



NIAD

# Automating Biostatistics Workflows for Bench Scientists Using R-based Web-tools

Jeff Skinner, Vivek Gopalan, Jason Barnett  
and Yentram Huyen

*useR! 2010 Conference*

*July 21-23, 2010*

*Gaithersburg, MD*

Office of Cyber Infrastructure and Computational Biology  
National Institute of Allergy and Infectious Diseases  
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES  
National Institutes of Health

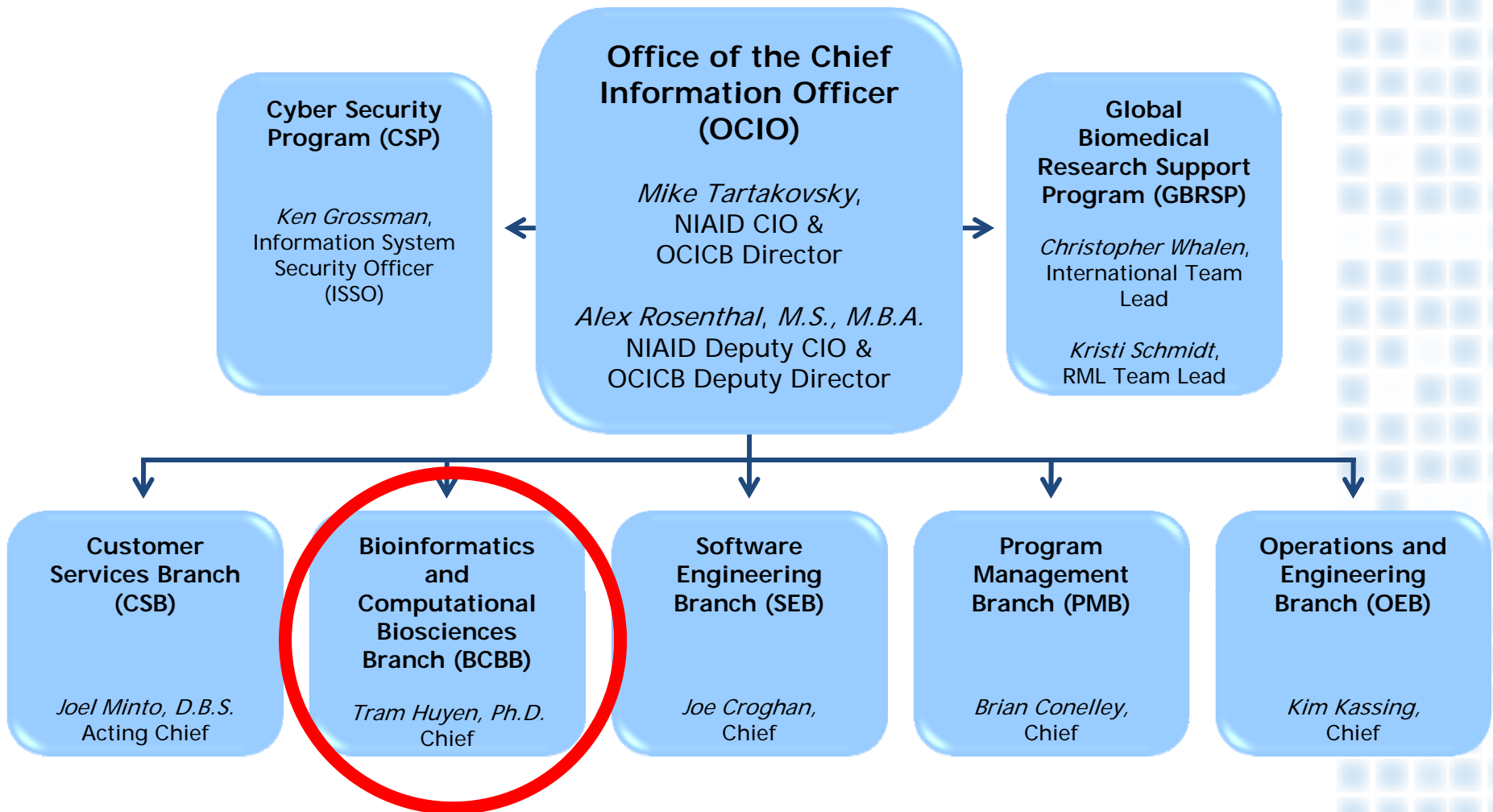


# NIAID Mission

- “The National Institute of Allergy and Infectious Diseases (NIAID) is one of the 27 Institutes of the NIH and conducts and supports basic and applied research to better understand, treat, and ultimately prevent infectious, immunologic, and allergic diseases.”



# OCICB Organization



# Commonly Encountered Problems

- Large complicated data files from biological instruments
  - Microarrays, Next-Generation Sequencing, 96-well plate readers, NMR and Mass Spectrometry
  - Arcane file extensions, ugly headers and footers, multiple tables per file
- Tedious data manipulation in MS Excel or Notepad
  - Simple formulas or cut-and-paste can add up to hours at the computer
- Critical analyses performed by legacy software
  - Many relevant software tools are no longer maintained, because they were created with outdated technology or the original developers have moved on to new careers



