

Hints for solving problems

Chapter 1.

Problem 1.1. (a) Note that the sixth column has all large non-polar residues (F, with one Y). Consider the distribution of positively-charged residues, such as R and K. Many columns have a few R and K residues but some, notably near the C-terminus, are rich in R and K.

(b) For example, L is absolutely conserved in the third column. Column 7 also has conserved L but with one exception. (c, d) Look for periodicities of approximately 4, suggesting an α -helix, or 2, suggesting a strand of sheet. (e) A cluster of positively-charged residues should bind a negatively-charged ligand. A nucleic acid would be a reasonable suggestion.

Problem 1.2. For example, 1.10a is helical, 1.10e and f are sheet.

Problem 1.3. Keep as much of the core of the code as possible. Apply it to the original sequence as in the original program, and to the original sequence with 1 character removed from the beginning, and with 2 characters removed from the beginning. Create the reverse complement by processing the original string character by character and repeat the steps in the previous sentence.

Problem 1.4. The straightforward way is to read the sequences in and process each pair on a column by column basis. In reading in be sure to eliminate the names of the species and the residue numbers from the text. If the two sequences are in variables \$a and \$b, one way to process them would be to use the chop command in perl:

```
$number_of_mismatches = 0;
while (length($a) > 0) { if (chop($a) ne chop($b) ) {$number_of_mismatches++;}}
```

Problem 1.5. Try them! Obviously (c) is the hardest.

Problem 1.6. The main difficulty is to understand why the given program does not work in cases where it fails. Do a flowchart of the program. The problem with example (c) is that the program is designed under the assumption that there is only one way to assemble partial overlaps at any stage.

Problem 1.7. Given a probe string \$motif and a sequence to be searched \$sequence, finding exact matches is easy: \$sequence =~ /\$motif/e where the e suffix means that the pattern to be matched is an expression. To find matches with one allowed error, replace each position in the \$motif with a 'wild card' ? character. (? matches any single character.) That is, if you want to find the word 'match' with one error, search for '?atch', 'm?tch', 'ma?ch', 'mat?h', and 'matc?'.

Problem 1.8. For example, the first two executable statements in the concise version read the data and create the array of fragments. They correspond to lines 4-9 of the long version. Reasonable comments might be:

```
# input of data.  create array each element of which contains one line of data.
$/ = "";          # signal to read in full paragraph at a time
@fragments = split("\n", <DATA>);  # split lines in paragraph into successive
# elements of an array
```

Chapter 2.

Problem 2.1. The original analysis leading to the conclusion is by Sturtevant in 1913.. Now it is possible to check the gene sequences to correlate positions in the genetic map with positions on a chromosome.

Problem 2.2. (a) For instance, for *M. genitalium*, the gene density is 468 genes/580.07 kb = 0.81 genes/kb.

Problem 2.3. Estimate the total number of bases required and compare with the size of the human genome.

Chapter 3.

Problem 3.1. Profit = 5800×price – (sum of costs) = 1.05×(sum of costs); add up costs and solve for price

Problem 3.2. Identify all amino acids that have volumes over the stated threshold. Identify all amino acids that have distal carboxy or amide groups on their sidechains. Identify which amino acids are common to both lists, and draw as Venn diagram.

Problem 3.3. (a, b) The mistakes made by the program are to count village as a noun rather than an adjective, and smithy as an adverb rather than a noun. The sense it makes of the second clause is that the village is standing, and smithy is an adverb describing how the village is standing. (c) It is probably easier to find examples of village used as an adjective than of smithy used other than as a noun. (d) Consider the sentence: ‘The village sleepy rests.’ This is understandable English, and in this case village is a noun and sleepy an adverb. This sentence does not mean quite the same thing as ‘The sleepy village rests.’ (e) Compound adjectives require a hyphen, as in ‘amino-acid sequence’ or ‘DNA-binding protein’. Therefore if ‘spreading’ modified ‘chestnut’ rather than tree, the sentence should be written ‘Under the spreading-chestnut tree ...’ Compare ‘Under the sugar-maple tree...’

Chapter 4.

Problem 4.1. If all ATP-dependent DNA helicases are DNA-binding proteins, there would have to be a connected path from DNA-binding proteins (the more comprehensive category) to ATP-dependent DNA helicases.

Problem 4.2. No alternative but just to count them!

Chapter 5

Problem 5.1. If you were to delete two t residues it would be a perfect palindrome.

Problem 5.2. (a) The difference is in the initialization. (b,c) The solution with no gaps internal to the motif atg would score the highest.

