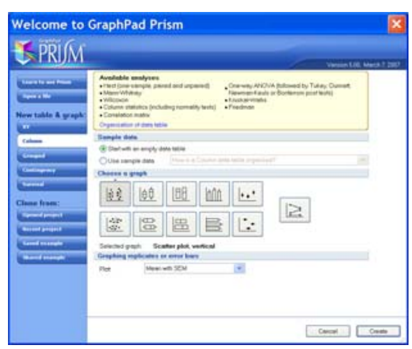
Curve-Fitting with GraphPad Prism

Presented by: Jeff Skinner, M.S.
Biostatistics Specialist
Bioinformatics and Computational Biosciences Branch
National Institute of Allergy and Infectious Diseases
Office of Cyber Infrastructure and Computational Biology

Bioinformatics for you...

GraphPad PRISM® at NIAID

- Free site license download
- Free training, consultation and support from BSIP ***
- Statistical results and help menus designed specifically for biological researchers
- Intuitive curve-fitting



*** Please contact:
ScienceApps@niaid.nih.gov

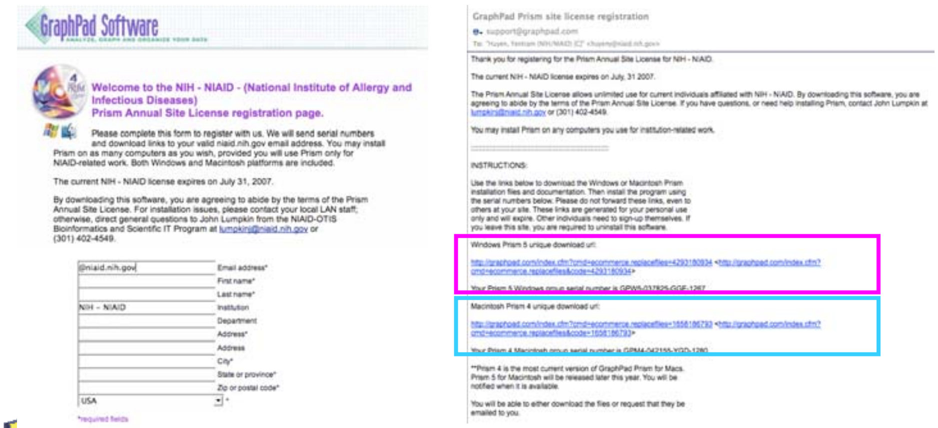
Bioinformatics for you...

BCBB

NIAID Site License

http://graphpad.com/paasl/index.cfm?sitecode=NIHNIAID

http://bioinformatics.niaid.nih.gov/

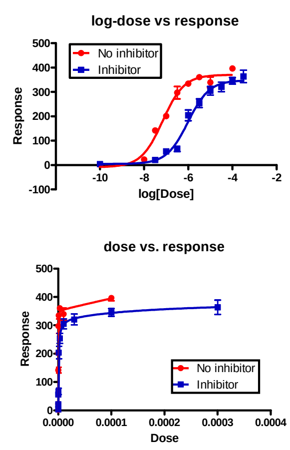


Bioinformatics for you...

BCBB

Nonlinear Regression

- Statistical technique commonly used for curve-fitting by biologists at NIAID
  - Dose-response studies
  - Protein binding experiments
- Nonlinear regression is rarely studied in statistics coursework
  - Sometimes mentioned briefly in a regression or linear models course
- Nonlinear regression is sometimes omitted from statistical software or poorly executed in software packages



Bioinformatics for you...

Linear Versus Nonlinear

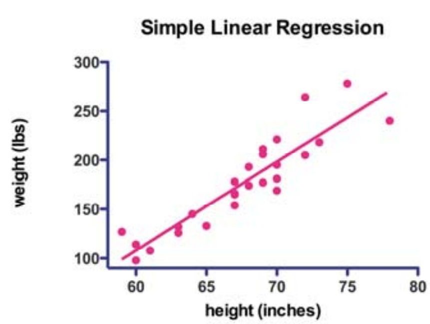
- There is an important distinction between a linear relationship and a linear statistical model
  - Linear or curved relationships describe how two variables might be visualized together on a graph
  - Linear models take the form $Y = mX + b$ or $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$
- Curved relationships can be fit with either linear or nonlinear statistical models
  - Polynomial regression is linear in its parameters
  - Sigmoidal dose-response models are not linear

Bioinformatics for you...