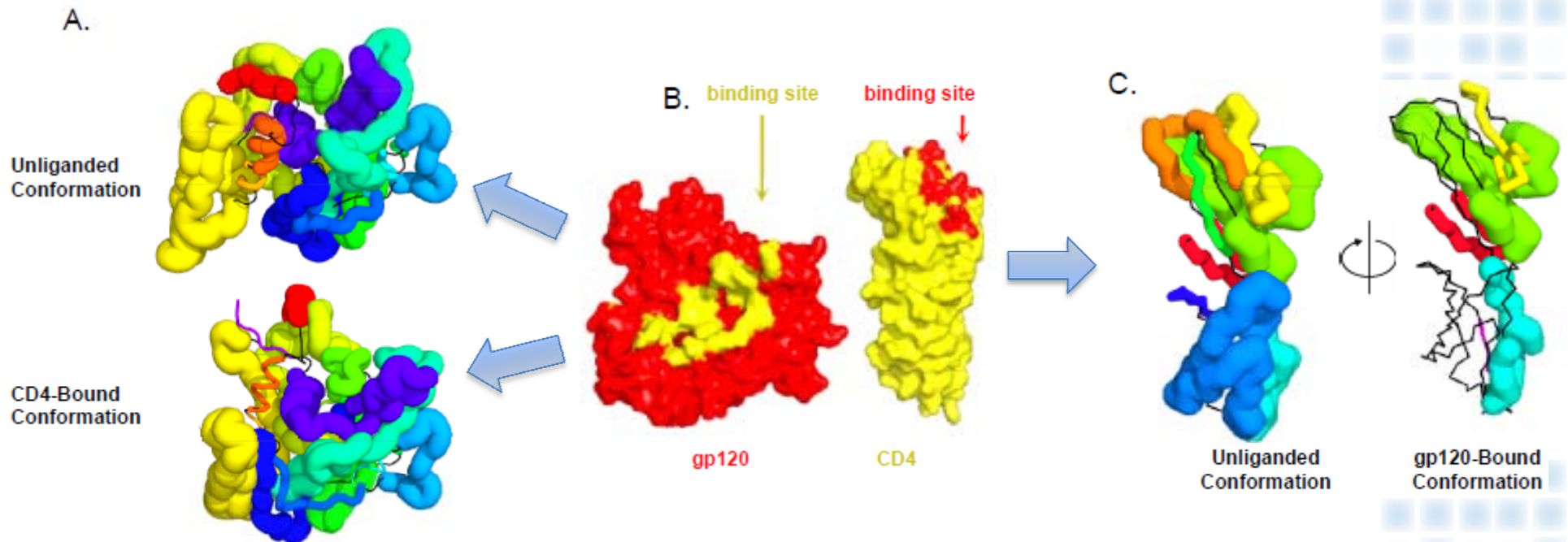
# Why Create a Webtool Using R?

- Advantages of using R
  - R scripting language is easy to use and will be long lived
  - Includes all necessary tools for data import, data manipulation, statistical analyses, graphing, generation of custom reports, etc.
- Advantages of building a webtool
  - Provides users an accessible graphical user interface (GUI)
  - Simplifies the distribution and maintenance of software
  - Agencies can link software to infrastructure resources like high performance computing clusters and databases, which may not otherwise be available to many end users





# HDX NAME

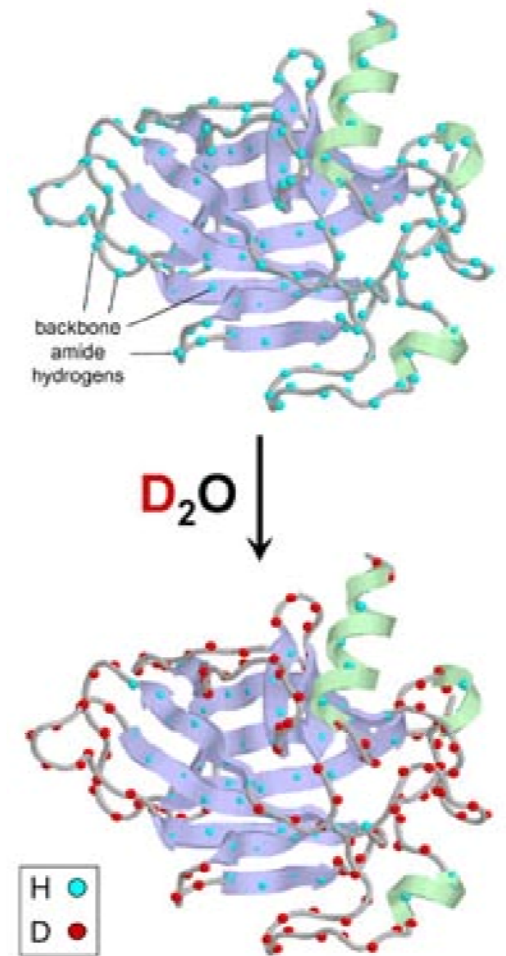


- Compute estimates of flexibility (i.e. protection factors) for multiple protein regions from hydrogen-deuterium exchange (HDX) data using the Maximum Entropy Method (MEM)
- Compare protection factors among two different groups
- Map protection factors on the protein surface



# Hydrogen-Deuterium Exchange

- Use changes in pH to force a protein to exchange hydrogen for deuterium
- Use nuclear magnetic resonance (NMR) spectrometry or mass spectrometry to detect the H/D exchange rates
- H/D exchange rates among different protein fragments reveal information about tertiary structure, folding, etc.



Source: [www.dac.neu.edu/barnett/Mem/engen.htm](http://www.dac.neu.edu/barnett/Mem/engen.htm)



# Maximum Entropy Methods

- Maximize function  $Q = S + \lambda C$  using LaGrange multipliers
- $S$  represents the Skilling entropy of HDX process

$$S = \sum_{k=k_1}^{k_2} f_k \left[ \ln \left( \frac{f_k}{A} \right) - 1 \right]$$

- $C$  represents the constraints on HDX imposed by the structure of the protein's tertiary structure

$$\chi^2 = \sum_i \frac{(D_i^{calc} - D_i^{exp})^2}{\sigma_i^2} \quad D_i^{calc} = N - \sum_{k=k_1}^{k_2} f_k \exp(-kt)$$





