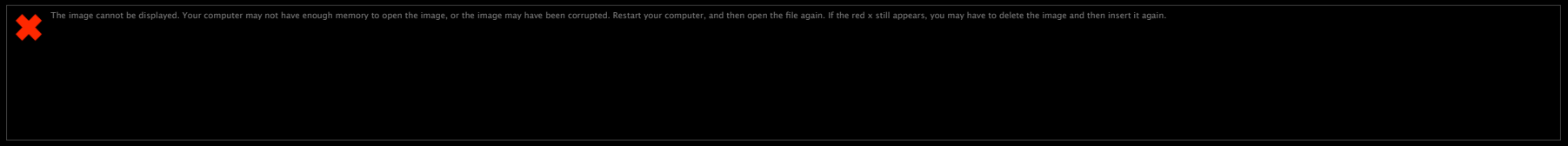
BIOISIS™ .net

Bio-SAXS database - Status & Workshop



John Tainer
Advanced Light Source
Berkeley CA
BL 12.3.1

http://bioisis.net

BIOISIS™ .net

Web-accessible database for storing SAS data and relevant analyses.

Conceived, motivated by needs of & named by John Tainer and Greg Hura

Programmed by Rob Rambo with help by Ivan Rodic using:

Web application created using the Ruby-on-Rails framework

Data storage is handled by MySQL open source database

Supported in part by:

DOE: Office of Science Integrated Diffraction Analysis Technologies

NIGMS: Macromolecular Insights Optimized by Scattering

Open Source Web Application Framework

Hulu
Twitter
FunnyorDie
Github
Groupon

Enforces good program practices such as:

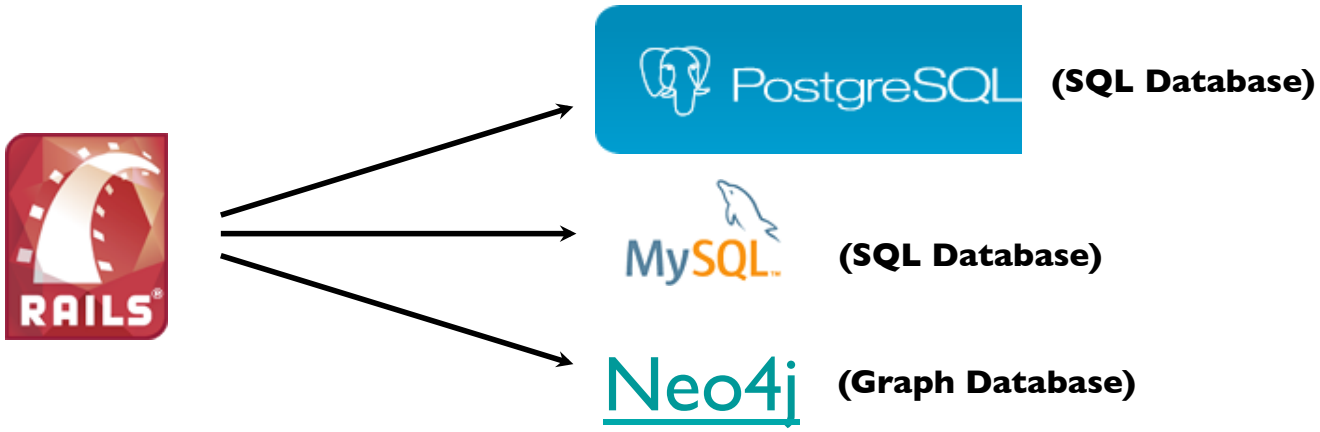
- **Model-View-Controller (MVC)**
- **Do not Repeat Yourself (DRY)**
- **Active Record (agnostic to database)**

Object Oriented:

- **database tables are treated as objects (Model)**
- **separates database logic from server logic from viewing logic**

BIOISIS™.net → basic operations

Create
Read
Update
Download



ActiveRecord is an object relational mapping (ORM)

- **makes web application independent of choice of database**
- **switch between databases single line of code**
- **choice and structure of database determined by complexity of queries**

Most of the time, search by keywords or by ID code (simple)

BIOISIS™.net

Database Considerations

At the time Bioisis was conceived and initiated:

CRYSOL was the only readily available SAXS calculator

- **Aqua-SAXS**
- **FAST-SAXS**
- **Zernicke-SAXS**
- **FOXS**

SAXS experiments were primarily used for:

- **bead modeling (DAMMIN/GASBOR/BUNCH)**
- **checking solution state of crystal structures**
- **Biophysical experiments including folding, flexibility. Rg**

For a given macromolecule, can have many SAXS experiments:

- **different concentrations**
- **different buffer conditions**
- **different temperature**

Want to capture the original data that supports the structural interpretation?