Simple Linear Regression

weight = -433.7 + 9.03 · height + error

- Linear relationship between a predictor, height, and a response variable, weight, is modeled by a straight line
  - Model: $Y = \beta_0 + \beta_1 X + \varepsilon$
  - Definition of a line: Y = mx + b
- The model is linear, so we can estimate slope and y-int using calculus, or ordinary least squares (OLS)

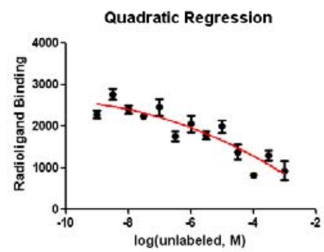


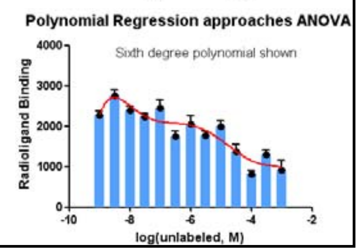
Bioinformatics for you...

Polynomial Regression

$$\text{binding} = \beta_0 + \beta_1 \cdot (\text{ligand}) + \beta_2 \cdot (\text{ligand}^2) + \ldots + \varepsilon$$

- Fits a curved relationship using a linear model
- Degree of polynomial indicates the number of curves
  - Quadratic (1 curve)
  - Cubic (2 curves)
- High degree polynomials are best fit by ANOVA models



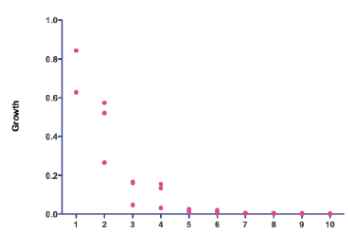


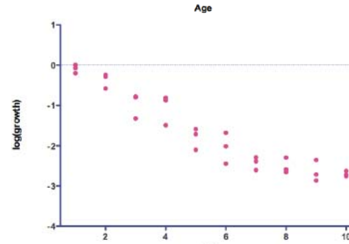
Bioinformatics for you...

Log-transformations

$$\text{growth} = \beta_0 \cdot \exp\{-\beta_1 \cdot \text{age}\} \cdot \varepsilon \quad \textbf{or} \quad \ln(\text{growth}) = \ln(\beta_0) - \beta_1 \cdot \text{age} + \ln(\varepsilon)$$

- Imagine growth modeled by an exponential equation with multiplicative error
  - I.e. errors get smaller as growth decreases with increasing age
- Log transformation is often used to "correct" a curved exponential relationship





Nonlinear Regression

- Describes a nonlinear relationship between predictor and response using nonlinear equations
  - E.g. $$\text{Response} = \text{Bottom} - \frac{\text{Top} - \text{Bottom}}{1 + 10^{(\log \text{EC50} - \log X)\cdot \text{Hillslope}}}$$
- Least squares equations do not have a closed form solution (i.e. cannot use calculus to find estimates)
- Estimates are found using iterative methods
  - Pick initial values and converge to solution step-by-step

Iterative Methods

- Linear descent method starts with an initial value, then "steps" away to minimize sums of squares (SS)
  - If first positive step increases SS, step backwards
  - If n^{th} step increases SS, take smaller steps
  - Linear descent is best for first steps from initial value
- Gauss-Newton uses matrix algebra to minimize SS, as if it were a change in "slope" from the initial value
  - Gauss-Newton is best for the last steps of minimization
- Levenberg-Marquardt method starts with linear descent and switches to Gauss-Newton

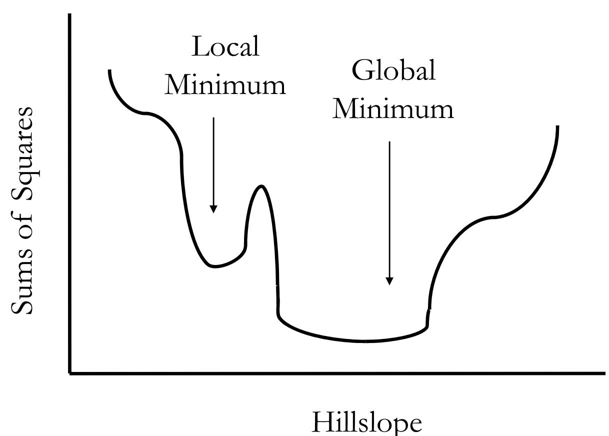
Bioinformatics for you...

