Introduction to Bioinformatics – Weblems

Weblems – Chapter 1

Weblem 1.1 What institution currently has the highest sequencing throughput power? How many Human genome equivalents of raw sequence per week can this institute produce?

Weblem 1.2 What are the differences between the standard genetic code and the vertebrate mitochondrial genetic code?

Weblem 1.3 Pick a topic from the list of Recommended Readings in this chapter. Prepare a list of web sites that cover the same material, at a tutorial level.

Weblem 1.4 The PERL program in Case Study 1.1 translated a DNA sequence into an amino acid sequence. Find on the web, download and install, a PYTHON program to carry out the translation, in all reading frames (as in Problem 1.1, and run it on the test data from the Case Study.)

Weblem 1.5 The PERL program in Case Study 1.1 translated a DNA sequence into an amino acid sequence. Find a web server that will carry out the translation, in all reading frames (as in Problem 1.1, and run it on the test data from the Case Study.

Weblem 1.6 Create a web site describing the background of the original discovery of the CRISPR/Cas system, the demonstration that it could be extended to mammalian systems, comments on the applicability to clinical practice in humans, animals, and plants, and to ecology (for instance, 'gene drive'). Include references to the original literature, and recent developments. Include also a section on the discussion of feasibility and ethics of applicability to human embryos, describing the work of He Jiankui, and public comments and governmental restrictions on use of CRISPR in humans. (An article that appeared as this book went to press is: Lander, E.S., Baylis, F., Zhang, F., Charpentier, E., Berg, P., Bourgain, C., Friedrich, B., Joung, J.K., Li, J., Liu, D., Naldini, L., Nie, J.B., Qiu, R., Schoene-Seifert, B., Shao, F., Terry, S., Wei, W. & Winnacker, E.L. (2019). Adopt a moratorium on heritable genome editing. Nature 567, 165–168.) Your target audience should be people with a scientific background but not necessarily specialists in genetic engineering.

Weblem 1.7 The same challenge as in the previous weblem, but for a target audience of ordinary newspaper readers not necessarily with any scientific background: 'the man on the Clapham omnibus'.

Weblems – Chapter 2

Weblem 2.1 Which of the following families have genes appearing in tandem arrays in the human genome and which have genes dispersed among several chromosomes? (a) Actin, (b) tRNA, (c) all globins, (d) HOX genes, (e) the major histocompatibility complex.

Weblem 2.2 The mutation causing sickle-cell anaemia is a single base change A→T, causing the change Glu→Val at position 6 of the β chain of haemoglobin. The base change occurs in the sequence 5′-GTGAG-3′ (normal) → GTGTG (mutant). What restriction enzyme has been used to distinguish between these sequences, to detect carriers? What is the specificity of this enzyme?

Weblem 2.3 What are the sequence specificities of the following restriction enzymes? Which are palindromes? Which produce cleavage products with blunt ends? sticky ends? (a) Esp4I (b) FnuCI (c) FnuDI (d) Bst28I (e) BmgBI (f) What would be the products of each of the 5 enzymes acting (separately) on:

5′-`tgctgccttaagcccgtgatccccaggcctcggctatcgatgtctcacgttcag`
3′-`acgacggaattcgggcactaggggtccggagccgatagctacagagtgcaagtc`

(g) What would be the product of a mixture of the 5 enzymes acting on the fragment?

Weblem 2.4 It is possible that you solved the previous problem using pencil-and-paper. Now write a program to accept as input (1) a series of cutting sites of restriction enzymes, and (2) a substrate sequence, and use it to determine the answers to parts (f) and (g) of the previous problem. Next extend the program so that it interfaces with the database of restriction sites `http://rebase.neb.com/rebase/rebase.html` so that the program could accept as input the names of one or more restriction enzymes, plus the substrate sequence, and print as output the same results. (Comment: the most straightforward way to do this is (1) to download the database file, which is easily parsable. As an alternative approach: (2) the command `wget --post-data "enzname=EcoRI" http://rebase.neb.com/cgi-bin/reb_get.pl` will return an HTML file corresponding to EcoRI, or any other enzyme name you insert at that position in the command. Conversion a text file with html2text will produce a file from which it is easy to extract the Recognition Sequence. Which procedure would you recommend? Under what circumstance – for a different problem – would approach (2) be preferable?)

Weblem 2.5 Show how you could generate the same output using a webserver; for instance, `http://tools.neb.com/REBsites/` or `http://rna.lundberg.gu.se/cutter2/` or `http://www.restrictionmapper.org/` or `http://www.justbio.com/index.php?page=cutter`

Weblem 2.6 Draw the generalized suffix tree for the sequences `atacgacgt` and `acgttcacg`. Show how you would deduce from the suffix tree that these sequences overlap:

```
atacgacgt
     acgttcacg
```

Proceed by (a) writing a program yourself, or (b) importing a program (from, for example, `https://pydigger.com/pypi/suffix-trees` or `https://bibiserv.cebitec.uni-bielefeld.de/mkesa/`) or (c) using the webserver at `http://guanine.evolbio.mpg.de/cgi-bin/drawStrees/drawStrees.cgi.pl`

Weblem 2.7 Write a web server that will accept an RNA sequence and return the longest predicted perfect hairpin (perfect = no mismatched bases). Make use of as much available software as you can.

Weblem 2.8 Harder version of previous problem: Write a web server that will accept an RNA transcript (no base modifications) and detect whether it looks as if it contains any regions that encode tRNAs.

Weblems – Chapter 3

Weblem 3.1 Of the taxa appearing in Figure 1.3, are there any groups that do *not* contain any species for which a full-genome sequence has been determined? Indicate which ones, and give an example of a species that might be sequenced to complete the picture.

Weblem 3.2 Find examples of additional completely sequenced eukaryotic genomes not listed in Table 3.1. Find at least two in each of the categories: Mammals, Other chordates, Higher plants, Other eukarya.

Weblem 3.3 What is the size of the genome of the aardvark, *Orycteropus afer?*

Weblem 3.4 Which insect has the largest genome? Which insect has the smallest genome? How many base pairs in each of these species?

Weblem 3.5 What mammal has the smallest genome?

Weblem 3.6 In the human, 1 cM $\sim 10^6$ bp. In yeast, approximately how many base pairs correspond to 1 cM?

Weblem 3.7 What chromosome of the cow contains a region homologous to human chromosome region 8q21.12?

Weblem 3.8 Find three examples of mutations in the CFTR gene (associated with cystic fibrosis) that produce reduced but not entirely absent chloride channel function. What are the clinical symptoms of these mutations?

Weblem 3.9 Find an example of a genetic disease that is: (a) Autosomal dominant (b) Autosomal recessive (other than cystic fibrosis) (c) X-linked dominant (d) X-linked recessive (e) Y-linked (f) The result of abnormal mitochondrial DNA (other than Leber's Hereditary Optic Neuropathy).

Weblem 3.10 Berardinelli-Seip syndrome is an inherited disease in which fatty tissue does not form. Macroscopic symptoms include a distinctive physical appearance characterised by absence of fatty tissue, but overgrowth of muscle containing the fat. High levels of fats circulating in the blood can lead to insulin resistance, causing diabetes. Buildup of fats in liver and heart damage these organs. There are several possible sources of Berardinelli-Seip syndrome. One is a defect in the gene AGPAT2. In Ensembl, search the human genome for the gene AGPAT2. (a) At what chromosomal location does it occur? (Which arm of which chromosome?) (b) What is the length in basepairs of the AGPAT2 gene? (c) What is its intron-exon structure? (d) Report the first 20 bases and the last 20 bases of the gene; report the first 20 bases of the first intron. (e) How many transcripts of the normal gene are known? (f) What is the length of the longest protein translated from the normal gene? (g) What is the function of this protein? (h) How was this function identified? What is the nature of the evidence for it? (i) What nucleotide sequence variations are observed in people showing Berardinelli-Seip syndrome? (j) What is the consequence, in the protein sequence, of these variations?

