Genetics and Genomics in Medicine Chapter 2

Questions & Answers

Multiple Choice Questions

Question 2.1

Regarding exons, which, if any, of the following statements is correct?

a) Some exons in protein-coding genes consist of noncoding DNA.
b) The first exon of a protein-coding gene always contains the translational start site.
c) The last exon of a protein-coding gene always contains the normal termination codon.
d) A coding exon is always translated in just one of the three possible forward reading frames.

Answer 2.1

a) Some exons in protein-coding genes consist of noncoding DNA.

Explanation 2.1

The 3' untranslated region often spans two or more exons, as occasionally does the 5' untranslated region. That means the translation initiation codon need not be in the first exon and the stop codon need not be in the last exon. Sometimes a coding exon can be simultaneously interpreted in two reading frames.

Question 2.2

Regarding the structures of amino acids, which, if any, of the following statements is incorrect?

a) The general formula for a nonionised amino acid is H_2N-CH(R)-COOH, that is, a central carbon atom is linked to an amino group, a carboxyl group, a hydrogen atom, and a side chain, R.
b) The identity of the amino acid is determined by the side chain that is connected to the central carbon atom.
c) The side chain of an amino acid is based on a branch that contains one or more carbon atoms
d) Proline is unique in having a side chain that is connected to both the central carbon and also to the nitrogen atom of the amino group.

Answer 2.2

c) The side chain of an amino acid is based on a branch that contains one or more carbon atoms

Explanation 2.2
Glycine has a side chain of just one hydrogen atom.

Question 2.3
Regarding polypeptide structure, which, if any of the following statements is incorrect.

a) A polypeptide is a polymer composed of a linear sequence of amino acids.
b) A polypeptide normally adopts a rod-like conformation, with the side chains orientated in the same direction relative to the polypeptide backbone.
c) The amino acids within a polypeptide are joined by a covalent bond known as a peptide bond.
d) Peptide bonds form by a condensation reaction between the amino group of an amino acid and the carboxyl group of its neighbor.

Answer 2.3

b) A polypeptide normally adopts a rod-like conformation, with the side chains orientated in the same direction relative to the polypeptide backbone.

Question 2.4
Regarding protein structure, which, if any, of the following statements is incorrect?

a) The primary structure is the linear sequence of amino acids.
b) The secondary structure is the path followed by the polypeptide backbone over its length.
c) The secondary structure of every protein contains an alpha-helix.
d) The structure of an alpha-helix is primarily determined by hydrogen bonding between chemical groups on the side chains.

Answer 2.4

c) The secondary structure of every protein contains an alpha-helix.
d) The structure of an alpha-helix is primarily determined by hydrogen bonding between chemical groups on the side chains.

Explanation 2.4
Not all proteins contain an alpha-helix, and the structure of such a component is primarily determined by hydrogen bonding between carbonyl and amino groups that form components of the peptide bonds.

Question 2.5
Which, if any, of the following statements is false?

a) Unlike DNA, RNAs are unmethylated, and only the four standard bases – A, C, G and U – are found in RNAs.
b) The primary structure of an RNA is the linear sequence of nucleotides.
c) The secondary structure of an RNA is dominated by hydrogen bonding between bases on the same strand.
d) Stem-loop structures are common in RNA and consist of complementary sequences that form stable base pairs, separated by a short sequence of unpaired bases.

Answer 2.5

a) Unlike DNA, RNAs are unmethylated, and only the four standard bases – A, C, G and U – are found in RNAs.

Explanation 2.5

RNAs often have multiple methylated bases as well as other minority bases such as pseudouridine, dihydrouridine, and so on (see Figure 2.4A).

Question 2.6

Which, if any, of the following statements is false?

a) A ribosome is a large ribonucleoprotein that contains multiple different types of RNA but just one type of protein.
b) During translation, a ribosome binds to the 5' end of a mRNA, slides along it until the initiator AUG codon is identified, and continues until reaching an in-frame termination codon.
c) The single type of ribosomal protein is a crucially important peptidyltransferase.
d) The peptidyltransferase catalyzes the condensation reaction that allows joining of amino acids using peptide bonds.