# HDX NAME Workflow

- Workflow inputs:
  - Protein structures (.pdb file): GP120 or CD4
  - Hydrogen exchange data (.txt file): fragment IDs and exchange rates
  - Additional data (.txt file): Temperature, pH, time series, replicates numbers, protein state (liganded or unliganded)
- Processing steps:
  - Compute number of deuterium exchanged per amide from the exchange rates, using differential equations for any liganded protein complexes
  - Normalize deuterium exchanged data for constant temperature and pH
  - Estimate average exchange rates using MEM (Laplace software)
  - Compute protection factors by normalization of average rates with intrinsic rates
  - Compute free energy from protection factors
  - Compare fragments from liganded and unliganded states with Student's T-tests
  - Map results to protein surface to explore conformational changes



# Development of Webtool

- Backend (Server)
  - Data import, processing and tests computed in R
    - HDXNAME : package library created for webtool
    - Bio3d : Extract sequences and structural properties from PDB files
    - Odesolve : Solving reaction kinetics for differential equations of liganded proteins
    - Rsolnp : Non-linear optimization tools for MEM computations
  - Perl used to visualize protein structures and run R from web
    - Bio::Perl : process fragment features from FASTA or PDB files
    - Bio::Structure : parse 3D coordinates of the protein structure
    - Bio::Graphics : generate 2D result images from the 3D structure
- Frontend (Client/Browser)
  - jQuery : Javascript library for AJAX implementation
  - Jmol : Browser plug-in to visualize results on protein structure



# HDX NAME Webtool

- Input Options
  - Structure data (FASTA or PDB)
  - HDX data (.txt from instrument)
  - Configuration file (.txt) stores user analysis and workflow settings
- Uploaded Files
  - List of all uploaded files
  - Buttons to run analyses
- Results
  - Displays jMol structure image
  - Displays protein sequence
  - Links to statistical result tables

The screenshot displays the HDX NAME Webtool interface. At the top, the title is "Hydrogen-Deuterium Exchange with Normalized Assessment of Maximum Entropy". Below this, the "Input Options" section contains three file upload fields: "Structure file - PDB format", "HDX data file - Tab delimited file", and "Configuration file - INI format". Each field has a "Browse..." button and a "Clear" button. There are also radio buttons for "Upload custom file" and "Use default file". At the bottom of the input section are "clear form" and "reset all" buttons, and a "Demo link".

The "Uploaded Files" section shows a session ID and a list of four files: "seq\_file", "struct\_file", "HDX\_file", and "params\_file", all with values of "<empty>". There are "run", "refresh", and "clear results" buttons.

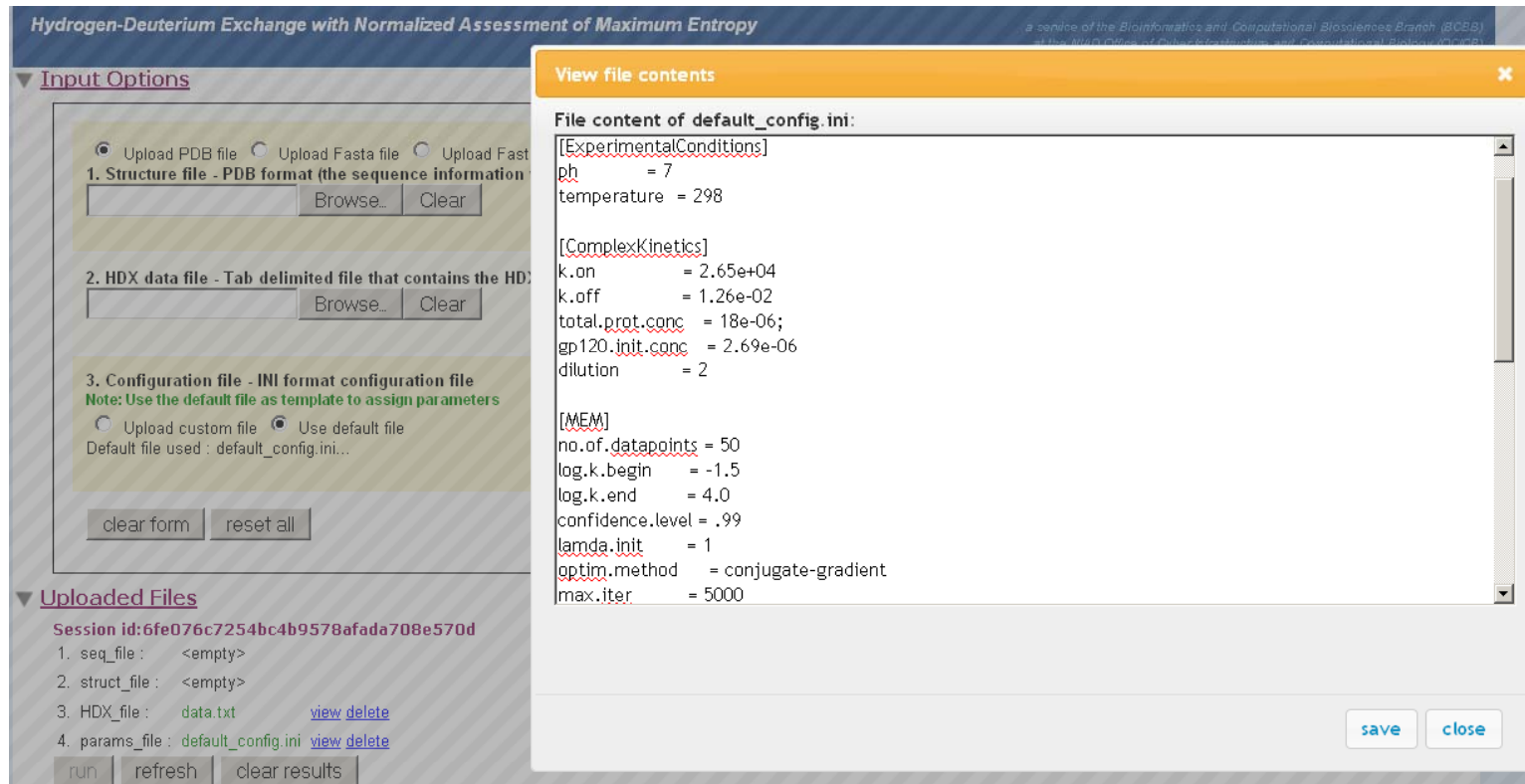
The "Results" section shows "Results refreshed." and a table of result files. The table has columns for "No.", "Process name", "Status", "Result file", and "Log file".