Weblem 3.11 Table 3.2 contains the statistics of current status of genome projects, as of March 2018. What are the current numbers?

Weblem 3.12 What is the chromosomal location of the human myoglobin gene?

Weblem 3.13 Box 3.4 shows the duplications and divergences leading to the current human α and β globin gene clusters. (a) In which species, closely related to ancestors of humans, did these divergences take place? (b) In which species related to ancestors of humans did the developmental pattern of expression pattern ($\zeta_2\epsilon_2$ = embryonic; $\alpha_2\gamma_2$ = foetal; $\alpha_2\beta_2$ = adult) emerge?

Weblem 3.14 (a) How many predicted ORFs are there on *Saccharomyces cerevisiae* chromosome X? (b) How many tRNA genes?

Weblem 3.15 (a) What is the normal function of the protein that is defective in Menke disease? (b) Is there a homologue of this gene in the *A. thaliana genome?* (c) If so, what is the function of this gene in *A. thaliana?*

Weblem 3.16 Duchenne muscular dystrophy (DMD) is an X-linked inherited disease causing progressive muscle weakness. DMD sufferers usually lose the ability to walk by the age of 12, and life expectancy is no more than about 20-25 years. Becker muscular dystrophy (BMD) is a less severe condition involving the same gene. Both conditions are usually caused by deletions in a single gene, dystrophin. In DMD there is complete absence of functional protein; in BMD there is a truncated protein retaining some function. Some of the deletions in cases of BMD longer than others that produce BMD. What distinguishes the two classes of deletions causing these two conditions?

Weblem 3.17 What mutation is the most common cause of phenylketonuria (PKU)?

Weblem 3.18 (a) Identify a state of the U.S.A. in which newborn infants are routinely tested for 2,4-dienoyl-coenzyme A reductase deficiency. (b) Identify a state of the U.S.A. in which newborn infants are not routinely tested for 2,4-dienoyl-coenzyme A reductase deficiency. (c) Identify a state of the U.S.A. in which newborn infants are routinely tested for Krabbe disease. (d) Identify a state of the U.S.A. in which newborn infants are not routinely tested for Krabbe disease. (e) What are the clinical consequences of failure to detect 2,4-dienoyl-coenzyme A reductase deficiency or Krabbe disease?

Weblem 3.19 Are language groups more closely correlated with variations in human mitochondrial DNA or Y chromosome sequences? Suggest an explanation for the observed result.

Weblem 3.20 Draw the equivalent of the Intron-Exon distribution diagram near the bottom of Box 3.4, for some plant globin.

Weblem 3.21 Genes distinguishing blood types A, B and O code for related proteins that add different saccharide units to an antigen on the surface of red blood cells. The enzyme in the O case is dysfunctional and adds no antigen to the cell surface. What would be the clinical application of an enzyme that would cleave the A and B antigens from red blood cell surfaces? Has such an enzyme been developed? To what extent is it effective?

Weblems – Chapter 4

Weblem 4.1. Retrieve the gene sequence of mitochondrial ATPase subunit 6 from Atlantic hagfish *(Myxine glutinosa).* Draw a dotplot against the homologous gene from sea lamprey *(Petromyzon marinus).* Comment on the similarity observed and compare with the similarity between the lamprey and dogfish sequences shown near the beginning of this chapter.

Weblem 4.2. A human protein has the following sequence:

```
MPRIDADLKLDFKDVLLRPKRSSLKSRAEVDLERTFTFRNSKQTYSGIPIIVANMDTVGTFEMAAVMSQHSMFTAIHKHY
SLDDWKLFATNHPECLQNVAVSSGSGQNDLEKMTSILEAVPQVKFICLDVANGYSEHFVEFVKLVRAKFPEHTIMAGNVV
TGEMVEELILSGADIIKVGVGPGSVCTTRTKTGVGYPQLSAVIECADSAHGLKGHIISDGGCTCPGDVAKAFGTGADFVM
LGGMFSGHTECAGEVIERNGRKLKLFYGMSSDTAMNKHAGGVAEYRASEGKTVEVPYKGDVENTILDILGGLRSTCTYVG
AAKLKELSRRATFIRVTQQHNTVFS
```

(a) What protein is it? (Easiest to use BLAST). (b) On what chromosome is its gene? (Easiest to use TBLASTN).

Weblem 4.3. Submit the amino acid sequence of papaya papain to a BLAST search and to a PSI-BLAST search. Which of the homologues appearing in Fig. 4.2 are successfully detected by BLAST? Which by PSI–BLAST?

Weblem 4.4. Submit the amino acid sequence of papaya papain to a PSI-BLAST search (see previous weblem). In the results for the match to human procathepsin L, indicate on a photocopy of the dotplot (Fig. 4.2b) the regions of local matches reported.

Weblem 4.5. Align the amino acid sequence of papaya papain and the homologues shown in Fig. 4.2 using CLUSTAL Omega or T-Coffee. Compare the results with the alignment table in Pfam based on Hidden Markov Models, and with the structural alignments in Fig. 4.2.

Weblem 4.6. Find structures of thioredoxins appearing in the alignment table in Figure 4.6(a), from organisms other than *E. coli.* On a photocopy of the alignment table, indicate the regions of helix and strands of sheet as assigned in the Protein Data Bank entries, and compare with the helices and strands of *E. coli* thioredoxin.

Weblem 4.7. The structure of the human PAX–6 protein is known (wwPDB code 6PAX). Using the sequence alignment between human PAX–6 and the *Drosophila* circadian clock protein, draw a picture of human PAX–6, indicating – perhaps by use of a different colour – which residues are alignable with the circadian clock protein.

Weblem 4.8. Find amino-acid sequences of the NADH-ubiquinone oxidoreductase chain 3 from several mammals not obviously closely related to elephants. Do these show that this catalytic activity does *not* require a sequence as similar to the homologous sequences in African elephants, Asian elephants, and mammoths as those three sequences are to one another?

Weblem 4.9. Four families of carbonic anydrases are known: α, β, γ, and δ. Finish the table 'Distribution of families of carbonic anhydrase' in the section Ancestral Sequence Reconstruction to show the species distribution of δ carbonic anhydrases.

Weblem 4.10. Can PSI-BLAST identify the homology between immunoglobulin domains and the domains of *Cellulomonas fimi* endogluconase C and *Streptococcus agalactiae* IgA receptor?

Weblem 4.11. (a) Can PSI-BLAST identify the relationship between *Klebsiella aerogenes* urease, *Pseudomonas diminuta* phosphotriesterase and mouse adenosine deaminase? (b) Compare the alignments of these three sequences produced by DALI or MUSTANG and by CLUSTAL Omega or T-Coffee.

Weblem 4.12. Which pair of species is more closely related (as measured by more recent time of divergence)? (a) human/bonobo or mouse/rat. (b) human/orangutan or whale/hippopotamus (c) human/kangaroo or *Arabidopsis thaliana*/wheat.

Weblem 4.13. The growth hormones in most mammals have very similar amino acid sequences. (The growth hormones of the alpaca, dog, cat, horse, rabbit, and elephant each differ from that of the pig at no more than 3 positions out of 191.) Human growth hormone is very different, differing at 62 positions. The evolution of growth hormone accelerated sharply in the line leading to humans. By retrieving and aligning growth hormone sequences from species closely related to humans and our ancestors, determine *where* in the evolutionary tree leading to humans the accelerated evolution of growth hormone took place.

