Lecture Title and Date

DATA - Proteomics II - 1/22

Objectives of the Lecture

By the end of this lecture, students should be able to:

1. Understand the fundamental concepts and uses of X-ray crystallography
2. Understand the developments in cryo-EM and its use cases
3. Explain the impact of AI, specifically AlphaFold, on solving the protein-folding problem.
4. Identify the resources and tools used in protein structure analysis, including databases like PDB.

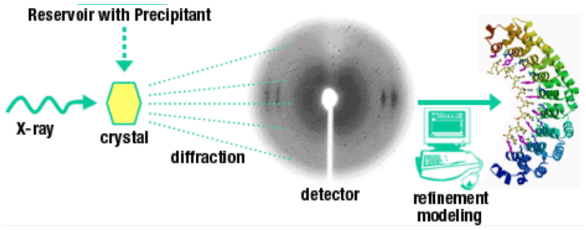
Key Concepts and Definitions

- **X-ray Crystallography**: Purified sample is crystallized and exposed to an x-ray beam obtaining 3d molecular structure from a crystal.
- **Cryo-EM**: A technique used to determine the 3D structure of biomolecules by flash-freezing samples and imaging them with an electron microscope.
- **Protein Data Bank (PDB)**: It is a global repository that stores three-dimensional structural data of biological macromolecules like proteins, DNA, and RNA
- **Diffraction Limit**: The smallest size resolvable by a given wavelength of light.
- **AlphaFold**: An AI-based system developed by DeepMind to predict protein structures with high accuracy.
- **Resolution**: A measure of the level of detail visible in a structure, typically in angstroms (Å).

Main Content/Topics

X-ray Crystallography

Purpose: Because there is a limit on how small an object can be seen under a light microscope, we need to use x-rays to more precisely detect positions of atoms. They can penetrate below the resolution of carbon bonds, with wavelength at or below 1.54 Å.
Crystals are used because x-rays are scattered by electrons, which makes it difficult to gather information from a single molecule → crystals present many molecules in the same orientation.
Method: A beam of x-rays is directed at a macromolecular crystal from multiple angles, generating multiple 2-dimensional images.



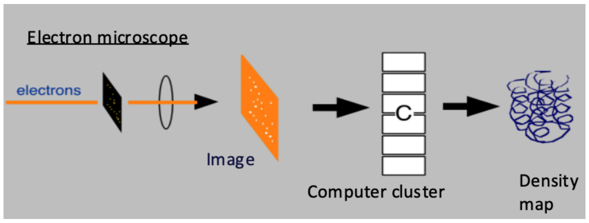
The Fourier transform method is used to combine the images to an electron density map, a picture of the crystal at the atomic level, which then helps produce a structural model.[1]

Results: Many online databases now exist containing protein structural information e.g. PDB, MMDB, FSSP, SCOP, and CATH. PDB contains information on more than 230,000 structures—42,668 from humans, though the majority are from other species.

Cryo-EM

Purpose: Some proteins can take months or years to crystallize, and even crystallized proteins sometimes produce messy diffraction patterns (e.g. membrane proteins), which makes x-ray crystallography difficult to use.[2] Single-particle cryo-EM requires no crystals and much fewer copies of protein (10^5 as opposed to 10^{12}).

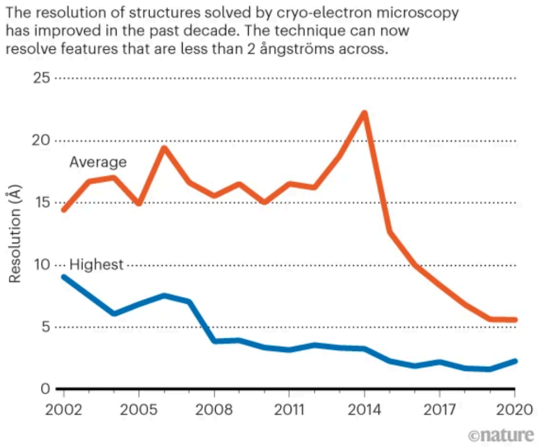
Method: Flash-freeze solutions of proteins (or other biomolecules), then place them into a transmission electron microscope, where electron beams travel through and cast shadows of the molecules onto a detector.



After multiple iterations, these images are then denoised, oriented, averaged, and reconstructed into a 3-D structure.[3]

Results: Although cryo-EM used to be less popular because it was lower-resolution, recent breakthroughs in electron microscopes (like the FEI Titan Krios Transmission Electron Microscope) and software have produced sharper images.

[1] Molinski and Morinaka, Ilari and Savino.
[2] Cheng.
[3] Callaway, Chui.


The resolution of structures solved by cryo-electron microscopy has improved in the past decade. The technique can now resolve features that are less than 2 ångströms across.