No.	Process name	Status	Result file	Log file
1.	HDX NAME workflow analysis	NOT_STARTED	1. features file 2. error1 file 3. error2 file 4. protection-factor file	log file

Below the table, it says "This table gets updated every 20 seconds." At the bottom of the page, there are navigation links: Home | Help | Accessibility | Privacy Policy | Disclaimers | Web Site Links & Policies | FOIA | Site Map | Contact Us, and logos for NIAID, NIH, HHS, and USA.gov.



# Configuration File



**Hydrogen-Deuterium Exchange with Normalized Assessment of Maximum Entropy**

**Input Options**

Upload PDB file  Upload Fasta file  Upload Fasta file

1. Structure file - PDB format (the sequence information)

2. HDX data file - Tab delimited file that contains the HDX data

3. Configuration file - INI format configuration file  
Note: Use the default file as template to assign parameters

Upload custom file  Use default file  
Default file used : default\_config.ini...

**Uploaded Files**

Session id: 6fe076c7254bc4b9578afada708e570d

1. seq\_file : <empty>

2. struct\_file : <empty>

3. HDX\_file : data.txt [view](#) [delete](#)

4. params\_file : default\_config.ini [view](#) [delete](#)

**View file contents**

File content of default\_config.ini:

```
[ExperimentalConditions]
ph = 7
temperature = 298

[ComplexKinetics]
k.on = 2.65e+04
k.off = 1.26e-02
total.prot.conc = 18e-06;
gp120.init.conc = 2.69e-06
dilution = 2

[MEM]
no.of.datapoints = 50
log.k.begin = -1.5
log.k.end = 4.0
confidence.level = .99
lamda.init = 1
optim.method = conjugate-gradient
max.iter = 5000
```

- Configuration file stores constants and parameters for all analyses
- Users can modify default configuration file to customize analyses and store custom settings for future use



Results tables are accessible using web links in table

**Results refreshed..**

session id: 05402c53f9e36ba9caec7bb56e8a9f0d  
 Start time: 2010-07-16 09:14:39  
 Current time: 2010-07-16 09:55:12

**Result files(s)**

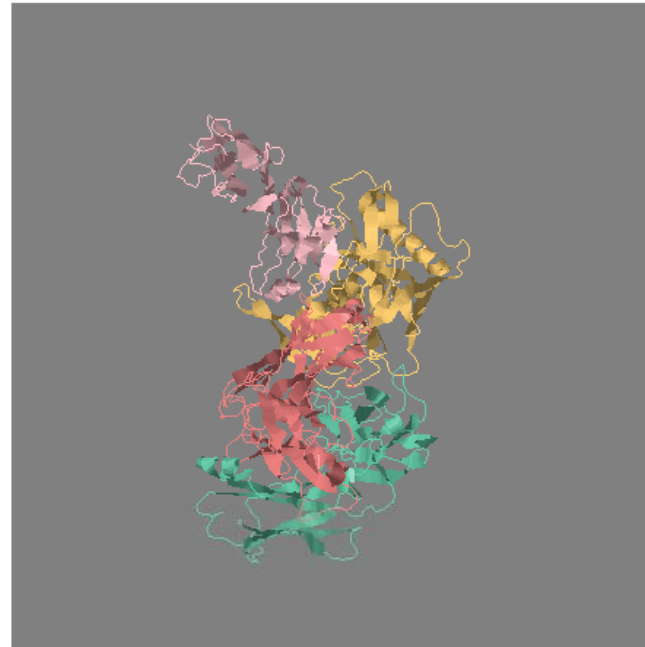
No.	Process name	Status	Result file	Log file
1.	HDX NAME workflow analysis	FINISHED	1. <a href="#">features file</a> 2. <a href="#">error1 file</a> 3. <a href="#">error2 file</a> 4. <a href="#">protection-factor file</a>	<a href="#">log file</a>

