SCIE2100/BINF6000 Learning Guide

Week 12: Protein bioinformatics

Outline

- Examples of the application of bioinformatics to study proteins
- Review: protein sequence
- Protein structure (primary, secondary, tertiary and quaternary structure)
- Mapping "secondary" structure (*defined* by 3D models, *predicted* from sequence)
- Method that predicts secondary structure from amino acid sequence
- Classification metrics: measuring how well bioinformatics work, and secondary structure prediction in particular (sensitivity, specificity, confusion matrix, Q_3)
- Proteins
    - have structure, function, interactions, and subcellular localisation
    - are part of a system
- Analysing biological components
    - in isolation
    - jointly, within system, across systems
- Random networks and scale free networks
- Types of biological networks involving proteins
    - Protein-protein interaction networks
    - Regulatory networks
    - Metabolic pathways/networks
- Examples of data integration, combining system-wide data sets
- Protein interaction and subcellular localisation
- Interaction and expression (multiple conditions)
-

Reflection

- *Why do we want to predict secondary structure, esp. if tertiary structure is unavailable? Use as a stepping stone for tertiary structure prediction; gain insights into biological function of protein by local structural context; improve prospects of identifying homology with structure being more conserved than sequence*
- *What descriptors are used to define structure at primary, secondary and tertiary levels?*
- *Identify secondary structure classes, and describe alpha-helix and beta-strand*

- *What is Chou-Fasman's approach to predict secondary structure? What is their propensity table and secondary structure prediction algorithm*
- *Describe the secondary structure prediction problem (sequence to structure) and the data that need to be presented; list approaches, tactics, limitations and difficulties of predicting secondary structure*
- *What metrics are helpful to evaluate predictions, specifically for secondary structure?*
- *Describe functions of proteins, such as localisation and molecular interaction, and how they could be predicted from sequence and structure*
- *Describe informally what a random network looks like; what a scale-free network looks like; draw examples of these networks and explain why one type is preferred by nature; your explanation should make reference to (a) scale to large number of nodes, and (b) tolerance to failures/robustness*
- *What is a protein-protein interaction network, and how can it be used to predict subcellular localization of proteins?*
- *How can a collection of gene expression data sets from multiple conditions and a protein interaction network (that is unspecific to condition) be used to predict which sets of proteins that interact simultaneously?*
- *Exam questions are appended below*

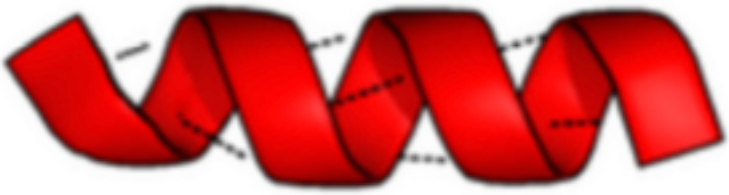
Resources

- Textbook: Zvelebil M & Baum JO (2008) Understanding Bioinformatics, Garland Science
  - Background protein structure and function: Zvelebil and Baum, ch. 2, sec. 2.1-2
  - Protein secondary structure: ch. 11, sec. 11.1-3, 11.4 and onwards for context only
  - Prediction methods: skim ch. 12, sec. 12.1-3, for technical details and context
  - Patterns and function: ch. 4, sec. 4.10 (incl. box 4.7; related material was covered also in L13)
  - Protein conservation: ch. 14, sec. 14.1
  - Protein binding: ch. 14, sec 14.3
- Wikipedia on "scale-free network" https://en.wikipedia.org/wiki/Scale-free_network

