Supplementary Information

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0

Solutions to Exercises

Chapter 1

Solution 1.1

DNA and RNA differ in the composition of their nucleotides. While in DNA deoxyribose is found as sugar residue, in RNA this is replaced by ribose. Furthermore, in RNA, the base uracil replaces thymine. DNA is present as a complementary double strand, whereas RNA is single-stranded.

Solution 1.2

In DNA, the base pairings A-T and C-G are seen. A purine ring is paired with a corresponding pyrimidine. Two hydrogen bonds are formed in the base pairing of A-T, whereas three such bonds are formed in the pairing of C-G.

Solution 1.3

Genome describes all genomic DNA, transcriptome all mature mRNA, and proteome all proteins in an organism.

Solution 1.4

The amino acid sequence of proteins is determined by the genetic code. There are 20 naturally occurring amino acids, but only 4 bases in the DNA to encode them. Consequently, amino acids must be encoded by combinations of bases. A base doublet of 4 bases allows for the encoding of 4^2 or 16 amino acids and is, therefore, insufficient to code for 20 amino acids. However, a base triplet allows for 4^3 or 64 combinations. Consequently, several triplets encode the same amino acid, and the genetic code is, therefore, referred to as being degenerate.

Solution 1.5

The name CRICK represents the amino acids cysteine, arginine, isoleucine, cysteine, and lysine. Cysteine is encoded by the base triplets UGU or UGC; arginine by CGU, CGC, CGA, or CGG; isoleucine by AUU, AUC, or AUA; and lysine by AAA or AAG. Thus, one possible genetic code encoding an amino acid sequence for which the one-letter sequence would be CRICK is UGU CGU AUU UGU AAA

Solution 1.6

The central dogma of molecular biology was coined by Francis Crick and describes the relation between DNA, RNA, and proteins. The information of DNA is transcribed into (messenger) RNA in the process of transcription, which is subsequently converted into proteins in the process of translation. This flow of information always proceeds in this direction in nature, with the exception of some RNA viruses that replicate RNA and transcribe RNA into DNA.

Solution 1.7

Splicing refers to the removal of introns from premature messenger RNA. The process of alternative splicing refers to varying possibilities for cutting and joining introns and exons. In this way one gene can code for several proteins. This is one explanation of why there is a smaller number of genes in the human genome relative to the number of proteins.

Solution 1.8

A Venn diagram (Fig. 1.7) can be used to display the properties of amino acids. The amino acids threonine

and cysteine are indicated as hydrophobic, polar, and small. Isoleucine, leucine, and valine are hydrophobic and aliphatic.

Solution 1.9
By definition the primary structure of proteins is read from the N-terminus to the C-terminus.

Solution 1.10
Three structural building blocks are found in the secondary structure of proteins: the helix, the β-strand (building up a β-sheet), and nonrepetitive turns. In addition, loops are often mentioned, which consist of turns and connect helices or β-strands.

Chapter 2

Solution 2.1
Go to the start page at NCBI (► http://www.ncbi.nlm.nih.gov/). Select the term Protein in the pulldown menu search in the top left. Then enter the search terms in the desired combination into the text entry field next to the pull-down menu on the right. To start the database search, click on the button Search on the right, next to the text entry field. Depending on the combinations of search terms, different results will be obtained. For example, using hydrolysis AND arabinofuranoside AND bacillus AND subtilis, six database records (as of March 2018) will appear. Because a plain text search was performed, not all six entries are coming from the organism *Bacillus subtilis*, the latter two entries are from *Bacillus halodurans* and from *Paenibacillus polymyxa*. If you want to restrict the search to a specific organism, you need to limit the search term *Bacillus subtilis* to the organism database field. In this case the query is Bacillus subtilis[ORGN] AND hydrolysis AND arabinofuranoside. The result is then only the first four database entries.

Solution 2.2
To find the nucleotide sequence of the corresponding gene for IABF2_BACSU open the second entry on the result page. IABF2_BACSU stands for Intracellular exo-alpha-L-arabinofuranosidase 2. If you scroll down the entry you will find the section *FEATURES* and within that section the keyword *gene*. There you can find two gene aliases *ASD* and *XSA*.

Now enter the gene name XSA into the text entry field on the NCBI start page. Check that the term Nucleotide is selected in the pull-down menu. In addition, combine this word with the term Bacillus subtilis and restrict this search term to the organism field. It should look as follows: XSA AND Bacillus subtilis [ORGN]. AND operators can be skipped – several terms are connected automatically with AND as long as no other operator is used.