BIOISIS™.net

Database Schema

Table attributes define the attributes of an object (OO).

Table: Experiments

1. id
2. pH
3. monovalent
4. divalent
5. title
6. description
7. R_g, d_{max}, I(0), V_{Porod}
8. wavelength
9. addititves
10. angular range
11. source
12. P(r) filename
13. I(q) filename

- has_many genes

- has_many experiments
- belongs_to experiments

Table: Genes

ORFs from SSO, VNG, and PF

1. id
2. Locus Name
3. molecular weight
4. pI
5. Sequence
6. gi
7. uniprot
8. gene coordinates
9. organism id

- has_many genes

- has_one organism
- belongs_to organism

Table: Organisms

1. id
2. Name
3. Abbreviation

Example Code:

```
experiment = Experiment.find (101)          <-- finds experiment 101 in BIOISIS
experiment.genes.size                       <-- tells me the number of genes in 101

for gene in experiment.genes do |name|
   gene.locus_name                          <-- list the associate locus name
end
```

BIOISIS™.net

Creating A Submission

Submit an email address & a submission link is emailed

Save and Exit

Registered as: robert_p_rambo@hotmail.com

Basic Information

Experimental Details

SAXS Parameters

Add Protein | RNA | Polymer sequence

Contributors

Upload Data Files

Add Models

Navigates to different subsections

Lists all the missing information

STATUS OF DEPOSIT | INCOMPLETE

To begin the deposition, please click on the links above. As you complete each section, errors will drop from the list below.

