Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

SKU: 9780199936991-SOLUTIONS

Concepts in Bioinformatics and Genomics

Solutions Manual

Jamil Momand / Alison McCurdy

Silvia Heubach / Nancy Warter-Perez

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Chapter 1

Review of Molecular Biology

EXERCISE 1

A good resource for information on disease-related molecules in the Online Mendelian Inheritance of Man (OMIM). OMIM is linked to the National Center of Biotechnology Information (NCBI) website. To use this website optimally, it is important to become familiar with chromosomes. Chromosomes are segments of DNA bound to protein found in the nucleus of eukaryotic cells. For many organisms, their genomes are distributed as chromosomes. Humans have 23 sets of chromosomes (46 total chromosomes). Each chromosome is divided into two major areas. The short arm (p-arm) is located above the centromere, and the long arm (q-arm) is located below the centromere. Telomeres are located at the very ends of the chromosome (Figure 1-17).

To determine the location of a gene on a chromosome, chromosomes can be stained with special dyes that create a series of bands on the chromosomes. If a gene is located at 8q21.3, that would mean chromosome 8, q-arm, band 21.3. The higher the band number, the farther away from the centromere the gene is located.

a. Give the chromosome location (in terms of banding position of the gene that causes sickle cell anemia.

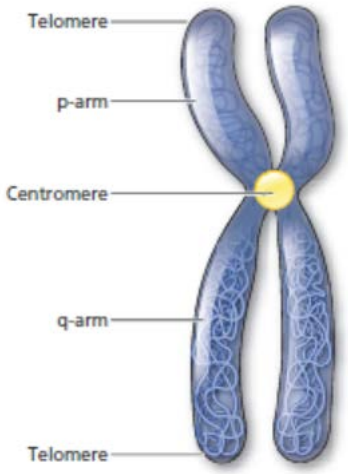


Figure 1-17. Human chromosome showing the centromere, telomeres, p-arm, and q-arm.

b. Give the name of the gene that causes sickle cell anemia.

c. Find the nucleotide change and amino acid change that leads to sickle cell anemia.

d. Explain the change in distribution of spots in the peptide map (Figure 1-16) based on properties of the side chain of the mutant amino acid in the disease.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

e. If sickle cell anemia is so devastating, why has it persisted in the population for such a long time? Give a molecular explanation.

Answer 1

a. chromosomal location 11p15.4

b. *HBB* (human beta globin)

c. A to T transversion in the third position of the seventh codon of the hemoglobin beta gene. E6V mutation-glutamate is changed to valine at position 6 in the mature protein. The mature protein sequence is highlighted in yellow below. The glutamate at position 6 is highlighted in pink.

```
1 MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK

61 VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG

121 KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

d. The mutant peptide containing the valine is more hydrophobic and less charged. It will stay closer to the neutral point on the paper.

e. Ongoing research suggests that sickle cells prevent the pathological effects of malaria. Red blood cells with at least one mutant copy of *HBB* are less prone to infection by the mosquito carrying the protozoan that causes malaria. Individuals heterozygous for the *HBB* sickle-cell allele do not have sickle-cell disease, and have increased malaria resistance compared to individuals homozygous for the normal *HBB* allele. The increased survival of heterozygous individuals promotes selection of the *HBB* sickle-cell allele. This is why the *HBB* sickle-cell allele persists. *Plasmodium* is the protozoan that causes malaria. There are several suggestions as to how the sickle human hemoglobin protein (Hb) operates at the molecular level to increase tolerance to *Plasmodium* infection. Here is one:

> "Sickle human hemoglobin (Hb) confers a survival advantage to individuals living in endemic areas of malaria, the disease caused by *Plasmodium* infection. As demonstrated hereby, mice expressing sickle Hb do not succumb to experimental cerebral malaria (ECM). This protective effect is exerted irrespectively of parasite load, revealing that sickle Hb confers host tolerance to *Plasmodium* infection. Sickle Hb induces the expression of heme oxygenase-1 (HO-1) in hematopoietic cells, via a mechanism involving the transcription factor NF-E2-related factor 2 (Nrf2). Carbon monoxide (CO), a byproduct of heme catabolism by HO-1, prevents further accumulation of circulating free heme after ***Plasmodium*** infection, suppressing the pathogenesis of ECM. Moreover, sickle Hb inhibits activation and/or expansion of pathogenic $CD8^+$ T cells recognizing antigens expressed by ***Plasmodium***, an immunoregulatory effect that does not involve Nrf2 and/or HO-1. Our findings provide insight into molecular mechanisms via which sickle Hb confers host tolerance to severe forms of malaria. (A. Ferreira, I. Marguti, I. Bechmann, et al. "Sickle Hemoglobin Confers Tolerance to Plasmodium Infection." *Cell* 145, no. 3 [April 2011]. 398-409. doi: 10.1016/j.cell.2011.03.049.PubMed PMID: 21529713)

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

EXERCISE 2

One important exercise a bioinformatician performs is to compare amino acid sequences. One reason to make comparisons is to determine the parts of the proteins that are critical for function. These regions are generally conserved within proteins that perform the same duties. Conserved regions are those that have nearly the same amino acid sequences. Proteins that perform the same duties are called homologs and can be found in different species. For example, p53 from humans and p53 from frogs perform the same functions. There are some regions within these proteins that will be similar in both humans and frogs. We call these regions conserved sequences. A multiple sequence alignment allows the bioinformatician to readily line up amino acid sequences of related proteins. The conserved regions are identified in the alignment. For this exercise, perform a multiple sequence alignment of three homologs of cytochrome C. The three homologs are from human, yeast, and dog, and the accession numbers for these sequences are AAA35732, NP_001183974, and 1YCC. The first two sequences can be found in the protein database at the NCBI. The last sequence can be found in the structure database at the Protein Data Bank. Print out your multiple sequence alignment result and attach a short paragraph explaining how the alignment gives you a clue as to which parts of the cytochrome C protein you would hypothesize are most important to its function. (The function is the same in all three organisms.)

For those of you unfamiliar with NCBI, here are specific instructions:

- Go to the NCBI website.
- Use the dropdown menu to search in the "Protein" database.
- Enter the accession number and click "Search."
- Change the format to FASTA.
- Copy the FASTA output into a sequence alignment window of a website that hosts a sequence alignment program. Make sure the header (the line with the > symbol) is placed on top of the sequence within the window.
- Repeat for each FASTA output of the remaining two proteins.
- Once all three sequences are pasted into the sequence alignment input window, run the program

Answer 2

Note that in the alignment output the asterisks (*) indicate perfect matches across the sequences, colons (:) indicate highly similar amino acids, periods (.) indicate slightly similar amino acids and no symbol indicate dissimilar amino acids.