Answer 2.6

a) A ribosome is a large ribonucleoprotein that contains multiple different types of RNA but just one type of protein.
c) The single type of ribosomal protein is a crucially important peptidyltransferase.

Explanation 2.6

The ribosome has many proteins but the crucial peptidyltransferase activity resides in the 28S rRNA, which is therefore a ribozyme (RNA enzyme).

Question 2.7

Which, if any, of the following statements are false?

a) During translation, the individual codons of an mRNA are "read" by transient hydrogen bonding to a complementary anticodon sequence on a transfer RNA.
b) According to their anticodon, transfer RNAs usually have a specific amino acid attached to their 3' end.
c) When the anticodon of a tRNA binds to a codon with a suitably complementary sequence, the amino acid is released and becomes part of a polypeptide.
d) Translation terminates at an in-frame termination codon (UAA, UAG, or UGA in the universal genetic code) because for these codons, there are no transfer RNAs with a complementary anticodon sequence.

Answer 2.7
None.

Explanation 2.7

All are true.

Question 2.8

Which, if any, of the following statements are false?

a) Stem-loop structures are formed by base pairing of complementary sequences that are separated by a short sequence of nucleotides with unpaired bases.
b) Stem-loop structures are common occurrences in both DNA and RNA.
c) Stem-loop structures are important for structural reasons only.
d) Stem-loop structures can be important functional elements.

Answer 2.8

b) Stem-loop structures are common occurrences in both DNA and RNA.
c) Stem-loop structures are important for structural reasons only.

Explanation 2.8

b) They are not normally found in DNA, which is double stranded
c, d) They can occasionally act as functional recognition elements (see iron-response elements on page 157).

Question 2.9

Which, if any, of the following statements are false?

a) A naked DNA double helix has a stiff rod-like structure.
b) An RNA usually has a stiff rod-like structure, but occasionally has elements of secondary structure caused by intrachain hydrogen bonding

c) Some proteins have stiff rod-like structures.
d) Some proteins have globular structures.

Answer 2.9

b) An RNA usually has a stiff rod-like structure, but occasionally has elements of secondary structure caused by intrachain hydrogen bonding

Explanation 2.9

Secondary structure is prevalent in RNAs causing them to adopt complex shapes.

Question 2.10

With respect to expression of an RNA gene, which, if any, of the following statements are false?

a) RNA processing sometimes includes RNA splicing
b) Regulatory antisense RNAs sometimes undergo splicing
c) Ribosomal RNAs and transfer RNAs are formed by cleavage of RNA precursors.
d) Ribosomal RNAs and transfer RNAs undergo base modification.

Answer 2.10

b) Regulatory antisense RNAs sometimes undergo splicing

Explanation 2.10

Antisense RNAs that regulate the expression of sense transcripts are not subject to RNA splicing

Question 2.11

Roughly what percentage of our genome is made up of coding sequence?

a) 1%
b) 5%
c) 10%
d) 25%

Answer 2.11

a) 1%

Question 2.12

In addition to coding sequence, which is generally very highly conserved, what additional percentage of our genome is highly to moderately conserved?

a) 0.4%
b) 4%
c) 14%

d) 24%

Answer 2.12

b) 4%

Question 2.13

Roughly what percentage of our genome is made up of constitutive heterochromatin?

a) 0.7%
b) 7%
c) 27%
d) 57%

Answer 2.13

b) 7%

Question 2.14

Roughly what percentage of our genome is made up of transposon repeats?

a) 0.45%
b) 4.5%
c) 25%
d) 45%

Answer 2.14

d) 45%

Question 2.15

Which, if any, of the following statements are false?

a) The broadest definition of a pseudogene is any gene that has received inactivating mutations and can no longer make a functional product
b) A retropseudogene is a non-functional cDNA copy of a processed RNA that has integrated elsewhere into the genome.
c) A non-processed pseudogene is an inactive copy of a gene that has arisen by some type of gene duplication.
d) An estimated 14,000 pseudogenes are found in the human genome.