The next series of weblems is designed to place the human species in its biological context by analysis of sequences from near and distant relatives, and to illustrate some of the variety of genetic information that has been used to investigate phylogenetic relationships.

Weblem 4.14. The living species most closely related to humans are apes and monkeys. Alu elements are a type of SINE (Short INterspersed Element) useful as species markers. Although part of the repetitive noncoding portion of the genome, some Alu elements function in gene regulation. On the basis of Alu elements that regulate the genes for parathyroid hormone, the haematopoietic cell-specific FcϵRI-γ receptor, the central nervous system-specific nicotinic acetylcholine receptor $\alpha3$, and the T-cell-specific CD8α, derive a phylogenetic tree for human, chimpanzee, gorilla, orangutan, baboon, rhesus monkey and macaque monkey.

Weblem 4.15. Humans are primates, an order that we, apes and monkeys share with lemurs and tarsiers. On the basis of the $\beta-$globin gene cluster (contents and gene order) of human, a chimpanzee, an old-world monkey, a new-world monkey, a lemur, and a tarsier, derive a phylogenetic tree of these groups.

Weblem 4.16. Primates are mammals, a class we share with marsupials and monotremes. Extant marsupials live primarily in Australia, except for the oppossum, found in North and South America. Extant monotremes are limited to animals from Australia and New Guinea: the platypus and echidna. Collect the nuclear genes for mannose 6-phosphate/insulin-like growth factor II receptor from mammalian species including placentals, marsupials and monotremes. From them draw an evolutionary tree, indicating branch lengths. Are monotremes more closely related to placental mammals or to marsupials?

Weblem 4.17. Mammals are vertebrates, a subphylum that we share with fishes, sharks, birds and reptiles, amphibia, and primitive jawless fishes (example: lampreys). For the coelacanth *(Latimeria chalumnae),* the great white shark *(Carcharodon carcharias),* skipjack tuna *(Katsuwonus pelamis),* sea lamprey *(Petromyzon marinus),* frog *(Rana pipens),* and Nile crocodile *(Crocodylus niloticus),* using sequences of cytochrome c oxidase subunit 1, derive evolutionary trees of these species.

Weblem 4.18. Tetrapods are gnathostomes, a superclass that we share with fishes. The traditional view of fish → tetrapod evolution is that jawed vertebrates split into one group containing cartilaginous fishes, *Chondrichthyes,* including sharks and rays; and another containing both ray-finned fishes, *Actinopterygii,* including modern bony fishes such as cod and salmon, and lobe-finned fishes, *Sarcopterygii,* including coelacanths, lungfishes and tetrapods (see Figure 1.4). Test this hypothesis using at least 12 mitochondrial protein-coding genes from at least 30 species including sharks, lungfish, bony fishes, amphibia, reptiles, birds, and mammals, using a lamprey as an outgroup. Of the three groups – cartilaginous fishes, bony fishes and tetrapods – which pair appears to be most closely related: bony fishes and tetrapods as in the traditional view, or a different pair? Consider specifically a phylogeny according to which the tetrapods split off first, and cartilaginous fishes, bony fishes, and even lungfishes are sister taxa (draw this tree).

Weblem 4.19. Vertebrates are chordates, a phylum that also includes lancelets (small fish-like marine animals; example, amphioxus), and jawless vertebrates (lacking a true vertebral column (example, lamprey). As in other organisms with bilateral symmetry (including insects), vertebrate HOX genes encode a family of DNA-binding proteins. The expression of these genes varies along the head-to-tail body axis, and controls the setting out of the body plan. Indeed there is a amazing mapping between the order of the genes on the chromosome, the order of their action along the body, and the relative times during development of the onset of their activity.

During the course of vertebrate evolution there have been large scale genomic duplications, associated presumably with the development of greater complexity of body architecture, as presciently suggested by S. Ohno in 1970. The genomes of insects and amphioxus have a single HOX cluster. Zebrafish have seven HOX clusters, interpretable in terms of a series of duplications: $1 \to 2 \to 4 \to 8$ followed by loss of one to reduce $8 \to 7$.

Find the number of HOX clusters in the human and the lamprey, perform a multiple sequence alignment to assign correspondences among the individual genes, and derive from the results a phylogenetic tree for amphioxus, lamprey, fishes, and mammals.

Weblem 4.20. Chordates are deuterostomes, a grouping we share with urochordates (example: sea squirts), hemicordates (example: amphioxus), and echinoderms (example: starfish). There are systematic differences between these four phyla in their mitochondrial genetic code. Determine, for examples of organisms in each phylum, the amino acids that correspond to the codons ATA and AGA. Derive from these results a phylogenetic tree of the four deuterostome phyla.

Weblem 4.21. Create a PYTHON program that accepts two sequences as input and draws a DOTPLOT, like the PERL program in Box 4.1, by downloading as much PYTHON code as possible from Web sources.

Weblem 4.22. (a) Create a PYTHON program that accepts the same input and produces the same output as the PERL program to draw trees in Box 4.9, by downloading as much PYTHON code as possible from Web sources (b) Create a PYTHON program that accepts as input the general Newick tree format which, for example, allows specifying branch lengths. An example of the input the program needs to deal with would be: (A:3,((B:5,C:9):3,D:2):5.3,(E:10,F:12):8); (To see what the output would look like, type the input string into a tree viewer available on the web.)

Weblems – Chapter 5

Weblem 5.1. Submit the sequences of bovine $\gamma-$chymotrypsin [8GCH] and and *S. aureus* epidermolytic toxin A [1AGJ] (Example 5.4) to a standard sequence alignment program (for instance, `http://www.ebi.ac.uk/Tools/psa/emboss_needle/`, but there are many many others.) Describe the differences between the sequence-based and structure-based alignment. How many pairs of residues are aligned the same way in both alignments? Which alignment contains more gaps? To what regions of the structures do the major differences correspond?

Weblem 5.2. Consider the three polypeptide chains in sperm whale myoglobin [1MBD] and human haemoglobin [1HHO] (α and β chains). Extract the amino acid sequences in 1-letter format. You can do this either by converting from information in the wwPDB entry, or by going back to UniProt `http://www.uniprot.org/`
(a) Align each pair of sequences using a standard sequence alignment server (for instance, `http://www.ebi.ac.uk/Tools/psa/emboss_needle/`, but there are many many others.)
(b) Align all three sequences using a multiple sequence alignment server (see `https://www.ebi.ac.uk/Tools/msa/`)

Compare these alignments. That is, the multiple sequence alignment induces three pairwise alignments (just delete each sequence separately from the multiple alignment). Compare these induced pairwise alignments with the original pairwise alignments in step (a) which were computed in igorance of the third sequence. How many residue correspondences are the same in both alignments?

Structural alignment, like sequence alignment, produces a set of residue-residue correspondences.
(c) Do a pairwise structural alignment of each pair of sequences. Compare these pairwise structural alignments with the corresponding pairwise sequences alignments derived in steps (a) and (b). One possibility: `http://ekhidna2.biocenter.helsinki.fi/dali/`
(d) Do a multiple structural alignment of the three sequences. One possibility: `http://lcb.infotech.monash.edu.au/mustang/` Compare the multiple structural alignment with the pairwise structural alignments. How many residue correspondences are the same?
(e) Compare the multiple sequence alignment with the multiple structural alignment. What general conclusions appear to emerge?

Weblem 5.3. The bacterium *Pseudomonas fluorescens* and the fungus *Curvularia inaequalis* each possesses a chloroperoxidase, an enzyme that catalyzes halogenation reactions. Do these enzymes have the same folding pattern?