Iofq file name Missing Required Field

Pofr file name Missing Required Field

Description Missing Required Field

Source location Missing Required Field

Io is not a number

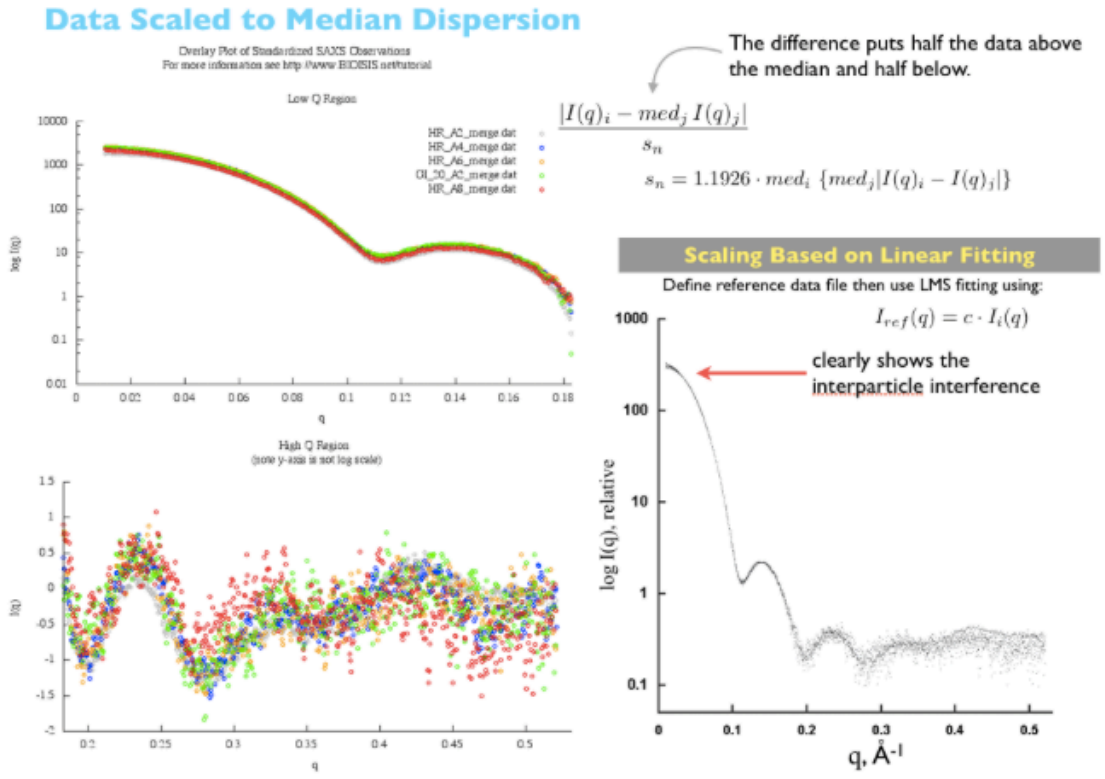
Io can't be blank

BIOISIS™ .net

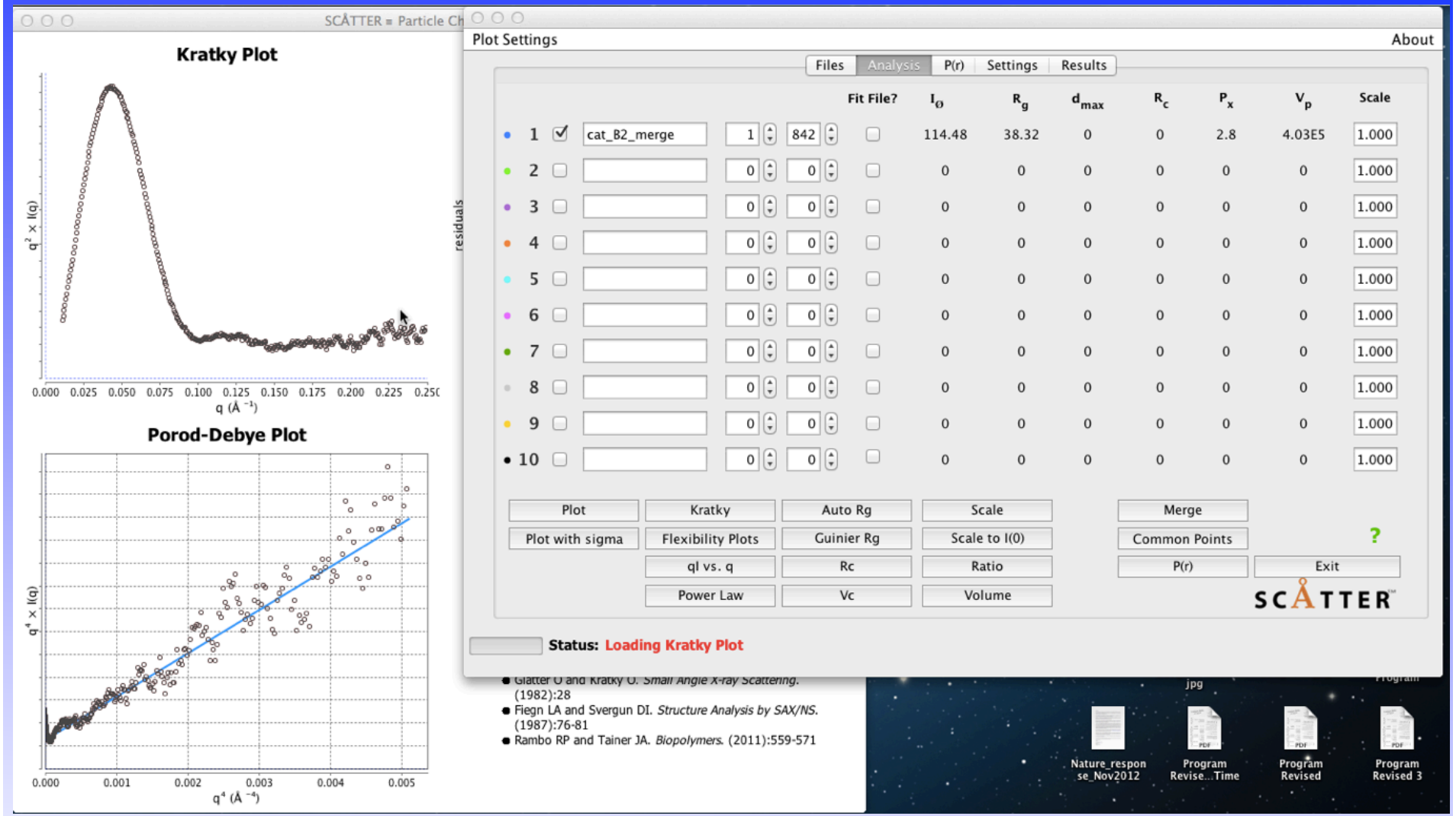
Tutorials are for SAXS analysis using:
•Scatter (Java-based application from BL12.3.1

scale invariant (Figure 1 left). One method for standardizing a dataset is to subtract each I(q) from the median I(q) value within a dataset and then divide by a scaling statistic. Since I favor the median for its robust properties, we need an alternative to the standard deviation and I used one defined by Rousseeux. Standardizing the dataset splits the dataset into halves centered at the median and can be used directly for model fitting. This is useful for examining features or trends at high q but requires each dataset in the collection to have the same number of points within the same q-range.