Answer 2.15

a) The broadest definition of a pseudogene is any gene that has received inactivating mutations and can no longer make a functional product

Explanation 2.15
The definition given in a) would include mutant alleles at a locus. A pseudogene has to be a *copy* of a functional gene.

Question 2.16

Which, if any, of the following statements are false?

a) Constitutive heterochromatin remains highly condensed throughout the cell cycle.
b) Unlike constitutive heterochromatin, facultative heterochromatin describes chromatin that can de-condense and behave as euchromatin under certain circumstances.
c) Most of the long arm of the Y chromatin is made up of constitutive heterochromatin.
d) In women one of the two X chromosomes in each diploid cell is heterochromatinised.

Answer 2.16
All statements are true.

Question 2.17

What is the total DNA content of (a) our genome; (b) an average human chromosome; (c) the human mitochondrial genome?

Answer 2.17

a) 3.2 Gb (= 3,200 Mb)
b) 140 Mb
c) 16.5 kb

Question 2.18

What is the approximate ratio between the DNA content of our nuclear genome and our mitochondrial genome?

a) 2,000:1
b) 20,000:1
c) 200,000:1
d) 2,000,000:1

Answer 2.18

c) 200,000:1

Explanation 2.18
Nuclear genome = 3.2 Gb = 3,200,000 kb.
Mitochondrial genome = 16.5 kb

Question 2.19

Of the various electronic resources listed below, which is (a) a genome browser; (b) a resource that identifies closely similar genes and proteins in different organisms; (c) a program that allows very fast searching of genome sequences for identity or similarity to a nucleotide or amino acid sequence query?

HOMOLOGENE
BLAT
ENSEMBL

Answer 2.19

a) ENSEMBL
b) HOMOLOGENE
c) BLAT

Question 2.20

Nearly half of our genome is composed of transposon repeats, some of which can actively transpose, occasionally causing disease by inserting into or close to genes, (causing gene inactivation or inappropriate expression of oncogenes). Many of them work by making copies that transpose and so can increase in copy number, and so there is a need to limit the number of actively transposing sequences in case they overwhelm the genome. Two types of small RNAs act to limit the spread of transposons. What are these RNAs and where do they work?

Answer 2.20

The two types of short RNA are piRNAs (Piwi-interacting RNAs) and endogenous short interfering RNAs. They work in germ cells.

Question 2.21

In what ways are the transcription and processing of our mitochondrial genes rather different from that of most of our nuclear genes?

Answer 2.21

Transcription. The two DNA strands of the mitochondrial genome are transcribed to give large multigenic transcripts that are then cleaved to release the respective rRNAs, tRNAs and mRNAs (see Figure 2.11 on page 43). In the nucleus most genes are individually transcribed, (but there are some exceptions, such as multigenic transcription of clustered 18S, 5.8S and 28S rRNA gene sequences).

RNA processing. The mitochondrial genome is a model of economy and there are no introns. Accordingly, RNA splicing does not apply to mitochondrial genes, but at least 90% of our nuclear genes undergo RNA splicing.

Fill in the Blanks Questions

Question 2.22

Fill in the blanks below with single words.

When a gene is expressed, the two DNA strands are locally unwound to allow access by the ____1_____ machinery. One of the DNA strands serves as a ____2____ for an RNA polymerase to synthesize a complementary RNA. The initial transcript, often called the ____3____ transcript, is identical in base sequence (except that U replaces T) to the sequence of the other DNA strand, which is known as the ____4____ strand (and so the opposing strand that serves as the ___2____ is also known as the ____5____ strand). The segment of genomic DNA that corresponds to the ____3____ transcript is known as the ___6___ ___7____.

Answer 2.22

1. transcriptional. 2. template. 3. primary. 4. sense. 5. anti-sense. 6. transcription. 7. unit.

Question 2.23

Fill in the blanks below.

