Structural biology meets data science: Does anything change?

Cameron Mura, Eli J. Draizen, and Philip E. Bourne

Survey by Eric Wang

2019 Spring @ https://qdata.github.io/deep2Read/

Introduction

- The term *Structural Biology* (SB) can be defined rather precisely as a scientific field, but *Data Science* (DS) is more enigmatic, at least currently. The intrinsic difference is two-fold.
- First, DS is a young field, so its precise *meaning*—based on what we practice and how we educate its practitioners—has had less time than SB to coalesce into a consensus definition.
- Second, and more fundamental, DS is interdisciplinary to an extreme; indeed, DS is not so much a field in itself as it is a way of *doing* science, given large amounts of diverse and complex data, suitable algorithms and sufficient computing resources.
- Such is the breadth and depth of DS that it has been described as a fourth paradigm of science, alongside the theoretical, experimental and computational. Because it is so vast and sprawling, a helpful organizational scheme is to consider four *V*'s and five *P*'s that characterize data and DS.

Four V’s and Five P's

- The four *V*'s describe the properties of data:
  - *volume*, *velocity*, *variety* and *veracity*.
- The *P*'s are the five disciplinary pillars (P-i through P-v) of DS:
  - (i) *data acquisition*
  - (ii) *data reduction*, *integration and engineering*
  - (iii) *data analysis* (often via machine learning)
  - (iv) *data visualization*,*provenance and dissemination*
  - (v) *ethical*, *legal*, *social and policy-related matters*.
- **What structural biology has to offer data science...**

Open Science

- SB has pioneered open science through the provision of the **PDB** and many derivative data sources. The complete corpus of structural information in the PDB is **free of copyright** and is available for unfettered use, non-commercial or otherwise.
- Moreover, community practices—such as virtually no journal publishing an article without its data deposited in the PDB—is a precedent that, if broadly adopted in other disciplines, would deepen the amount and diversity of data available for DS-like approaches in those other scientific and technical domains.
- The creation and free distribution of **software tools** has echoed this trend, as epitomized by the *Collaborative Computational Project 4* (*CCP4*) since 1979, has been a mainstay of the crystallographic structure-determination process.
- To succeed, we believe that any DS must abide by the 'FAIR' principles, enabling researchers to *Find*, *Access*, *Interoperate* and *Reuse* data and analytics. SB has exercised this for decades, and is thus positioned to lead the way.

Reproducibility

- In principle, reproducibility is the bedrock of the scientific enterprise. And, as a byproduct of open science, reproducibility has been central in SB, though often less so in other realms of DS.
- Cultural differences across various disciplines, often driven by (perceived) competitive pressures, have dampened what could be the norm. In SB, the systematic, pipelined nature of many structure-determination approaches has facilitated reproducibility.
- A notable example is the effort, spurred by structural genomics, to annotate large-scale macromolecular crystallization experiments and to conduct careful target tracking; in principle, such efforts afford a rich source of data, exploitable by DS via data mining and machine learning methods.

Workflows, High-Performance Computing

- Reproducibility, in turn, is facilitated by workflows. Some workflow management systems (WMS) are domain-specific (e.g., Galaxy for genomics), while others are more generic or monolithic (e.g., KNIME)
- Structural genomics and other data-rich areas have prompted the development of WMS solutions. Closely related to workflows, recent technologies that have become best practices in DS—such as **Jupyter** notebooks (as a user interface) and **Docker** 'containers' —likely will be adopted more broadly in SB, as research questions become more **quantitative** and as **data-intensive** computational steps are pursued via distributed computing and other modes of HPC.
- Cloud computing and related approaches, such as the MapReduce paradigm, rapidly entered genomics and bioinformatics early on and are becoming more widely adopted in other biosciences too, including SB; other examples include large-scale biomolecular modeling for virtual screening and drug design and, more recently, cryo-EM pipelines for structure determination.

HPC continued...

- A recent example using HPC involves the phasing of diffraction data. Recognizing the wealth of structural information in the PDB, and that molecular replacement (MR) can be used for all these structures, the *BALBES* pipeline leverages all known 3D structures to create and then use MR search models in an automated manner.
- This approach was recently extended to fitting 3D models into cryo-EM maps . Somewhat similar in spirit, *PDB_REDO* endeavors to automatically improve all PDB structures by re-refining 3D models against the original X-ray data, utilizing established refinement approaches (e.g., TLS) and grid computing.

The Phase Problem

Molecular Replacement

(In case of availability of coordinates of similar protein)