Figure 1:



ScÅtter

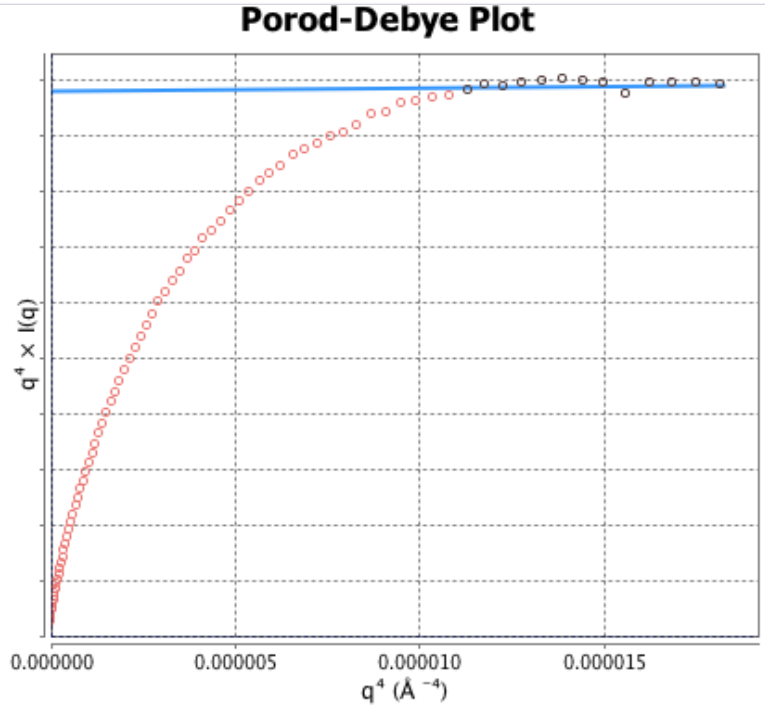


JAVA-based, co-developed on Mac & Windows & Linux
Performs data reduction, analysis and fitting
Released as open-source license free

BIOISIS™.net

ScÅtter

Catalase



Start	End	Q	V (Å³)	P
76	88	0.007	313244	4.0

Use arrows to manually define linear region in Porod Plot (left).

Fit determines the Porod Invariant, Q, and Exponent, P.

- Slope of the line must be positive
- Distribution of residuals should be unbiased
- Must be repeated if I(zero) changes

For more information, please see:

- Glatter O and Kratky O. *Small Angle X-ray Scattering*. (1982):28
- Fiegn LA and Svergun DI. *Structure Analysis by SAX/NS*. (1987):76-81
- Rambo RP and Tainer JA. *Biopolymers*. (2011):559-571

Theory matches Porod's law : get P = 4 for well folded particle.
Volume consistent with compact particle with density 1.37

Questions:

1.Can this be used to follow flexibility quantitatively?

2.Can volume changes be monitored for changes in state?

BIOISIS™.net

Uploading

I(q) files:

1. Files must be 3 column, space or tab delimited

P(r) files:

1. Files can be 3 column, space or tab delimited
2. GNOM file

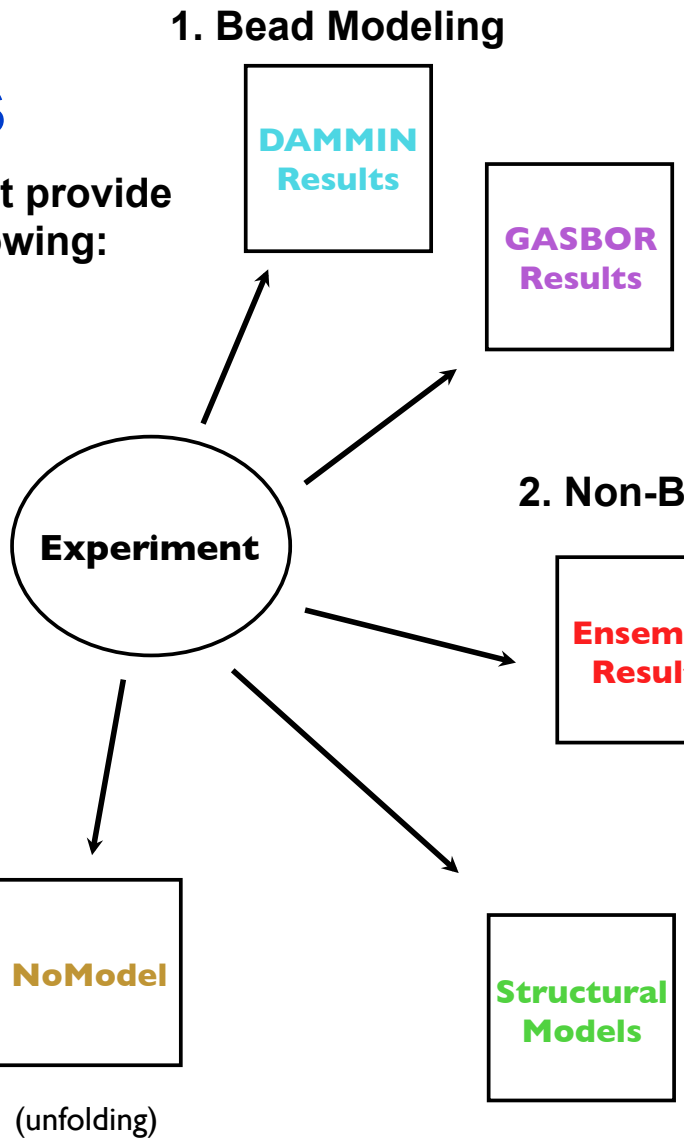
Can add more sophisticated formats:

- create specialized methods for each format in DataModel

An experiment can have many I(q) files and likewise P(r) distributions

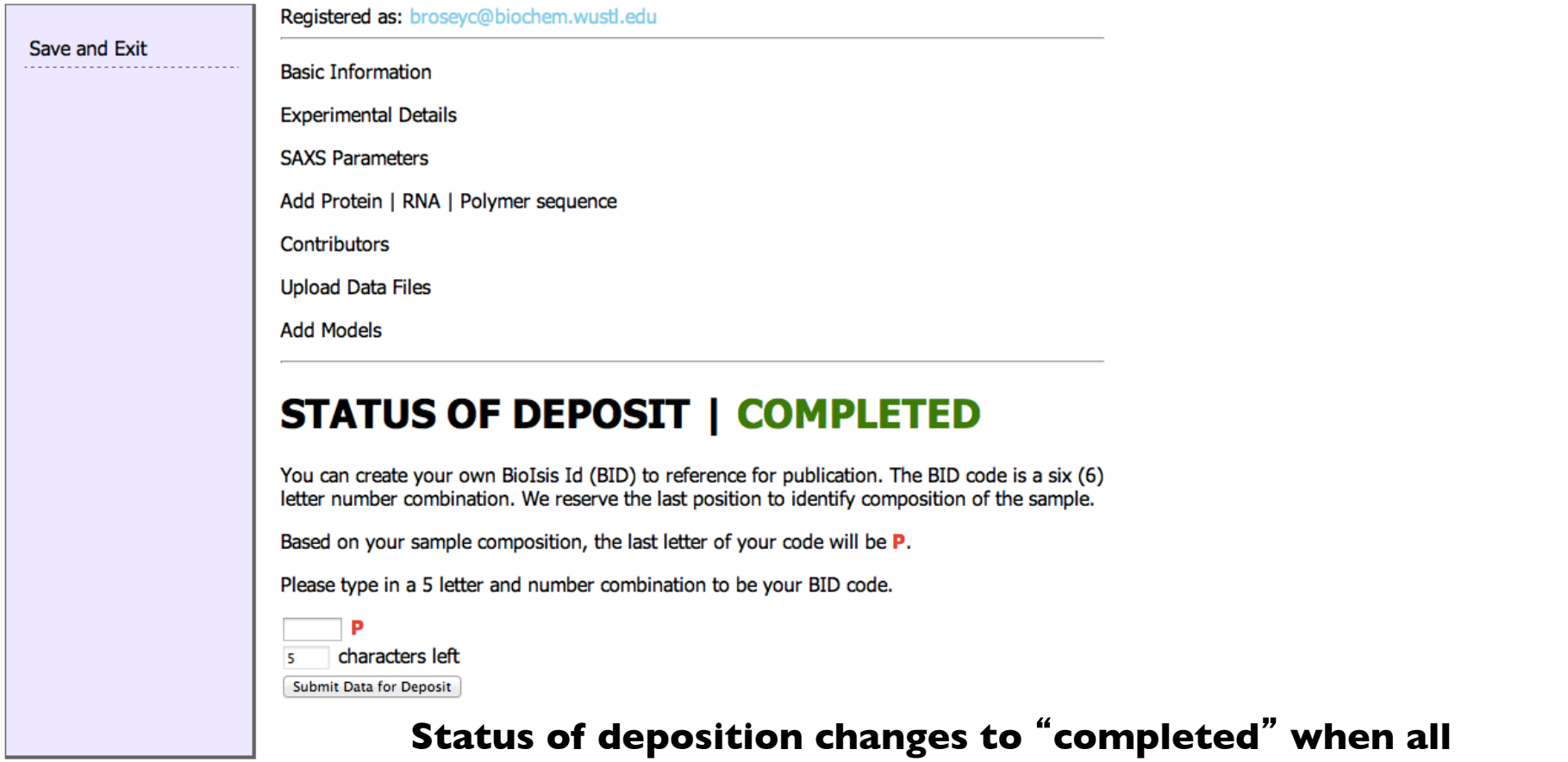
Analyses

Per Experiment, Must provide 1 or more of the following:



BIOISIS™.net

Completing A Submission

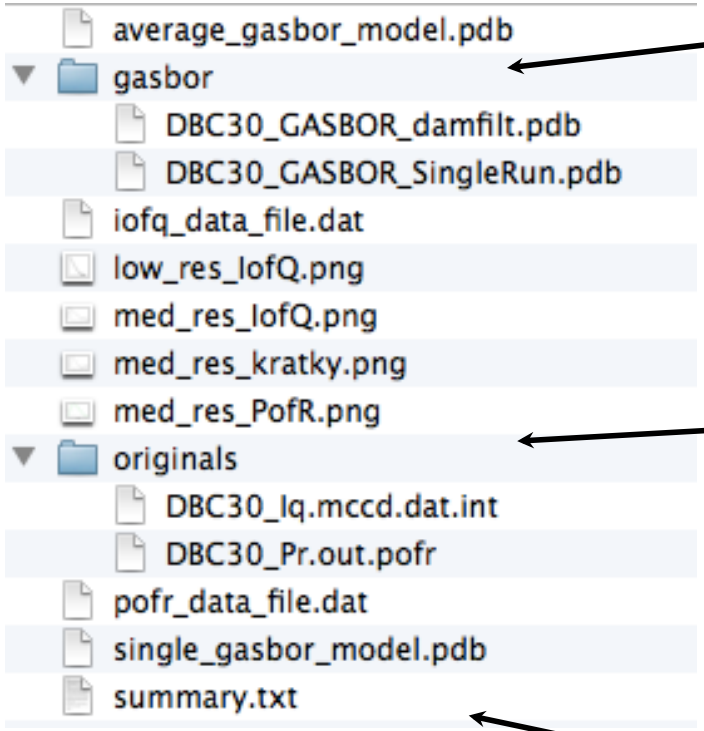


Status of deposition changes to "completed" when all required information has been completed

User is allowed to choose BID (BioIsis ID)

BIOISIS™.net

Directory Structure



Model(s) are given unique subdirectories

- **original filenames are preserved**

Holds original data files, un-altered

- **data is parsed and used to create new files and images**
- **original filenames are preserved**

summary file contains all the information required to recreate database entry

BIOISIS™.net

BIOISIS™.net

Goals (pre-workshop):

Expand database to include data types from other sources:

- **commercial instruments**
- **next generation synchrotron experiments (XFEL)**
- **SANS**

Refine Metadata required for data deposition

- **architecture of the database tables aids in database searching**
- **too much design puts burden on depositor**

Maybe paragraph describing experimental conditions is all that is required versus several drop down choices?

Get feedback from workshop on these and other ideas

BIOISIS™.net

Bioisis workshop at the ALS 2014 Berkeley CA

Attending:
25 attendees, presentations from PDB (John Westbrook), UCLA, UCSF, and LBNL

Representatives from NIST, SSRL, Australian Synchrotron, ORNL and Rigaku & Bruker

Focuses:
how to develop a sustainable Bio-SAXS database, include SANS data, and best work with the PDB

BIOISIS™.net

Bioisis workshop: role for the PDB

Possible role of the PDB in a SAS database?
Presentations examples of SAS experiments that were not used for structure determination.

PDB does not store structures from hybrid methods (e.g. Integrative Modeling Platform)

What meta data would be important in storing small-angle scattering (SAS) datasets and data formats?

For NMR and EM derived structures, PDB associates deposited structure with empirical data deposited in an independently maintained database. BioIsis database will likely hold a similar role where deposited structures in the PDB will be associated with datasets in BioIsis.

BIOISIS™.net

Bioisis workshop: suggestions for Bioisis

Major suggestions:

The SAXS database will store both published and unpublished data and should store sufficient information to:

1. reproduce the analysis proposed by the depositor
2. participate in the editorial process
3. reproduce the actual scattering experiment
4. be useable for computational and methods development

BIOISIS™.net

Format for Bioisis

1. Format of the deposited dataset will be expanded to include either a buffer subtracted & un-subtracted dataset.
2. Un-subtracted datasets will require additional information required for proper subtraction such as transmission factor corrections.
3. BioIsis will remain agnostic to data file formats and will build separate methods for extracting data from uploaded file formats.
4. All original data will be maintained but parsed for the relevant SAS data and attempts will be made to use and contribute to the PDB SAXS dictionary.