Finished running all the steps

jMol plug-in provides interactive 3D image of protein structure

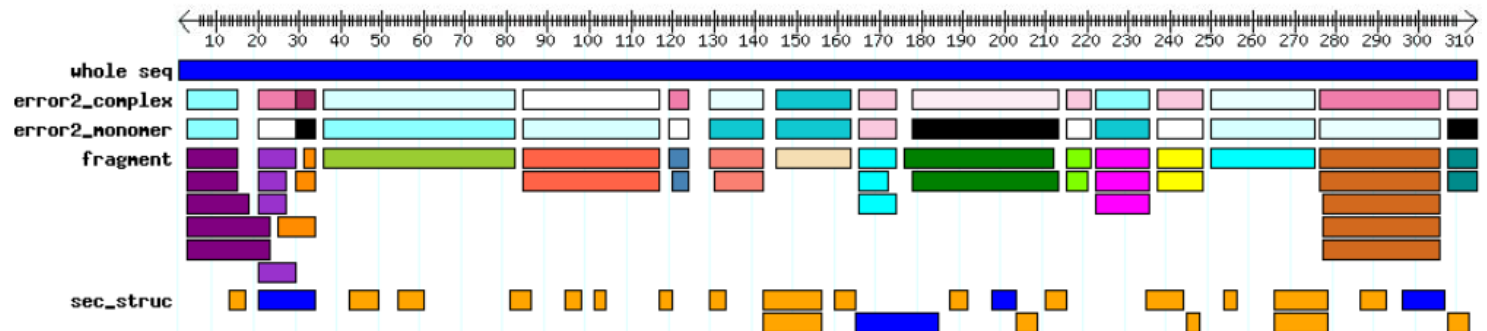
Image can be rotated by point-and-click

Links allow users to zoom, change colors or animate figure



- Selection [all](#) [Input chain:G](#) [center selection](#) [restrict selection](#)
- Display [cartoon](#) [trace](#) [wireframe](#) [spacefill](#) [ball & stick](#) [dots \(slow\)](#) [iso surface \(slow\)](#)
- Color [chains](#) [secondary structure](#) [amino acid](#)
- Background [black](#) [grey](#) [custom](#) [white](#)
- Animate [animate](#)
- View [zoom +](#) [zoom -](#) [reset](#)
- Viewer [size +](#) [size -](#) [reset](#)

Fragment lengths, sec structure and errors mapped on protein sequence



# Dose-Response Analysis Pipeline (DRAP)

- Need to fit logistic dose-response curves to data from dozens or hundreds of 96-well plates
  - Plates can be organized in countless ways
  - One factor per plate or multiple factors
  - Dilutions on columns or rows
- Want to view the curve-fits and export summary statistics
  - Want to compare EC50s with statistical tests
  - Want to export EC90s for use in QTL analyses

The screenshot shows the Dose-Response Analysis Pipeline (DRAP) web interface. At the top, there is a blue header with the text "Dose-Response Analysis Pipeline (Beta version)" and two buttons: "BCBB Home" and "Support". Below the header, there is an "Input" section with a text box for "Response files in ZIP format" and a "Browse..." button. The "Workflow Options" section includes a checkbox for "Autorun" and buttons for "Stop", "Reset", and "Parameters". The "Workflow Status" section displays a progress bar with icons for "Not Started", "In Progress", "Error", "Complete", and "Disabled". Below this, a list of workflow steps is shown, each with a "run" button and a status indicator (blue circle for "Not Started", green circle for "In Progress", red circle for "Error", green checkmark for "Complete", and grey triangle for "Disabled"). The steps are: "Upload Zip File", "UnZip and Organize Files", "Process Response File Names", "Map Dosage to Response", "Read Dosage & Response Files", "Extract Experiment Design", "Process Dosage & Response Data", "Analyze Dosage-Responses", "Generate Report", and "Run QTL Analysis". At the bottom, there is a "Results Link" section with the text "Results (inactive)" and an "Information panel" at the very bottom.





# Logistic Dose-Response Curves

- Captures the “S” shape of many types of biological assays
  - Drug dose-response experiments
  - ELISA experiments
- Unknown model parameters are estimated using iterative Levenberg-Marquart methods
  - Top and Bottom parameters estimate maximum and minimum response
  - LogEC50 parameter estimates the location of the curve on X-axis
  - Hillslope parameter estimates rate of increase or decrease per unit X
- Slopes or EC50 estimates can be used to compare effectiveness of different vaccines, drugs, etc.

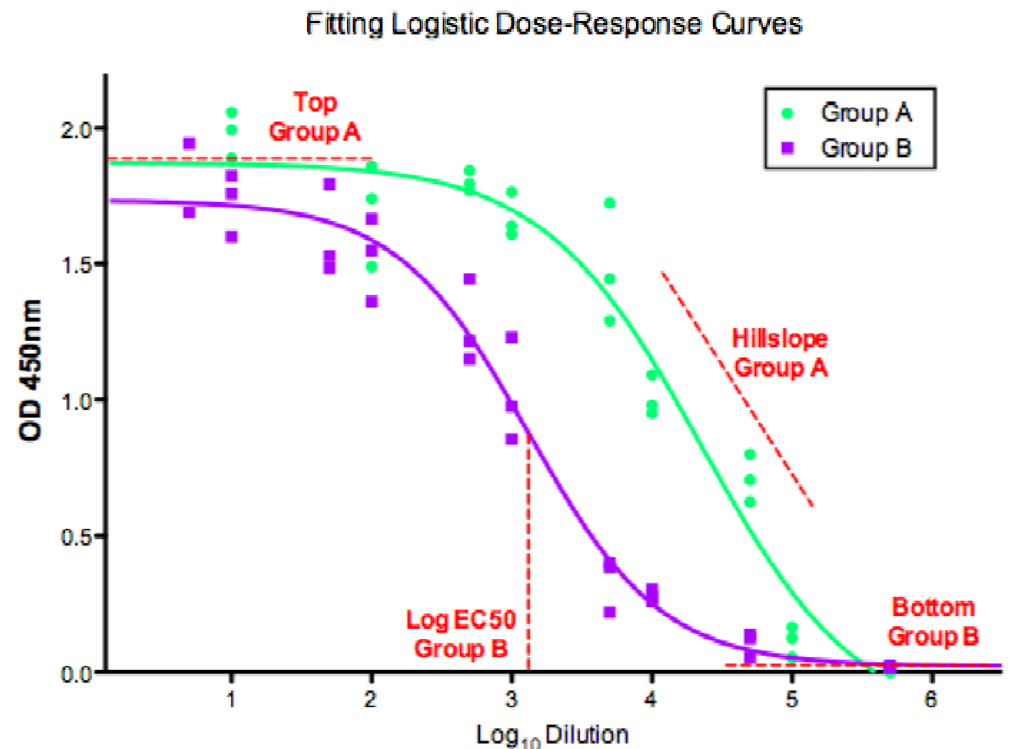


Image created using GraphPad Prism v. 5.03



# DRAP Workflow

- Data from 96-well plates (.dat files) processed in MS Excel
  - Remove headers and footers, record positive and negative controls
  - Identify data from multiple groups, noting that some groups may occur within a single plate while other groups occur between plates
- Logistic curve-fits computed in commercial Prism software
  - Data from each plate must be imported into Prism separately
  - Data need to be reorganized in Prism to create appropriate graphs and statistical tests, which may require data from multiple plates
- Summary statistics from Prism pasted into MS Excel or PowerPoint to summarize, reorganize and present results from multiple tests in a single report