Weblem 5.4. An inventory of the structures common to all three domains (bacteria, archaea, and eukarya) showed that the five most common folding patterns of domains are (1) the P-loop-containing NTP hydrolase fold, (2) the NAD-binding domain, (3) the TIM-barrel fold, (4) the flavodoxin fold, and (5) the thiamin-binding fold. (Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274,** 562-576.) (a) Using facilities available in SCOP, draw pictures of an example of each type of structure. (b) Draw simplified topology diagrams analogous to the one in Figure 5.1(e) for the structures chosen.

Weblem 5.5. Using standard pairwise sequence-alignment tools, align the sequences of human neutrophil elastase and *C. elegans* elastase. (for instance, `http://www.ebi.ac.uk/Tools/psa/emboss_needle/`, but there are many many others.) (a) In the optimal alignment, how many identical residues are there? (b) Would it be reasonable to build a model of *C. elegans* elastase starting from the structure of human neutrophil elastase?

Weblem 5.6. Does the human θ globin gene encode an active globin? Or is it a pseudogene? Send the amino-acid sequence of human θ globin to SWISS-MODEL, including a request for a WhatCheck report on the result. What can you conclude from the result about the status of human θ globin?

Weblem 5.7. S. Chakravarty and K.K. Kannan solved the structures of carbonic anhydrase with a benzenesulphonamide ligand (Protein Data Bank entry 1CZM.) Draw pictures of the binding site showing the nature of the interactions between the protein and ligands. Describe the nature of the interactions in terms of the conclusions drawn from the QSAR analysis.

Weblem 5.8. MARCOIL predicts with high confidence that residues 164–191 of chicken c-fos form a coiled-coil. Residues in coiled-coil structures are assignable to positions `abcdefg` in successive 7-residue oligopeptides (see section ‘Coiled-coiled proteins’). What assignment of residues 164–191 in chicken c-fos does MARCOIL predict?

Weblem 5.9. Submit the sequence of chicken c-fos to MARCOIL `http://bcf.isb-sib.ch/webmarcoil/webmarcoilC1.html` and COILS `https://embnet.vital-it.ch/software/COILS_form.html` Compare the predictions on a residue-by-residue basis. For how many residues do the predictions agree (to within 10% probability)?

Weblem 5.10. Chicken c-fos was used as an example of coiled-coil prediction by MARCOIL. In order to get some idea of how challenging the prediction was: (a) Identify the protein of known structure with amino-acid sequence most similar to that of chicken c-fos. (b) Determine from inspection of the structure the residues in that structure that form a coiled-coil. (c) Draw a dotplot of the two sequences. Indicate on the dotplot the region in the protein of known structure that form a coiled-coil, and the region in chicken c-fos that are predicted by MARCOIL to form a coiled-coil. (d) Align the two sequences and report the % identical residues in the two complete sequences. (e) Report the % identical residues in the region that forms a coiled-coil in the protein of known structure.

Weblem 5.11. The structure of the SH3 domain from the cytoskeletal protein $\alpha-$spectrin from chicken has been determined both by X-ray crystallography [1SHG] and by NMR [1AEY]. The crystallographic result contains one set of coordinates, containing residues 6–62. The NMR result contains 15 sets of coordinates, each containing residues 5–62. For every pair of structures (15 NMR structures and 1 X-ray crystal structure), compute and tabulate the root-mean-square deviation of (a) The Cα atoms. (b) All mainchain (N, Cα, C and O) atoms, and (c) All atoms. (d) Describe your results. Do you see any systematic differences between X-ray and NMR structures?

Weblem 5.12. M. Vihinen curates a database of elastase mutants, ELA2base (`http://structure.bmc.lu.se/idbase/ELA2base/index.php` see also Thusberg, J. & Vihinen, M. (2006) Bioinformatic analysis of protein structure-function relationships: case study of leukocyte elastase (ELA2) missense mutations. Human Mutat. 27, 1230–1243.) J. Moult, P. Yue, E. Melamud & M. Zuhl created a database to evaluate the effects on protein structure of human SNPs (`http://www.snps3d.org`, see also Yue, P. & Moult, J. (2006). Identification and analysis of deleterious human SNPs. J. Mol. Biol. 356, 1263–1274 and P. Yue, E. Melamud & J. Moult (2006). SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 7:166.)
(a) For the single amino-acid replacements in ELA2base, draw a picture of the human neutrophil elastase structure [1HNE]) showing the positions of the residues. How would you describe the distribution in the structure of the disease-causing substitutions? (b) Submit each single amino-acid replacement in ELA2base to SNPS3d, and tabulate the scores returned. (c) Can you correlate the scores with the distribution of residues in the structure? (d) Can you correlate the scores with the severity of the symptoms? (e) The following elastase mutations are associated with cyclic neutropenia: A32V, L177F,

R191Q. The following elastase mutations are associated with severe congenital neutropenia (non-cyclic): P110L, V72M, S97L. Plot these residues on the structure. Can you see a systematic difference between the locations within the structure of the two different phenotypes? (The residue numbers in this part (e) refer to the following sequence:)

```
   1  MTLGRRLACL FLACVLPALL LGGTALASEI VGGRRARPHA WPFMVSLQLR    50
  51  GGHFCGATLI APNFVMSAAH CVANVNVRAV RVVLGAHNLS RREPTRQVFA   100
 101  VQRIFENGYD PVNLLNDIVI LQLNGSATIN ANVQVAQLPA QGRRLGNGVQ   150
 151  CLAMGWGLLG RNRGIASVLQ ELNVTVVTSL CRRSNVCTLV RGRQAGVCFG   200
 201  DSGSPLVCNG LIHGIASFVR GGCASGLYPD AFAPVAQFVN WIDSIIQRSE   250
 251  DNPCPHPRDP DPASRTH   267
```

Weblem 5.13. From the multiple sequence alignment of ETS domains (see Problem 5.6), available in computer-readable form from the Web Resource Centre, make a sequence logo.

Weblem 5.14. Many molecular graphics programs for display of protein structures are freely available:
`https://www.rcsb.org/pages/thirdparty/molecular_graphics`
Some need installation on your computer; others are available via browsers. The component institutions of the wwPDB provide albums of still pictures, or slide shows, of each entry, plus links to three-dimensional viewers.

Install, if necessary, one of the graphics programs, and use it to produce pictures similar to Figure 5.1, for wwPDB entry [1OCX] chain A, *E. coli* maltose O-acetyltransferase. Do not include ligands and water molecules.

From the picture (preferably) or from the text in the wwPDB file (if necessary) describe the order of the secondary structure elements of this chain.

Draw a suitable picture of the trimer (chains A, B and C) and by inspection of the picture determine which secondary structure elements are involved in the quaternary structural interaction.

Weblem 5.15. This is the first of a series of problems aimed at identifying the protein illustrated in Figure 5.1.
(a) Count the number of helices and strands.
From the PDB search site `http://www.rcsb.org/pdb/search/advSearch.do`
find a list of proteins with this secondary structure content.
How many are there? Can you identify the subject of Figure 5.1? If not, what fraction of wwPDB entries have you eliminated from contention?

Weblem 5.16. Another try to identify the subject of Figure 5.1. Download the file: `https://cdn.rcsb.org/etl/kabschSander/ss.txt.gz`
This file gives the secondary structure assignments for wwPDB entries. To see the format, this file begins:

```
>101M:A:sequence
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGA
ILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKEL
GYQG
>101M:A:secstr
    HHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHH GGGGGG TTTTT  SHHHHHH HHHHHHHHHHHHHHHH
HHTTTT  HHHHHHHHHHHHHTS   HHHHHHHHHHHHHHHHHH GGG SHHHHHHHHHHHHHHHHHHHHHHHHT
```

Write a program to read in the secondary structure assignment. For each wwPDB entry, print a line containing the code (*e.g.* 101M) and a character string containing the order of appearance of helices (H) and strands (E). Recognize a helix as a string of consecutive H and a strand as a string of consecutive E. In the entry for 101M there are 7 sets of consecutive H so you would produce a string reporting seven helices: HHHHHHH.