BIOISIS™.net

PDB-inspired features

BioIsis will not police data deposits

Bioisis will provide user feedback similar to the PDB validation report that accesses quality of
1. data
2. analyses (such as Guinier analysis, P(r) determination)
3. model generation and fits

Initial set of quality tests will be focused on auto-generation & testing using the P(r) distribution.

BIOISIS™.net

Uploaded I(q) data should recapitulate the uploaded P(r) distribution

User inputs of the real-space R_g, and d_{max} , suggest use of several methods for performing an indirect transform & testing how will the uploaded I(q) data can recapitulate the uploaded P(r) distribution.

The disagreement should be quantitated and used as a fit metric.

We will implement a cross-validation scheme to test how well the uploaded P(r) distribution agrees with the experimental I(q) datasets.

This will provide a visual & numerical assessment of how much of the data lies within 3 standard deviations of the empirical variance.

If too much of the data lies outside of the rejection criteria, this will inform the user or reviewer that data quality or processing may be problematic.

BIOISIS™.net

Data Format Changes

BioIsis currently parses data uploads for 3 columns but will be adapted for SANS experiments by incorporating a fourth column describing the uncertainties in the moment transfer vector.

We will expand the database to hold SANS specific information such as SANS profiles of the buffer, cell and sample.

BioIsis will not be adapted at this point to store XFEL SAXS datasets.

BIOISIS™.net

Deposition

BioIsis captures experiment-centric views of the SAS experiment

The database will collect, in a single deposit, SAXS and SANS datasets that are relevant to a specific analysis or publication. For example, a deposited contrast variation experiment can include both SAXS and SANS datasets encoded by a single BioIsis ID.

However, the database will be expanded to tag each SAS curve independently with the meta data that specifies buffer condition, polymer type, etc., as outlined in the BioIsis Experimental Description.

For deposits with multiple curves relating to a concentration series, this will reduce the repetitive burden on User by allowing prior conditions to be selected and applied to a newly uploaded SAS curve.

Immediate improvements to BioIsis will focus on making the deposition process efficient and informative to the user.

BIOISIS™.net

Participating in the Editorial Process

BioIsis will participate in the editorial process by providing anonymous platform for reviewers to interact with data (download).

This will require users to deposit data and obtain a BioIsis ID prior to manuscript submission.

Similar to the PDB, we will make a validation report of the data available but do not allow direct download of the data until the data is officially released.

BIOISIS™.net

Linking deposits

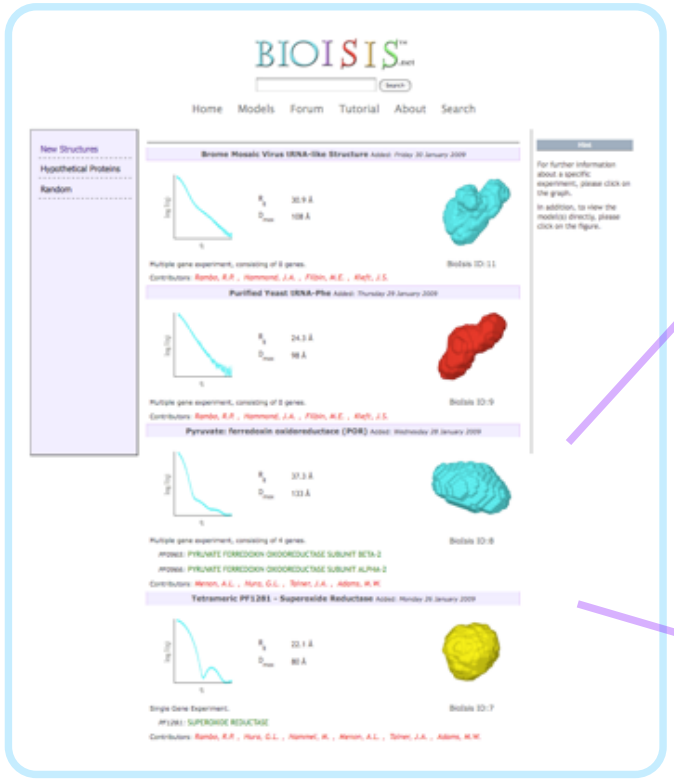
SAS datasets can be reused in other experiments and analysis (such as contrast matching)

a mechanism to search and link related BioIsis deposits was proposed to enhance existing deposits.