```
   CLUSTAL O(1.2.1) multiple sequence alignment

yeast    TEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDANIKK
human    ----MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK
dog      ----MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNK
              *..:** .:*  :* ********* **.******:***::*** *:*** ** :*



yeast    NVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKACE-
human    GIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
dog      GITWGEETLMEYLENPKKYIPGTKMIFAGIKKTGERADLIAYLKKATKE
          : * *:.: *** *********** * *:**  :* ***:***** :
```

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Sequence alignment shows that these three proteins are well conserved, especially sequences QCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDAN and EYLTNPKKYIPGTKM. These may be regions that have important features for activity and structure of the protein.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Chapter 2

Information Organization and Sequence Databases

EXERCISE 1

For this exercise, you will need to access GenBank by going to the NCBI website and using the dropdown menu to search "Nucleotide." Note that the definition of the coding strand is the strand of DNA within the gene that is identical to the transcript (for genetic code, see Figure 1.11). On the other hand, the template strand is the strand that is complementary to the coding strand.

a. Use the following accession number to access the nucleotide sequence in GenBank: CU329670

b. Go to the FEATURES section of the record.

c. Link to the CDS to gain access to the first 5662 nucleotides of the sequence.

d. Name the protein product of the CDS.

e. Write the first four amino acids (starting from the N-terminus).

f. Write the nucleotide sequence of the coding strand that corresponds to these amino acids.

g. Write the nucleotide sequence of the template strand that corresponds to these amino acids.

h. Using the sequence shown in the record, give the nucleotide number range that corresponds to these amino acids.

Answer 1

d. RecQ type DNA helicase

e. MVVA

f. 3'tcgctgctggta5'

g. 5'agcgacgaccat3'

h. 5651–5662

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

EXERCISE 2

Genes in eukaryotes are often organized into exons and introns, which require splicing to produce an mRNA that can be translated. The gene organization is the order of the DNA segments that comprise the gene starting with the promoter, the first exon, the first intron, the second exon, and so on. The interspersed introns can make gene identification difficult in eukaryotes—particularly in higher eukaryotes with many introns and alternative spliced mRNAs. Prediction of many genes and their organization has been based on similarity searches between genomic sequence and known protein amino acid sequences and genomic sequence and the corresponding full-length cDNAs. cDNAs are reverse-transcribed mRNAs and therefore generally do not contain intron sequences. cDNAs (i.e., copied DNA) can be considered mRNAs. A comparison of a genomic sequence (with introns) to its corresponding cDNAs will reveal where introns begin and end. GenBank will contain the genomic sequence and the cDNA sequence. To find out the structure of the gene (i.e., the arrangement of the exons and introns) we simply need to perform a sequence comparison between the genomic sequence and the cDNA sequence. Shown below is a genomic sequence from the species *C. elegans*. The Basic Local Alignment Sequence Tool (BLAST) can be used to elucidate part of the gene organization (arrangement of exons and introns) of a genomic sequence. BLAST can be used to compare genomic DNA sequence with all RNA sequences (i.e., cDNA sequences) in GenBank. The top hit of the output will be a sequence comparison between your sequence (the query sequence) and the most similar sequence in the database (subject sequence). Subsequent hits will display sequence comparisons between the query sequence and subject sequences that are increasingly less similar. If all hits have 100% identity, use the hit with the most extensive percentage coverage to report on. Use the nucleotide BLAST tool and appropriate databases to construct a schematic diagram that shows the arrangement of introns and exons in the genomic sequence. Remember that the species source of genomic sequence is *Caenorhabditis elegans*.

