New Online Curriculum: The PDB Pipeline & Data Archiving

Cathy Lawson, Rutgers University

July 22, 2018 / ACA-Toronto
BD2K Call for Proposals

- Develop open educational resources for sharing, annotating and curating "Biomedical Big Data"
- Target audience: librarians/instructors, for training biomedicine students and researchers
Enabling Data Science in Structural Biology (eDSB) Project

- Unique opportunity to document RCSB’s development and management practices
Project Goals are Consistent with RCSB’s Educational Mission

- Promote understanding of biomolecules and PDB
- Provide a structural view of biology and medicine

http://pdb101.rcsb.org
Project can Support Expanding Experimental Data Archives in Structural Biology

- Trained scientists are needed to develop federated data archives supporting new methods/model types (e.g. FRET, Mass Spectrometry ...)

Hybrid Methods Task Force EMBL-EBI, Hinxton 2014
Eight Curriculum Modules Follow the Data Pipeline

1. Enabling Data Science in Biology: Overview

Data Creation → Deposition → Processing → Archiving → Distribution

2. Creating Archive Requirements
3. Designing a Data Structure
4. Assembling Data Sets
5. Curating Deposited Data
6. Ensuring Data Consistency
7. Creating & Maintaining a Data Archive
8. Sharing Data with End Users
# Modules

<table>
<thead>
<tr>
<th>1. Overview</th>
<th>5. Curating the Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Creating Archive Requirements</td>
<td>6. Ensuring Data Consistency</td>
</tr>
<tr>
<td>3. Designing the Infrastructure</td>
<td>7. Creating and Maintaining an Archive</td>
</tr>
<tr>
<td>4. Data Deposition</td>
<td>8. Data Distribution</td>
</tr>
</tbody>
</table>
Learning Objectives/Skills

- Recognize what is involved in designing and maintaining an archive for shared data
- Identify key stakeholders
- Develop requirements for what data to include
- Understand how to develop a data dictionary with the appropriate level of granularity
- Construct a deposition and annotation workflow based on a data dictionary
Lectures

- Lectures were developed and delivered by RCSB PDB group members according to their expertise
- 3-5 video segments per Module
- Transcripts were carefully curated to support closed-captioning
Exercises/Homework

- Students are guided step-by-step to design, create, and query a database on a topic of their own interest
- Exercises introduce tools needed to complete assignments
- Worked example included in all assignments
## Homework Flow

<table>
<thead>
<tr>
<th>Module</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Select set of PDB entries on topic of interest (50-100)</td>
</tr>
<tr>
<td>2</td>
<td>Create PDB data reports, get primary citations</td>
</tr>
<tr>
<td>3</td>
<td>Define questions about your topic, create new data terms</td>
</tr>
<tr>
<td>4</td>
<td>Create a deposition form for your new terms and fill it in</td>
</tr>
<tr>
<td>5</td>
<td>Review validation reports for your PDB entries</td>
</tr>
<tr>
<td>6</td>
<td>Check filled data for errors</td>
</tr>
<tr>
<td>7</td>
<td>Create a database combining PDB data and your new data</td>
</tr>
<tr>
<td>8</td>
<td>Perform queries to answer the questions about your topic</td>
</tr>
</tbody>
</table>

Tools used:
- RCSB PDB website search/browse/reports, simple text editor, Excel or equivalent, Google Forms, MySQL Server and MySQL Workbench
Worked Example

- Recent E. coli ribosome cryoEM structures (61)

Example Questions:
- How many structures have both ribosomal subunits?
- Which structures include messenger RNA?
- What type of tRNA is bound in the P (peptidyl) site?
- Do ribosome structures with bound antibiotics have good model quality?

Distribution of tRNA types in the peptidyl site of recent *E. coli* ribosome structures:

<table>
<thead>
<tr>
<th>COUNT(pdb_id)</th>
<th>p_site_trna_aa_type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glycine</td>
</tr>
<tr>
<td>2</td>
<td>Aspartate</td>
</tr>
<tr>
<td>3</td>
<td>Proline</td>
</tr>
<tr>
<td>4</td>
<td>Unknown</td>
</tr>
<tr>
<td>17</td>
<td>none</td>
</tr>
<tr>
<td>34</td>
<td>Initiator Methionine</td>
</tr>
</tbody>
</table>
Initial Implementation

- The curriculum was pilot-tested at Rutgers in Spring 2016, and then again in Spring 2018

- Students included:
  • Rutgers Graduate Students (Chemistry, Mol Bio)
  • Information Specialists from Rutgers Libraries
  • International Scientists interested in developing data archives
Use in a Flipped Classroom

The Flipped Classroom

IN CLASS

Students practice applying key concepts with feedback

DURING

Learn to Search/Browse PDB

Module 1 Class Exercise

OUT OF CLASS

Students prepare to participate in class activities

BEFORE

GOAL

Students check their understanding and extend their learning

AFTER

GOAL

Module 1 Homework: Identify Your Research Topic

Image Source: U. Texas at Austin Faculty Innovation Center
Dissemination

- All materials will be accessible via PDB-101 and http://edsb.rcsb.org
  - Lectures: Slides, Transcripts, Videos
  - Exercise, Homework Slides
  - Links to Additional Resources
  - Licensing: Attribution-NonCommercial-ShareAlike 4.0 International

- Coursera MOOC: to be developed
Project Personnel

Catherine Lawson  
Project PI

Helen M. Berman  
Pilot 1 Lead

Maggie Gabanyi  
Video Production Lead

John Westbrook

Jasmine Young

Shuchismita Dutta

Brian P. Hudson

Ezra Peisach

Peter Rose

Jose Duarte

Stephen K. Burley

Amy Sarjeant  
CCDC

Interested in using this curriculum? Let us know: edsb@rcsb.org