# Development of Webtool

- Backend (Server)
  - All data processing and curve-fitting performed in R
    - drc : Core library to process dose reponse analysis
    - R2HTML : Generate HTML output
  - Perl CGI used to run R from the web
    - CGI::Application library for handling CGI requests
    - Methods to handle Workflow functionalities
- Frontend (Client)
  - Google Web Toolkit
    - Interactively build plate through web interface
    - Create all the widgets and controls in the web interface
    - Process JSON data from server and updates the widgets
    - User interface for CRUD operation on plate data.



# Dose-Response Analysis Pipeline (Beta version)

[BCBB Home](#) | [Support](#)

a resource of the *Bioinformatics and Computational Biosciences Branch* (BCB) at the *NIAD Office of Cyber Infrastructure and Computational Biology* (OCIB)

running map\_dosage...



Select zip file with input data

**Input**  
Response files in ZIP format

Buttons to start or stop analysis

**Workflow Options**  
 Autorun

Symbols display status of the workflow steps

**Workflow Status**  
● Not Started In Progress Error Complete Disabled

Log info links provide R info and diagnostics

- 2 Upload Zip File run [log info](#)
- 2 UnZip and Organize Files run [log info](#)
- 2 Process Response File Names run [log info](#)
- 2 Map Dosage to Response run [log info](#)
- 2 Read Dosage & Response Files run ●
- 2 Extract Experiment Design run ●
- 2 Process Dosage & Response Data run ●
- 2 Analyze Dosage-Responses run ●
- 2 Generate Report run ●
- 2 Run QTL Analysis run ▲

Info panel shows diagnostics and provides link to final results

Results Link [Results \(inactive\)](#)

### Information panel

File 'dbf\_format.zip' successfully uploaded...  
1 dosage file(s) 12 Response files present in the zip file  
1 dosage file(s) 12 Response files present in the zip file

User Manual     
Test files: [dbf\\_format.zip](#), [raw\\_fluorostar.zip](#)

User manual and sample data

**Dosage Files**

delete	#	Name
<input type="checkbox"/>	1	dosage.txt

**Response Files**

delete	#	Name
<input type="checkbox"/>	1	2C7-1-030108.dbf
<input type="checkbox"/>	2	2C7-1-24feb08.dbf
<input type="checkbox"/>	3	2C7-1-26feb08.dbf
<input type="checkbox"/>	4	2C7-1-28feb08.dbf
<input type="checkbox"/>	5	2C7-1-3mar08.dbf
<input type="checkbox"/>	6	2C7-1-5mar08.dbf
<input type="checkbox"/>	7	2C7-2-030108.dbf
<input type="checkbox"/>	8	2C7-2-24feb08.dbf
<input type="checkbox"/>	9	2C7-2-26feb08.dbf
<input type="checkbox"/>	10	2C7-2-28feb08.dbf
<input type="checkbox"/>	11	2C7-2-3mar08.dbf
<input type="checkbox"/>	12	2C7-2-5mar08.dbf

Browse and edit input data files

Rainbow icon for "dosage designer"

Long lists of assay response files are loaded interactively like Google Maps

Files can be edited in browser, then saved to computer

[Home](#) | [Help](#) | [Accessibility](#) | [Privacy Policy](#) | [Disclaimer](#) | [Web Site Links & Policies](#) | [FOIA](#) | [Site Map](#) | [Contact Us](#)



National Institute of Allergy and Infectious Diseases (NIAID)



National Institutes of Health (NIH)



Department of Health and Human Services (HHS)



NIAID Office of Cyber Infrastructure and Computational Biology



# Editing Files

The screenshot shows a software interface with a 'Parameters' tab and a 'Dosage Files' window. The 'Dosage Files' window contains a list of files with columns for 'Name' and 'delete #'. A dialog box titled 'File Name : dosage.txt' is open, showing a table with the following data:

well	drug	condition	sample	block
dosage				
AU1	CQ	-	1	125U.U
A02	CQ	-	1	625.0
A03	CQ	-	1	312.5
A04	CQ	-	1	156.3
A05	CQ	-	1	70.1
A06	CQ	-	1	39.1
A07	CQ	-	1	19.5
A08	CQ	-	1	9.8
A09	CQ	-	1	4.9
A10	CQ	-	1	2.4
A11	CQ	-	1	1.2
A12	CQ	-	1	0.0
B01	MDCQ	-	2	20000.0
B02	MDCQ	-	2	10000.0
B03	MDCQ	-	2	5000.0
B04	MDCQ	-	2	2500.0
B05	MDCQ	-	2	1250.0
B06	MDCQ	-	2	625.0
B07	MDCQ	-	2	312.5

The screenshot shows the 'Dosage Designer' interface. It features a grid of dosage data for different drugs (A-H) across 12 wells. The interface includes a 'Plate Editor' section with 'Selection' and 'Dose types' options.