Final exam 2020: PROTEIN BIOINFORMATICS

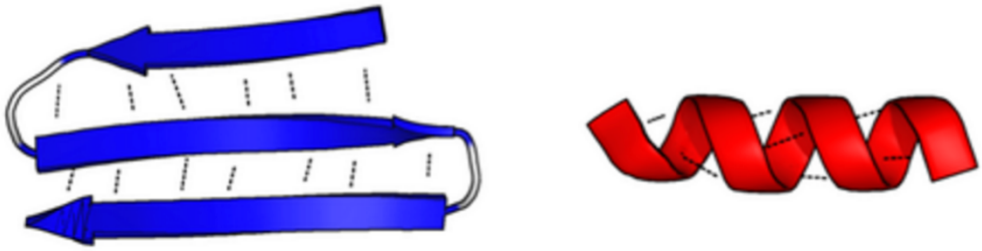
Proteins can be described in terms of their primary, secondary and tertiary structure. The Figure below contains a visualisation of the tertiary structure of a protein fragment. The Table below contains the propensity of amino acids to form part of different secondary structure classes (here labelled S1 and S2).

a. Give the three-letter code for three amino acids with the highest propensity to be part of the fragment (3 marks).
b. Give the three-letter code for the amino acid that has the greatest propensity to introduce a kink in the fragment (1 mark).
c. Spell out the conventional names of S1 and S2, in that order (1 mark).



Amino acid	S1	S2
Ala	1.42	0.83
Cys	0.70	1.19
Asp	1.01	0.54
Glu	1.51	0.37
Phe	1.13	1.38
Gly	0.61	0.75
His	1.00	0.87
Ile	1.08	1.60
Lys	1.16	0.74
Leu	1.21	1.30
Met	1.45	1.05
Asn	0.67	0.89
Pro	0.57	0.55
Gln	1.11	1.10
Arg	0.98	0.93
Ser	0.77	0.75
Thr	0.83	1.19
Val	1.06	1.70
Trp	1.08	1.37
Tyr	0.69	1.40

Final exam 2019: Protein bioinformatics.



A. The schematic structure on the left (in the figure above) is a transmembrane domain, consisting of three alpha strands. True or false? (1 mark)

B. The schematic structure on the right is an alpha helix. True or false? (1 mark)

C. The dotted lines in both structures represent hydrogen bonds. True or false? (1 mark)

D. An alpha helix must be at least 3 residues long. True or false? (1 mark)

E. A beta strand must be at least 10 residues long. True or false? (1 mark)

F. Both proline and glycine tend to support the formation of beta strands; therefore their absence cause kinks in beta strands. True or false? (1 mark)

G. Proline tends to disrupt the formation of alpha helices; therefore it causes kinks in alpha helices. True or false? (1 mark)

H. Hydrogen bonds are strong non-covalent bonds that link amino acids, to form (and define) secondary structure. True or false? (1 mark)

Final exam 2018: Protein bioinformatics

The Chou-Fasman propensity (*P*) values for each amino acid to form an α-helix or a β-strand, respectively is shown in the table below. A larger value denotes higher propensity.

Amino acid	α	β
Ala	1.42	0.83
Cys	0.70	1.19
Asp	1.01	0.54
Glu	1.51	0.37
Phe	1.13	1.38
Gly	0.61	0.75
His	1.00	0.87
Ile	1.08	1.60
Lys	1.16	0.74
Leu	1.21	1.30
Met	1.45	1.05
Asn	0.67	0.89
Pro	0.57	0.55
Gln	1.11	1.10
Arg	0.98	0.93
Ser	0.77	0.75
Thr	0.83	1.19
Val	1.06	1.70
Trp	1.08	1.37
Tyr	0.69	1.40

Based on your understanding of the formation of protein secondary structure and the propensity table, predict the secondary structure class (**α-helix**, **β-strand** or **coil**) at the highlighted position of the following amino acid sequences. You do not need to use Chou-Fasman's algorithm, but similar principles should apply in determining your answers.

(a) 1 mark

His – Lys – Glu – Ile – Cys – Leu – Pro – Ile – Val – Phe – Lys – Asp

(b) 1 mark

Arg – Pro – Met – Ala – Lys – Thr – Gln – Ala – Phe – Cys – Gly

(c) 1 mark

Pro – Gly – Cys – His – Pro – Ser – Tyr – Ala

(d) 2 marks: Justify the predictions for (a)-(c) by explaining what the table is based on and the strategy with which predictions were made. Calculations are not required.