During gene expression, the initial RNA transcript needs to undergo processing to make a mature RNA, either a ____1____ RNA or a ____2___ RNA. For many of our genes, the initial RNA transcript needs to be cleaved into pieces. Some of the pieces, called ____3____, are discarded, but other alternating pieces called ___4____ are retained and fused in the same linear order as their order when transcribed. The junctions between ___4____ and ____3____ contain some highly conserved nucleotides, notably a ____5____ dinucleotide at the beginning of ___3____ and an ___6____ dinucleotide at the ends of ____3____. For ____4_____ and _____3_____ the original definitions have been broadened to include the corresponding segments of _____7_____ ____8_____

Answer 2.23

1. messenger. 2. noncoding. 3. intron(s). 4. exon(s). 5. GU. 6. AG. 7. genomic. 8. DNA.

Question 2.24

Fill in the blanks below.

Two important RNA processing events lead to specialized end sequences in most human mRNAs: ____1____ at the 5’ end, and ____2____ at the 3’ end. The altered sequences protect

the RNA from attack by cellular ____3_____ and confer a measure of stability. In ____1____ the most distinctive change is a specialized end nucleotide, ____4_____ _____5_____, that is joined to its neighbor using a distinctive ____6_____ bond. In this case, the ____7____ carbon atom of the end nucleotide is joined to the ____7_____ carbon atom of its neighbor. In ____2_____ a sequence of about 200 ____8____ is enzymatically added to the 3' end by a dedicated enzyme called _____9_____ ____10_____.

Answer 2.24
1. capping. 2. polyadenylation. 3. exonucleases. 4. 7-methylguanosine. 5. triphosphate. 6. phosphodiester. 7. 5'. 8. adenines. 9. poly(A). 10. polymerase.

Question 2.25
Fill in the blanks below.

During evolution duplication of a gene produces two copies. The sequence of one copy may continue to be conserved (because it remains subject to ____1_____ ____2____; the other copy is free to mutate. The latter will most likely acquire deleterious mutations and degenerate to become a ____3_____. If duplication occurs at the genome level, the _____3_____ will often be located close to the parent gene. It may contain copies of the full length sequence of the parent gene (including the promoter, exons, and introns), and is known as a ____4_____ _____3_____ . Sometimes, however, the duplication involves making a cDNA copy of an mRNA after which the cDNA copy integrates into a new locus that is often very distant from the parent gene. Because the cDNA copy lacks promoter sequences, it is usually not expressed and will acquire inactivating mutations and degenerates. This type of _____3____ is known as a _____5____ ___3_____ or a ____6_____. Sometimes, the cDNA copy integrates close to a promoter sequence and is expressed, and if so, on rare occasions, the expression of this gene copy becomes an asset to the cell so that it becomes subject to ____1____ ____2____ and is a conserved functional gene. Such a cDNA copy is known as a ____7____.

Answer 2.25
1. purifying (or negative). 2. selection. 3. pseudogene. 4. non-processed. 5. processed. 6. retropseudogene. 7. retrogene.

Question 2.26
Fill in the blanks below.

Our genome has numerous identical or similar copies of certain DNA sequences. Some of these are ____1____ ____2_____, neighboring duplicated segments that are more than 1 kb in length (and often much larger), and that show more than 90% sequence identity, having duplicated very

recently during evolution. Many of our genes are present in multiple copies that are collectively known as ____3_____ ____4_____ (and often contain both functional gene copies and ____5_____). They arose by a slow process of intermittent gene duplication over sometimes long periods of evolutionary time. Extremely similar gene copies, such as the two human ___6____ - globin genes that make identical proteins, arose by evolutionarily recent gene duplications. More distantly related gene copies generally arose from comparatively ___7___ gene duplications.

Answer 2.26
1. segmental. 2. duplications. 3. gene. 4. families. 5. pseudogenes. 6. alpha. 7. ancient.

Question 2.27

In some gene families the genes are clustered in defined chromosomal regions as a result of ___1____ gene duplication. That often occurs as a result of misalignment of chromatids: over a limited chromosomal region, the DNA sequences are paired but out of register. Subsequent ___2_____ in the mispaired region can generate chromatids with two copies of a gene. Successive gene duplications results in a cluster of highly related genes. Not all the gene copies are functional: some acquire inactivating mutations to become a type of ____3_____ known as a _____4_____ ____3____. In other gene families there may be up to many hundreds of more members scattered across the genome. They often have large numbers of ____5____ ____3_____ , also known as ____6_____ that arose by copying the RNA transcripts of a functional gene using a ____7____ _____8_____ to make cDNA copies that integrated into the genome at other locations but subsequently acquired deleterious mutations.