```
ATTTTTAAAAATGTACAAAATCAAACGCCCTACAAATCATGTGTGTGAAGAAGAATAATAACTACAT
ATCTATTTATATTTACCGAATAAATATATATTCATCAATTAACCTGAAGAACAAACGAATTCGGCTAC
AGGCGTCGATCAGTCTCGAATCTAGTAACAACAAGAGAGCAATACGAAAACCGGTAAATCAATAGG
GGGAAGCGAAACAGTAGGTACAAATTGGAGGGGAAGCACCAATACATTAGGTGGGGGGTACGACTTG
AAAAATGAGCTGATTTTCGAATAGTTAAAGCGATGATCGTGTCCGAAAAACAGTTCATTTTTCAAG
ACAACATTGAGACTGGGAGTACGGGGAAGCTCATTTACGGTGAGAGGAATTGGTGAGATCTTTAGA
ATATGCTTAAGGAGTTGGGGTGGCTGGAGAAGTTCCTGTAGCCTCCGTGCCGGGATTCGATGGAGA
AGTCGTTGCGGCTGGTCCCTTTTCCTTCACTGGTGCTGGATCCTTGGCTGGAAGACATATGCGTGGC
TTGACAGTCGATGAGGTGCGAGCCGACGAGTCCTTGTGAACTTCGTATCTGGAAATATTTTACTTAGA
TAGCAAATACTAAAATTGTAAAATTACCTCAAAATCTCAGTATCCGGAATGCTCAATTTCTGCTTCA
AAACCTGTCCGATGCGAAGATTGACATCATCGCGAGTAGCATCACGAGTCCACAAGGAAACCTTGT
CACCCTTTTGACGAACATTCACGACAGCTCCGCAGATGTAGTCTCCGTACTCGTCGAATTGCTCTC
CAACAATAGCCATCAACAGCTCCAACCAGTAGTGATCGAGCAATTGCGTTCTTCTCTGAAGCTTCTA
TGATTCATTGAATAAAATATATTTCTCAAAACGTACTTGCTTATCGACAACAACCAACCAACGTCCA
CCTTGAACGTTGTTGACGTCCTCCCACATTGGCTTGATTCCTTCCTTGAACAAGTAATAATCGGATCC
CCAGTTCAATCCTCCGGCAGACTGAATGTGATTGTACAGCGACCAGAAGTCCTCGACAGTGTCGAAA
AGTGAAACCATCTGGAAAAAATCGATAAAAGACGTATTTAAAAATCTTCTACCTTCAGACAATCCTC
CCATTCCTTGTTACGGTCAGCTTTCAAGTACCAGAGAGCCCAGCGATTCTGGAGGGGGTGTCTGGTGA
GAAGCTCTGGAGGAACTGAAGCATCGGACGCATTCACATCGCCGGAAGCTGACAATGCTTTGTTTTCC
GCTACGGATGTGCTCATTTAGCTGAAAATAGGTAATATTATATACGATTAGAGCTCGGAAAACGATA
AAATAGAGAAGAGTATGAATTTGGTTCAAATAACTCGGATTTTATAGGAAATTTTGTTTTACTGCAC
ATTTTCGGCTAGTTTCCAAGCTTTTTAGATTTTTCAAGTGTAATTGGTAACATCGGGCACAATAAAT
TGATATTAAAGCTTGGAAAACAATAA
```

Note: The sequence can be found on the publisher's website.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

In addition to construction of the schematic diagram, answer the following:

a. Give the name and accession numbers of each distinct mRNA produced from this gene.

b. Give the names and accession numbers of the protein product(*s*).

c. Note the numbering of the sequences in the alignments. Does the database genomic sequence progress in the same direction as the database mRNA? In other words, is it the same orientation (see below):

```
  1 ..... 114   = query
 61 ..... 98    = subject or opposite orientation (below)

