BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2921–2923
doi:10.1093/bioinformatics/btr489

*Data and text mining*

Advance Access publication August 24, 2011

# SDRS—an algorithm for analyzing large-scale dose–response data

Rui-Ru Ji[1,*], Nathan O. Siemers[1], Ming Lei[2], Liang Schweizer[2] and Robert E. Bruccoleri[3,*]

[1]Department of Applied Genomics, [2]Department of Lead Evaluation and Mechanistic Chemistry, Bristol-Myers Squibb, Princeton, NJ 08543 and [3]Congenomics, PO Box 1422, Glastonbury, CT 06033, USA

Associate Editor: Jonathan Wren

## ABSTRACT

**Summary:** Dose–response information is critical to understanding drug effects, yet analytical methods for dose–response assays cannot cope with the dimensionality of large-scale screening data such as the microarray profiling data. To overcome this limitation, we developed and implemented the Sigmoidal Dose Response Search (SDRS) algorithm, a grid search-based method designed to handle large-scale dose–response data. This method not only calculates the pharmacological parameters for every assay, but also provides built-in statistic that enables downstream systematic analyses, such as characterizing dose response at the transcriptome level.

**Availability:** Bio::SDRS is freely available from CPAN (www.cpan.org).

**Contacts:** ruiruji@gmail.com; bruc@acm.org

**Supplementary Information:** Supplementary data is available at *Bioinformatics* online.

Received on April 7, 2011; revised on August 18, 2011; accepted on August 19, 2011

## 1 INTRODUCTION

Dose–response assays are routinely used in today's pharmaceutical development. Mechanistically, compound:target binding occurs at a single site and follows the law of mass action that is reflected by the sigmoidal dose–response pattern seen in many assays (Balakrishnan, 1991). In statistics, sigmoidal dose-responses can be identified by non-linear regression, a form of regression analysis where the model function is a non-linear combination of the model parameters (Seber and Wild, 1989). Non-linear regression methods such as the well-known Levenberg–Marquardt algorithm involve successive approximations that aim at minimizing an error function (Marquardt, 1963). Despite the general applicability of the iterative non-linear regression methods, there are a couple of limitations in their application to large-scale dose–response screening data. First, the iterative methods do not impose a boundary on the model parameter values, and thus the output model may contain unrealistic or uninterpretable values such as a negative EC50. Second, these methods only calculate the parameter values and fitting statistic for the best model, but do not provide a means that can be integrated in downstream analyses such as the characterization of transcriptome response (Ji *et al.,* 2009).

*To whom correspondence should be addressed.

Ji *et al.* (2009) recently described a grid search-based algorithm, Sigmoidal Dose Response Search (SDRS), for identifying transcripts that exhibited sigmoidal dose-response to the treatments of kinase inhibitors. Since the SDRS algorithm is generic and can be expanded to identify other dose–response patterns in different sources of quantitative data, we have implemented the method as a Perl module with C inline codes (Bio::SDRS). We demonstrated the general utility of the method using a dataset from high content screening (HCS).

## 2 METHOD AND IMPLEMENTATION

Our implementation of the SDRS algorithm includes a typical sigmoidal dose–response model for one-site compound:target interaction,

$$Y = A + (B - A)/(1 + (X/C)^D)$$

where Y is the assay readout value, X is the dose and the four unknown parameters correspond to minimal response (A), maximal response (B), EC50 (C) and the Hill slope (D).

In essence, the SDRS algorithm tests a series of candidate EC50 values (i.e. search doses) across the experimental dose range. Therefore, at every search dose it is a three point grid search for the one-site model. For transcription profiling data, every probeset on the array is treated as an independent assay for the response of its corresponding transcript and its expression values at the experimental doses constitute the assay data.

We assume that every assay generates a positive readout. For every assay, the range for the parameter A is determined based on the six (default, or per user defined) lowest readout values, and is set to be the mean value plus or minus two multiples of the standard deviation (SD). If the lower boundary is less than zero, it is reset to the minimal of the readouts. The search step for A is one-fourth of the SD. Similarly, the range for the parameter B is determined using the six (default, or per user defined) highest values and the step is also one-fourth of the SD. The parameter D can vary between $-6.3$ and 6.3, with a step of 0.3. (In reality, D can vary from $-\infty$ to $+\infty$. However, when the absolute value of D is $>6$, additional increments have only marginal impact on the estimates of the other three parameters.) Placing data-driven limits on parameter values allows SDRS to exclude unusable parameters such as negative EC50 values.