Answer 2.27
1. tandem. 2. crossover. 3. pseudogene(s). 4. non-processed. 5. processed. 6. retropseudogene. 7. reverse. 8. transcriptase.

Question 2.28

Fill in the blanks below.

More than half of our genome is composed of families of highly repetitive DNA sequences. Close to 15% of these are composed of tandem repeats of short sequences that are found predominantly at or close to ____1____ and in other regions of ____2____ heterochromatin (which includes much of the long arm of the ___3____ chromosome, and much of the short arms of the five acrocentric chromosomes). The remaining 85% or so of the highly repetitive DNA is made up of interspersed repetitive DNA sequences that are scattered across the genome and belong to families of ____4____ repeats. Only a small fraction of the ____4____ repeats are based on DNA ____4___ (which transpose by a ___5___-and-paste mechanism). The great

majority are ____6____ repeats, some of which can actively transpose by a ___7___-and-paste mechanism (which involves using a _____8____ ____9_____ to make cDNA copies of RNA transcripts, with the copies integrating elsewhere in the genome).

Answer 2.28
1. centromeres. 2. constitutive. 3. Y. 4. transposon(s). 5. cut. 6. retrotransposon. 7. copy. 8. reverse. 9. transcriptase.

Question 2.29
Fill in the blanks below.

Retrotransposon repeats account for just over ___1___ % of our genome and are classified into three broad families. One family resembles a class of RNA virus, known as a ___2____. Like a ___2____, they contain direct repeats at their ends, known as ___3____ ____4____ repeats, and full-length family members have the same gene structure as a simple ____2____. A second family of retrotransposon repeats, known as ____5____ , has some full-length copies with sizes of 6-8 kb, and like the retrovirus-like family some of them are able to make a specialized DNA polymerase known as _____6____ ____7_____. A third family of retrotransposon repeats, known as ___8____, have short full-length sequences of between 100 and 300 bp, and are exemplified by ___9____ repeats, the most prolific DNA sequence in the human genome, with a copy number of more than ____10____ ____11_____ repeats. Only a small fraction of the retrotransposon repeats can actively transpose (most are truncated copies or have ____11_____ mutations). ___8____ repeats and other _____8_____ are unable to make a ____6____ ____7_____ but very occasionally do transpose using a ____6____ ____7_____ produced by another retrotransposon.

Answer 2.29
1. 40. 2. retrovirus. 3.long. 4. terminal. 5. LINEs. 6. reverse. 7. transcriptase. 8. SINEs. 9. Alu. 10. one. 11. million. 12. inactivating.

Essay Questions

Question 2.30

Illustrate, with examples, how noncoding RNAs are more than ubiquitous general regulators of transcription or protein synthesis.

Answer 2.30

Many noncoding RNAs are not ubiquitous, but show tissue-specific or developmental-stage-specific expression. Many of them regulate expression of a specific gene or a subset of our genes, or play very defined roles.

Individual tiny microRNAs, for example, regulate expression of certain target genes (usually by binding to target sequences in the untranslated regions of mRNAs and so specifically down-regulate the expression of the target genes). Certain other ncRNAs, including abundant circular RNAs and functional pseudogene ncRNAs, regulate specific miRNAs by competing with them for binding to their target sequences.

Many long ncRNAs are also specific regulators. Some of them are important in epigenetic regulation, such as the developmentally regulated *XIST* RNA that is crucial for X-inactivation and various long ncRNAs that are important in imprinting. Various antisense RNAs work by binding to a complementary sense transcript at the same locus in order to regulate that RNA, such as the *TSIX* RNA that regulates *XIST* expression.

Some other RNAs have specialized role such as the piRNAs that damp down excess transposon activity in germline cells.

Question 2.31

Four different levels of protein structure are recognized. What are they? Illustrate your answer with examples, wherever possible.

Answer 2.31

The primary structure is the linear sequence of amino acids in constituent polypeptides.