The number of entries in the Electron Microscopy Data Bank is growing exponentially, reflecting cryo-EM's development; EMICSS also launched in 2022 to provide cross-reference information and annotations to EMDB data.

The protein-folding problem

Explanation of the problem: In protein folding prediction, the challenge lies in determining how a protein's three-dimensional structure forms solely from its amino acid sequence. This is crucial because a protein's function is primarily dictated by its structure. Developing a computational algorithm to accurately predict protein folding can significantly enhance our understanding of biological processes and facilitate advancements in drug discovery and bioengineering.[4]

Solutions: The challenge of predicting a protein's 3D structure from its amino acid sequence has been addressed through AI deep-learning models trained on structural databases like PDB. These repositories, which contain a vast collection of experimentally determined protein structures, have been crucial in advancing AI-driven predictions.

Examples: AI-driven models like AlphaFold and RoseTTAFold have transformed protein structure prediction by leveraging deep learning to infer 3D structures from amino acid sequences.[5] [6] Trained on existing experimental structural data and resources like PDB, these models predict residue-residue distances and protein folding patterns with high accuracy

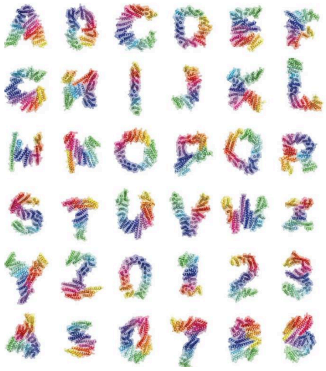
Applications: Beyond prediction, we can also use generative models to *design* proteins to accomplish a specific function.[7] (The below image demonstrates just how flexible that capability is: proteins that can be made to look like letters or numbers.)

[4] W. Senior et. al.

[5] Ibid.

[6] M. Baek et al.

[7] Kauffman.



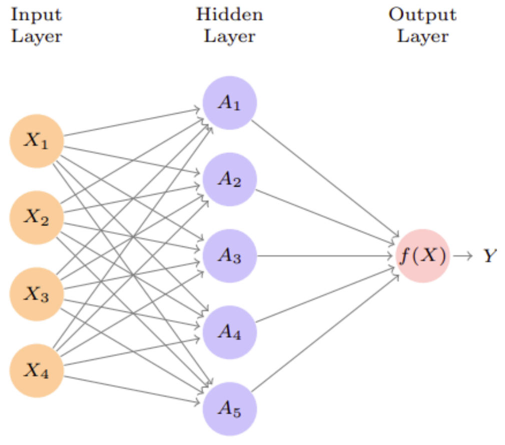
Additionally, by turning gene sequences into structures, we can more easily identify evolutionary relationships across species, creating a more comprehensive picture of how our proteomes intertwine.

References ISL/ESL

In the lecture, deep learning was highlighted as a transformative approach that solves the long-standing protein-folding problem. In Chapter 10 of ISL and Chapter 11 of ESL, the books introduce relevant deep learning topics on neural networks, optimization challenges, and model architecture, which underpin modern AI-driven protein structure prediction.

ISL (An Introduction to Statistical Learning, with Applications in Python)

- Relevant Chapter: Chapter 10 – Deep Learning (*James et al., 2023*)
- Supplementary content:
  - A neural network is a machine learning model designed to recognize patterns and make predictions by mimicking the structure of the human brain. It consists of interconnected layers of neurons that process input data through weighted connections and activation functions.
  - As shown in the figure below, a basic neural network comprises three types of layers:
    - Input Layer – Receives raw data (e.g., features X_1, X_2, …, X□).
    - Hidden Layer(s) – Transforms input data using nonlinear activation functions. Each neuron in this layer computes activations A□ = h□(X), which are learned during training rather than fixed beforehand.
    - Output Layer – Produces final predictions based on the transformed inputs. The output function f(X) maps the hidden activations to the desired output.



  - Neural networks learn by adjusting the weights of connections between neurons using algorithms such as backpropagation and gradient descent. Modern implementations, including deep learning models, use libraries like Python's torch package to train and apply neural networks efficiently.
- Citation:
  James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Chapter 10 – Deep Learning. First Printing, Springer.

