11: Catchup II

Machine Learning and Real-world Data (MLRD)

Andreas Vlachos
(based on slides by Ann Copestake)

Lent 2022

Last session: HMM in a biological application

- In the last session, we used an HMM as a way of approximating some aspects of protein structure.
- Today: catchup session 2.
- Very brief sketch of protein structure determination: including **gamification** and **Monte Carlo methods**.
- Related ideas are used in many very different machine learning applications …

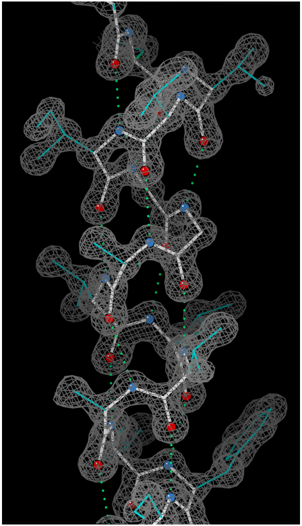
What happens in catchup sessions?

- Lecture and demonstrated session scheduled as in normal session.
- Lecture material is non-examinable.
- Time for you to catch-up in demonstrated sessions or attempt some starred ticks.
- Demonstrators help as usual.

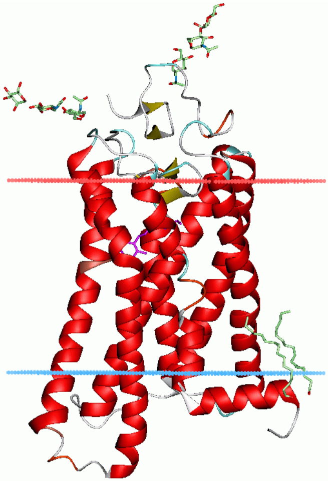
Protein structure

- Levels of structure:
  - Primary structure: sequence of amino acid residues.
  - Secondary structure: highly regular substructures, especially α-helix, β-sheet.
  - Tertiary structure: full 3-D structure.
- In the cell: an amino acid sequence (as encoded by DNA) is produced and folds itself into a protein.
- Secondary and tertiary structure crucial for protein to operate correctly.
- Some diseases thought to be caused by problems in protein folding.

Alpha helix

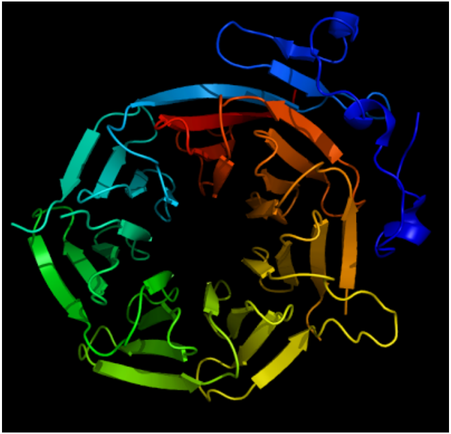


Bovine rhodopsin



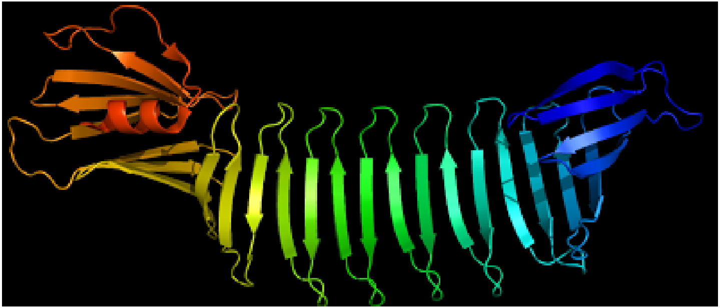
By Andrei Lomize - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=34114850

7-bladed propeller fold



http://beautifulproteins.blogspot.co.uk/

Peptide self-assembly mimic scaffold: an engineered protein



http://beautifulproteins.blogspot.co.uk/

Protein folding

- Anfinsen's hypothesis: the structure a protein forms in nature is the global minimum of the free energy and is determined by the animo acid sequence.
- Levinthal's paradox: protein folding takes milliseconds — not enough time to explore the space and find the global minimum. Therefore kinetic function must be important.

