00.904 AST Jan 2008

This topics here are examples of what might be asked in the exam.

* I have bacteria growing in the laboratory. I want to estimate the rate of mutation per generation. I count mutations after 1000 generations. I divide by the number of sites (base-pairs) and divide by 1000. Why is this fundamentally wrong ?

* Write down an expression for calculating evolutionary time
based on the number of observed mutations
using a k for the constant of proportionality
based on the Jukes-Cantor model

* I claim that you can estimate evolutionary time from $t = -k \ln\left(1 - \frac{4}{3}\frac{n_{mut}}{n_{res}}\right)$ where k is some constant, n_{mut} is the number of mutations seen within n_{res} base-pairs.

* This applies to DNA, but what would I have to change to have a similar expression for proteins ?

* Population b has evolved from a for a time t_b and shows n_b mutations in n_{res} base-pairs. Population c has evolved for time t_c and shows n_c mutations. What is the ratio, $\frac{t_c}{t_b}$?

* Describe a probabilistic measure for deciding if two aligned protein (or DNA) sequences are really related or if the similarity is purely a matter of chance.

* I want to compare drug-like molecules, but they have different shapes and different number of atoms. Describe a method for measuring the similarity of any two molecules.

* I have collected information on organisms. My descriptors are
* number of (legs + arms) N_l
* length of digestive tract in meters L_d
* mass of body hair (kg) m_h
* average number of children per generation N_c

(a) write down a formula for Manhattan distance measure for comparing two kinds of organism
(b) write down a formula for a Euclidean distance measure comparing two kinds of organism

* I have a collection of data points and a similarity measure between every pair of data points. Write down in words and pseudocode how a k-means clustering method would put the objects into clusters. Use a diagram if it is helpful.

* Describe in words, pseudocode and a diagram if necessary an agglomerative cluster method.

* You have points in a two-dimensional space. Draw a diagram of points which would not be suitable for a joining cluster method.

* Some cluster algorithms require that you can measure the similarity of clusters as well as individual points. Describe two sensible measures of cluster similarity.

* Name two criteria one could use to recognize secondary structure. Draw a diagram.

* I predict the secondary structure of a region of polypeptide to be

```
---EEEHHEEE--
```

Where E means β-strand and H means α-helix. What is obviously wrong with the prediction I have made ?

* What is the fundamental reason that one might expect secondary structure prediction in proteins to be successful ?

* In a simple forwards-feed, backwards propagation neural network, the response ($output_i$) of a unit is given by $output_i = \frac{1}{1+e^{-input_i}}$ Draw a diagram which shows the shape of the output as a function of input. Label the axes. On the *y*-axis, add numbers so one can see the range of the output.

* A neural network usually has numbers as its input. A protein sequence is a series of characters (amino acid types). How would you encode a protein sequence so it is suitable for a neural network ?

* My neural network has 10^7 hidden nodes and, during training, appears to produce better results than a network with 100 hidden nodes. What is the danger of too many hidden nodes ?

* Neural networks were originally used to predict protein structure using only the sequence of the protein in which one is interested. What is a major improvement that could be made to this input information ?

* Give a piece of empirical evidence that neural networks using only protein sequence will probably never be completely successful in predicting protein secondary structure.

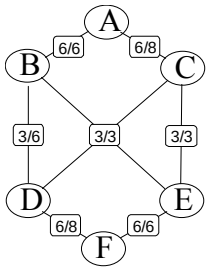
* Describe and use a diagram to explain how protein domains may be recognised using sequence information, even when the structure is not known.

* Protein domains mean different things to different people. Give two different and reasonable definitions.

* One example domain recognition technique uses a hierarchical clustering method. What is the similarity measure used ?

* Early protein domain recognition methods tried to cut a protein structure at a single position in order to split a protein into two compact domains. Draw a diagram of a protein structure where this would not be adequate.

* I would like to implement a domain recognition program based on the network flow



algorithm. I have a graph where the flow goes from A to F. Each edge is marked with numbers *x*/*y*, where *x* is the flow through the edge. *y* is the capacity of the edge. I propose to cut the graph at the edges DF and EF. How can you quickly say that this is wrong.

* Draw a diagram of a protein. Say what edge node is and explain in words and pseudocode how the Taylor / Ising spin approach could be used for finding protein domains.

* What would be a typical size for a protein domain in terms of the number of amino acid residues ?

* Would you say that the "gene ontology" method for classifying protein function is more flexible than using "EC" enzyme classification numbers ? Why ?

* I have discovered a sequence motif `ACDXX[WY]` which is present in all enzymes of the sulphur-happyase class. I claim I can look for this motif and recognise more proteins of the same type. What should I do to see if this is a reasonable claim ?

* I have two proteins. One has a well-defined function and is well studied. Nothing is known about the function of the second protein. I want to make a functional prediction for the second protein, based on the first. Would it be more helpful to find sequence or structural similarity ? Explain why.

* Maybe one would like to find functional similarities between proteins by recognising 3-dimensional motifs. Describe a method which would be able to recognise similar arrangements of chemically important entities. You do not have to repeat any method from the lectures. You can describe any reasonable approach. You must specify some tractable method for representing relationships within the protein. You must specify a practical method for finding similarities between proteins.

* Why might 3-dimensional motifs let you find functional similarities which would not be recognised by sequence-based motifs ?

* In the problem of protein sequence optimisation, how large is the potential search space, as a function of protein size (N_{res}) and the number of types of amino acid (N_{type}) ?

* What might make the search problem easier than considering all possible sequences ?

* You have a function which tells you how well a protein sequence fits to a structure. It acts as if it were an energy function $E(S)$ of the sequence S so a more negative energy is more favourable. Write in pseudocode a Monte Carlo method for simulating a walk over possible sequences at constant temperature.

* You have a quasi-energy energy function $E(S)$ of the sequence S. It contains two terms, $E^1(S)$ and $E^2(S)$ which correspond to the interactions of an amino acid with its fixed environment (E^1) and pairwise interactions (E^2).
Describe a method with pseudocode to find the best possible score a residue of type *a* could have at a position *i* in the sequence.

* Why is it not adequate to use an energy or free energy function as the score for protein sequence optimisation ?

* Explain why protein sequence optimisation is a discrete optimisation problem. Give an example of a continuous optimisation problem.

* I develop a method for finding the best energy sequence for a protein structure. Why may it not be important to find the single best sequence ?

* I have successfully described an optimisation method (maybe Monte Carlo) which lets me find a good sequence for a given protein structure. How would you generalise the method to handle arbitrary rotamers ?

* Describe in words and a diagram why the methods used in the program "psi-blast" are able to find more remote homologues than those used in "blast".

* Explain why a good amino acid substitution matrix is more important for comparing remote protein homologues than for similar proteins.

* I want to align a protein sequence to a protein structure. Why can I not use a simple scoring scheme as in sequence alignments ?