BIOISIS™.net

…an online database for macromolecular SAXS

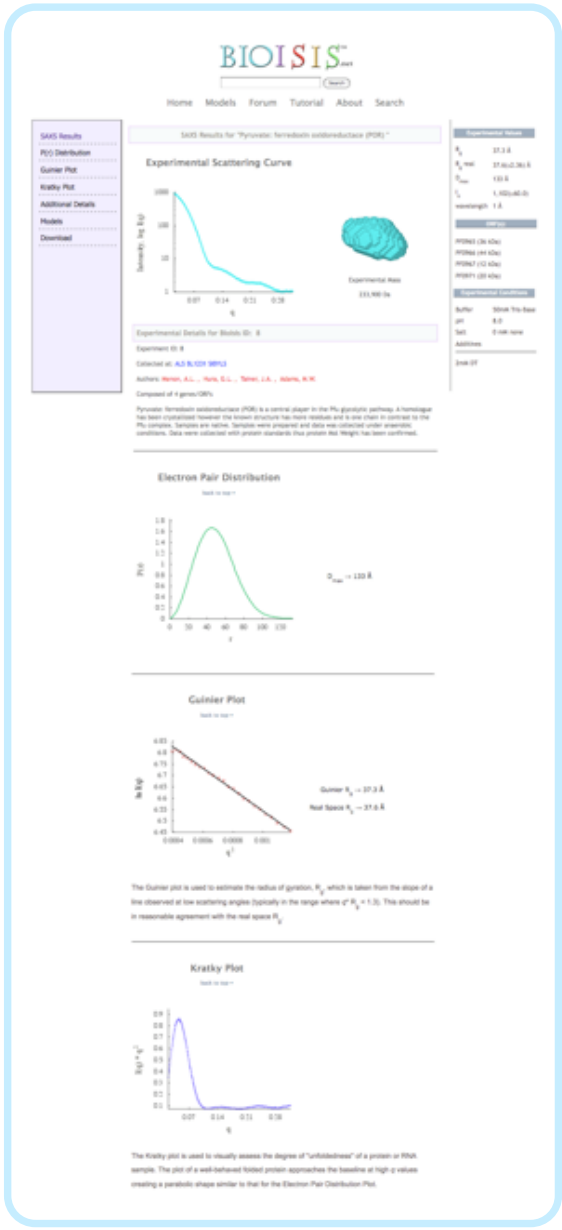
- raw scattering data
- transformed data
- experimental conditions
- SAXS derived models
- linked to related experiments (e.g. from different conditions)



P(r) Plot

Guinier Plot

Kratky Plot



conditions

Data is available for scrutiny and to encourage software development.

BIOISIS™.net

What next?

Diamond will use Bioisis for their SAXS data

Would wish to implement Workshop suggestions:

- SANS data – reformatting to include
- reproduce the analysis proposed by the depositor
- participate in the editorial process
- reproduce the actual scattering experiment
- be useable for computational and methods development
- Alternative methods for analyzing SAS data
- Other Detailed methods, e.g. BilboMD, EOM

ACKNOWLEDGEMENTS

ADVANCED LIGHT SOURCE

LAWRENCE BERKELEY NATIONAL LAB, CA

Robert Rambo

FRANCIS REYES

CAMILLE SCHWARTZ

JANE TANAMACHI

GREG HURA

MICHAL HAMMEL

KEVIN DYER

SCOTT CLASSEN

JOHN TAINER

FUNDING -

US Department of Energy (1 year)

- Novel Technology for Structural Biology
- Integrated Diffraction Analysis Technologies

WWW.BIOISIS.NET/TUTORIAL

BIOISIS™ .net

New SAXS Analysis Tools & the Q problem for flexibility

Modeling SAXS solution structures - ab initio &/or with added info

SAXS Invariants (structural parameters directly derived from SAXS)

Q, Porod Invariant

$$Q = \int_0^\infty q^2 \cdot I(q)\, dq$$

Directly related to mean square electron density of scattering particle

Requires convergence in Kratky plot

V_p, Porod Volume

$$V_p = 2\pi \cdot \frac{I(0)}{Q}$$

l_c, correlation length

$$l_c = \pi \cdot \frac{\int_0^\infty q \cdot I(q)}{Q}$$

Requires a folded particle, otherwise Q won't converge properly

Q acts as a normalization constant and corrects for:

1. concentration
2. contrast, $(\Delta\rho)^2$

R_g, radius-of-gyration

$$R_g^2 = \frac{1}{2}\frac{\int r^2 \cdot P(r)\, dr}{\int P(r)\, dr}$$

Does not require Q

Concentration independent,

Contrast independent (as long as structure does not change)

Essentially normalized to I(0)

V_c, volume of correlation

$$V_c = \frac{V}{2\pi\, l_c}$$

Creating A Submission

Must be registered

Submit an email address and a submission link is emailed

Active for 30 days

DEMO

BIOSIS™.net

Demo

Demo