Protein structure determination and prediction

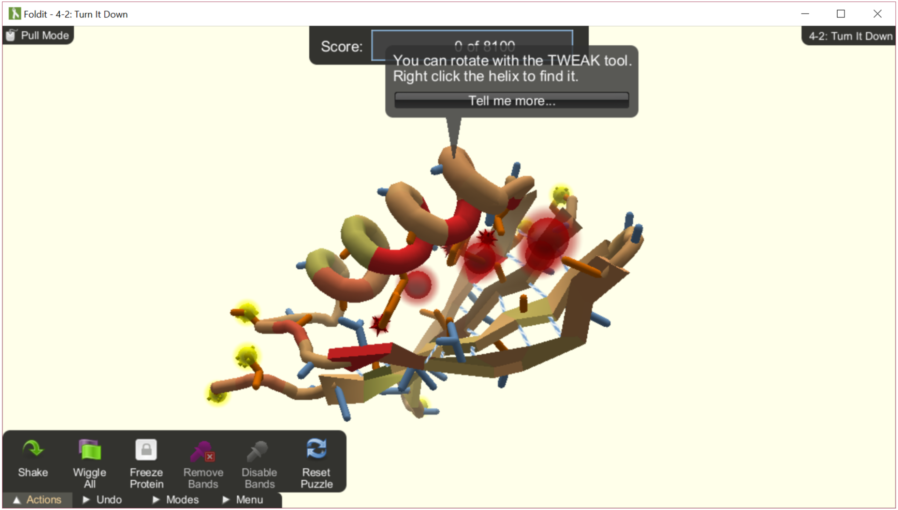
- Primary structure may be determined directly or from DNA sequencing: relatively easy.
- Secondary and tertiary structure can be determined by x-ray crystallography and other direct methods, but difficult, expensive, time-consuming.
- Given amino acid sequence, can we predict the structure? i.e., determine how the protein will fold.
- Secondary structure prediction is relatively tractable: various prediction methods, including HMMs (cf last session).
- Tertiary structure prediction is very difficult.

Protein tertiary structure prediction

- Modelling protein structure fully is hugely computationally expensive.
- Ideally, should model all the water molecules too . . .
- Several approaches, including:
  1. Molecular Dynamics (MD): modelling chemistry. folding@home: use home computers to run simulations.
  2. Foldit: get lots of humans to work on the problem (an example of **gamification**).
  3. Use **Monte Carlo methods** (repeated random sampling) to explore possibilities.
  4. Specialised neural networks have recently achieved state-of-the-art (AlphaFold, Bumper, Nature 2021)

     https://www.nature.com/articles/s41586-021-03819-2Sec10

Foldit: combined human-computer intelligence



Monte Carlo methods in protein structure prediction

- Objective: find lowest energy state of protein.
- Idea: start with secondary structure, try (pseudo)random move, see if result is lower energy and repeat.
- Problem: **local minima** — locally good move may not be part of best solution.
- So: also sometimes accept a move that increases energy.
- Specific approach **Metropolis-Hastings**: a type of **Markov Chain Monte Carlo** method.

Monte Carlo methods

- Using random sampling to solve intractable numerical problems.
- Physicists developed modern Monte Carlo methods at Los Alamos: programmed into ENIAC by von Neumann.
- Bayesian statistical inference not until 1993 (Gordon et al): essential for many modern machine learning approaches.
- Gibbs sampling is a special case of Metropolis-Hastings.
- More about this in later modules: Advanced Data Science, Machine Learning and Bayesian Inference, Bioinformatics.
- Practical introduction by Geyer in

  http://www.mcmchandbook.net/HandbookTableofContents.html