Several database entries on the bacterium will be found, including the complete genome of *B. subtilis*. Clicking the corresponding hyperlink will display the complete genome of the bacterium. The information for the corresponding gene is again found in the *features* section. It's best to use the text search function of your browser to look for the gene name XSA. Above the

gene name, next to the keywords for the subsection (*gene* and *CDS*), are found the number of the first and last bases of the nucleotide sequence. If the keyword *complement* is indicated next to the numbers of the first and last bases, then the gene is localized on the complementary DNA strand.

Solution 2.3

Entrez is the database query system at NCBI. Therefore, go to the start page at ▶ http://www.ncbi.nlm.nih.gov/Entrez. Querying the system is done as in ▶ Exercise 2.1. Enter the accession number P94552 into the text entry field and then click *Search*. Be sure that *Protein* is selected in the *Search* pull-down menu. Alternatively, on the NCBI start page follow the hyperlink *Proteins* in the light blue panel on the left-hand side and then the hyperlink *Protein Database* to the Entrez system. Enter the accession number P94552 into the text field and click *Search*. Both alternatives produce the database entry for the protein IABF2_BACSU.

Solution 2.4

Go to the start page of the EBI (▶ http://www.ebi.ac.uk), enter the AN P94552 into the text entry field *Find a gene, protein or chemical*, and press the button with the magnifying glass. Among other entries, the database record of the protein IABF2_BACSU will be found. At first sight, the entry appears different from the corresponding entry at NCBI. As mentioned earlier in ▶ Chap. 2, the standard view at the EBI server for the Uniprot database, from which this record arises, is graphical. The original database record can be seen upon following the hyperlink *Format:Text*, which is located in the blue bar directly above the database record. There one can also find hyperlinks to representations of the information in other formats.

Solution 2.5

In the graphically formatted view, the database record is divided into 16 sections. In the first section, the protein name, the corresponding gene name, the organism name, and the entry's status are listed. The status concerns the origin of the entry; it can either be reviewed, meaning the entry originates from UniProtKB/SwissProt, or unreviewed, which means the entry originates from UniProtKB/TrEMBL. The second section has to do with function and gives relevant references. It is followed by sections about taxonomy, expression, interactions, and structure, followed by information on protein families and protein structure.

The section *Cross-references* (left-hand side) lists hyperlinks to other databases that contain entries for this protein. A mouse click on one of these hyperlinks queries the relevant database and displays the database record. The *Publications* section lists relevant publications, and the section *Entry information* has information on the entry history, for example, the date of the last modification. The last two sections contain hyperlinks to useful documents and database entries for this entry.

Quite a bit of this information is available in the raw-text entry too, but not all of it, or at least not explicitly. This information is generated or retrieved from other databases on an ad hoc basis in the graphical view.

Solution 2.6
Go to the graphical view of the database record from Exercise 2.5 and follow one of the hyperlinks of reference 1. Depending on which one you clicked, the hyperlink provides a bibliography and a summary of the corresponding publication. For some references, a hyperlink to the complete paper is available in addition, for example, reference 2.

Solution 2.7
Two genes, arf1 and arf2, of an unknown species that are homologous to the α-L-arabinofuranosidase 1 or 2 of *Bacillus subtilis* are sought. To solve the problem, a short literature search will be performed. Return to the start page of NCBI and query PubMed by choosing *PubMed* in the pull-down menu on the left beside the text field. Enter the search terms into the text field. Using a combination of the terms bacillus subtilis AND arabinofuranosidase several publications are found. The solution to the question is hidden in the publication of Kim et al. (1998) (see references section of ► Chap. 2). Arf1 and arf2 are from *Cytophaga xylanolytica* and are homologous to proteins in *Bacteroides ovatus* and *Clostridium stercorarium*. You can further restrict the query, for example, with bacillus subtilis AND arabinofuranosidase AND arfI. With this query, only the Kim et al. publication is found.

Solution 2.8
One can search for a publication by an author in different ways. The simplest way is to type the last name of the author into the text entry field on the NCBI start page and then click Go. Because a full text search is performed, all publications that contain the author's name in the text will be displayed. To restrict the search to authors only, upon typing the name, specify the database field to be searched. To do this, enter the *identifier* of the appropriate database field in square brackets (without any blanks) immediately after the search term. For this example, the search string is Blobel[au], and only those publications that contain the name Blobel in the author list are found. However, there are many authors with the last name Blobel besides Günther Blobel. To retrieve only Günther Blobel's publications, Blobel G can be entered as a search term. Using this syntax, Entrez automatically recognizes the search for an author's name and restricts the search accordingly. For several author names, their first two initials must be written without spaces after their surnames (e.g., Edison TA for Thomas Alva Edison). To restrict the search to the author field, [au] can be added again. In the tutorial for the PubMed database (► http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html), additional useful information about how to restrict search results can be found.