Iterative Methods (continued)

- Strength of the iterative methods
  - Can be used to fit any model
- Weaknesses of the iterative method
  - Computationally slower than calculus method
  - May find the local min, instead of global min
  - May get “stuck” if model choice is poor

Bioinformatics for you...

Local vs. Global Minima

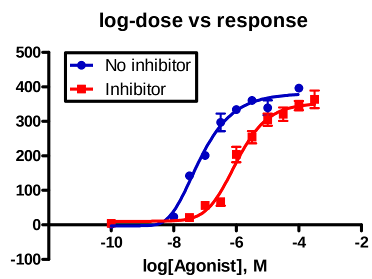


Finding a local minimum implies you have not found the best possible fit

Bioinformatics for you...

Dose-Response Models

- Most commonly used model among NIAID researchers
- Up to five parameters
  - TOP and BOTTOM
  - Log(EC50)
  - Hillslope (default = 1)
  - Symmetry (default = 1)
- Concepts presented for dose-response models will apply to all other nonlinear models

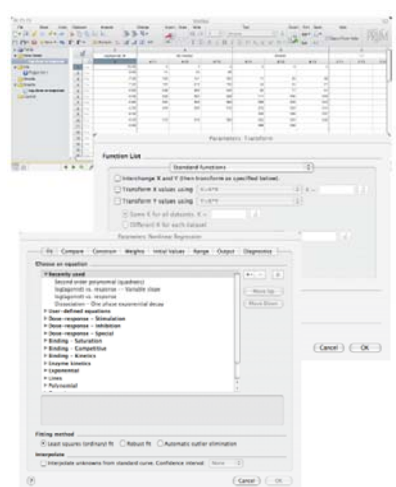


$$\text{Response} = \text{Bottom} - \frac{\text{Top} - \text{Bottom}}{1 + 10^{(\log \text{EC50} - \log X)\cdot \text{Hillslope}}}$$

Bioinformatics for you...

Getting Started in PRISM

- Open a XY data table and enter data
- Transform and normalize data if necessary
- Open the nonlinear fit menu in PRISM

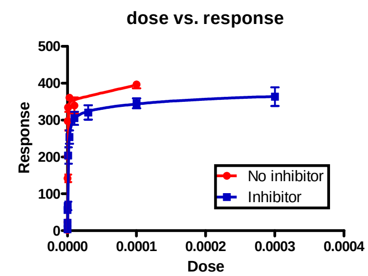


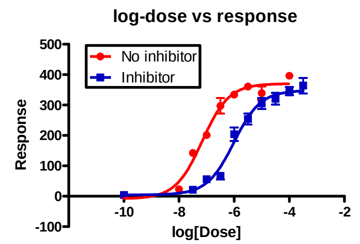
Bioinformatics for you...

BCBB

Why Use log(X) For Sigmoid Dose Response?

- Raw X values produce a hyperbolic curve
  - Large errors at the ascent
  - Small errors at the plateau
- Log(X) values produce a sigmoid “esse” curve
  - Moderate errors throughout
  - Best estimates of EC50