  1 ..... 114   = query
 98 ..... 61    = subject
```

d. Consider the alignment of the query sequence and the subject sequence. What does the orientation of the sequences relative to each other tell you about the sequence that was used as the query sequence?

e. Give the amino acid sequences separately translated from each exon sequence of the longest transcript.

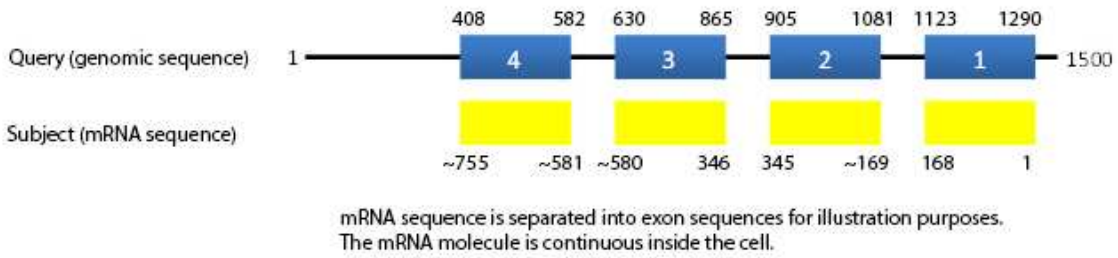
f. How many alternative splice variants are associated with this genomic sequence? List their accession numbers.

g. Give the chromosome position numbers that denote the start and end of the *TP53* gene. The position number is the base number on chromosome 17. Calculate the length of the primary transcript. Give the lengths, in base pairs, of each exon and intron that is used for the transcription of *TP53* into mRNA isoform a. Cite the sources you used to gather your information.

Answer 2

A schematic diagram of the genomic sequence is shown below on the first line with exons shaded in blue (these exons correspond to isoform b transcript). The exon numbers at the top of diagram proceed from right to left because the given genomic sequence is the complement (also known as the negative strand or non-coding strand). These numbers mark the boundaries of the exons within the genomic sequence. A schematic diagram of the mRNA sequence is shown on the second line. The numbers at the bottom of the diagram mark the approximate nucleotide boundaries of the exons sequences (shaded yellow) where nucleotide 1 is the first nucleotide of exon sequence 1 (from BLAST analysis against RefSeq_genome sequence database—see Accession number NC_003283). Note that the numbers at the bottom may not match the output from BLAST because some nucleotides within the intron sequences at exon/intron junctions in the genomic sequence will be identical to nucleotides at the exon/exon junctions in the mRNA.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual



mRNA sequence is separated into exon sequences for illustration purposes.
The mRNA molecule is continuous inside the cell.

a. 1. Caenorhabditis elegans Eukaryotic translation initiation factor 4E-3 (ife-3), partial mRNA (longest mRNA) accession number: NM_070722; 2. Caenorhabditis elegans Eukaryotic translation initiation factor 4E-3 (ife-3), partial mRNA mRNA accession number: NM_171920; 3. Caenorhabditis elegans Eukaryotic translation initiation factor 4E-3 (ife-3), partial mRNA mRNA accession number: NM_070723 (shortest mRNA)

b. 1. Eukaryotic translation initiation factor 4E-3 [Caenorhabditis elegans] accession number: NP_503123 (251 amino acids); 2. Eukaryotic translation initiation factor 4E-3 [Caenorhabditis elegans] accession number: NP_741502 (250 amino acids); 3. Eukaryotic translation initiation factor 4E-3 [Caenorhabditis elegans] accession number: NP_503124 (248 amino acids)

c. Opposite orientation

d. The query sequence matches the template strand of the gene.

e. 1–168 Exon 1: **Met** S T S V A E N K A L S A S G D V N A S D A S V P P E L L T R H P L Q N R W A L W Y L K A D R N K E W E D C L K

167–345 Exon 2: **Met** V S L F D T V E D F W S L Y N H I Q S A G G L N W G S D Y Y L F K E G I K P **Met** W E D V N N V Q G G R W L V V V D K Q

346–581 Exon 3: K L Q R R T Q L L D H Y W L E L L **Met** A I V G E Q F D E Y G D Y I C G A V V N V R Q K G D K V S L W T R D A T R D D V N L R I G Q V L K Q K L S I P D T E I L

408–580 Exon 4: R Y E V H K D S S A R T S S T V K P R I C L P A K D P A P V K E K G P A A T T S P S N P G T E A T G T S P A T P T P **Stop**

f. 3 splice variants: NM_070722 , NM_171920, NM_070723.

g. This answer may change slightly depending on the version of the human genome used.

Chromosome position number at end of TP53 gene: 7590863
Chromosome position number at beginning of TP53 gene 7571720
Primary strand length = 7590863–7571720 + 1 = 19,144 bases

The coding strand of TP53 is on the complementary (negative) strand.

Gene segment	Ex.1	In.1	Ex.2	In.2	Ex.3	In.3	Ex.4	In.4	Ex.5	In.5	Ex.6
(# of base pairs)	(169)	(10,754)	(102)	(117)	(22)	(109)	(279)	(757)	(184)	(81)	(113)

Gene segment	In.6	Ex.7	In.7	Ex.8	In.8	Ex.9	In.9	Ex.10	In.10	Ex.11
(# of base pairs)	(567)	(110)	(343)	(137)	(92)	(74)	(2819)	(107)	(919)	(1289)

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Primary strand length = 19,144

How to answer this question: Obtained reference sequence information for p53 isoform a mRNA from Entrez gene with keywords TP53 and *Homo sapiens* (accession number: NP_000537.3). Used annotations in Refseq record of isoform a to obtain exon positions and lengths in the mRNA. Used BLAST against RefGenome database *Homo sapiens* with each exon sequence to obtain chromosome positions that correspond to beginning and end of each exon. The absolute value of the difference of the chromosome position corresponding to end of one exon and beginning of next exon minus 1 gives the number of bases in the intron. One final note: There are likely to be several approaches to arriving at the correct answer to this question.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Chapter 3

Molecular Evolution

EXERCISE 1

Obtain the human p63 isoform 1 protein sequence and the human p53 isoform a protein sequence. Use the Needleman-Wunsch program (http://www.ebi.ac.uk/Tools/psa/) to align the two proteins. Use the following parameters when you run the Needleman-Wunsch program:

Matrix: BLOSUM62; Gap Open Penalty: 10; Gap Extend Penalty: 0.5; End Gap Penalty: False; End Gap Open Penalty: 10; End Gap Extend Penalty: 0.5