Problem 5.3. First identify the positions of the turns by inspection of Fig. 5.6. An example of a turn not corresponding to a region containing insertions or deletions appears near residue 30.

Problem 5.4. Two possibilities are (a) determine the sequence similarities of all pairs of sequences, and retain only one example of any set of sequences for which all pairs have high mutual similarity. (b) determine the sequence similarities of all pairs of sequences, and weight each sequence by 1 divided by the number of sequences that have high mutual similarity.

Problem 5.5. Basic steps: (1) Read alignment table (2) Process alignment table column by column, making inventory, (3) To score a distribution of amino acids in the inventory against a single amino acid in a query sequence, either take the minimum score of the query amino acid with any amino acid that appears in that column, or take a weighted average of the scores.

Problem 5.6. (a) One way is to take a running 5-character window from the first sequence and make an associative array with the five-character sequence as argument. Then take a running 5-character window from the second sequence and check whether each 5-character sequence appears as an argument in the associative array. (b) For this you will have to record where in the first sequence each 5-character substring appeared.

Problem 5.7. You are being asked to produce something like Figure 5-3, with specific character strings.

Problem 5.8. Produce individual frames, convert to gif format, and collect into a movie.

Problem 5.9. Draw a structure analogous to a dotplot, except that the original sequence should appear along the rows, and the reverse complement should appear down the columns. A diagonal series of matches in this plot corresponds to a pair of complementary regions.

Problem 5.10. For each position of each die, there are only four possible successor positions. Use the perl rand function to choose the successor, and keep track of the 'trajectory' of the system.

Problem 5.11. Basic idea: if there are two distinct ways to get from node A to node B, then go from A to B along one path and back from B to A along the reverse of the other.

Problem 5.12. (a) Just draw them out. Not as bad as it seems. Remember that you can't retrace any step. (b) For instance, there are 3 paths from Start to B (Northeast-Northeast-Southeast, Northeast-Southeast-Northeast, and Southeast-Northeast-Northeast). (d) There are $6 \times 5 \times 4$ ways to make successive choices that assign 3 left turns to six steps, For each assignment, it could have been arrived at by choosing those three steps in six different orders.

Problem 5.13. The example in question is Ex. 5.6 on page 289. There are several possibilities. For instance, the immediate ancestor of ATCC and ATGC could be either ATCC or ATGC. The choice at this point will influence choices at other nodes.

Problem 5.14. The left tree in part (a) would be written: $(A(((EC)D)(BF)))$

Problem 5.15. Give both an introductory statement of what is going on and what the method is. Also end each line with a comment, beginning with #, stating what that line accomplishes.

Problem 5.16. The program should create strings representing the tree in parenthetical form (see problem 5.14).

Problem 5.17. For A—B, UPGMA tree give distance: $2.66 + 2.66 + 1.25 + 0.5 = 7.07$. The split decomposition diagram gives 4.

Problem 5.18. Generalize the answers to Problem 5.16 and 5.14, keeping a number associated with each left and right parenthesis.

Chapter 6

Problem 6.1. (a) Count the number of identical positions in each pair and report minimum and maximum. The first two sequences have 23 identical positions. (b) Use 50% identical amino acids to the first sequence as the threshold.

Problem 6.2. You could do this with a network with 8 nodes in the input layer, all feeding into one node on the second layer, the output of which is the output from the network.

Problem 6.3. Consider (a) allowing mismatches from the PPHPPHH pattern, (b) looking for periodicities of 4 in the hydrophobicity.

Problem 6.4. (a) The network at the bottom of page 341 is an example of a network that produces a logical AND. (b) You could change that network to OR logic by changing the criterion for output from sum of inputs > 2 to sum of inputs > 0 . (c) Cannot be done with one layer.

Problem 6.5. To change colours in postscript print the statement `r g b setrgbcolor` where `r g` and `b` are the relative values of red, green and blue. Thus `100 = red, 010=green, 001=blue, 011=cyan, 101 =magenta.` (There must be a space or spaces between the `r, g, b` values and before the word `setrgbcolor.`) Each time the protein prints an amino acid one-letter abbreviation, look up the appropriate colour and print the appropriate `setrgbcolor` statement..

Problem 6.6. Map the sequence into a two-dimensional set. That is, associate `x` and `y` coordinates with each residue, by converting cylindrical polar coordinates to `x` and `y`. In fact, associate two sets of `x` and `y` coordinates with each residue corresponding to the two copies of the helix surface network. Connect neighbouring hydrophobic residues. Find large connected subsets. One way to do this is to assign a different ‘colour’ to each residue, and whenever two residues are connected, change the colour of one so that the entire connected set has one colour. Then for each colour count the number of residues with that colour.