ESL (The Elements of Statistical Learning, Second Edition)

- Relevant Chapter: Chapter 11 – Neural Networks (*Hastie et al., 2017*)
- Supplementary content:
  - Issues in Training Neural Networks
    - Starting Values: Proper initialization is crucial. Weights near zero make the model behave like a linear model, while very large weights can lead to poor optimization.
    - Overfitting: Neural networks tend to memorize training data, requiring regularization techniques such as weight decay, which is analogous to ridge regression.
    - Scaling of Inputs: Standardizing inputs to zero mean and unit variance ensures even weight updates and improves training stability.
  - Model Architecture Considerations
    - Number of Hidden Units and Layers:
      - Too few hidden units reduce model flexibility, while too many can lead to overfitting.
      - A moderate number of units with regularization is recommended for optimal performance.
      - Multiple hidden layers enable the hierarchical learning of complex patterns.
    - Multiple Minima in Optimization:

    - The error function of neural networks is nonconvex, leading to many local minima.
    - Using techniques like bagging (averaging over multiple trained networks) helps stabilize predictions
- Citation:
  Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Chapter 11 – Neural Networks. Corrected 12th Printing, Second Edition, Springer.

Suggest references for many of the key concepts

X-ray Crystallography:

- Ilari, Andrea, and Carmelinda Savino. "Protein structure determination by x-ray crystallography." *Bioinformatics: Data, Sequence Analysis and Evolution* (2008): 63-87.
- Molinski, Tadeusz F., and Brandon I. Morinaka. 2012. “Integrated Approaches to the Configurational Assignment of Marine Natural Products.” *Tetrahedron* 68 (46): 9307–43. https://doi.org/10.1016/j.tet.2011.12.070.
- Shi, Yigong. 2014. “A Glimpse of Structural Biology through X-Ray Crystallography.” *Cell* 159 (5): 995–1014. https://doi.org/10.1016/j.cell.2014.10.051.

Cryo-EM

- Callaway, Ewen. 2020. “Revolutionary Cryo-EM Is Taking over Structural Biology.” *Nature* 578 (7794): 201–1. https://doi.org/10.1038/d41586-020-00341-9.
- Cheng, Yifan. "Membrane protein structural biology in the era of single particle cryo-EM." *Current opinion in structural biology* 52 (2018): 58-63.
- Chui, Glennda. n.d. “Cryogenic Electron Microscopy (Cryo-EM): Amazing Views of Life’s Machinery.” SLAC National Accelerator Laboratory. https://www6.slac.stanford.edu/research/slac-science-explained/cryo-em.

The protein-folding problem

- Callaway, Ewen. 2022. “What’s next for AlphaFold and the AI Protein-Folding Revolution.” *Nature* 604 (7905): 234–38. https://doi.org/10.1038/d41586-022-00997-5.
- Kauffman, M. 2024 "From Sequence to Structure to Function: De Novo Protein Design, the Role of AI and Structure Prediction Neural Networks" Preprints. https://doi.org/10.20944/preprints202404.0220.v1
- Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, et al. 2020. “Improved Protein Structure Prediction Using Potentials from Deep Learning.” *Nature* 577 (7792): 706–10. https://doi.org/10.1038/s41586-019-1923-7.

Discussion/Comments

- Limitations of X-ray Crystallography:
  X-ray crystallography requires well-ordered crystals, which can be a bottleneck, particularly for membrane proteins or transient complexes. The emergence of Cryo-EM was highlighted as a method that bypasses the need for crystals and directly observes molecular complexes.
- Impact of Cryo-EM:
  Cryo-EM was recognized as a revolutionary tool, especially for its ability to resolve structures of large protein assemblies like ribosomes, electron transport chains, and other dynamic macromolecules. However, there are some challenges in computational resources, such as phase contrast image processing and 3D reconstruction.
- Computational Methods and AI:
  The lecture acknowledged AlphaFold as a groundbreaking AI tool that has significantly advanced the protein structure field. Integrating experimental data like X-ray crystallography and Cryo-EM for model validation was also emphasized. Ethical questions in designing therapeutic proteins and synthetic biology were likely raised.
- Role of Structural Databases:
  The rapid increase in database entries was noted, emphasizing the need for standardized validation of both experimental and computational models to maintain scientific integrity.
- Useful Resources:
  - Protein Structure Information:
    PDB: http://www.rcsb.org/pdb
    MMDB: https://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
    FSSP: https://www.ebi.ac.uk/msd-srv/ssm/
    SCOP: https://www.ebi.ac.uk/pdbe/scop/
    CATH: https://www.cathdb.info/
  - Tools for Viewing Structures:
    Jmol: http://jmol.sourceforge.net/
    PyMOL: http://pymol.sourceforge.net/
    Swiss PDB viewer: http://www.expasy.ch/spdbv
    Mage/KiNG: http://kinemage.biochem.duke.edu/software/mage/
    http://kinemage.biochem.duke.edu/software/king/
    Rasmol: http://www.umass.edu/microbio/rasmol/

- Advances and Limitations in Protein Structure Prediction

  AlphaFold and RoseTTAFold have greatly improved protein structure prediction using deep learning, enabling high accuracy and accelerating biomedical research. However, they have limitations in modeling protein dynamics, interactions, and cellular context, highlighting areas for future improvement.