a. Which amino acid sequences within p53 are conserved in p63? (Hint: Find a region longer than 50 fifty amino acids with no gaps spanning longer than 5 five amino acids.)

b. A second paralog of p53 is p73. Which amino acid sequences within p53 isoform a are conserved in p73 isoform a?

c. Is the particular domain common within p53, p63, and p73 associated with a specific function?

Answer 1

a. p53 and p63 are conserved from aa 97 through aa 312 within human p53.

b. p53 and p73 are conserved from aa 97 through aa 291 within human p53.

c. This domain binds DNA in a sequence-specific fashion.

EXERCISE 2

Retinoblastoma is a rare childhood cancer of the retina. The retina is located in the back of the inside of the eye and is responsible for detecting light. Alfred Knudson, at the University of Texas at Houston, observed that retinoblastomas fall into two classes. In Class I, the retinoblastoma is first diagnosed in children at the mean age of three, and the tumors are often found in both eyes. In Class II, the retinoblastoma is diagnosed at the mean age of five, and a single tumor is detected in one eye. Given what you know about germ cell mutations and somatic cell mutations, give a plausible explanation for the observation of two classes of retinoblastoma.

Answer 2

Class I cases are characterized by a germ cell mutation in one allele of the retinoblastoma gene (Rb). As the retinal cells of the infant proliferate during the first few years of life, mutations in the second Rb allele occur in more than one cell. This is a somatic cell mutation. The independent somatic cell mutations in separate cells cause several retinal cells to form tumors. In other words, there are multiple tumors in both eyes. This accounts for the observation of tumors in both eyes of the patients

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

and for their relative early onset. Class II cases are characterized by two somatic cell mutations in one retinal cell. Because it takes longer for two independent mutations to occur (rather than one mutation) in the retinal cell, only one tumor is created per patient and thus only one eye shows the cancer. This requirement for two independent mutations also accounts for the relatively late onset of the disease. These observations caused Knudson to formulate the "two-hit hypothesis" and led to the prediction for the existence of tumor suppressor genes.

EXERCISE 3

Peyton Rous (1879–1970) was a relatively young man when, in 1911, he discovered a virus that causes sarcomas in chickens. The virus was named Rous sarcoma virus (RSV). Later, it was found that RSV is a retrovirus that contains an oncogene, *v-src*, in its RNA genome. Perform pairwise global alignments with the Needleman-Wunsch algorithm between v-src protein and the following proteins: chicken c-src, human c-src, and mouse c-src proteins. The "c-" prefix is short for cellular. Sometimes proto-oncogenes are distinguished from viral oncogenes with the prefix "c-" and "v-," respectively. Report the identities for each pairwise alignment. Given what you know about the origin of v-src, does the result match your expectations? Use the Schmidt-Ruppin A strain of RSV as your source for the v-src sequence.

Answer 3

Chicken: 94% identity
Human: 88% identity
Mouse: 87% identity

The high identity shared between v-src and chicken c-src (94%) is consistent with the idea that RSV was first generated in a chicken and not a human or a mouse. Here is a possible scenario: a retrovirus captured a chicken c-src proto-oncogene in a mechanism similar to what is described in Figure 3-9. The resulting virus, called RSV, was able to out-compete the parental retrovirus for survival. RSV confers a growth advantage onto cells it infects because the c-src promotes cell growth. As the virus continues to regenerate itself, it creates a few mutations that make v-src a more powerful cell growth inducer than its c-src counterpart.

EXERCISE 4

In the Li-Fraumeni syndrome *TP53* is often mutated. Compare the mRNAs from a Li-Fraumeni patient to transcript variant 1 of wild-type p53 with the Needleman-Wunsch global alignment software program.

> Matrix: DNAfull; Gap Open Penalty: 10; Gap Extend Penalty: 0.5; End Gap Penalty: False; End Gap Open Penalty: 10; End Gap Extend Penalty: 0.5

Classify the mutations in the areas that overlap. List the nucleotide location (using the wild-type p53 nucleotide numbering system as the reference) of the mutation. Indicate whether it is a point mutation or indel, transversion or transition, missense, or other type of DNA mutation. Give the amino acid substitution(s), if there is any, that occurs in the p53 protein. The p53 transcript sequence from a Li-Fraumeni patient can be obtained from GenBank with accession number BT019622.1. For a codon table, please see Chapter 1.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Answer 4

Accession number of *Homo sapiens* tumor protein p53 (TP53), transcript variant 1, mRNA: NM_000546.5

Location(nt)	Nucleotide change	Point or indel	Transition/ transversion	Missense, silent, or other	Amino acid change
417	C→G	Point	Transversion	Missense[a]	P72→R72
764	C→A	Point	Transversion	Missense[b]	L187→M187
1384	– →A	Indel		Silent	None

[a]Polymorphism

[b]Missense mutation in the DNA binding domain leads to loss of tumor suppressor function.

Below are segments from sequence alignment with nucleotide changes in boldface. "Li" is mRNA from Li-Fraumeni individual and "wt" is wild-type tumor protein p53 transcript variant 1.

