008	0.256
	10	2	0	7	11.21	2.04E+04	0.1812	-0.135
	10	3	0	12	17.65	6.06E+05	0.5637	-0.058
	10	4	0	12	25.25	1.89E+04	0.07925	-0.176
	20	0	0	0	0	0.09	<2.2e-16	1.000
	20	2	1	4	2.19	2.61E+03	0.872	0.016
	20	4	0	6	4.44	2.01E+04	0.6786	0.042
	20	6	0	5	5.85	9.80E+04	0.7925	0.027
	20	8	0	6	5.9	4.93E+07	0.2669	-0.112

No1: number of data points; No2: number of randomly perturbed data points; No3: number of curves without fit; No4: number of curves with absolute differences in *EC50*, y_{min} and y_{max} > 2.0, 200.0% and 200.0%, respectively and maximum difference in ratio of *slope* > 2.0 or < 0.5; Median Dif: median difference; Max Dif: maximum difference; pvalue: significance of Pearson correlations; and r: Pearson correlation coefficient.

Table 4. Differences Between Fitted and Actual Values of *EC50, Slope,* y_{min}, and y_{max} by XLfit without Prefit

Parameter	No1	No2	No3	No4	Median Dif	Max Dif	pvalue	r
EC50	10	0	0	8	0	3.58	<2.2e-16	0.775
	10	1	0	22	0.13	37.17	0.2758	0.11
	10	2	0	39	1.41	10.58	0.02382	0.226
	10	3	0	49	1.93	10.4	0.01612	0.24
	10	4	0	40	1.56	24.79	0.009522	0.258
	20	0	0	6	0	3.61	<2.2e-16	0.929
	20	2	0	14	0.09	11.73	<2.2e-16	0.709
	20	4	0	11	0.19	8.3	1.23E-12	0.635
	20	6	0	13	0.22	10.42	<2.2e-16	0.73
	20	8	0	19	0.34	12.78	4.53E-08	0.514
slope	10	0	0	32	0	4.82	8.28E-07	0.47
	10	1	0	46	1.13	6.84E+15	0.0933	-0.169
	10	2	0	75	2.3	189.41	0.9782	0.003
	10	3	0	78	2.02	120.51	0.4397	0.078
	10	4	0	76	3.21	253.95	0.878	0.016
	20	0	0	15	0	4.48	<2.2e-16	0.746
	20	2	0	41	0.89	66.5	0.9688	0.004
	20	4	0	55	1.34	56.63	0.4695	0.073
	20	6	0	57	1.56	135.52	0.02047	0.232
	20	8	0	67	1.9	104.98	0.4246	0.081
y_{min}	10	0	0	3	0	1.79E+03	0.06308	0.187
	10	1	0	13	5.73	2.75E+10	0.3146	0.102
	10	2	0	5	13.25	2.77E+03	0.4586	0.075
	10	3	0	5	13.46	4.79E+04	0.8188	0.023
	10	4	0	9	14.21	1.99E+03	0.7425	0.033
	20	0	0	4	0	976	0.534	0.063
	20	2	0	3	2.96	762.23	0.3416	0.096
	20	4	0	1	3.85	358.8	0.01858	0.235
	20	6	0	4	4.79	7.93E+03	0.325	0.099
	20	8	0	5	6.42	6.71E+03	0.6725	0.043
y_{max}	10	0	0	3	0	2.23E+03	0.1889	-0.132
	10	1	0	1	5.49	1.86E+03	0.01008	0.256
	10	2	0	19	48.94	1.01E+06	0.1667	-0.139
	10	3	0	30	75.75	7.36E+07	0.895	0.013
	10	4	0	30	75.74	1.69E+08	0.5535	0.06
	20	0	0	0	0	99.51	<2.2e-16	0.81
	20	2	0	7	3.82	3.12E+04	0.6598	-0.045
	20	4	0	9	6.57	4.59E+07	0.6877	-0.041
	20	6	0	8	6.12	1.66E+07	0.4453	0.077
	20	8	0	5	7.02	2.77E+04	0.3871	-0.087