Problem 6.7. For example, residues 27 and 28 are predicted correctly because they are within the experimental helix 27-29. Residue 29 is not predicted correctly because it is in fact helical.

Problem 6.8. Bonneau, Tsai, Ruczinski and Baker also predicted a helix starting at residue 22. They got residue 29 right.

Problem 6.9. Assign a variable to each node. Given any input (`x` and `y` coordinates), write a succession of test statements that assign output values to each node. What is crucial is the order in which these tests are executed; they must be in order of the layers of the network.

Problem 6.10. (a) There are $M - N + 1$ positions at which to align the first residue of the shorter sequence with a position of the longer. If the first residue of the shorter sequence is aligned with the `k`th residue of the longer, it is possible to insert $(N - k - M)$ gaps into the shorter sequence. How many ways are there to do this? (b) Repeat calculation in (a), recounting the number of possible gaps.

Problem 6.11. Use the perl `rand` function. Run the calculation several times, and plot the accuracy against the number of points.

Problem 6.12. As $X \rightarrow \infty$, $\exp(-X) \rightarrow 0$, so the $\exp(-X)$ terms ‘drop out’. (b) Replace ‘output = 1 if $x > 0$, else output 0’ with ‘output $1/(1 + \exp(-x))$ ’. Replace ‘output = 1 if $x > 2$, else output 0’ with ‘output $1/(1 + \exp(-(x-2)))$ ’. Ask: will every point that is accepted as inside the triangle by the original network also be accepted by the new one? Consider points on the boundary.

Problem 6.13. (a) Count the number of identical residues. Is the percentage of identical residues above the 'twilight zone'? (b) The termini are at the upper right of the picture. Which is which? The C-terminus has a cysteine adjacent to the terminal residue. It forms a disulphide bridge, shown in this representation by a double-lollypop.

Chapter 7

Problem 7-1. (a) Number of positions = number of rows \times number of columns. (b, c) Count number of red and green spots. (d) Count number of yellow spots. (e) Count positions that are not red and not yellow and not green.

Problem 7-2. (a) $k_{\text{off}} = k_{\text{on}} \times K_D$. (b) $t_{1/2} = 0.693/k$

Problem 7-3. Strains (a) and (c) do not feel effects of isoniazid. (c) in strain c the drug is not activated.

Problem 7-4. (a) $k_{\text{off}} = k_{\text{on}} \times K_D$. (b) You should draw conclusions from the values calculated in (a). Note however that the k_{on} rates are all equal to within a factor of 10, but the differences in K_D are much greater.

Problem 7.5. For instance, the number of neighbours of Oxford Circus is 6.

Problem 7.6. No alternative but to count them.

Problem 7.7. (a) The sum of $C \exp(\alpha k) = C[\exp(\alpha) + \exp(\alpha)^2 + \exp(\alpha)^3 + \dots] = C \exp(\alpha) / [1 - \exp(\alpha)] = 1$. Solve for C.

(b) Knowing C from part (a), compute successive values of $C \exp(\alpha k)$ for $k = 1, 2, \dots$ and stop when the value becomes < 0.01 (d) Hint given in text. (e) Substitute $\alpha = 0.8$ into answer to (d) (f) If M is the median, then the sum of $C \exp(-\alpha k)$ from 1 to M is ≤ 0.5 but the sum from 1 to M+1 is ≥ 0.5 . The sum of $C \exp(-\alpha k)$ from 1 to M is $C \exp(-\alpha)(1 - \exp(-\alpha M)) / (1 - \exp(-\alpha))$.

Problem 7.8. Two forks in succession would do it.

Problem 7.9. With 5 yes-or-no questions one could distinguish 32 different items. So 5 is enough to distinguish 24 letters, but 4 yes-or-no questions would not be sufficient.

Problem 7.10. For instance, CII activates CI. Therefore add a node corresponding to CII and an arrow from it to CI.

Problem 7.11. For instance, if mutant cro failed to bind OR3, the repression indicated by the link between cro and CI would be lost.

Problem 7.12. Glycolysis and gluconeogenesis are largely but not entirely reverse sequences. Therefore the dotplots will show a diagonal signal going from lower left to upper right. That is opposite to the usual upper left to lower right signal one sees in dotplots of similar sequences.

Problem 7.13. For instance, from the entry 7 5 in row 2 (second and third columns) it would be possible to generate 9 2, 5 7, or 4 8. To generate 5 7 would not be in the shortest path because it generates a combination that already occurs.