```
Li      199 CCAGAGGCTGCTCCCC**G**CGTGGCCCCTGCACCAGCAGCTCCTACACCGGC  248
            ||||||||||||||||.||||||||||||||||||||||||||||||||||
wt      401 CCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCAGCTCCTACACCGGC  450

Li      549 AGATAGCGATGGT**A**TGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAA  598
            |||||||||||||.||||||||||||||||||||||||||||||||||||
wt      751 AGATAGCGATGGTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAA  800

Li     1149 CATGTTCAAGACAGAAGGGCCTGACTCAGACT**A**G----------------  1182
            |||||||||||||||||||||||||||||||| |
wt     1351 CATGTTCAAGACAGAAGGGCCTGACTCAGACT-GACATTCTCCACTTCTT  1399
```

Below are segments from sequence alignment with amino acid changes in boldface. Query is protein sequence is from Li-Fraumeni individual and subject is wild-type tumor protein p53 isoform a.

```
Query 61   DEAPRMPEAAP**R**VAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK  120
           DEAPRMPEAAP VAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
Sbjct 61   DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK  120
Query 181  RCSDSDG**M**APPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS  240
           RCSDSDG+APPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
Sbjct 181  RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS  240
```

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

EXERCISE 5

The sequence of two different forms of a gene starting with ATG codon is shown below. Which of the base differences in the second sequence are synonymous changes, and which are non-synonymous changes?

```
Form 1: ATGTCTCATGGACCCCTTCGTTTG
Form 1: ATGTCTCAAAGACCACATCGTCTG
```

This exercise is adapted from Hartwell et al. (2008).

Answer 5

Recall that synonymous changes are those nucleotide changes that do not change the amino acids. Nonsynonymous changes are nucleotide changes that do change the amino acids. The sequences are rewritten with corresponding amino acids.

```
Form 1:  ATG  TCT  CAT  GGA  CCC  CTT  CGT  TTG
         Met  Ser  His  Gly  Pro  Leu  Arg  Leu