From inspection of Figure 5.1(c), determine the *order* of the secondary structure elements, and express this as a string containing H and E. Search the output of your perl program for this string, and report how many matches were found.

Are you able to identify uniquely what protein appears in Figure 5.1? If not, what fraction of wwPDB entries have you eliminated from contention? Are there few enough possibilities that you can look at them all and recognize the right one (say, within 20 minutes)?

Note that there is a potential problem in inferring the individual secondary structure elements from the file ss.txt If there are two consecutive helices (for example) without an intervening residue in a non-helical conformation, you will be unable to distinguish the two consecutive helices from one long helix. Check the PDB entry for 101M, chain A, inspecting the lines beginning HELIX to see whether there are indeed only seven helices, or whether there are any pairs of consecutive helices that *appear* single in the file ss.txt.

A safer way to infer the order of secondary structure elements is therefore to extract the lines beginning HELIX and SHEET from the full wwPDB entries, and to assemble the results from them. This requires access to an online version of the PDB, and also might involve a non-trivial amount of computer time in executing the program you write.

An alternative would be to use the program SST: `http://lcb.infotech.monash.edu.au/sstweb2/Submission_page.html` In what respect is the output from SST more convenient than the format of ss.txt for the purposes of this problem?

Weblem 5.17. Still another try at identifying the protein shown in Figure 5-1. Interrogate a database that contains information about the geometric relationships between the elements of secondary structure. The Protein Topology Graph Library (PTGL) is such a database. `http://ptgl.uni-frankfurt.de/`
To use it, carry out the following steps:
(a) Draw a horizontal row of dots to represent the elements of secondary structure in the protein shown in Figure 5.1. From left to right, these represent the elements of secondary structure in order in the chain.
(b) Draw an arc between every pair of secondary-structure elements that are in contact in the structure. These contacts include both lateral interactions within the $\beta-$sheet, possibly contacts between helices, and packings of a helix against one or more strands of sheet. Label each arc with the relative geometry of the two secondary structure elements in contact: p = parallel, a = antiparallel, m = mixed (*i.e.,* neither)
(c) Encode this graph in one dimension by following these rules: Create a sequence of fields separated by commas. Choose one secondary structure element to start. This could be the one corresponding to the leftmost dot in your row. The first entry in the output sequence is either e or a depending on whether the first secondary structure element chosen is a strand (e) or a helix (a).

Subsequent fields contain three elements: (1) a number, which gives the displacement of the dot to which an arc is connected, (2) the relative orientation of the two secondary structure elements connected by an arc: p = parallel, a = antiparallel, m = mixed (*i.e.,* neither), or z = no contact (3) e or h depending on the nature of the secondary structure element to which the arc connects its predecessor. For instance, a $\beta-$hairpin would appear as a diagram with two dots, one arc connecting them, represented by the character string: {e, 1ae}. (What would be the representation of a $\beta - \alpha - \beta$ unit, assuming each pair of secondary structure elements is in contact?)

It is necessary to include additional arcs specified by z (no contact) in order to represent the entire graph by a one-dimensional character string. The fact that it is not always possible to draw a single connected path tracing all secondary structure contacts is a consequence of Euler's famous analysis of the paths across the bridges of *Königsberg.* (See Chapter 2.)

Given that these are rather tricky operations, check that you have come out with the character string:

{e, 3ae, -1ae, -1mh, 3mh, -4pe, 5ae, -1ah, -1me, -2zh, 5me, -4pe}

(d) Select Alpha-Beta ADJ in the search form `http://ptgl.uni-frankfurt.de/SearchFields.html` and enter the string into the Graphs field in the search form.
How many hits did you get? Are you able to identify the molecule?
(e) If you did not derive the correct character string on your own, reconstruct the graph from the correct result and compare it with the contact graph that you constructed. Are you able to see what went wrong?

What if you were not given the checkpoint? If you got everything right, the search would find the protein illustrated in Figure 5.1, and possibly others. If not, the search might either return nothing, or return proteins with different topology.

What could you do to 'troubleshoot' if the procedure didn't return anything, implying that your character string was incorrect? A complete list of topologies represented by the character strings appears in `ptgl.uni-frankfurt.de/statistics/AlbeADJ.html`. Assuming that you believe at least in the representation of the first few arcs – say e,3ae,-1ae – you could search the file for that fragment, and, if that didn't reduce the possibilities sufficiently, screen further for the right number of secondary structure elements.

Weblem 5.18. Prior to its weekly release of new structures, wwPDB posts the amino-acid sequences of the structures a few days earlier, at `https://www.rcsb.org/pages/search_features#search_unreleased-and-new`
This creates an opportunity to test prediction methods.
For individuals: pick a few sequences. For classrooms: distribute sequences to students.
(1) Submit the sequences to a BLAST search:
`http://www.ebi.ac.uk/Tools/sss/ncbiblast/` or `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins`
Choose Protein Structure Sequences (at EBI) or Protein Data Bank Proteins (pdb) (at NCBI), and determine whether there is any close homologue of known structure.
(2) No later than Monday, send the sequences to selected protein structure prediction servers:
`http://swissmodel.expasy.org/`
`http://www.predictprotein.org/`
`http://robetta.bakerlab.org/`
and any others that appear appropriate
Collect the results of the predictions.

(3) On the following Wednesday, superpose the predicted and experimental structures.
`http://wishart.biology.ualberta.ca/SuperPose/`
This server will superpose the structures based on the alignment implied by the fact that the sequences of predicted and the experimental structures are the same.

If the prediction contains only parts of the structure, or if a misalignment is suspected, try a more general structural alignment (see: `http://lcb.infotech.monash.edu.au/mustang/`)
(4) Report on the successful and unsuccessful aspects of each of the predictions.

Weblem 5.19. You are working towards a Ph.D. in a research group headed by a professor heavily into consulting for the pharmaceutical industry. Your assigned thesis topic is to design a drug to knock out the activity of a specified target enzyme. The plan is that you will begin by purifying, crystallizing, and solving the structure of the target protein. You will then have to study the structure, and check aligned sequences for conservation, to find the active site. Then you need to identify a ligand that might bind and inhibit the activity. And that will take you only as far as a 'lead' compound. All this is quite resource- and (of more concern to you) labour-intensive.

You ask yourself, why not just determine the gene sequence and design a short RNA that interferes with translation? Could this be a useful 'short-cut' to a drug?

Discuss the advantages and problems of this idea. Has this approach succeeded in any practical cases?

Weblems – Chapter 6

Weblem 6.1. From PubMed, determine how many papers related to Breast Cancer were published in 2018. Read the abstract of one of them. How long did it take you? Multiply the time required by the number of papers. If you wanted to be an expert on the subject of Breast Cancer, how much time would it take you to read all the abstracts of all the papers on the subject published in 2018?

Weblem 6.2. How many open access journals are there categorized as in (a) the Health Sciences, (b) in Biology and Life Sciences, (c) in Agriculture and Food Sciences, and (d) Bioinformatics (see the Directory of Open Access Journals: `http://www.doaj.org/`)

Weblem 6.3. When will the works of Dylan Thomas go out of copyright in the UK?

Weblem 6.4. The home page of the RCSB `http://www.rcsb.org` includes links to query facilities that permit identification and retrieval of particular macromolecular structures from its contents, by specifying desired features. Find at least two other sites that permit identification of structures from the Protein Data Bank. There will be substantial overlap in the features of these query systems. What if any unique facilities does each have?

Weblem 6.5. (a) Give examples of three types of information about fruit flies that appear in *both* the European Nucleotide Archive (ENA) and FlyBase. ENA is based at the European Bioinformatics Institute (EBI). (b) Give examples of three types of information about fruit flies that appear in FlyBase but *not* in the ENA.

