ECS129: Quiz 8
Answers

A. Predicting Secondary structures

Using the Chou and Fasman probability values presented at the right, predict the secondary structure of the following peptide sequences: (Secondary structures are given in one-letter code, and only three states are considered: E = Strand, H = Helix and O = other)

The Chou and Fasman method proceeds in a few steps:

a) build a table in which you write the propensities of all amino acids to be in helices or strands
b) Separately, predict helices and strand:
    a. For helices:
        i. Find a potential nucleation site: a stretch of 6 amino acids with at least 4 having a propensity to be in a helix greater than 1
        ii. Expend the nucleation site in both directions: add an amino acid X, and check if the average helical propensity over a window of 4 amino acids ending at X is greater than 1; if it is, add X to the current prediction, otherwise stop
        iii. Check that the average propensity over the whole region predicted to be helical is greater than 1 (if not, prediction is discarded)
    b. For strands
        i. Find a potential nucleation site: a stretch of 5 amino acids with at least 3 having a propensity to be in a strand greater than 1
        ii. Expend the nucleation site in both directions: add an amino acid X, and check if the average strand propensity of a window of 4 amino acids ending at X is greater than 1; if it is, add X to the current prediction, otherwise stop
        iii. Check that the average propensity over the whole region predicted to be expanded is greater than 1 (if not, prediction is discarded)

In cases in which the same region is predicted to be both helical and strand, pick the prediction with the greatest overall average.

1) **WHGCITVYWMTV**
    A) OOOOHHHHHHOO
    B) EEEEEEEEEEEE
    C) HHOOHHHHHHOH
    D) EHOEEEEEEEEE

We use the Chou and Fasman method: first we build a table for the alpha and beta propensities for each amino acid of the peptide:

We start by writing the propensities:

	W	H	G	C	I	T	V	Y	W	M	T	V
P(helix)	1.14	1.24	0.53	0.77	1.0	0.82	1.14	0.61	1.14	1.20	0.82	1.14
P(strand)	1.19	0.71	0.81	1.30	1.60	1.20	1.65	1.29	1.19	1.17	1.20	1.65

a) Predicting helices

There are multiple possible initiation sites for helices. We can pick the Cter of the sequence: VYWMTV. We try to prolong on the left side:
-TVYW: sum P(alpha) = 0.82+1.14+0.61+1.14=3.71 <4
We can't!
Finally, we compute the average P(alpha) over the peptide VYWMTV:
Sum = 1.14+0.61+1.14+1.20+0.82+1.14=6.05
Average = 6.05 / 6 = 1.01 > 1
The fragment VYWMTV could be helical

b) Predicting strand

We try now to see if it could be a strand. There are multiple initiation sites for strands. We pick again YWMTV. It is easy to see that we can prolong it at least up to C: CITVYWMTV. To prolong further on the left:
- GCIT: sum P(beta) = 0.81+1.30+1.60+1.2 = 4.91 > 4
- HGCI: sum P(beta) = 0.71 + 0.81+1.3+1.6 = 4.42 > 4
- WHGC: sum P(beta) = 1.19+0.71+0.81+1.30 = 4.01 > 4
Finally, we compute the average P(beta) over the whole peptide:
Sum = 1.19+0.71+0.81+1.30+1.60+1.20+1.65+1.29+1.19+1.17+1.20+1.65=14.96
Average = 14.96/12=1.24 > 1
The whole peptide can be a strand

c) Combining the results

Since the average for beta is > average for alpha, the peptide is predicted to be fully extended!

2) **CAENKLDHVRGP**

A) HHHHHHHHHHOO
B) OOHHHHHHHHHH
C) HHHHHHHHHHHH
D) OOHHHHHHHHOO

We use the Chou and Fasman method: first we build a table for the alpha and beta propensities for each amino acid of the peptide:

We start by writing the propensities:

	C	A	E	N	K	L	D	H	V	R	G	P
P(helix)	0.77	1.45	1.53	0.73	1.07	1.34	0.98	1.24	1.14	0.79	0.53	0.59
P(strand)	1.30	0.97	0.26	0.65	0.74	1.22	0.80	0.71	1.65	0.90	0.81	0.62

a) Predicting helices

There are multiple possible initiation sites for helices. We can pick the Nter of the sequence: CAENKL. We try to prolong on the right side:
-NKLD: sum P(alpha) = 0.73+1.07+1.34+0.98 =4.12 > 4

- KLDH: sum P(alpha) = 1.07+1.34+0.98+1.24 > 4
- LDHV: sum P(alpha) = 1.34+0.98+1.24+1.14 > 4
- DHRV: sum P(alpha) = 0.98 + 1.24 + 1.14 + 0.79 = 4.15 > 4
- HRVG: sum P(alpha) = 1.24+1.14 + 0.79 + 0.53 = 3.7

The longest we can go is CAENKLDHVR!
Finally, we compute the average P(alpha) over this peptide:
Sum = 0.77+1.45+1.53+0.73+1.07+1.34+0.98+1.24+1.14+0.79=11.04
Average = 11.04/10 = 1.104 > 1
The fragment CAENKLDHVR could be helical

b) Predicting strand
We try now to see if it could be a strand. There are no initiation sites for strands.

c) Combining the results
The fragment CAENKLDHVR is then helical, and the last two residues are "others", i.e. O, hence response A.