Take similar protein

↓

Coordinates

↓

Rotate and Translate in our new crystal system

↓

Good Match found

↓

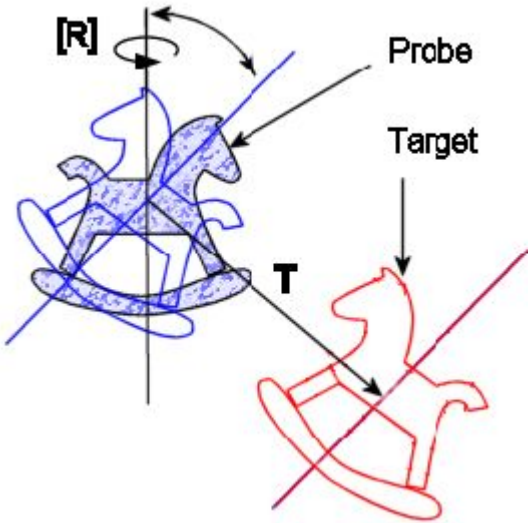
Measure Phase Angle & Amplitudes

↓

Using Fourier Transform

↓

Produce Electron Density Map



Visualization

- Recent advances have occurred in web browser-embedded, hardware-accelerated tools for interactive molecular visualization, such as the NGL Viewer. To transcend how molecular renderings are usually communicated (as **static images**), we suspect that much could be gained by comparing visualization techniques in DS and SB. Though iconic and highly informative, beware the "curse of the ribbon": macromolecules are dynamic, multifaceted entities, and static renditions are a starting point.
- There is a need for molecular visualization platforms that transcend **facile, flexible and extensible** integration of other forms/modalities of data and novel visualization techniques
- We believe that DS tools can address this need. Ideas and methods from beyond SB—such as "chord diagram" layouts in genomics, termed "hierarchical edge bundles" in computer graphics—can be applied in SB, for instance to visualize data associated with hierarchical clustering of protein structural differences.

Further Overlap

- (i) database-related issues, including
  - structured versus unstructured data,
  - relational versus non-relational databases and query languages
- (ii) systems and network biology
- (iii) ontologies and formal knowledge representation systems

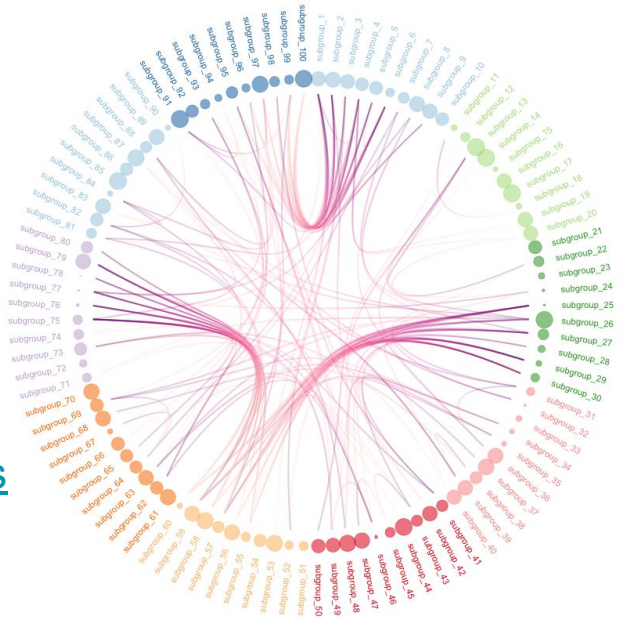


Figure of hierarchical edge bundles

What data science analytics has to offer structural biology

- Focus on two machine learning approaches
  - Deep learning **[DL]**
  - Natural language processing **[NLP]**
- DL methods have been applied to model and predict protein•ligand and protein•protein interactions (PLI, PPI)
- Accurately predicting and modeling PLIs would advance many areas, both basic (e.g., evolutionary analyses of ligand-binding properties) and applied (e.g., drug design and discovery).

Deep Learning in Biomolecular Interactions

- Two distinct methodological approaches:
  - Quantitative structure-activity relationships (QSAR)
    - **Virtual screening**, wherein one docks against large libraries of small compounds
    - Workflow-based approaches to high-throughput crystallographic fragment screening
    - Analyze human **kinome** by integrating ligand-binding data with protein-ligand "interaction fingerprints" and a sequence order-independent profile–profile alignment method; useful for determining specificity among similar ligand-binding sites.
  - *In silico* docking.
    - recognizing that a protein exists as an ensemble of thermally-accessible conformational states in solution, simulations have been combined with docking in the "relaxed complex" scheme to capture **receptor flexibility**
    - Workflow to discover **druggable binding sites** was developed by integrating comparative structural analyses, pocket-detection algorithms, fragment docking, molecular simulations, and an ML classifier

Natural language processing applied to biomolecular assemblies

- NLP-like approaches have been applied to detect the subcellular localization of proteins and to predict structures of protein complexes.
- Notably, ML-enhanced NLP, versus a purely text-mining–based NLP approach, was found to significantly improve the structural predictions of complexes. Note that both sorts of problems—subcellular localization and structural modeling—are distinctly spatial, or image-based, as opposed to textual.
- We expect that a relatively new and highly-generalized approach to NLP, termed topic modeling (TM), holds great promise in the biosciences.

Topic Modeling

- In TM, ‘topics’ are extracted over a corpus of unstructured data (e.g., a set of books) using a probabilistic machine learning framework; fundamentally, this is achieved by examining the distributions of words ("bag of words") under a generative statistical model, such as the latent Dirichlet allocation (LDA).
- To extend TM to other areas—including even the learning of topics from non-textual data like protein structures—the basic issue is one of defining a **suitable mapping** of one’s problem to TM’s core framework of *document* ↔ *topic* ↔ *word*.
- TM may be applicable to the analysis of protein folds and other biomolecular structures. Such an application of NLP to what is a fundamentally geometric problem finds precedent in the pioneering development of a generative Bayesian hierarchical model for scene classification from raw image data.

End of Survey

- Several human protein complex databases have been developed to date, including CORUM [14, 15] and disease-related complex [16].
- The protein complexes in CORUM were collected only from literature. The database does not provide information about many uncharacterized proteins whose interactions are supported by PPI data.
- The disease-related complex database [16] is focused on disease complexes, using information on proteins known to be involved in similar disorders. Accordingly, it contains a relatively small number of complexes (506) and lacks many other important complexes.