Solution 2.9
Go to the Prosite Web page (► http://prosite.expasy.org/) and enter the sequence (Raw or FASTA format) into the text entry field in the section *Quick Scan mode of ScanProsite*, by cutting and pasting. Alternatively, the Swissprot AN P94552 or the Swissprot-ID ABF2_BACSU can be entered. Click on the *Scan* button to start the search.

Unless the box *Exclude motifs with a high probability of occurrence* is checked, 30 hits will be found with the following four motifs: N-myristoylation site, CK2 phosphorylation site, N-glycosylation site, and PKC phosphorylation site (as of July 2016). All four motifs carry the warning *pattern with a high probability of occurrence*, which means that they frequently occur in sequences and might lead to an incorrect functional annotation. Next to each motif is placed a hyperlink to the corresponding entry in the Prosite database.

Solution 2.10

Go to the start page of the PRINTS server (► http://bioinf.man.ac.uk/dbbrowser/PRINTS/index.php) and follow the hyperlink *FPScan* in the section *PRINTS search*. On the following page in the text entry field, enter the sequence of the entry ABF2_BACSU (raw format) by cutting and pasting. After clicking *Send Query*, the results page should not show any significant hits for the chosen sequence. Repeat the same search with the sequence ADA1B_HUMAN in the UniprotKB/SwissProt database. To do this, load the relevant database record from UniprotKB/SwissProt and enter the sequence in raw format by cutting and pasting. The results page should show the three highest scoring fingerprints in the first section. The following two sections list the ten best fingerprints. Each of the three highest scoring fingerprints has three links, one each to the PRINTS database, to a graphical representation of the motif's distribution along the sequence and to a 3D representation of the motif in the protein structure. The example sequence is a human adrenergic G-protein-coupled receptor, which is confirmed by the three fingerprints.

Solution 2.11

Go to the start page of the Blocks Web server (► http://blocks.fhcrc.org/), and follow the hyperlink *Blocks Searcher*. P35368 is the AN of the sequence A1AB_HUMAN from Exercise 2.10. In case the corresponding browser window has been closed, download the sequence again from Swissprot and enter the sequence by cutting and pasting into the corresponding text entry field of the *Blocks Searcher*. Also, enter your e-mail address into the appropriate field to receive the search results by e-mail. Then submit the query by clicking on *Perform Search*. After a few minutes, an e-mail in HTML format will arrive. If the e-mail program cannot display HTML, it can be saved and opened from within a browser.

The actual result of the search is found below a short description of the organization of the results page. The first section contains a summary of the search followed by a list of the possible hits. For ADA1B_HUMAN, nine possible hits should be listed. The first hit (*alpha-1B adrenergic receptor signature*) with an E Value of 3.2e-123 can be regarded as statistically significant, and all seven corresponding motifs are found. The E Value is a measure of the chance of finding a hit of the same quality within a random amino acid sequence. The value should be as small as possible in accordance with its mathematical definition (see also ► Chap. 3). The next four hits show decreasing statistical significances.

Also, not all motifs of the corresponding class are found in most sequences; this suggests, therefore, that these receptors are part of a superfamily. The remaining hits are not statistically significant and can be disregarded. The lower part of the results page contains detailed information on each of the possible hits.

Solution 2.12

Go to the start page of the Pfam Web server (► http://pfam.xfam.org/), click on *SEQUENCE SEARCH*, and enter the sequence in FATSA format by cutting and pasting to the search text field. Start the search by clicking *GO*.

After a few seconds the results of the query are shown. The most probable hit will be the Pfam protein family *7tm_1*. This designation stands for the rhodopsin family of G-protein-coupled receptors with seven transmembrane helices.

If you want to access the precalculated result, enter the AN or ID in the text field in the *Jump* section. Both result pages (calculated and precalculated) contain hyperlinks to annotations of the protein family.

Solution 2.13

Go to the start page of the Interpro Web server (► https://www.ebi.ac.uk/interpro/search/sequence-search), and enter the sequence (FASTA format) into the text field by cutting and pasting. Start the search by clicking on *Search*.