Drug	1	2	3	4	5	6	7	8	9	10	11	12
A	CQ	CQ	CQ	CQ	CQ	CQ	CQ	CQ	CQ	CQ	CQ	CQ
B	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ	MDCQ
C	AQ	AQ	AQ	AQ	AQ	AQ	AQ	AQ	AQ	AQ	AQ	AQ
D	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ	MDAQ
E	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ	PPQ
F	QN	QN	QN	QN	QN	QN	QN	QN	QN	QN	QN	QN
G	MF	MF	MF	MF	MF	MF	MF	MF	MF	MF	MF	MF
H	ART	ART	ART	ART	ART	ART	ART	ART	ART	ART	ART	ART

- Users can click on “notepad” icons to edit and save dosage or response data in an interactive text file environment
- Dosage data can also be edited in the interactive Dosage Designer



# Interactive Results Report

## Results of DRA

[Click here](#) for results in CSV format 1 (for viewing in Excel)

[Click here](#) for results in CSV format 2 (for viewing in Excel)

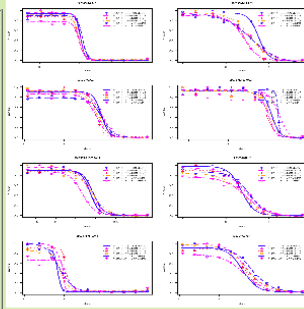
## Results Summary:

Sample name : 2C7

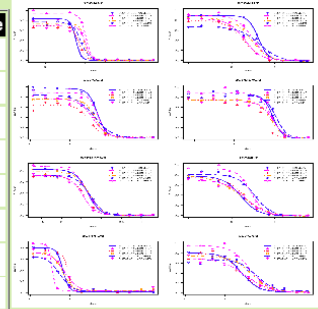
Name : Condition = 1

Name : Condition = 2

	IC50_mean	IC90_mean	HillSlope_mean	IC50_se	IC90_se	HillSlope_se
AQ	10.6	14.7	6.9	0.36	0.41	0.37
ART	3.7	11.3	2.4	0.45	2.11	0.40
CQ	76.2	150.6	3.5	5.04	14.37	0.42
MDAQ	135.0	224.7	6.1	16.96	29.93	2.22
MDCQ	727.9	1523.1	3.3	59.12	191.06	0.38
MF	2.8	8.5	2.2	0.17	1.10	0.23
PPQ	8.8	13.5	8.4	0.64	1.40	2.55
QN	48.1	154.3	1.9	3.92	15.17	0.11



	IC50_mean	IC90_mean	HillSlope_mean	IC50_se	IC90_se	HillSlope_se
AQ	8.6	12.6	6.1	0.78	1.07	0.497
ART	3.1	8.0	2.5	0.44	1.48	0.286
CQ	42.3	92.6	3.0	2.80	8.57	0.302
MDAQ	88.6	180.2	3.2	7.68	17.37	0.175
MDCQ	436.7	993.7	2.7	27.63	71.43	0.043
MF	2.3	6.8	2.1	0.28	0.78	0.221
PPQ	6.5	11.0	5.6	0.75	1.61	1.187
QN	39.1	125.0	1.9	5.53	15.78	0.098



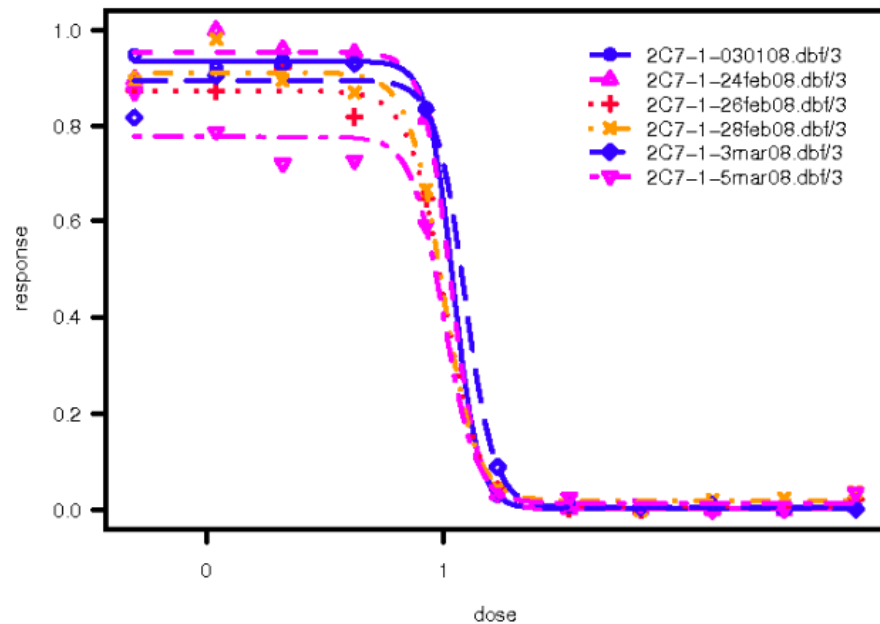
Generated on: Tue Jun 1 15:43:06 2010 - R2HTML

- Interactive report includes tables to organize results according to groups within plates (e.g. drugs) or between plates (e.g. condition)
- Click on logistic curve graphs for higher resolution images