At every search dose, the SDRS algorithm evaluates all possible combinations of parameter values and calculates the deviation

**Table 1.** Summary of SDRS and XLfit outputs

| Compound | Assay | SDRS output | | | | | XLfit output | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *P*-value | A | B | C (EC50, nM) | D | Fitted? | A | B | C (EC50, nM) | D |
| Compound1 | Caspase 3 | 7.6E-08 | 4.4 | 92.6 | 9068.3 | −6 | Ok | 3.7 | 113.6 | 11991.5 | −2.2 |
| Compound1[a] | Caspase 8 | 5.3E-11 | 2.6 | 100.9 | 1528.3 | −1.8 | Ok | 2.5 | 102.5 | 1570.5 | −1.7 |
| Compound1[b] | Cytochrome C | 1.7E-04 | 11.5 | 88.4 | 9688.3 | −6 | NoFit | | | | |
| Compound2 | Caspase 3 | 5.7E-03 | 1.8 | 3.2 | 1608.3 | −1.5 | Ok | 1.8 | 3.2 | 1775.7 | −1.6 |
| Compound2 | Caspase 8 | 8.1E-03 | 2.1 | 4.2 | 1588.3 | −6 | Ok | 2.1 | 4.2 | 1515.9 | −52.0 |
| Compound2 | Cytochrome C | 5.2E-01 | | | | | NoFit | | | | |
| Compound3 | Caspase 3 | 9.2E-02 | | | | | NoFit | | | | |
| Compound3 | Caspase 8 | 5.7E-06 | 1.8 | 5.3 | 1568.3 | −1.8 | Ok | 1.8 | 5.4 | 1628.1 | −1.5 |
| Compound3 | Cytochrome C | 2.9E-05 | 6.7 | 31.0 | 5108.3 | −1.2 | Ok | 6.5 | 40.1 | 10833.1 | −0.9 |
| Compound4[c] | Caspase 3 | 9.9E-07 | 1.7 | 47.4 | 13268.3 | −6 | Ok | 1.9 | 756.0 | 44125.9 | −4.7 |
| Compound4[b] | Caspase 8 | 1.1E-06 | 1.7 | 89.4 | 13308.3 | −6 | NoFit | | | | |
| Compound4[b] | Cytochrome C | 2.7E-06 | 5.1 | 79.0 | 13608.3 | −6 | NoFit | | | | |
| Compound5 | Caspase 3 | 5.3E-04 | 1.6 | 6.9 | 548.3 | −6 | Ok | 1.6 | 6.5 | 371.4 | −20.5 |
| Compound5 | Caspase 8 | 1.3E-03 | 1.7 | 6.9 | 748.3 | −3.3 | Ok | 1.6 | 7.2 | 847.0 | −1.9 |
| Compound5[a] | Cytochrome C | 1.1E-07 | 3.8 | 67.6 | 1288.3 | −1.5 | Ok | 4.1 | 66.4 | 1241.8 | −1.7 |
| Compound6 | Caspase 3 | 6.6E-01 | | | | | NoFit | | | | |
| Compound6 | Caspase 8 | 3.7E-01 | | | | | NoFit | | | | |
| Compound6 | Cytochrome C | 9.8E-02 | | | | | NoFit | | | | |
| Compound7 | Caspase 3 | 1.2E-02 | 2.2 | 5.2 | 5768.3 | −1.5 | Ok | 2.2 | 7.3 | 17804.1 | −0.8 |
| Compound7 | Caspase 8 | 1.4E-02 | 1.8 | 4.7 | 3468.3 | −1.2 | Ok | 1.8 | 5.8 | 8804.0 | −0.9 |
| Compound7 | Cytochrome C | 7.4E-03 | 6.7 | 16.6 | 1028.3 | −6 | Ok | 6.7 | 16.7 | 1001.0 | −8.4 |
| Compound8 | Caspase 3 | 2.8E-02 | 1.6 | 2.8 | 1108.3 | −6 | Ok | 1.6 | 6.2 | 54354.5 | −0.7 |
| Compound8 | Caspase 8 | 8.7E-01 | | | | | NoFit | | | | |
| Compound8[c] | Cytochrome C | 1.0E-05 | 7.3 | 56.8 | 10828.3 | −6 | Ok | 6.8 | 4099.0 | 475017.8 | −1.5 |
| Compound9[a] | Caspase 3 | 2.3E-10 | 2.7 | 100.8 | 388.8 | −2.4 | Ok | 3.1 | 99.3 | 390.5 | −2.6 |
| Compound9[a] | Caspase 8 | 3.8E-12 | 1.6 | 99.0 | 77.7 | −3.6 | Ok | 1.7 | 99.6 | 76.6 | −3.3 |
| Compound9[a] | Cytochrome C | 4.4E-08 | 8.2 | 94.9 | 1848.3 | −2.1 | Ok | 8.2 | 97.9 | 1946.7 | −1.9 |
| Compound10 | Caspase 3 | 6.2E-01 | | | | | NoFit | | | | |
| Compound10 | Caspase 8 | 6.9E-01 | | | | | NoFit | | | | |
| Compound10 | Cytochrome C | 4.2E-04 | 8.7 | 27.0 | 9208.3 | −6 | Ok | 8.6 | 39.2 | 18268.6 | −1.6 |