The results page displays each hit from the different member databases of Interpro in graphical format. The result is identical to the results of the previous exercises, i.e., querying Interpro can frequently replace the searches of the individual databases.

Solution 2.14

Go to the start page of the RCSB PDB database (► http://www.rcsb.org), and type in the search term Bovine Rhodopsin into the text field at the top of the page; then click *Go*. The search will return 32 hits in PDB (as of July 2016). However, because a full text search was performed, not all hits represent the 3D structure of bovine rhodopsin. The structure with the highest crystallographic resolution of 2.2 Å has the PDB ID 1 U19. Clicking on the title or the picture of the structure will display the database record. The *Structure Summary* lists the relevant reference, some information about the experimental method, the crystallographic unit cell, biological function, and cocrystallized ligands. Also, a ribbon representation of the biological unit is shown. In the menu beneath, several hyperlinks to different representation methods are given. The button *Download Files* allows for downloading the structure file in different file formats. Detailed information regarding the preceding individual points can be found on the tabs View, Annotations, Sequence Details, Structure Similarity, Experiment, and Literature. Thus, the crystallization temperature of 283 K is found in the section *Experiment*, and the graph in the section, *Sequence Details*, details a cysteine bridge (yellow line between two cysteine symbols).

Solution 2.15

Go to the start page of Entrez (► http://www.ncbi.nlm.nih.gov/entrez/), and in the *Search* menu select the database *PubChem BioAssay*. Then type HERG channel activity into the text field on the right and

click *Search*. For the development of a potential drug, knowledge of its hERG activity is paramount. Therefore, it is not surprising that quite a few hERG assays are used, currently 6047 (as of July 2016). Position 10 lists an assay that corresponds exactly to our query: HERG Channel Activity, Assay ID (AID) 376. It reports that 1960 compounds were tested so far using this assay, 252 of which were active.

Solution 2.16

Go to the start page of PubChem (► http://pubchem.ncbi.nlm.nih.gov) and select the tab *Compound*, or go to Entrez (► http://www.ncbi.nlm.nih.gov/entrez/) and select the database *PubChem Compound* in the drop-down menu at the top. Enter the search term fenbendazole in the text field and click on *Search*. Ten compound entries are found (as of March 2018). The first entry is on fenbendazole, while the remaining entries are about derivatives of fenbendazole. If you click on *Bioactivity Analysis* on the right, you can see in the overview that fenbendazole and its derivatives were tested in 1608 bioassays and that fenbendazole was found active in 93 cases.

Repeat the search with the search term Albendazole. Its chemical structure is rather similar to that of fenbendazole; the difference lies in the substitution pattern of the thioether. Albendazole and its derivatives have been tested in 1525 bioassays and were active in 117 cases.

Information on the usage of the compounds can be found in the section *Pharmacology and Biochemistry* of both database entries. Both compounds are nematocidal drugs that are used in the veterinary environment.

Solution 2.17

Go to the start page of PhenomicDB (► http://www.phenomicDB.de), and enter the search term coproporphyria into the text field at the top of the page. In the menu *Select Organisms*, choose the term *All* or restrict your search to humans with the term *Human*. Using the *Shift* and *Alt* keys in Windows, one can select several terms in this and the other selection menu, *Select data fields to show*. For the other parameters leave the standard settings untouched and click *Search*. Six genotypes (three humans, two mouse, and one rat) and seven phenotypes should result. The first phenotype bears the name *Coproporphyria* and is causally associated with a defect in the gene CPOX. Click the button *Orthologies* on the left next to the corresponding genotype. For *D. melanogaster,* ten genotype entries in FlyBase are shown, of which some entries in the field *Phenotypic class* contain the entry *lethal*. Therefore, a similar genotype-phenotype relationship also exists in the fruit fly.

Chapter 3

Solution 3.1

Open the Needle application and enter the two sequences. Enter a *Gap open penalty* and a *Gap extend penalty* of 1.0 and the desired matrix at *more option*s. Start the analysis by clicking *submit*. The results will be shown directly after calculation. The scores for the best alignment of the two sequences will be 31.0, 29.0, and 48.0 with the BLOSUM62, PAM250, and PAM30 matrices, respectively. The calculated alignments are quite different,

however. For example, with PAM30 the introduction of several gaps is suggested. This shows that the choice of a similarity matrix is important for the assessment of an alignment.

