E1. A subtractive cDNA library contains cDNAs that are derived from mRNAs that were produced only under one set of conditions but not under another set of conditions. For example, as described in Figure 21.1, a subtractive cDNA library may only contain cDNAs that were derived from mRNAs that were made when cells were exposed to a hormone. This provides a way to identify genes that are induced by the presence of the hormone.

E2. As described in solved problem S1, one reason for making a subtractive DNA library is to determine which RNAs are produced when environmental conditions change. You want to load a small amount of cDNA on the column that was derived from cells that had been exposed to mercury. You want all of the cDNA made in both the presence and absence of mercury to bind to the column. Remember that the cDNA, which is derived from mRNA that is made in the absence of mercury, is already bound to the column. You want the cDNA that is made in the presence of mercury to bind to this cDNA, if it is complementary. If too much of this cDNA is loaded, all of the cDNAs will have complementary cDNA bound to them, and some of them will not bind to the column, even though they may be complementary to the cDNAs. In other words, if you load too much cDNA (derived from the mercury-exposed cells), you will have saturated the binding sites for cDNAs that are made in both the presence and absence of mercury. You do not want this to happen, because you want only the cDNAs that are derived from mRNAs that are specifically expressed in the presence of mercury to flow through the column. These cDNAs are not complementary to any cDNAs that are attached to the column.

E3. A. A DNA microarray is a small slide that is dotted with many different fragments of DNA. In some microarrays, DNA fragments, which were made synthetically (e.g., by PCR), are individually spotted onto the slide. The DNA fragments are typically 500 to 5,000 bp in length, and a few thousand to tens of thousands are spotted to make a single array. Alternatively, short oligonucleotides can be directly synthesized on the surface of the slide. In this process, the DNA sequence at a given spot is produced by selectively controlling the growth of the oligonucleotide using narrow beams of light. In this case, there can be hundreds of thousands of different spots on a single array.

B. In most cases, fluorescently labeled cDNA is hybridized to the microarray, though labeled genomic DNA or RNA could also be used.

C. After hybridization, the array is washed and placed in a scanning confocal fluorescence microscope that scans each pixel (the smallest element in a visual image). After correction for local background, the final fluorescence intensity for each spot is obtained by averaging across the pixels in each spot. This results in a group of fluorescent spots at defined locations in the microarray.

E4. The cDNA that is labeled with a green dye is derived from mRNA that was obtained from cells at an early time point, when glucose levels were high. The other samples of cDNA were derived from cells collected at later time points, when glucose levels were falling, and when the diauxic shift was occurring. These were labeled with a red dye. The green fluorescence provides a baseline for gene expression when glucose is high. At later time points, if the red:green ratio is high (i.e., greater than one), this means that a gene is induced as glucose levels fall, because there is more red cDNA compared to green cDNA. If the ratio is low (i.e., less than one), this means that a gene is being repressed.

E5. This is a way to analyze DNA microarray data. Using a computer, the data are analyzed to determine if certain groups of genes show the same expression patterns under a given set of conditions. This is illustrated in the data of Figure 21.3. Certain groups of genes form clusters that show very similar patterns of transcription. This is useful because it may identify genes that participate in a common cellular function.

E6. In the first dimension (i.e., in the tube gel), proteins are separated according to the isoelectric point. This is the pH at which a given protein's net charge is zero. In the second dimension (i.e., the slab gel), proteins are coated with SDS and separated according to their molecular mass.

E7. Yes, two-dimensional gel electrophoresis can be used as a purification technique. A spot from a two-dimensional gel can be cut out, and the protein can be eluted from the spot. This purified protein can be subjected to tandem mass spectroscopy to determine peptide sequences within the protein. It should be mentioned, however, that two-dimensional gel electrophoresis would not be used to purify proteins in a functional state. The exposure to SDS in the second dimension would denature proteins and probably inactivate their function.

E8. In tandem mass spectroscopy, the first spectrometer determines the mass of a peptide fragment from a protein of interest. The second spectrometer determines the masses of progressively smaller pieces that are derived from that peptide. Because the masses of each amino acid are known, the molecular masses of these smaller fragments reveal the amino acid sequence of the peptide. With peptide sequence information, it is possible to use the genetic code and produce putative DNA sequences that could encode such a peptide. These sequences, which are degenerate due to the degeneracy of the genetic code, are used as query sequences to search a genomic database. This will (hopefully) locate a match. The genomic sequence can then be analyzed to determine the entire coding sequence for the protein of interest.