[a]Representative dose responses identified by both SDRS and XLfit, shown in Supplementary Figure S2.
[b]Dose responses identified by SDRS but not XLfit, shown in Supplementary Figure S1.
[c]Dose responses where XLfit generated EC50 values larger than the highest experimental dose, shown in Supplementary Figure S3.

of expected values based on the dose–response model from the observational data. The goodness of fit is measured by an $F$-statistic: $F = \frac{\text{MSR}}{\text{MSE}}$, where MSR is the mean square of the variance explained by the model and MSE is the mean square of error (Supplementary Table S1). Assuming that the residuals are normally distributed, the $F$-statistic follows an $F$-distribution, $F(p-1,\ n-p)$, where $n$ is the number of experimental dose points and $p$ is the number of parameters in the model. For every assay, at every search dose tested, the (local) maximal $F$-statistic and the corresponding parameter values are recorded.

At the end of the grid search, every assay is associated with a series of $F$-statistic. An assay is designated as fitted to a dose–response model if its global maximal $F$-statistic (i.e. best fit) is larger than a predefined critical $F$ value, for example, at $P < 0.05$. For each assay, the parameter values that gave rise to the global maximal $F$-statistic define the optimal model. The 95% confidence interval for C (i.e. EC50) is defined as from the lowest search dose where the local maximal $F$-statistic is larger than the critical value to the highest search dose that meets the same criteria. Confidence intervals for other model parameters can be found similarly.

One output of SDRS is qualitatively similar to that of an iterative algorithm: each assay is associated with a predicted EC50, $P$-value and fold-change (i.e. the ratio of B to A). However, SDRS also generates an $F$-statistic for every assay at each search dose. This output, which is unique to the grid search method, allows for a global characterization and comparison of dose responses (Ji *et al.*, 2009). For example, the $F$ scores at a search dose can be fed to a multiple test correction procedure (such as FDR) to calculate the number of 'true responses' at the dose. Repeating this procedure for every search dose across the dose range can uncover peak(s) of response. The $F$ score output also allows for pathway impact mapping and dose–response comparison at the transcriptome level across the dose range.

## 3 RESULTS AND DISCUSSION

Herein, we present the SDRS algorithm, which is implemented as a Perl module with C inline codes (Bio::SDRS). We applied the algorithm to a dataset from HCS assays that measured programmed cell death using caspase 3, caspase 8 and cytochrome C as readouts

in the ovarian cancer cell line, OVCAR-4. The SDRS outputs were compared with those generated by XLfit, a software that implements the Levenberg–Marquardt algorithm (Table 1). XLfit identified 19 dose responses in these assays. In contrast, SDRS identified three dose responses in addition to those identified by XLfit. The three additional dose responses identified by SDRS appear to be real (Supplementary Figure S1). There is a gradual increase in cytochrome C readouts as the Compound1 concentration increases. In the case of Compound4, it is likely that the compound also has a dose response since both the caspase 8 and cytochrome C assay produced high readouts at the highest dose. When both response plateaus are present, the parameter values generated by SDRS are almost identical to those generated by the iterative method (Table 1 and Supplementary Figure S2). However, when one of the curve plateaus is not present, i.e. where A or B are not well defined, the output is dependent on the behavior of the algorithm utilized. For example, when the high plateau is missing, extreme values for B and C are generated, with C often larger than the maximal experimental dose (Table 1 and Supplementary Figure S3). Similarly, when the low plateau is missing, iterative methods may generate negative estimates for A and C. Although there is no 'right' solution in these cases, as the data are not sufficient for parameter estimation, SDRS generates more 'realistic' estimates because it imposes constraints on the parameter values based on assay data and experimental dose range (Table 1).

Although SDRS was initially developed to handle genomic scale transcriptional dose–response data, it can be used to analyze all other types of dose–response data from qPCR and lead evaluation where it performs as efficiently as iterative non-linear regression methods (Ji *et al.*, 2009, and Rui-Ru Ji). For large datasets, SDRS can be run in parallel very efficiently across a multicore system (Supplementary Table S2). SDRS is robust to the naturally occurring variability in large-scale screening data, where the assays are not necessarily 'optimized'. Importantly, only SDRS provides a full set of $F$-statistic across the dose range that can be utilized in downstream system level analyses and comparisons.

## ACKNOWLEDGEMENTS

We are very grateful to Michael G. Neubauer, Petra Ross-Macdonald and Karl-Heinz Ott, for their insightful comments and discussions.

*Funding*: Bristol-Myers Squibb, the present and past employer of the authors.

*Conflict of Interest*: none declared.

## REFERENCES

Balakrishnan,N. (1991) *Handbook of the Logistic Distribution.* CRC Press, Boca Raton, Florida, p. 601.

Ji,R. *et al.* (2009) Transcriptional profiling of the dose response: a more powerful approach for characterizing drug activities. *PLoS Comput. Biol.,* **5**, e1000512.

Marquardt,D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, **11**, 431–441.

Seber,G.A. and Wild,C.J. (1989) *Nonlinear Regression*. John Wiley and Sons, New York.