Form 2:  ATG  TCT  CAA  AGA  CCA  CAT  CGT  CTG
         Met  Ser  Gln  Arg  Pro  His  Arg  Leu
```

Underlined nucleotides are changed. Those in blue font are synonymous changes (no change in amino acid) and those in red font are non-synonymous changes (change in amino acid).

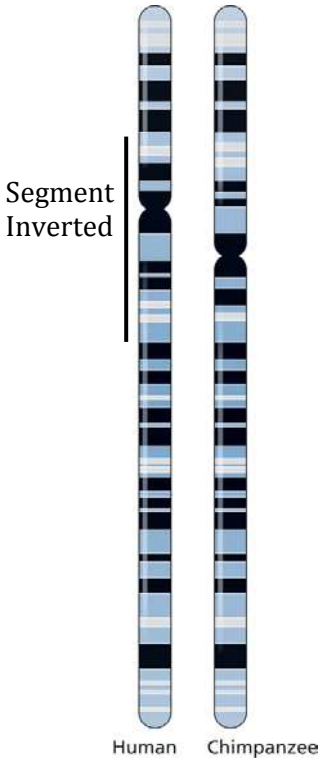
EXERCISE 6

Synteny is the conservation of the physical order of genetic loci in the genomes of two species. For example, a segment of mouse chromosome 2 is similar to the entirety of human chromosome 20 (chromosome 20 is one of the smaller chromosomes in humans). Cytogeneticists use chemicals to stain chromosomes to produce banding patterns. The banding patterns mark where sections of genes are located. Similar banding patterns denote regions of synteny. The figure shows the banding patterns on human chromosome 5 and the chimpanzee chromosome 5. Describe the region(s) where an inversion event occurred. Exercise adapted from Lesk (2008).

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual



Answer 6



The region surrounding the centromere is inverted. Interestingly, many inversion mutations can account for the differences between humans and other primates.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Chapter 4

Substitution Matrices

EXERCISE 1

Give two reasons why Dayhoff used closely related sequences to create her matrix of amino acid substitutions.

Answer 1

One reason is that it was easy to infer ancestral sequences from phylogenetic trees. Another is that consideration of closely related sequences reduced the possibility of counting double substitutions, because the Markov model only depends on the current states of the amino acids.

EXERCISE 2

Use the value for $\mathbf{S}_{W,W}$ in the PAM250 log-odds substitution matrix to calculate the value of $\mathbf{M}_{W,W}$ in the PAM250 mutation probability matrix.

Answer 2

$\mathbf{S}_{W,W} = 1.7$

$\mathbf{S}_{W,W} = 10\log_{10}(\mathbf{R}_{W,W})$

$\mathbf{R}_{W,W} = 50.1187$

$\mathbf{R}_{W,W} = \mathbf{M}_{W,W} / f_W$

$\mathbf{M}_{W,W} = (\mathbf{R}_{W,W})(f_W) = (50.1187)(0.01) = 0.501187$

Note that the $\mathbf{M}_{W,W}$ value is much lower in the PAM250 mutation probability matrix than in the PAM1 mutation probability matrix. The $\mathbf{M}_{W,W}$ in the PAM250 mutation probability matrix was obtained by multiplying the PAM1 mutation probability matrix by itself 250 times.

EXERCISE 3

Download the PAM250 and PAM30 substitution matrices and place them side by side on your computer screen. Describe the differences between the two matrices. Explain why you see these differences.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Answer 3

PAM30 substitution matrix

#	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	6	-7	-4	-3	-6	-4	-2	-2	-7	-5	-6	-7	-5	-8	-2	0	-1	-13	-8	-2	-3	-3	-3	-17
R	-7	8	-6	-10	-8	-2	-9	-9	-2	-5	-8	0	-4	-9	-4	-3	-6	-2	-10	-8	-7	-4	-6	-17
N	-4	-6	8	2	-11	-3	-2	-3	0	-5	-7	-1	-9	-9	-6	0	-2	-8	-4	-8	6	-3	-3	-17
D	-3	-10	2	8	-14	-2	2	-3	-4	-7	-12	-4	-11	-15	-8	-4	-5	-15	-11	-8	6	1	-5	-17
C	-6	-8	-11	-14	10	-14	-14	-9	-7	-6	-15	-14	-13	-13	-8	-3	-8	-15	-4	-6	-12	-14	-9	-17
Q	-4	-2	-3	-2	-14	8	1	-7	1	-8	-5	-3	-4	-13	-3	-5	-5	-13	-12	-7	-3	6	-5	-17
E	-2	-9	-2	2	-14	1	8	-4	-5	-5	-9	-4	-7	-14	-5	-4	-6	-17	-8	-6	1	6	-5	-17
G	-2	-9	-3	-3	-9	-7	-4	6	-9	-11	-10	-7	-8	-9	-6	-2	-6	-15	-14	-5	-3	-5	-5	-17
H	-7	-2	0	-4	-7	1	-5	-9	9	-9	-6	-6	-10	-6	-4	-6	-7	-7	-3	-6	-1	-1	-5	-17
I	-5	-5	-5	-7	-6	-8	-5	-11	-9	8	-1	-6	-1	-2	-8	-7	-2	-14	-6	2	-6	-6	-5	-17
L	-6	-8	-7	-12	-15	-5	-9	-10	-6	-1	7	-8	1	-3	-7	-8	-7	-6	-7	-2	-9	-7	-6	-17
K	-7	0	-1	-4	-14	-3	-4	-7	-6	-6	-8	7	-2	-14	-6	-4	-3	-12	-9	-9	-2	-4	-5	-17
M	-5	-4	-9	-11	-13	-4	-7	-8	-10	-1	1	-2	11	-4	-8	-5	-4	-13	-11	-1	-10	-5	-5	-17
F	-8	-9	-9	-15	-13	-13	-14	-9	-6	-2	-3	-14	-4	9	-10	-6	-9	-4	2	-8	-10	-13	-8	-17
P	-2	-4	-6	-8	-8	-3	-5	-6	-4	-8	-7	-6	-8	-10	8	-2	-4	-14	-13	-6	-7	-4	-5	-17
S	0	-3	0	-4	-3	-5	-4	-2	-6	-7	-8	-4	-5	-6	-2	6	0	-5	-7	-6	-1	-5	-3	-17