Weblem 6.6. Find, in the Protein Data Bank, two structures of the same protein, determined by X-ray crystallography at different resolution: one at low resolution ($\geq$ 2.9 Å) and the other at high resolution ($\leq$ 1.9 Å). Remember: the lower the number specifying the resolution, the higher the potential quality of the result. Check the PDBREPORT entries for the two entries, in `http://swift.cmbi.ru.nl/gv/pdbreport/`. Compare the evaluations of the two structures by the software underlying the reports.

Weblem 6.7. (a) Search PubMed for the article: Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. Nature 261, 552–558. Can you get the full text of the article via PubMed? Print a copy of the page that provides the most detailed information about this article that is accessible to you from PubMed.
(b) Do a Google search for this article. (It is enough to enter search terms Levitt Chothia Structural Patterns). Try the first two or three links returned. Did any of those give you access to the full text of the article?
(c) Search Google Scholar for the same article. Under this entry, select Web Search. Can you easily find a site containing a complete text of the article? If so, print the page containing the link to the site that provides access to the full text, and highlight the link to the full-text site.
(d) Compare the ease of getting to the full text through PubMed, Google, and Google scholar.
(e) What conclusions can you draw?
(f) Now suppose you knew or suspected that if the full text were available on the web it would be as a pdf file (= Adobe public document format). Do a google search for keywords:

Levitt Chothia Structural patterns pdf

Do the results suggest that you should reconsider your conclusions?

Weblem 6.8. On 2 May 2006, US Senators John Cornyn (Texas) and Joseph Lieberman (Connecticut) proposed the Federal Research Public Access Act (S. 2695). The bill was considered and reintroduced in several sessions of congress, but not passed. The successor proposal, Fair Access to Science and Technology Research Act (FASTR) was introduced in both the Senate and the House, most recently in 2017, but has not been passed. (a) Summarise the salient provisions of the FASTR proposal. (b) How does it differ from the most recent version of its predecessor proposal, the Federal Research Public Access Act (FERPA). (c) What groups support FASTR, and what groups oppose it?

Weblem 6.9. Submit to `https://www.bioinformatics.org/textknowledge/genetag.php` (a) the title + abstract and (b) the entire text of the following paper: Zhu, Y., Culmsee, C., Klumpp, S. & Krieglstein, J. (2004). Neuroprotection by transforming growth factor-beta1 involves activation of nuclear factor-kappaB through phosphatidylinositol-3-OH kinase/Akt and mitogen-activated protein kinase-extracellular-signal regulated kinase signaling pathways. Neuroscience 123, 897–906. What gene and protein names does the program find in the title + abstract that were not found in the title? What gene and protein names does it find in the full paper that were not found in the title + abstract?

Weblem 6.10. CbiT is a protein the function of which was originally hypothesized, on the basis of homology to proteins of known function, to be a precorrin-8w decarboxylase in the biosynthesis of vitamin B12. The crystal structure, solved in 2002, showed it to be a methyltransferase. (Keller, J.P., Smith, P.M., Benach, J., Christendat, D., deTitta, G.T., & Hunt, J.F. (2002). The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase is a methyltransferase. Structure 10, 1475–1487.) In many databases, the annotations were corrected. How many sites with incorrect annotation as a precorrin-8w decarboxylase can you still find?

Weblem 6.11. Your research director is scheduled to give a 1-hour lecture to beginning graduate students at your institution on 'Information retrieval from the scientific literature'. A month before the lecture, he or she receives an invitation to a prestigious international conference that conflicts with the lecture, and asks you to prepare and deliver the lecture. What sources on the web do you find useful for preparation? What specific examples would you include? What sites would you recommend to your audience for following up on the lecture? Create a fairly detailed outline of the lecture you propose to give.

Weblem 6.12. Same as previous problem, but change 'A month' to 'Two days'. (Perhaps your supervisor is replacing someone who cancelled.)

Weblem 6.13. Find a web site containing ambiguous sentences. (For instance: `http://www.fun-with-words.com/ambiguous_headlines.html` or `http://www.ling.upenn.edu/~beatrice/humor/headlines.html`) For a few selected examples, submit them to a syntactic analyser (for example: `http://www.link.cs.cmu.edu/link/` and decide whether it parsed the sentence correctly. (If in a classroom context, for this and the following problem, it would be possible for different students to take different examples, and present the results together.)

Weblem 6.14. Find a web site containing amusing ambiguous headlines. (Examples: `http://www.fun-with-words.com/ambiguous_headlines.html` or `http://www.ling.upenn.edu/~beatrice/humor/headlines.html`) Pick a foreign language, and for a few selected examples, submit them to Google translate, and back-translate the result. Did you return to the original?

Weblems – Chapter 7

Weblem 7.1 Devise a method for assigning GO classifications to proteins from sequence and/or structure. Collect a data set of proteins of known structure, covering different structural types and different functions. Record the GO assignments of molecular function, cellular component, and biolgoical process. Divide the data set into a training set and a test set. Train your method using the first set, then test it using the second.

Weblem 7.2 Derive from the wwPDB database of known protein structures an assignment of $\beta-$hairpins in structures determined by X-ray crystallography to a resolution ≤ 2.2 Å. Use secondary structure and connectivity assignments within the entry. A $\beta-$hairpin is defined as a region between two successive strands of antiparallel $\beta-$sheet, such that the two strands are adjacent in the same sheet, and the region between the two strands does not contain any other secondary structure element. Take 8 residues as the maximum length of such a region to include in your database.

Weblem 7.3 Using the database created in the previous weblem, design a method to identify $\beta-$hairpins from amino-acid sequences alone. Split the database created in the previous problem into training and test sets. Use the entire sequence of the wwPDB entry to optimize the method using the training set. Then test with the test set.

Weblem 7.4 Download the Schneider *et al.* program for classifying RNA transcripts `https://github.com/hugowschneider/longdist.py`. Run the progrm on one of author-described datasets and verify that you get the same answer. Run on some new dataset and report performance.

Weblem 7.5 The procedure in Case Study 7.1 does quite a good job at distinguishing protein-coding and long non-coding RNA transcripts. How could the result be improved? Several other tools have been published – see Schneider *et al.*, cited in Case Study 7.1, for references. Pick several of them, including the Schneider *et al.* program `https://github.com/hugowschneider/longdist.py` and run them on some appropriate dataset. Do they all give the same pattern of correct and incorrect predictions? If so, then they are, in effect, redundant. However, if the pattern of correct and incorrect predictions is different, it might be possible to achieve improved performance by combining the predictions of several methods. This could be done by simple vote, or by feeding them into a neural network that could be optimized by training. How well can you do?

Weblem 7.6 ASTRAL: `http://scop.berkeley.edu/downloads/scopeseq-2.07/`
`astral-scopedom-seqres-gd-sel-gs-bib-40-2.07.fa`
has collected sets of domains extracted from the wwPDB such that any two entries have < 40% sequence identity. This provides an unbiased set of structures for analysis.
The goal of this weblem is to create a database of sequences of secondary structure elements – helices H and strands of sheet E – from the entries of ASTRAL. Do not retain the residue numbers of the helices and sheets, but just the order of their appearance in the chain.
Download the latest version of ASTRAL, and run each entry through SST:
If the pdb file is `d1xyza_.ent`, execute the command:

```
curl -s -F "pdb_file=d1xyza_.ent"
http://lcb.infotech.monash.edu.au/sstweb2/formaction.php
```

Then extract the sequence of H and E, and concatenate them into a character string to characterize the dataset `d1xyza_.ent`

Weblem 7.7 Repeat the previous weblem, but retaining the residue numbers of the start and end of the helices and sheets. Write a web server based on your result that allows the user to type in the identifier of any ASTRAL entry and returns the secondary structure assignment.