The secondary structure is the path that a polypeptide backbone follows within local regions of the primary structure. Common elements of secondary structure include the α-helix, the β-sheet, and the β-turn.

The tertiary structure is the overall three-dimensional structure of a polypeptide. It is the combination of all of the secondary structures.

Because some proteins are multimeric (composed of two or more polypeptide subunits that may be of more than one type) the quaternary structure refers to the aggregate structure of a multimeric protein. Note that the focus in the above has been on polypeptides, but many proteins have significant carbohydrate components (glycoproteins, proteoglycans) or attached lipids.

Question 2.32

What is the purpose of RNA splicing? Why do some of our genes not undergo RNA splicing?

Answer 2.32

RNA splicing is important in eukaryotic cells, but especially prevalent in complex multicellular organisms. A major justification is an evolutionary argument. By splitting the genetic information within genes into different little exons, it becomes possible to create new genes by recombining exons from one gene with exons from another. Thus, for example, various genetic mechanisms allow individual exons to be duplicated or swapped from one gene to another on an evolutionary timescale. See Figure 2.15 for an example. An additional source of complexity comes from using different combinations of exons to make alternative transcripts from the same gene (alternative splicing).

Question 2.33

What are the different natural ways in which proteins are chemically modified in cells and why do they need to be modified?

Answer 2.33

Proteins are naturally modified by having different chemical groups attached to the side chains of certain amino acids. The attached chemical groups can be quite small, such as an hydroxyl group, a carboxyl group, a methyl group, an acetyl group, or a phosphate group. Or they can be quite complex groups such as complex carbohydrate or lipid structures. There is specificity regarding where the chemical group is attached: that is, it is attached to a preferred type of amino acid while taking into account the local sequence context. For example, when a carbohydrate group is attached (glycosylation), it is normally attached to the hydroxyl group of a serine or threonine (*O*-glycosylation), or to the amino group of an asparagine (*N*-glycosylation). Sequence context is clear in the latter case because the asparagine that is selected to be *N*-glycosylated is almost always one that occurs in the sequence: **Asn** – X – Ser or Thr (where X is any amino acid). Chemical modification of proteins can sometimes be important in organizing the structure of some proteins (an important example would be hydroxylation of prolines in collagens), but often the modifications occur for functional reasons. such as collagens. For example, attaching small chemical groups often allows some change in the activity status of a protein, as in the case of phosphorylating proteins that are used in cell signaling. It can also allow functional changes in chromatin that are important in epigenetic control of gene regulation, such as by adding acetyl groups, methyl groups and phosphates to (usually) lysine residues in histone proteins. Sometimes, proteins are modified by attaching specific small proteins, as in the case of adding a chain of ubiquitin proteins to a protein that may be damaged or harmful in some way so that the cell knows that it is time to degrade the protein.

Question 2.34

What roles do snRNA, snoRNA and scaRNA have in RNA maturation? Do any of them participate in other functions?

Answer 2.34

snRNAs are integral components of the spliceosome and so are required for the process of RNA splicing. Both snoRNA and scaRNA are required for modifying bases in certain RNAs during RNA processing: snoRNAs allow base modification in ribosomal RNA, and scaRNAs are responsible for base modification in the case of transfer RNA.
Certain snRNAs and snoRNAs are known to have other functional roles. For example, U1 and U2 snRNAs are ubiquitous transcriptional regulators, and the HBII-52 snoRNA regulates splicing of specific target genes.

Question 2.35

What is the type of natural selection that is responsible for strong evolutionary conservation of functionally important DNA sequences, and how does it work?

Answer 2.35

A type of natural selection called *purifying* selection (or *negative* selection) acts to conserve functionally important sequences, whereas unimportant sequences are not conserved during evolution. If a sequence is functionally important, mutations that stop the sequence from working normally can contribute to disease or infirmity. A person with the harmful mutation has a significantly reduced chance (through early death, illness, infirmity) of transmitting his or her genes to the next generation. The mutant alleles are quickly removed from the population and the sequence is comparatively conserved. However sequences that are not functionally important are free to pick up many mutations, so that their sequence changes much more rapidly during evolution.