T	-1	-6	-2	-5	-8	-5	-6	-6	-7	-2	-7	-3	-4	-9	-4	0	7	-13	-6	-3	-3	-6	-4	-17
W	-13	-2	-8	-15	-15	-13	-17	-15	-7	-14	-6	-12	-13	-4	-14	-5	-13	13	-5	-15	-10	-14	-11	-17
Y	-8	-10	-4	-11	-4	-12	-8	-14	-3	-6	-7	-9	-11	2	-13	-7	-6	-5	10	-7	-6	-9	-7	-17
V	-2	-8	-8	-8	-6	-7	-6	-5	-6	2	-2	-9	-1	-8	-6	-6	-3	-15	-7	7	-8	-6	-5	-17
B	-3	-7	6	6	-12	-3	1	-3	-1	-6	-9	-2	-10	-10	-7	-1	-3	-10	-6	-8	6	0	-5	-17
Z	-3	-4	-3	1	-14	6	6	-5	-1	-6	-7	-4	-5	-13	-4	-5	-6	-14	-9	-6	0	6	-5	-17
X	-3	-6	-3	-5	-9	-5	-5	-5	-5	-5	-6	-5	-5	-8	-5	-3	-4	-11	-7	-5	-5	-5	-5	-17
*	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	1

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

PAM250 matrix

#	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

Both matrices are symmetrical. PAM30 corresponds to an amino acid difference of 25% ($\lambda = 0.25$) in two aligned sequences (percentage identity of 75%). It is expected that most of the $\mathbf{S}_{i,j}$ values will be negative because radical differences in side chain structure will be heavily penalized. Excluding amino acid pairs with X, B and Z (recall that these are non-standard amino acids), there are only six $\mathbf{S}_{i,j}$ pairs that have positive values. PAM250 corresponds to an amino acid difference of 80% in two aligned sequences (percentage identity of 20%). For PAM250, it is expected that, compared to PAM30, more $\mathbf{S}_{i,j}$ values will be positive. Excluding X, B and Z, there are 34 $\mathbf{S}_{i,j}$ pairs with positive values.

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

EXERCISE 4

Show how to calculate the value for $A_{L,I}$ in the amino acid replacement matrix A shown in Figure 4.12. Use the data in Figure 4.11 for this calculation.

Answer 4

There are two positions where residues L and I are aligned: positions 2 and 5.
Position 2 calculation for contribution to $\mathbf{A}_{L,I}$ is 5 × 2 = 10.
Position 5 calculation for contribution to $\mathbf{A}_{L,I}$ is 2 × 3 = 6.
$\mathbf{A}_{L,I}$ = 10 + 6 =16

EXERCISE 5

Download BLOSUM30 and BLOSUM80 substitution matrices and place them side by side on your computer screen. What are the differences between the two matrices? Why do you see these differences?

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual

Answer 5

BLOSUM30

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z	*
A	4	0	-3	0	0	-2	0	-2	0	0	-1	1	0	-1	1	-1	1	1	1	-5	0	-4	0	-7
B	0	5	-2	5	0	-3	0	-2	-2	0	-1	-2	4	-2	-1	-2	0	0	-2	-5	-1	-3	0	-7
C	-3	-2	17	-3	1	-3	-4	-5	-2	-3	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	-6	0	-7
D	0	5	-3	9	1	-5	-1	-2	-4	0	-1	-3	1	-1	-1	-1	0	-1	-2	-4	-1	-1	0	-7
E	0	0	1	1	6	-4	-2	0	-3	2	-1	-1	-1	1	2	-1	0	-2	-3	-1	-1	-2	5	-7
F	-2	-3	-3	-5	-4	10	-3	-3	0	-1	2	-2	-1	-4	-3	-1	-1	-2	1	1	-1	3	-4	-7
G	0	0	-4	-1	-2	-3	8	-3	-1	-1	-2	-2	0	-1	-2	-2	0	-2	-3	1	-1	-3	-2	-7
H	-2	-2	-5	-2	0	-3	-3	14	-2	-2	-1	2	-1	1	0	-1	-1	-2	-3	-5	-1	0	0	-7
I	0	-2	-2	-4	-3	0	-1	-2	6	-2	2	1	0	-3	-2	-3	-1	0	4	-3	0	-1	-3	-7
K	0	0	-3	0	2	-1	-1	-2	-2	4	-2	2	0	1	0	1	0	-1	-2	-2	0	-1	1	-7
L	-1	-1	0	-1	-1	2	-2	-1	2	-2	4	2	-2	-3	-2	-2	-2	0	1	-2	0	3	-1	-7
M	1	-2	-2	-3	-1	-2	-2	2	1	2	2	6	0	-4	-1	0	-2	0	0	-3	0	-1	-1	-7
N	0	4	-1	1	-1	-1	0	-1	0	0	-2	0	8	-3	-1	-2	0	1	-2	-7	0	-4	-1	-7
P	-1	-2	-3	-1	1	-4	-1	1	-3	1	-3	-4	-3	11	0	-1	-1	0	-4	-3	-1	-2	0	-7
Q	1	-1	-2	-1	2	-3	-2	0	-2	0	-2	-1	-1	0	8	3	-1	0	-3	-1	0	-1	4	-7
R	-1	-2	-2	-1	-1	-1	-2	-1	-3	1	-2	0	-2	-1	3	8	-1	-3	-1	0	-1	0	0	-7
S	1	0	-2	0	0	-1	0	-1	-1	0	-2	-2	0	-1	-1	-1	4	2	-1	-3	0	-2	-1	-7
T	1	0	-2	-1	-2	-2	-2	-2	0	-1	0	0	1	0	0	-3	2	5	1	-5	0	-1	-1	-7
V	1	-2	-2	-2	-3	1	-3	-3	4	-2	1	0	-2	-4	-3	-1	-1	1	5	-3	0	1	-3	-7
W	-5	-5	-2	-4	-1	1	1	-5	-3	-2	-2	-3	-7	-3	-1	0	-3	-5	-3	20	-2	5	-1	-7
X	0	-1	-2	-1	-1	-1	-1	-1	0	0	0	0	0	-1	0	-1	0	0	0	-2	-1	-1	0	-7
Y	-4	-3	-6	-1	-2	3	-3	0	-1	-1	3	-1	-4	-2	-1	0	-2	-1	1	5	-1	9	-2	-7
Z	0	0	0	0	5	-4	-2	0	-3	1	-1	-1	-1	0	4	0	-1	-1	-3	-1	0	-2	4	-7
*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	1

Concepts in Bioinformatics and Genomics 1st edition Momand Solutions Manual