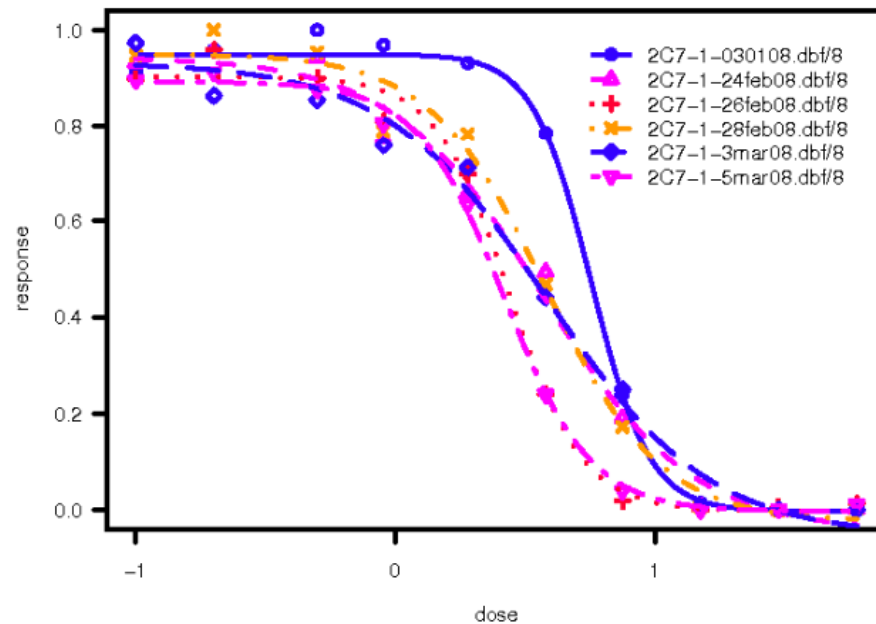




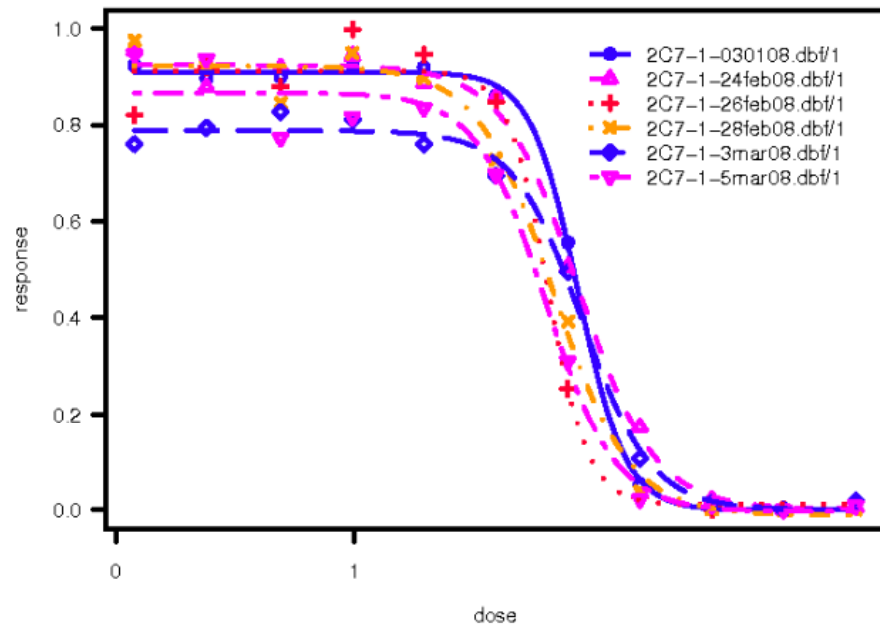
2C7/AQ/1



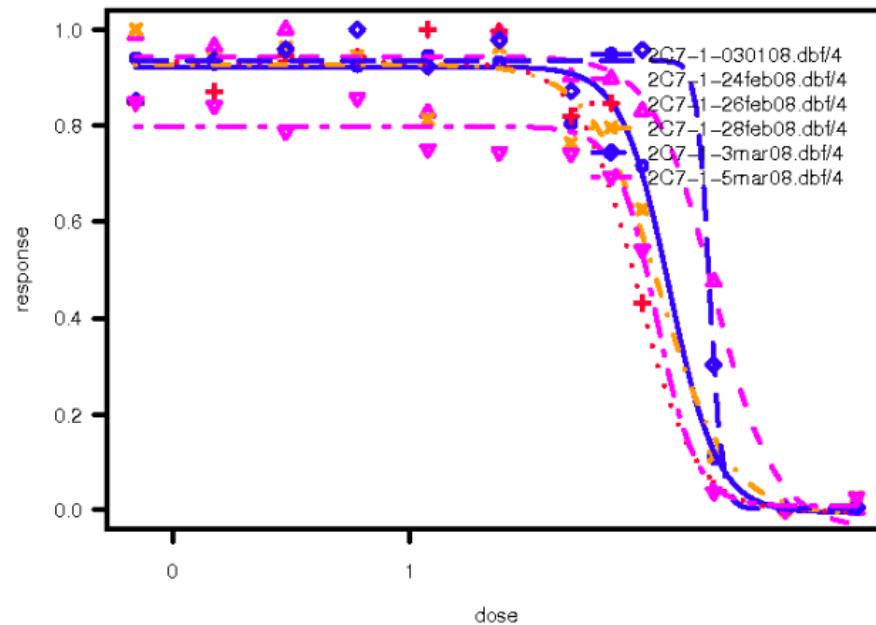
2C7/ART/1



2C7/CQ/1



2C7/MDAQ/1



# Acknowledgements

- Vivek Gopalan: R programming and web development
- Jason Barnett: Interface design on DRAP tool
- Tram Huyen and Mike Tartakovsky: Funding and oversight
- Leo Kong and Peter Kwong: HDX NAME experiments
- Juliana Sa, Olivia Twu, Hongying Jiang, Thomas Wellems and Xin-zhuan Su: DRAP experiments

Websites:

<http://exon.niaid.nih.gov>

<http://exon.niaid.nih.gov/drap/>

[http://exon.niaid.nih.gov/HDX NAME/](http://exon.niaid.nih.gov/HDX_NAME/)



# Literature Cited

- Kong *et al.* 2010. Hydrogen-deuterium exchange mass spectrometry of HIV-1 gp120 in unliganded and CD4-bound states. **J. Virology**. *in press*.
- Sa *et al.* 2009. Geographical patterns of *P. falciparum* drug resistance distinguished by differential responses to amodiaquine and chloroquine. **PNAS**. 106(45): 18883-18889
- Yuan *et al.* 2009. Genetic mapping of targets mediating differential chemical phenotypes in *P. falciparum*. **Nature Chemical Biology**. 5:765-771