No1: number of data points; No2: number of randomly perturbed data points; No3: number of curves without fit; No4: number of curves with absolute differences in *EC50*, y_{min} and y_{max} > 2.0, 200.0% and 200.0%, respectively and maximum difference in ratio of *slope* > 2.0 or < 0.5; Median Dif: median difference; Max Dif: maximum difference; pvalue: significance of Pearson correlations; and r: Pearson correlation coefficient.

searches the best solution of each parameter at a predefined interval in the biologically reasonable domain, it is designed to avoid local minimum trap, and it does not need any starting points. The Grid algorithm does not evaluate any derivatives. The DRC and XLfit programs use first and second order derivatives to find the direction for downward movement. For data of poor quality, the derivatives could be numerically unstable, and the DRC and XLfit programs could be trapped in local minimum to produce unreasonable solutions.

In this study we focus on high throughput fitting of large amount of dose-response data and manual interaction with individual curves is not allowed. For the simulated data with noise, manual interaction, such as trying different starting points and identifying outlier data points, should improve the fitting results for the DRC and XLfit programs. Such manual interaction, however, is time-consuming, subjective and error-prone, and it is not feasible for fitting large number of dose-response data.

3.5. Benchmarking and Practical Application of Grid Program

Exact benchmarking and comparison of speed for the Grid, DRC and XLfit programs is difficult. For processing 500 20 data points dose-response curves, the Grid program took about 200 seconds on Mac Pro with 3GHz Intel Xeon micro processor using the default settings. The 200 seconds include outlier detection, fitting, statistical evaluation, and possible refitting. On the same machine, DRC took 470 seconds. But this comparison is not fair: In the R script for fitting the simulated data, the DRC library was loaded for each dose-response curve; the loading process could take more time than the fitting itself. For XLfit, we imported data into Excel on a windows XP machine and used a macro to perform the curve fitting. Since invoking the macro is a manual process, exact timing is not practical. Excluding manual interactions, the speed of XLfit seems comparable to the Grid program or faster.

As demonstrated in analyzing large amount of dose-response data at NCGC, the Grid program is very productive, robust, and scalable, and it routinely processes hundreds of thousands of dose-response curves with minimum manual intervention. The Grid program has a prescreening module to detect and skip those biologically insignificant curves. For 100,000 7-point dose-response curves, the Grid program is able to accomplish curve fitting, statistical testing, and database transaction within one CPU hour on Linux workstation with 3GHz Intel multi-core Xeon microprocessor. The Grid program supports multiple threads, and it is scalable to millions of dose-response curves.

4. CONCLUSION

The Grid program is robust, automated, and scalable for high throughput analysis of large amount of dose-response data, it consistently reproduces the actual values for both complete and partial curves with or without noise, and it will help in speeding drug development and reducing the cost. For very large data set, it can be indispensible. For data set of small and medium size, the Grid program will enhance the productivity and avoid human errors. The Grid algorithm is also applicable to many other curve fitting problems in biological and medical sciences.

REFERENCES

[1] Liu B, Li S, Hu J. Technological advances in high-throughput screening. Am J Phamacogenomics 2004; 4: 263-76.

[2] Inglese J, Auld DS, Jadhav A, *et al.* Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc Natl Acad Sci USA 2006; 103: 11473-8.

[3] Motulsky HJ. Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting. USA: Oxford University Press, 2004.

[4] Motulsky HJ, Ransnas LA. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. FASEB J 1987; 1: 365-74.

[5] Levenberg K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. Q Appl Math 1994; 2: 164-8.

[6] Ritz C, Streibig JC. Bioassay analysis using R. J Stat Softw 2005; 12: 1-22.

[7] R Development Core Team. R: A language and environment for statistical com- puting. R Foundation for Statistical Computing, Vienna, Austria 2005. URL http://www. R-project.org. ISBN 3-900051-07-0.

[8] Hill AV. The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. J Physiol 1910; 40: 4-7.

Received: July 13, 2010 Revised: August 04, 2010 Accepted: August 30, 2010