Solution 3.2

Go to the NCBI page [ncbi] and select the protein database with the term *Protein* in the pull-down menu in the top left besides the search field. Then enter the search string 5-hydroxytryptamine 2A receptor into the text field next to the pull-down menu on the right. Click *Search* next to the text field on the right. To limit the search further, you can combine the search string homo sapiens with AND. Several entries for the human serotonin receptor are found. Check the box to the left of the Swiss-Prot database record (Swiss-Prot AN P28223; ID 5HT2A_HUMAN). Then select the data format FASTA by clicking *FASTA* and save it via the pull-down menu *Send To*. Alternatively, examine the sequence in the browser and copy and paste it for Exercise 3.3.

Solution 3.3

Go to the NCBI-BLAST page [ncbi-blast]. You have to use the blastp software since the query sequence is that of a protein and the search should be executed in the non-redundant protein database of NCBI. Click on *Protein Blast* [blastp]. Then copy and paste the sequence from Exercise 3.2 into the search text field. Rather than the sequence text, only the AN (P28223) or the NCBI identifier (gi|543727) may be used. However, this is a distinctive feature of the NCBI-BLAST server and not available for all servers on the Web. Explanations about this text field and other fields or menu items can be found by following the respective hyperlink next to the entry field (e.g., Search). Start the search by clicking on *BLAST*. All additional settings can be used for a refinement of the BLAST search, but this needs some practice. Upon sending the query, a confirmation page is displayed that includes a multidigit request ID. This ID allows the later retrieval of the result. Often the BLAST analysis takes a little time, e.g., owing to a large number of concurrent queries of the server, but a self-updating status page will display until the analysis is finished.

More than 250 hits are returned from the database (as of December 2016), whereas only 100 are displayed using standard settings. The graphical overview provides a summary of the position and length of the hits with respect to the query sequence. The quality of the hits (alignment score) is color-coded.

Solution 3.4

The blastn program is found on the NCBI-BLAST Web site [blast] using the hyperlink *Nucleotide BLAST* or in the blastx program using the hyperlink *blastx*. You can also switch between both programs on each search site using their named tabs. Execute both programs with the same nucleotide sequence (AB037513). Either the sequence can be downloaded from the server, as detailed in Exercise 3.2, or its accession number can be entered into the search text field *Enter Query Sequence* (Exercise 3.3). Select *Database Others (nr etc.)* and *Reference genomic sequences* for blastn and *Nonredundant pro-*

tein sequences for blastx. Limit the analysis to the organism *Drosophila melanogaster* by entering Drosophila melanogaster under *Organism*.

The blastn search stops with the information that no significant similarity can be found. With tblastx, however, more than 100 database records are found, and some of these show a high significance. The discrepancy between the results is due to differences in how blastn and blastx execute searches and the codon usage between the two species. While blastn performs a simple comparison at the nucleotide level, blastx works at the protein level by first translating the query sequence into all six reading frames and then comparing these six theoretical proteins against a protein database. Because the genetic code is degenerate, an amino acid can be encoded by different codon triplets. The codon usage between *Drosophila melanogaster* and *Homo sapiens* is so different that no good agreement was found at the nucleotide level.

Solution 3.5

Go to the Global Align program at NCBI. Enter the two ANs into the corresponding text fields in the sections Enter Query Sequence and Enter Subject Sequence. Before starting the analysis, select the appropriate program. Click *Align*.

The result shows that in the two sequences, two regions with an identity of over 40% are present. In the human serotonin receptor, the two regions lie close together, whereas they are separated by more than 200 amino acids in the sequence of *Drosophila melanogaster*. The graphical overview displays very well the spatial arrangement of these sequence regions. However, this overview should not be considered definitive because it contains little information regarding the alignment quality.

Solution 3.6

Enter the protein sequences in FASTA format in the text field *STEP1 – Enter your input sequences* or open a text file with all sequences in FASTA format. Click *Submit*. The result page shows four different tabs. The standard view *Alignments* shows the final multiple-sequence alignment. On *Show Colors* the single amino acids are colored, which supports an easy analysis. The *Phylogenetic Tree* tab shows a tree representation, whereas the distances between the sequence represent the sequence similarity.

The multiple alignment of the three proteins shows a low number of matches, whereas two sequences show identical amino acids in wide areas and even more when conserved amino acids are considered. Identical amino acids in all three sequences occur quite rarely. In the Phylogenetic Tree view this can be clearly seen since all sequences have similar distances.