Question 2.36

Sequence conservation analyses often use computer-based alignment of the nucleotide sequences of equivalent genes in different organisms, or of the amino acid sequences of the corresponding proteins. The alignment below shows a BLAST alignment of the first 100 amino acids of the human CFTR (cystic fibrosis transmembrane receptor) protein (shown as the Query) and the equivalent sequence in the corresponding mouse protein (shown as Sbjct, an abbreviation of subject). The intervening middle line shows whether at the same position in the two sequences the amino acids are identical or chemically similar.

```
Query   1    MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLSEKLEREWDRE  60
             MQ+SPLEKAS +SKLFFSWT PILRKGYR  LELSDIYQ PS DSAD+LSEKLEREWDRE
```

```
Sbjct   1    MQKSPLEKASFISKLFFSWTTPILRKGYRHHLELSDIYQAPSADSADHLSEKLEREWDRE  60


Query  61    LASKKNPKLINALRRCFFWRFMFYGIFLYLGEVTKAVQPL  100
              ASKKNP+LI+ALRRCFFWRF+FYGI LYLGEVTKAVQP+
Sbjct  61    QASKKNPQLIHALRRCFFWRFLFYGILLYLGEVTKAVQPV  100
```

Calculate (a) the degree of sequence identity for the aligned sequences (b) the degree of sequence similarity.

Answer 2.36

a) Sequence identity = the number of perfect matches/the total number of positions
= 85/100 or 0.85

b) Similarity = the number of perfect matches + the number of chemically similar amino acids/the total number of positions
= 92/100 or 0.92.

Question 2.37

Following the completion of the Human Genome Project the ENCODE Project was developed as a major follow-up project. What were the aims, and what the outcome?

Answer 2.37

The ENCODE Project gets its name from the acronym Encyclopedia of DNA elements. After the Human Genome Project was completed, a major priority was to understand how our genome works. To define the functional elements in our genome and to assess how much of our genome is functionally significant, a second large international project was needed. The aim was to use various assays to identify functional DNA elements in the human genome. Operationally, a functional DNA element was defined as a discrete genome segment that makes a defined product (protein or ncRNA) or displays a reproducible biochemical signature (such as a protein-binding capacity or a specific chromatin structure).

The project included genomewide mapping of RNA transcripts in 15 human cell lines (75% of the genome was reported to be transcribed in at least one of the cell types studied). In addition, genomewide binding sites for various transcription factors and genomewide patterns of chromatin structure were defined in various human cell lines.

A principal conclusion of ENCODE was that a large part of our genome is functionally significant: 80.4% of the human genome was claimed to participate in at least one RNA-associated or chromatin-structure-associated event in at least one cell type. That conclusion has been strongly resisted by evolutionary biologists. Part of the difficulty in interpreting the ENCODE data is that much of the 80.4% figure comes from the observed representation of RNA transcripts, but many RNAs are produced at such low levels that they might alternatively represent transcriptional background ‘noise.’ the proportion of our genome that is subject to

purifying selection is now thought to be of the order of 10%; on that basis, most of our genome does not seem to have a valuable function.

Question 2.38

The endosymbiont hypothesis can explain why we have two very different genomes in our cells. What does it propose?

Answer 2.38

It proposes that our two genomes originated when a type of aerobic prokaryotic cell was endocytosed by an anaerobic eukaryotic precursor cell, at a time when oxygen started to accumulate in significant quantities in the Earth's atmosphere. Over a long period, much of the original prokaryote genome was excised, causing a large decrease in its size, and the excised DNA was transferred to the genome of the engulfing cell. The latter genome increased in size and went on to undergo further changes in both size and form during evolution, developing into our nuclear genome; the much reduced prokaryotic genome gave rise to the mitochondrial genome.

The theory explains why mitochondria have their own ribosomes and their own protein-synthesizing machinery and why our mitochondrial DNA closely resembles in form a reduced (stripped-down) bacterial genome. Our current nuclear genome is much larger than that of the ancestral eukaryotic precursor cell because of mechanisms that copied existing DNA sequences and added them to the genome. After some considerable time, the copies acquired mutations to make them different from the parent sequences and led to the formation of new genes, new exons, and so on.