E9. The two general types of protein microarrays are antibody microarrays and functional protein arrays. In an antibody microarray, many different antibody molecules are spotted onto the array. Since each antibody recognizes a different peptide sequence, this microarray can be used to monitor protein expression levels. A functional protein microarray involves the spotting of many cellular proteins onto a slide. This type of microarray can be analyzed with regard to substrate specificity, drug binding, and/or protein-protein interactions.

E10. The first strategy is a search by signal approach, which relies on known sequences such as promoters, start and stop codons, and splice sites to help predict whether or not a DNA sequence contains a structural gene. It would try to identify a region that contains a promoter sequence, then a start codon, a coding sequence, and a stop codon. A second strategy is a search by content approach. The goal is to identify sequences whose nucleotide content differs significantly from a random distribution, which is usually due to codon bias. A search by content approach attempts to locate coding regions by identifying regions where the nucleotide content displays a bias. A third method to locate structural genes is to search for long open reading frames within a DNA sequence. An open reading frame is a sequence that does not contain any stop codons.

E11. A motif is a sequence that carries out a particular function. There are promoter motifs, enhancer motifs, and amino acid motifs that play functional roles in proteins. For a long genetic sequence, a computer can scan the sequence and identify motifs with great speed and accuracy. The identification of amino acid motifs helps a researcher to understand the function of a particular protein.

E12. By searching a database, one can identify genetic sequences that are homologous to a newly determined sequence. In most cases, homologous sequences carry out identical or very similar functions. Therefore, if one identifies a homologous member of a database whose function is already understood, this provides an important clue regarding the function of the newly determined sequence.

E13. In a comparative approach, one uses the sequences of many homologous genes. This method assumes that RNAs of similar function and sequence have a similar structure. Computer programs can compare many different 16S rRNA sequences to aid in the prediction of secondary structure.

E14. The basis for secondary structure prediction is that certain amino acids tend to be found more frequently in *a* helices or *ß* sheets. This information is derived from the locations of amino acids within proteins that have already been crystallized. Predictive methods are perhaps 60 to 70% accurate, which is not very good.

E15. The three-dimensional structure of a homologous protein must already be solved before one can attempt to predict the three-dimensional structure of a protein based on its amino acid sequence.

E16. The backtranslate program works by knowing the genetic code. Each amino acid has one or more codons (i.e., three-base sequences) that are specified by the genetic code. This program would produce a sequence file that was a nucleotide base sequence. The backtranslate program would produce a degenerate base sequence because the genetic code is degenerate. For example, lysine can be specified by AAA or AAG. The program would probably store a single file that had degeneracy at particular positions. For example, if the amino acid sequence was lysine–methionine–glycine–glutamine, the program would produce the following sequence:

5′–AA(A/G)ATGGG(T/C/A/G)CA(A/G)

The bases found in parentheses are the possible bases due to the degeneracy of the genetic code.

E17. The advantages of running a computer program are speed and accuracy. Once the program has been made, and a sequence file has been entered into a computer, the program can analyze long genetic sequences with great speed and accuracy.

E18. A. To identify a specific transposable element, a program would use sequence recognition. The sequence of P elements is already known. The program would be supplied with this information and scan a sequence file looking for a match.

B. To identify a stop codon, a program would use sequence recognition. There are three stop codons that are specific three-base sequences. The program would be supplied with these three sequences and scan a sequence file to identify a perfect match.

C. To identify an inversion (of any kind), a program would use pattern recognition. In this case, the program would be looking for a pattern in which the same sequence was running in opposite directions in a comparison of the two sequence files.

D. A search by signal approach uses both sequence recognition and pattern recognition as a means to identify genes. It looks for an organization of sequence elements that would form a functional gene. A search by content approach identifies genes based on patterns, not on specific sequence elements. This approach looks for a pattern in which the nucleotide content is different from a random distribution. The third approach to identify a gene is to scan a genetic sequence for long open read frames. This approach is a combination of sequence recognition and pattern recognition. The program is looking for specific sequence elements (i.e., stop codons) but it is also looking for a pattern in which the stop codons are far apart.

E19. A sequence element is a specific sequence (i.e., a base sequence or an amino acid sequence). Two examples would be a stop codon (i.e., UAA), which is a base sequence element, and an amino acid sequence that is a site for protein glycosylation (i.e., asparagine–any amino acid–serine or threonine), which is an amino acid sequence element or motif. The computer program does not create these sequence elements. The program is given the information about sequence elements, and this information comes from genetic research. Scientists have conducted experiments to identify the sequence of bases that constitute a stop codon and the sequence of amino acids where proteins are glycosylated. Once this information is known from research, it can be incorporated into computer programs, and then the program can analyze new genetic sequences and predict the occurrence of stop codons and glycosylation sites.