Weblem 7.8 CATH is a database containing a hierarchical classification of protein structures. Its highest-level categories are:

Class the overall secondary-structure content of the domain.

Architecture sets of proteins with high structural similarity but no evidence of homology

Topology/fold a large-scale grouping of topologies which share structural features

Homologous superfamily proteins related by evolution

Using the results of Weblem 7.7, devise a procedure for classifying proteins based on the sequence of secondary-structure elements (helices and strands of sheet). Divide a reasonably large randomly-chosen subset of ASTRAL entries into training and test sets. (a) How well can you predict CATH Class? (b) How well can you predict CATH Architecture? (c) How well can you predict CATH Topology/fold? (d) Once you have a working procedure, take the entries from the test set, extract their amino-acid sequences, and use one or more web servers to *predict* their secondary structures (for example, `https://www.predictprotein.org/`). Repeat the calculations for parts (a), (b) and (c). Compare the accuracy of the results.

Weblem 7.9 Create a database of protein complexes for which structures are known, both of the components separately and of the complex. These complexes include multimeric proteins in which the association is relatively permanent, and transient complexes that form and dissociate during some process, for instance, an antibody binding an antigen. Note that one subunit may be a part of several different complexes.

For each entry determine the root-mean-square-deviation after optimal superposition of the Cα atoms, of each subunit determined separately and within a complex, and use this as an index of the structural change upon complex formation. (Not really the best index: Can you think of a better one?) What is the distribution of the structural changes? Determine the residues involved in the interfaces between components in the complex, and compare the extents of their structural changes to the overall structural change.

Using one or more protein docking web servers, try to reconstruct the complexes, first from the artificially separated components simply extracted from the complex, and second from the structures of the components determined separately. Compare the results.

Weblems – Chapter 8

Weblem 8.1 Find a fragment of the genealogy of the Royal families of Europe containing a family tree that is not a ‘tree’ in the graph-theoretic sense.

Weblem 8.2 One model for the growth of a scale-free network suggests that if new nodes and edges are added according to the rule that the probability of adding a new edge is proportional to the number of edges that a node already has, the network will remain scale free and retain the same exponent. Using historical maps of the London Underground, check whether earlier networks are scale-free. Test whether addition of edges to the network has followed the rule that edges have been added preferentially to more highly-connected nodes.

Weblem 8.3 Create a graph of the London Underground, in which the nodes are the stations and edges connect stations such that one or more trains run between them with no intermediate stops. The result can be applied to answer certain Problems from this chapter in the text, and other weblems. See: `https://lasttrain.co.uk/tube-train-lines/london-underground-tube-line-names/` or `https://commons.wikimedia.org/wiki/London_Underground_geographic_maps/CSV`

Weblem 8.4 Modify the program in problem 8.5 to take into account current closings and substantial delays. (See `http://www.tfl.gov.uk/tfl/livetravelnews/realtime/tube/default.html`)

Weblem 8.5 Modify the program in problem 8.5 for persons in wheelchairs. (See `http://www.tfl.gov.uk/assets/downloads/avoiding-stairs-tube-guide.pdf`) It may not be possible to begin and/or end at the desired station. Suggest alternative entry and/or exit points that are the fewest stations away.)

Weblem 8.6 For Weblems 8.4 and 8.5, can you write a web server that people can see on their portable phones?

Weblem 8.6 A tourist interested in classical music visits London. TimeOut offers a list of sites for concerts: `https://www.timeout.com/london/music/classical-music`, and a calendar of classical concerts: `https://www.timeout.com/london/music/the-best-classical-concerts-in-london-this-month` Write a web server that would present a choice of events, allow selection of an event, accept a Tube Stop from which the tourist would like to get to the event, and propose an underground route to the event venue.

Weblems – Chapter 9

Weblem 9.1. (a) What are the Enzyme commission and Gene Ontology identifiers of the enzyme triose phosphate isomerase? (b) What are the Enzyme commission and Gene Ontology identifiers of the ATP- and GTP-linked succinyl-coA synthases?

Weblem 9.2. The Figure in box 9.4 shows the amino acid biosynthesis leading from aspartate to methionine, threonine, and lysine. On a photocopy of this figure, write in the names of the omitted intermediates at the unlabelled positions between consecutive arrows.

Weblem 9.3 The genes encoded by metC and malY in *E. coli* convert cystathione to L-homocysteine. Each has another function in addition. What are these other functions?

Weblem 9.4 Compare the pathways for biosynthesis of chorismate in *E. coli, M. jannaschii,* and *Aeropyrum pernix.* What is the earliest common intermediate in these pathways? What are its precursors in the three species?

Weblem 9.5 Compare the methionine biosynthesis pathway from asparate to methionine in *E. coli* and yeast. (a) Are there any differences in the series of intermediates? (b) Are the enzymes that catalyse similar transformations homologous? Show alignments of the amino acid sequences where possible. Use EcoCyc for *E. coli* (`http://ecocyc.org`) and the Saccharomyces Genome Database (`http://www.yeastgenome.org`) for yeast, searching in each case for ‘methionine biosynthesis’.

Weblem 9.6. Find the page in EcoCyc corresponding to the Figure at the end of Box 9.4. Choose different levels of detail and list what types of information are presented at each level.

Weblem 9.7. An enzyme with a function related to peptidylglycine monooxygenase (EC 1.14.17.3), and linked to it in the ENZYME DB, is 1-aminocyclopropane-1-carboxylate oxidase. (EC 1.14.17.4). (a) What is the lowest common ancestor of the two reactions in the EC classification? (b) What is the lowest common ancestor of the two reactions in the Gene Ontology molecular function classification? (c) Are these two enzymes closely related in the Gene Ontology classification?

Weblem 9.8. What identifiers does Gene Ontology associate with *E. coli* asparate aminotransferase, in the molecular function category? Arrange them in a directed acyclic graph, indicating the parent-child relationships between these identifiers.

Weblem 9.9. According to EcoCyc, what reactions can orotidine 5′-monophosphate undergo? What enzymes catalyse these reactions? What genes encode these enzymes?

Weblem 9.10 In phenylketonuria, dysfunction of the enzyme that converts phenylalanine to tyrosine causes a buildup and excretion of phenylpyruvic acid. (a) What reaction converts phenylalanine to phenylpyruvic acid? What other reactants and products participate in this reaction? (b) What cofactor is required for normal conversion of phenylalanine to tyrosine? What change in the cofactor is produced by the normal conversion of phenylalanine to tyrosine? How is the original state of the cofactor normally restored? Suppose that there is a defect in the enzyme that regenerates the cofactor. Would you expect this to produce PKU-like symptoms? Would dietary control of phenylalanine be sufficient for health in such a patient?

Weblem 9.11 What is meant by the P1 position of a protease substrate? From the MEROPS database (`https://www.ebi.ac.uk/merops/` find a protease that has a strong preference for cleavage of substrates with a Leu at the P1 position. To what protease structural family does this enzyme belong?

Weblem 9.12 It is asserted in the text that the *Methanococcus jannischii* shikimate kinase '. . . has no sequence similarity to bacterial or eukaryotic shikimate kinases. A protein from a different family has been recruited for the archaeal pathway.' This statement appears to assume that the absence of detectable sequence similarity implies different structural family. Is this true for this case? Find the amino acid sequence corresponding to this gene (*MJ1440*) and run a PSI-BLAST search against a database of known protein structures. Did this turn up any sequences similar enough to justify confident homology modelling of the *M. jannaschii* sequence? If so, determine a homology model (try `https://swissmodel.expasy.org/`). Compare the folding pattern of the result with the structure of known bacterial and eukaryotic shikimate kinases. Is the archaeal enzyme a member of a different family of protein folds?