Question 2.39

During evolution, as multicellular organisms became ever more complex, there has been a relentless drive to duplicate DNA sequences. As a result, our genome contains many examples of duplicated exons, duplicated genes, plus duplications of large chromosomal regions. What kinds of advantages might DNA duplication events confer that could enable ever greater functional complexity?

Answer 2.39

Duplication of genes allows more gene product to be made. Increased gene dosage is an advantage for genes that make products needed in large amounts in cells—we have hundreds of virtually identical copies of genes that make individual ribosomal RNAs and individual histone proteins, for example. Exon duplication might also be an advantage when an exon (or group of exons) encodes a structural motif that can be repeated, allowing proteins such as collagens to extend the size of structural domains during evolution.

Gene and exon duplication also allows functional divergence. After duplication, there are initially two copies with identical sequences. The constraints of purifying selection to conserve the sequence may effectively be applied to one of the two sequences; the other sequence becomes free to diverge in sequence over many millions of years to produce a different but related genetic variant. Multigene families could develop with related but different genes with slightly different or significantly different functional properties. Divergent exons allowed the formation of different but related protein domains and the possibility of alternative splicing to produce transcripts with different exon combinations. Additionally, retrotransposons allow the copying of exons from one gene to another (exon shuffling) to produce novel combinations of exons (Figure 2.15).

Question 2.40

Explain what is meant by a functional pseudogene and illustrate your answer with an example.

Answer 2.40

The term pseudogene means a defective copy of a functional gene that resides at a different locus (which can be very close to the functional gene locus or distant from it). The most well-studied pseudogenes are copies of protein-coding genes (it is easier to identify a pseudogene in this case because it can be seen to contain inactivating mutations in the sequence that corresponds to the coding DNA of the protein-coding gene). Some of these "pseudogenes" are nevertheless functional because although they cannot make a protein like the parent gene, they can make a functional RNA that regulates the expression of the parent protein-coding gene.
A good example of a functional pseudogene is the *PTENP1* pseudogene that makes a regulatory RNA to control the expression of the closely related *PTEN* gene and functions as a tumor suppressor (see Box 2.4 and Section 6.1).

Question 2.41

Exon shuffling has been thought to have occurred periodically during the evolution of me. What advantages might it have, and how might it have arisen?

Answer 2.41

The main evolutionary advantage of exon shuffling is to create hybrid genes that might result in novel gene functions. The most general mechanism of exon shuffling would involve retrotransposons copying part of a gene sequence containing an exon at the RNA level using a reverse transcriptase to produce a cDNA that integrates into a gene. For example, a LINE-1 repeat element located within the intron of a gene and with a functional capacity for transposition may sometimes produce extended RNA transcripts that extend past its own polyadenylation signal to also include an exon of the host gene. When the extended transcript is converted into

cDNA and integrates into another gene a new hybrid gene is obtained with a copy of the exon from the original gene.

Explanation 2.41
See Figure 2.15 on page 52.

Question 2.42

Describe the DNA composition of the centromeres of our chromosomes. To what extent are these DNA sequences conserved between different chromosomes, and to what extent do they resemble the sequences of centromeres in other organisms?

Answer 2.42
The DNA of our centromeres is composed of highly repetitive satellite DNA sequences that consist of tandem repeat sequences that are based on repeats units which can be of moderate length or sometimes be oligonucleotides. Thus, all human chromosomes have alphoid satellite DNA sequences that are based on a specific α repeat unit that is 171 bp long. But human centromeres also show differences with some types of satellite DNA present in the centromeres of some chromosomes, but not others. For example, satellite DNA based on the 68 bp β repeat unit is present at centromeres of the acrocentric chromosomes plus those of chromosomes 1, 9, and Y only.
Despite being vitally important for chromosome function, the DNA sequences at centromeres are not evolutionarily conserved in sequence. Instead, they are among the fastest-evolving sequences in the genome. Divergence of centromere sequences leads to reproductive isolation and the development of new species.