E20. A.The amino acids that are most conserved (i.e, the same in all of the family members) are most likely to be important for structure and/or function. This is because a mutation that changed the amino acid might disrupt structure and function, and these kinds of mutations would be selected against. Completely conserved amino acids are found at the following positions: 101, 102, 105, 107, 108, 116, 117, 123 (Note: Asp or Glu are found here, and these two amino acids are very similar), 124, 130, 134, 139, 143, and 147.

B. The amino acids that are least conserved are probably not very important because changes in the amino acid does not seem to inhibit function. (If it did inhibit function, natural selection would eliminate such a mutation.) At one location, position 118, there are five different amino acids.

E21. A.Since most family members contain a histidine, this is likely to be the ancestral codon. The histidine codon mutated into an arginine codon after the gene duplication occurred that produced the ζ-globin gene. This would be after the emergence of primates or within the last 10 or 20 million years.

B. We do not know if the ancestral globin gene had a glycine or proline at codon-121. The mutation probably occurred after the duplication that produced the α–globin family and *β*-globin family, but before the gene duplications that gave rise to the multiple copies of α– and *β*-globins on chromosome 16 and chromosome 11, respectively. Therefore, it occurred between 300 million and 200 million years ago.

C. All of the *β*-globins contain glutamic acid at position 103, and all of the α–globins contain valine, except for *θ* -globin. We do not know if the ancestral globin gene had a valine or glutamic acid at codon 121. Nevertheless, a mutation, converting one to the other, probably occurred after the duplication that produced the α–globin family and the *β*-globin family, but before the gene duplications that gave rise to the multiple copies of α– and *β*-globins on chromosome 16 and chromosome 11, respectively. Therefore, it occurred between 300 million and 200 million years ago. The mutation that produced the alanine codon in the *θ*-globin gene probably occurred after the gene duplication that produced this gene. This would be after the emergence of mammals (i.e., sometime within the last 200 millions years).

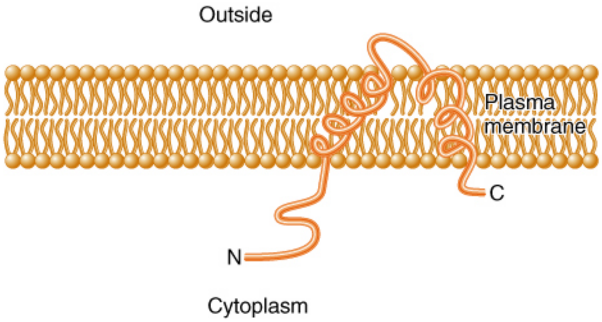
E22. As described in part C of solved problem S2, a serine codon was likely to be the ancestral codon. If we look at the codon table, an AGU or AGC codon for serine could change into an Asn, Thr, or Ile codon by a single-base change. In contrast, UCU, UCC, UCA, and UCG codons, which also code for serine, could not change into Asn or Ile codons by a single-base change. Therefore, the two likely scenarios are shown next. The mutated base is underlined. The mutations would actually occur in the DNA, although the sequences of the RNA codons are shown here.

$$\begin{array}{ccccc} & & \underline{\text{Ancestral codon}} & & \\ \text{A}\underline{\text{C}}\text{U (Thr)} & \leftarrow & \text{AGU (Ser)} & \rightarrow & \text{A}\underline{\text{A}}\text{U (Asn)} \\ & & & & \\ & & \text{A}\underline{\text{U}}\text{U (Ile)} & & \end{array}$$

$$\begin{array}{ccccc} & & \underline{\text{Ancestral codon}} & & \\ \text{A}\underline{\text{C}}\text{C (Thr)} & \leftarrow & \text{AGC (Ser)} & \rightarrow & \text{A}\underline{\text{A}}\text{C (Asn)} \\ & & & & \\ & & \text{A}\underline{\text{U}}\text{C (Ile)} & & \end{array}$$

E23. A. This sequence has two regions that are about twenty amino acids long and very hydrophobic. Therefore, it is probable that this polypeptide has two transmembrane segments.

B.



E24. RNA secondary structure is based on the ability of complementary sequences (i.e., sequences that obey the AU/GC rule) to form a double helix. The program employs a pattern recognition approach. It looks for complementary sequences based on the AU/GC rule.

E25. A and C are true. B is false because the programs are only about 60 to 70% accurate.