Weblem 9.13 Figure 9.10 shows the reductive carboxylate cycle, and the EC numbers of the enzymes that catalyse the individual steps. Find the corresponding information for the tricarboxylic acid cycle (or Krebs cycle), and for the glyoxylate cycle. Does an alignment of the metabolites participating in these cycles, display the EC numbers of the enzymes that correspond to different reactions in different cycles. Report what is common to pairs or to all three of these pathways, in terms of (a) metabolites, (b) links between metabolites, corresponding to reactions, (c) enzymes that catalyse the reactions.

Weblem 9.14. Go to the page showing biosynthesis of ascorbate in KEGG: `http://www.genome.jp/kegg-bin/show_pathway?org_name=hsa&mapno=00053&mapscale=&show_description=hide` Do this in two windows, side by side on your screen. In one of the windows, select *Homo sapiens* as the species. In the other, select *Bos taurus.* Indicate on copies of both screens, how KEGG reports which enzyme cows have that we do not; that requires us, but not cows, to include vitamin C in our diet?

Weblem 9.15. Can the carrot (*Daucus carota sativus*) convert 16-hydroxypalmitate to 16-feruloyloxypalmitate? If so what cofactors are involved? What is the biological role of this reaction? Can any human enzyme catalyse this reaction?

Weblem 9.16. Consider the standard Krebs (tricarboxylic acid) cycle as a metabolic network. (a) Extract from KEGG the representation of the Krebs cycle as a graph, with metabolites as nodes, and reactions as edges. (b) Draw another graph with the enzymes as nodes, connect two enzymes by an edge if the product of the reaction catalysed by the first can serve as the substrate of the reaction catalysed by the second. (c) Compare the structures of these graphs. (d) Is it true that if the first graph is connected the second must also be connected?

Weblem 9.17. (a) Retrieve the amino-acid sequences of enolase, mandelate racemase, and muconate lactonizing enzyme. Align the sequences. Describe the common patterns in the sequences. (b) Retrieve the structures. Create a multiple structural superposition of the three, and draw intelligible pictures of the superposed structures. What common features do you see?

Weblem 9.18. Find examples of metabolic pathways linked to specific types of cancers.

Weblem 9.19. The drug sapropterin is used to treat phenylketonuria. What is the mechanism of action of this drug?

Weblem 9.20. What is the metabolic effect of gain-of-function mutations in cytosolic and mitochondrial isocitrate dehydrogenase? Why would you expect such mutations to increase risk of cancer?

Weblem 9.21. Given a protein of known structure but unknown function, a logical beginning to the analysis would be to try to identify the active site. Finding a crevice in the surface is much the same problem that presents itself in the early stages of drug design (see Chapter 5). Try to identify a crevice in the structure [2MCP] (small shallow site) or [1GS5] (large deep site). Delete the ligand from your calculation, and reintroduce it at the end to check the results of other methods. (a) One possibility is to use available resources for computing the accessible surface area of the protein (for instance, `http://curie.utmb.edu/getarea.html`). What would you expect to be the difference in which residues are surface-exposed if you vary the radius of the probe object? (b) Try some of the tools offered at `https://bioinformatictools.wordpress.com/tag/active-site-prediction/`. Is there consensus? Which if any methods gave the correct answer?

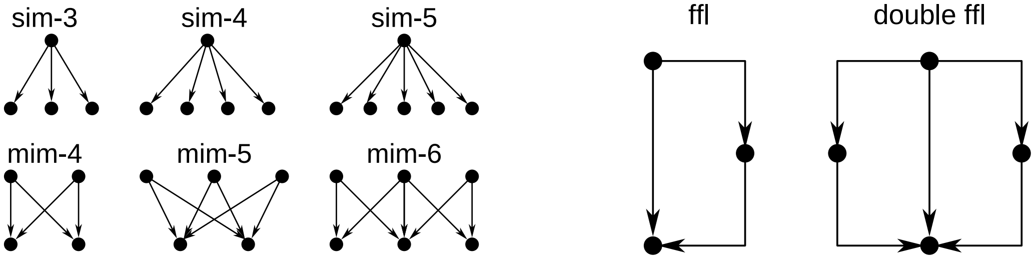
Weblems – Chapter 10

Weblem 10.1. Define the following terms: (a) interactome (b) metabolome (c) signalome. (d) More difficult: can you think of, and define, a reasonable 'ome' that has not yet been proposed?

Weblem 10.2. The catalase-peroxidase KatG of *M. tuberculosis* activates the drug isoniazid. The most common mutation in KatG associate with resistant strains is S315T. From a model of *M. tuberculosis* KatG based on the structure of the homologue from *Burkholderia pseudomallei* (PDB entry [1MWV]), suggest a mechanism for the reduced activity of the mutant to activate isoniazid while retaining activity against smaller substrates.

Weblem 10.3. Search the Database of Interacting Proteins DIP for proteins that bind actin. How many did the search return? List five of them and report what evidence for the interaction was given.

Weblem 10.4. Some simple combinations of the basic network motifs (mim, sim, ffl) appear in this diagram. Create a data structure from them suitable as input to SOM-PAK. What is the output of SOM-PAK? What clusters does it report? Into what classes could they easily be separated?



Weblem 10.5. GRNsight is web server for visulizing models of networks. (`http://dondi.github.io/GRNsight/`) For an example of its use, Camacho *et al.* have derived a small network involved in controlling the coldshock response in yeast.[1] Download Table 9 from `https://link.springer.com/article/10.1007%2Fs11538-015-0092-6`. Display this network using GRNsight. On a copy of the result, indicate with a highlighter the answers to the following questions. (a) Which node has greatest number of incoming edges? (b) Which node has greatest number of outgoing edges? (c) What is longest cycle? (d) Which genes would be affected by a deletion of HSF1? (e) Does one gene or a cluster of genes seem to function as a hub?

Weblem 10.6. STRING `https://string-db.org/cgi/input.pl` is a database of protein-protein interactions. Search the database for human thrombin. Note that it is possible to select degree of neighbours, and clusters. By clicking on the node: (a) What protein is Fga? (b) What is the nature of the interaction of Fga with thrombin? (c) What is Serpind1? (d) What is the nature of the interaction of Serpind1 with thrombin?

Weblem 10.7. Pyruvate is a branch compound in intermediary metabolism. Find pathways involving pyruvate in KEGG (`http://www.genome.jp/kegg/`) (a) What compounds are substrates in reactions that produce pyruvate? (b) To what compounds can pyruvate be converted directly? (c) Which of these reactions are active in humans?

Weblem 10.8. Identify a protein for which a structure of reasonable quality is available in the wwPDB for the amino-acid sequence reflecting all the exons in the gene, but for which at least the amino-acid sequences of one or more splice variants are known. Draw a picture of the protein, indicating in a separate colour those regions that would be absent in a splice variant. What can you say about the structural role of these missing segments – that is, does it look as if the splice-variant might be structurally similar to the corresponding substructure of the full protein? Can you find examples of splice variants for which the experimental structures were indendently determined? If so, what if any significant structural differences do you see in the common regions of the structures of the splice variants?

Weblem 10.9. Interference with subcortical dopamine D2 receptor *(DRD2)* signaling is a causative factor in drug addiction, schizophrenia, and Parkinson's disease. What mutations affect *DRD2* splicing? What are the clinical consequences of these mutations?

[1] Camacho, E.T., Entzminger, S.D. & Wanner, N.C. (2015). Parameter estimation for gene regulatory networks from microarray data: cold shock response in *Saccharomyces cerevisiae Bull. Math. Biol.* **77,** 1457-1492.