

11: Catchup II

Machine Learning and Real-world Data (MLRD)

Ann Copestake

Lent 2019

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** (and a little about AlphaFold).
- Related ideas are used in many very different machine learning applications . . .

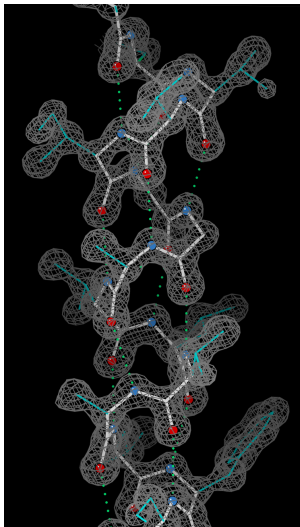
What happens in catchup sessions?

- 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



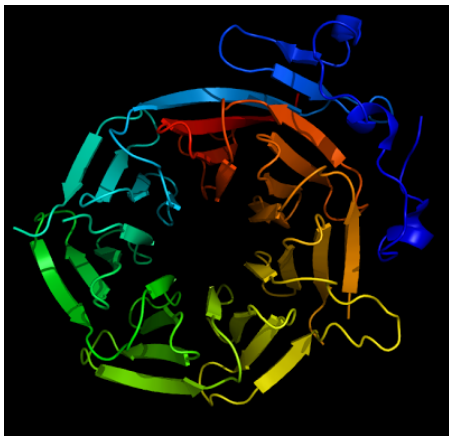
Bovine rhodopsin



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

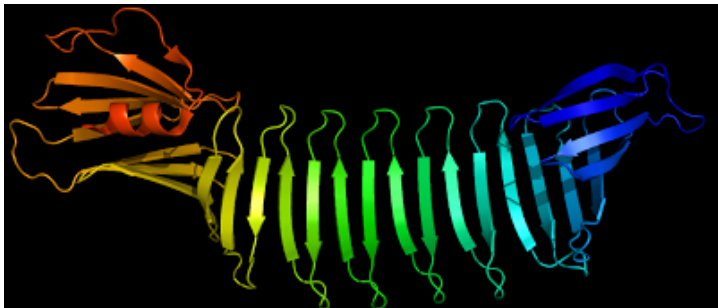
- found in the rods in the retina of the eye
- a bundle of seven helices crossing the membrane (membrane surfaces marked by horizontal lines)
- supports a molecule of **retinal**, which changes structure when exposed to light, also changing the protein structure, initiating the visual pathway

7-bladed propeller fold (found naturally)



<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 amino 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 Use additional information either a) previously determined structures or b) **evolutionary coupling** (e.g., DeepMind's AlphaFold)

2: Foldit: combined human-computer intelligence

The screenshot displays the Foldit game interface. At the top left, the window title is "Foldit - 4-2: Turn It Down". Below it, a "Pull Mode" button is visible. In the top right corner, there are window control icons and the text "4-2: Turn It Down".

At the top center, a score display shows "Score: 0 of 8100". Below the score, a tooltip box contains the text: "You can rotate with the TWEAK tool. Right click the helix to find it." and a "Tell me more..." button.

The main area of the interface features a 3D protein puzzle. The protein is represented by a complex structure of colored bands (red, green, blue, yellow) and sticks, set against a light yellow background. The puzzle is partially assembled, with some bands highlighted in red and yellow.

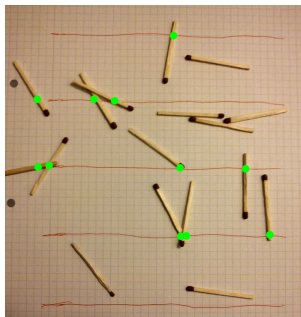
At the bottom of the interface, there is a toolbar with several icons and labels: "Shake", "Wiggle All", "Freeze Protein", "Remove Bands", "Disable Bands", and "Reset Puzzle". Below the toolbar, there are navigation buttons labeled "Actions", "Undo", "Modes", and "Menu".

3: 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 (e.g., **Rosetta**).

Monte Carlo methods in general

- Using random sampling to solve intractable numerical problems.
- Buffon's needle problem used for estimating π ('experiment' by Lazzarini 1901).



Monte Carlo methods

- 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 courses: Mathematical Methods, Machine Learning and Bayesian Inference, Bioinformatics.
- Practical introduction by Geyer in

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

4: Using additional information in protein folding

- 1 use previously determined structures of similar proteins.
- 2 **evolutionary couplings**: databases of proteins in an evolutionary relationship, mutations tend to be correlated if amino acids are physically close in folded protein:
 - generate likely contacts (nowadays using deep learning), feed info into folding program;
 - Deep Mind's AlphaFold: produce full probability distribution of distances, statistical potential function which is directly minimized by **gradient descent** (L-BFGS???)

<https://deepmind.com/blog/alphafold/>

<https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/>

Conclusions

Protein structure prediction is still unsolved.

Possible approaches involve several techniques used elsewhere in machine learning:

- gamification: an example of human-computer collaboration
- Monte Carlo methods
- using additional information sources . . .

General discussion in deep learning: constraints/priors vs **tabula rasa** approaches (also question as to what counts as tabula rasa . . .)