Secondary Structure Preferences

- Amino-acids have distinct preferences for secondary structure (alpha-helix/beta-sheet)
- Secondary structure prediction algorithms are biased towards the over-prediction of secondary structure

Information Theory

- The GOR secondary structure prediction algorithm was based on formal information theory

- This model considered an input sequence of symbols which in various combinations contained information about some output property, in this case a secondary structure symbol sequence

- Note that whatever method is used to extract information from DNA or protein sequences they are all limited to the same total information present, i.e. all programs use the same sequence database

Information Theory Model

$R_1R_2R_3 \ldots\ldots\ldots\ldots\ldots\ldots R_{n-1}R_n$ e.g. MKYW...

↓

$S_1S_2S_3 \ldots\ldots\ldots\ldots\ldots\ldots S_{n-1}S_n$ e.g. HHHH....

- The amino-acid sequence is transformed by some unknown function to a secondary structure sequence

Information Theory

- References:

  B. Robson, Biochem. J., 141, 853-867, 1974
  B. Robson, Biochem. J., 141, 869-882, 1974
  B. Robson, Biochem. J., 141, 883-897, 1974
  B. Robson, Biochem. J., 141, 899-904, 1974

  J. Garnier, D.J. Osguthorpe, B. Robson, J. Mol. Biol., 120, 97-120, 1978

INFORMATION THEORY

$I(x; y_1, y_2, y_3 \ldots.)$

where y_1, y_2, y_n are events that contribute to information about event x

Specifically for protein secondary structure prediction:

$I(S_j; R_1, R_2, \ldots. \ R_n)$

where R_1 is the type of residue 1 etc.

i.e. the amino-acid sequence AGDE...

S_j is the secondary structure state i.e. H (helix), E (strand), C (coil) of residue j

Information Theory

Types of information:

$I(S_i; R_j)$ - The information residue R_j carries about residue i being in state S alone

If j = i this is the information a residue itself carries about its secondary structure

NOTE j can be any index – it does not have to equal i

e.g. I(H;A) – the information Alanine carries about the Alanine residue being in the helical state

Information Theory

$I(S_i; R_j\ R_k)$ - The information two residues carry about residue i being in state S (pair information) – again note j does not have to equal k or equal i or both, and j and k can be different by any integer - pair information

$I(S_i; R_j R_k R_l)$ - The information three residues carry about residue i being in state S - triple information

Information Theory

Strictly all we really know in protein folding is:

$I(S_i; R_1R_2R_3..........R_{n-1}R_n)$ for i = 1 to n

i.e. the secondary structure state of residue i depends on information from all residues in the protein sequence

Information Theory – GOR

The GOR algorithm is simply given as follows:

$I(S_i; R_j)$ $\quad$ $j = i - 8$ to $i + 8$

j = i is the basic information a residue carries about itself

One of the crucial features of the GOR algorithm was the use of something we called directional information, the j = i-8 to i+8

This introduced information from neighbouring residues – but notice this information does NOT depend on the type of residue at position i

INFORMATION THEORY

Breaking down the information:

$I(x; y_1\, y_2) = I(x; y_1) + I(x; y_2 \mid y_1)$

where $I(x; y_1)$ is the information event y_1 carries about event x alone

$I(x; y_2 \mid y_1)$ is the information event y_2 about the event x given y_1

Determination of Information

$$I(x;y) = \ln \frac{P(x|y)}{P(x)}$$

where I(x; y) is the information event y carries
about x
P(xly) is the conditional probability of event
x given y
P(x) is the probability of event x

ESTIMATION OF INFORMATION

P(xly) can be estimated in many ways:

Simple frequency:

P(HlA) = f(Alanine is helical) / f(Alanine)
P(H) = f(Helical) / f(all residues)

f denotes frequency of occurrence

For low frequencies Bayesian statistical theory was used to estimate the probabilities in the GOR algorithm

Prediction algorithm

- Estimate information for each secondary structure state at each residue position independently – possibly use smoothing to reduce fluctuations

  $I(S_i) = I(S_i; R_i) + I(S_i; R_j) + I(S_i; R_k\ R_l)$ ....

  $I(S_i) = (1/n_s) \sum I(S_{i+k})$ $\qquad$ $k = -n_s/2$ to $+n_s/2$

- For each secondary structure state choose a decision constant (arbitrary parameter) at which value if the information value from above is greater than this we predict state S

  $\mathrm{MAX}(I(S_i) - S_D) > S_i$ prediction

Secondary Structure Accuracy Measures

- Fraction correct:
  Count each correct assignment for each state to give fraction correct each state (1.0 would be perfect prediction)
  F_H = count H correct / count H

  $F_{correct} = (1/N_S) \Sigma F_S$
- SOV – Segment Overlap

$$SOV(S) = \frac{1}{N(S)} \sum_{S_S} \frac{MINOV(S1;S2) + DELTA(S1;S2)}{MAXOV(S1;S2)} * LEN(S1)$$

Secondary Structure Accuracy Measures

- SOV – Segment Overlap

```
Observed: CCCHHHHHHHCCCCC.....
Predicted: CCCCCCCHHHHHHHHHHHCC......
S1:                     |                |
S2:                              |                         |
MINOV:                           |       |
MAXOV:                  |                                  |
```

where S defines current secondary structure
    state (H, E etc.)
    S1, S2 are observed and predicted segments of the sequence that overlaps
    LEN(S1) is length of observed segment
    MINOV(S1;S2) is the length of residues that overlap with identical secondary structure
    MAXOV(S1;S2) is the total length of overlap segments

Secondary Structure Accuracy Measures

- SOV – Segment Overlap
  where DELTA(S1;S2) is a fudge factor for allowing errors at the N and C termini of segments

$$\mathrm{MIN}\{(\mathrm{MAXOV}(S1;S2)-\mathrm{MINOV}(S1;S2));\mathrm{MINOV}(S1;S2);\\ \mathrm{INT}(\mathrm{LEN}(S1)/2);\mathrm{INT}(\mathrm{LEN}(S2)/2)\}$$

$$N(S) = \sum_{S_S} \mathrm{LEN}(S) + \sum_{S_S'} \mathrm{LEN}(S')$$

where S_S is count of residues in segments which overlap in secondary structure S

S_S' is count of residues in segments observed in secondary structure S that do not overlap with any predicted segment

Secondary Structure Accuracy Measures

- SOV – Segment Overlap

  Note that the measure described here is NOT symmetrical to switching observed and predicted

  Why use such a complicated measure?

  Because fraction correct measures can be confused by simple prediction algorithms such as all residues are helical – for helical proteins in which 80-90% of the residues are helical this algorithm is 80-90% correct!!

Problem V

- To create your own version of the GOR algorithm
- Choose what type of states to include:
  two (secondary/no secondary)
  three (H/E/C)
  four (H/E/T/C)
- Choose what type of information to be included:
  single/pair/triple residue
  directional information
  your own idea

Problem V

- Using the supplied database create values for the information measures to be included Check the information measures are statistically valid or are derived using an estimator suitable for low frequency data if necessary
- Compute the SOV measure of secondary structure correctness to check how well your prediction algorithm has done.
(Split database estimate + test, jackknife)