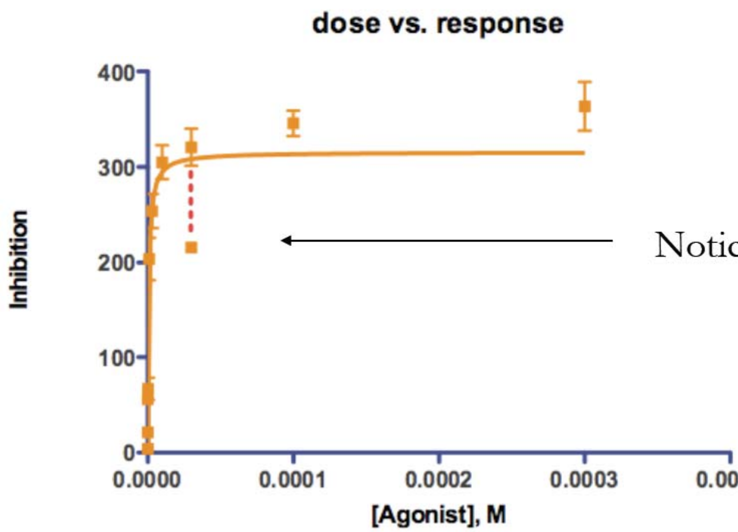




Bioinformatics for you...

BCBB

Importance of Log Transform

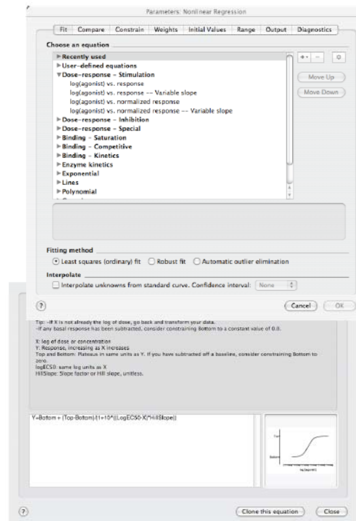


Notice large errors

Bioinformatics for you...

Choose a Model

- Choose an equation from the *Fit* tab of the nonlinear menu
- Equations are grouped by experiment types
  - E.g. dose-response, binding, …
- Click the details button to view the actual equation and a sample graph of the model



Bioinformatics for you...

Fitting Methods

- Use Least Squares (LS) whenever possible
  - LS produces the correct tests using all data
- Use Automatic Outlier Elimination or Robust Fit when outliers influence your results
  - Robust Fit uses all of your data, but cannot produce confidence intervals or tests
  - Automatic Outlier Detection removes data and applies least squares methods

Compare Groups or Models?

- Compare the fit of two different equations on the same data set(s)
  - Compare dose-response models with and without a variable slope parameter
  - Compare different sets of constraints, etc.
- Compare best fit parameters between two or more groups in your experiment
  - E.g. Compare EC50 between 2 or more groups

F-test or AIC?

- Use extra sums of squares F-test whenever the two equations are nested:
  - Dose-response vs Dose-Response (variable slope) is nested, because one equation is a simplification of the other equation
  - Dose-response versus binding equations is not nested, because the equations are unrelated
  - Comparing 2 or more groups is always nested
- Use AIC if you are unsure about nesting

Bioinformatics for you...

Statistical Output

1	Comparison of Fits		
2	Null hypothesis	log(agonist) vs. response	log
3	Alternative hypothesis	log(agonist) vs. response -- Variable slope	log
4	P value	0.4435	0.5
5	Conclusion (alpha = 0.05)	Do not reject null hypothesis	Do
6	Preferred model	log(agonist) vs. response	log
7	F (DFn, DFd)	0.6081 (1,23)	0.3

1	Comparison of Fits	
2	Null hypothesis	LogEC50 same for all data sets
3	Alternative hypothesis	LogEC50 different for each data set
4	P value	< 0.0001
5	Conclusion (alpha = 0.05)	Reject null hypothesis
6	Preferred model	LogEC50 different for each data set
7	F (DFn, DFd)	64.86 (1,48)

- PRISM provides you the null and alternative hypotheses, p-values and conclusions
- Interpretation of the results is often spelled out for you in plain language

Bioinformatics for you...

Model Diagnostics

- Goodness-of-fit tests are used to compare two equally valid models
- Normality tests evaluate the model assumption of normal errors
- Runs tests determine if errors are random
- Residual plots identify non-constant variance, correlated errors, etc.

Bioinformatics for you...

BCBB

Goodness of Fit Tests

- Coefficient of determination (R^2) is the percent of variation explained by the model
  - Remember R^2 is only meaningful if the model assumptions have been met
  - Typically used to compare valid models
- Sums of squares (SS) and standard deviation of the residuals ($s_{y.x}$) are more descriptive
  - Bigger SS implies model is more informative
  - Smaller $s_{y.x}$ implies model has smaller errors

Bioinformatics for you...