3) **TSPTAELMRSTG**
   A) HHHHHHHHHHOO
   B) OOHHHHHHHHHH
   C) HHHHHHHHHHHH
   D) OOHHHHHHHHOO

We use the Chou and Fasman method: first we build a table for the alpha and beta propensities for each amino acid of the peptide:

We start by writing the propensities:

	T	S	P	T	A	E	L	M	R	S	T	G
P(helix)	0.82	0.79	0.59	0.82	1.45	1.53	1.34	1.20	0.79	0.79	0.82	0.53
P(strand)	1.20	0.72	0.62	1.20	0.97	0.26	1.22	1.17	0.90	0.72	1.20	0.81

a) Predicting helices
There are multiple possible initiation sites for helices. We can pick AELMRS. We try to prolong first on the right side:
-MRST: sum P(alpha) = 1.20+0.79+0.79+0.82 < 4: we cannot prolong on the right side.
We try to prolong now on the left side:
- TAEL: sum P(alpha) = 0.82+1.45+1.53+1.34 > 4
- PTAE: sum P(alpha) = 0.59+0.82+1.45+1.53 > 4
- SPTA: sum P(alpha) = 0.79+0.59+0.82+1.45 < 4: we cannot include the S

The longest we can go is PTAELMRS
Finally, we compute the average P(alpha) over this peptide:
Sum = 0.59+0.82+1.45+1.53+1.34+1.2+0.79+0.79=8.51
Average = 8.51/8 = 1.06 > 1
The fragment PTAELMRS could be helical

b) Predicting strand
We try now to see if it could be a strand. There is one possible nucleation site for strand: LMRST.
It cannot be elongated on the right or on the left. The average P(beta) over LMRST is:

(1.22+1.17+0.9+0.7+1.2)/5=5.3/5=1.03. The fragment could be a strand.

c) Combining the results
Since the average for alpha is > average for strand, the peptide is predicted to be :
OOHHHHHHHHOO

A. Scoring secondary structure prediction

Consider the mini protein of sequence: ALHEASGPSVILFGSDVTVPPASNAEQAK. The actual secondary structure of this protein is known: HHHHHCCCCEEEECCCEEECCCCHHHHH. In the following questions, we assume a 3-state secondary structure definition, with E for strand, H for helix, and C for coil (i.e. not strand nor helix).

4) What would be the Q3 for a fully random secondary structure prediction?
   A) 66%
   B) 33%
   C) 76%
   D) 95%

At each position, a random choice has only 1/3 chance to be correct. So, overall, the Q3 is 33%.

5) A first method for protein secondary structure prediction gives this assignment:
   CHHHCCCCEEEECCCCCEEECCCHHHHHH Give the Q3 value for this prediction:
   A) 66%
   B) 33%
   C) 76%
   D) 95%

Let us write the actual secondary structure on top of the prediction:
HHHHHCCCCEEEECCCEEECCCCCHHHHH
CHHHCCCCEEEECCCCCEEECCCHHHHHH

There are 22 amino acids correctly predicted out of 29: Q3=22/29=76%

6) A first method for protein secondary structure prediction gives this assignment:
   HHHHHCCCCHHHHCCCHHHCCCCCHHHHH Give the Q3 value for this prediction:
   A) 66%
   B) 33%
   C) 76%
   D) 95%

Let us write again the actual secondary structure on top of the prediction:
HHHHHCCCCEEEECCCEEECCCCCHHHHH
HHHHHCCCCHHHHCCCHHHCCCCCHHHHH

There are again 22 amino acids correctly predicted out of 29: Q3=22/29=76%

7) The first of the 2 predictions given above (qeustion 5) is useful, while the second (question 6) is terrible. Based on your answers to the two preceding questions, you would say that:
   E) Q3 is not a good measure of the usefulness of a secondary structure prediction
   F) One of my answers must be wrong, as Q3 is known ot be a useful measure of the quality of a secondary structure prediction method
   G) Q3 only depends on the length of the protein considered and therefore cannot be discriminative when the proteins have the same lengths
   H) There must something wrong in this question, as the prediction given in question 6 is closer to the real solution

Q3 is an overall measure and does not indicate if secondary structures have been correctly predicted locally.

8) Which of the following statement is more likely the reason that Chou and Fassman is not more successful?
   A) Chou and Fassman is too old a method
   B) Chou and Fassman does not take into account the solvent
   C) Chou and Fassman is too simple
   D) Chou and Fassman does not take into account well enough non local interactions

D is the most likely answer.