BCBB

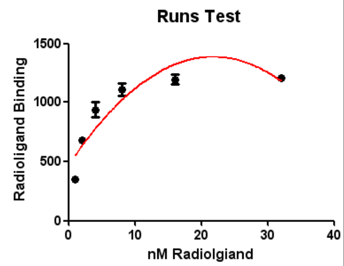
Normality Tests

- D'Agostino-Pearson Omnibus test
  - Tests the magnitude of skewness and kurtosis statistics to determine if errors are normal
- Shapiro-Wilk test
  - Uses ranks to compare errors to normal distribution
- Kolmogorov-Smirnov test (*historical use only*)
  - Tests the magnitude of the largest difference between two empirical cdf plots

Bioinformatics for you...

Runs Tests and Replicates Tests

- Does the curve systematically differ from observed points?
- Use replicates test for replicated curve-fit data
  - Compare error between points and curve to error among reps
- Use runs test for non-replicated curve-fit data
  - Tests the largest “run” of points above or below the curve

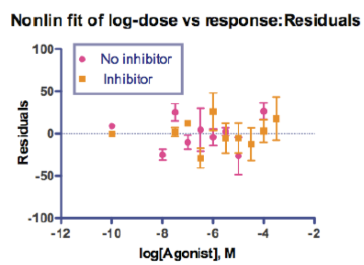


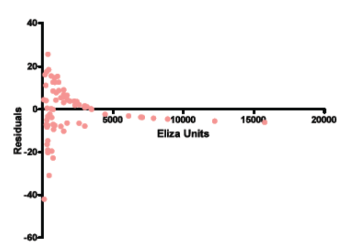
Notice the largest run of 3 points above the curve suggests model is incorrect

Bioinformatics for you...

Residual Plots

- Residuals are the errors between points and curve
- Want to find independent and identically distributed random normal errors
- Look for non-random trends and non-constant variance





Improving Model Fit

- Change the initial values to aid convergence
- Limit the range of X or Y to eliminate outliers
- Add constraints to force the model to meet certain criteria that improve model fit
  - E.g. Bottom = 0%, Top = 100%, …
- Use a weighting scheme to combat non-constant variance or emphasize parts of the curve

Bioinformatics for you...

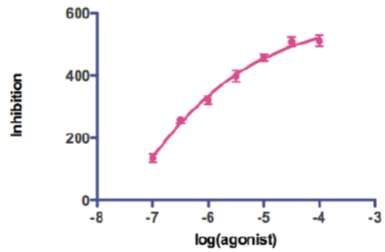
Initial Values and Range

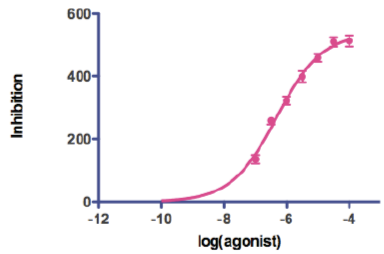
- Change the initial values of the parameters, if the model does not converge
  - Try the “Don’t Fit Curve” option on the diagnostics tab to find better initial values
- Select the range of X values used curve-fit to eliminate outliers or invalid data points
- Select the range of the fitted curve to complete the graphs and figures, if necessary

Bioinformatics for you...

Model Constraints

- Add hard or soft constraints to a model parameter
  - Bottom = 0
  - Top ≤ 100 %
- Improve model fit if doses are poorly chosen, etc.
  - Not enough low doses
  - Doses too low for saturation





Bioinformatics for you...

Weights

- Use 1 / SD^2 weights for non-constant variance
  - Requires many replicates per X
- Use other weight schemes to apply unequal weights to one part of the curve
  - Use 1 / Y^2 for larger errors at the highest responses
  - Use 1 / X to emphasize values on the left of the graph

Bioinformatics for you...

Summary

- Remember these same basic techniques can be applied to any nonlinear regression
- Contact ScienceApps@niaid.nih.gov for additional help if needed
- Additional PRISM training is available
  - Statistical Applications in Prism – April 8th, 1-3 pm