The alignment can be stored as a simple text file with the extension. clustal so that the alignment can be viewed with other software. Click on *Download Alignment File* above *Alignment*. Other visualization tools include, for example, SeaView [seaview] and BioEDIT [bioedit], which is unfortunately not updated but still creates good results. Use the *Open File* dialogue to open a saved file, change to file format *All Files (*.*)* if your file is not shown. The correct format is usually recognized when the file is opened.

You will find other useful tools on the expasy home page. An in-depth study of this page is therefore recommended.

Solution 3.7

The multiple alignment makes it obvious that the sequences are very similar. The amino acids are either identical or conservatively exchanged over wide regions. The sequence NP_640355.1 has an insertion of approx. 10 amino acids. Because of the high identity, one can assume that they are homologous sequences. Indeed, the sequences are proteases of the cathepsin family from different species:

Q28944.1 Cathepsin L precursor *Sus scrofa* (pig)

P25975.3 Cathepsin L precursor *Bos taurus* (cattle)

NP_081182.2 Cathepsin 3 precursor *Mus musculus* (mouse)

NP_640355.1 Cathepsin Q *Rattus norvegicus* (rat)

NP_001903.1 Cathepsin L preproprotein *Homo sapiens* (human)

AAH12612.1 similar to Cathepsin L *Homo sapiens* (human)

The phylogenetic tree indicates the relationship between the six sequences. A close relationship between the two human sequences, as well as between the sequences from cattle and pig, is calculated. According to this analysis, therefore, the sequences from mouse and rat are more distantly related.

Solution 3.8

Enter AC012088 into the search field of the NCBI server. Copy-paste the FASTA sequence of the eukaryotic cosmid into the input field of the Genscan server [genscan]. If the sequence has been saved in FASTA format to the hard disk, the file can be sent to the Genscan server by *File upload*. Before starting the analysis, the organism from which the sequence is derived must be selected in the pull-down menu *Organism*. Because AC012088 is a human sequence, *Vertebrate* must be chosen. Click *Run GENSCAN* afterwards. Optionally, a name for the sequence can be given; however, it will only be used for identification in the report. Depending on the settings (pull-down menu *Print options*), either just the proteins predicted to be present in the query sequence or the predicted proteins together with the corresponding nucleotide sequences will be displayed. It is also possible to display a graphic identifying the position of the predicted coding nucleotide sequences along the query sequence. For cosmid AC012088, two proteins are predicted, one of which is encoded by a single exon gene, meaning that the gene consists of a single exon without introns.

Chapter 4

Solution 4.1

Go to the dbEST home page at ► https://www.ncbi.nlm.nih.gov/dbEST/index.html and follow the hyperlink *Number of ESTs* in the lower part of the page. The dbEST contains more than 74 million ESTs; 13.5 million come from humans and mice. Therefore, about 20% of all EST sequences come from these two organisms (dbEST release 130101).

Solution 4.2

Go to the home page of dbEST and enter Magnifera indica in the search field for *Search EST*. The search results in 1714 hits. A search for Mangifera indica [ORGANISM] results in 1690 hits (dbEST release 130101). The difference between the two queries lies in the database fields that are searched. In the first case, all database fields are searched for the term *Mangifera indica*. For instance, if a database entry says gene A is similar to gene B of *Mangifera indica*, this database entry would be reported even if gene A came from a totally different organism. In the second query, only the database field *Organism* is searched. Only database entries that come from *Mangifera indica* are reported in this case.

Solution 4.3

Click on the small triangle at the top or bottom of the page beside the keyword *Send to* and select the option *File* in the drop-down menu. Then select the option *FASTA* in the *Format* field and click on *Create File*. Save the file on your computer's hard disk. The file can be viewed using any editor installed on your computer, e.g., Notepad in Windows.

If you do not want to save the FASTA sequence but just display it, you can click on the keyword *Summary* at the top or bottom of the page and then select the option *FASTA* or *FASTA (Text)* in the drop-down menu. You are then presented with a preselected number of sequences. In the standard setup, it is 20 sequences.

Solution 4.4

Go to the home page of the CAP3 sequence assembly program of the PRABI-Doua Institute (► http://doua.prabi.fr/software/cap3). Copy the first 75 EST sequences of *Magnifera indica* into the search field and start the program by clicking on the *Submit* button. Inspect the results in the files *Contigs*, *Single Sequences*, and *Assembly Details*, and save the results on your computer's hard disk. In total, the sequence assembly of the 75 *Magnifera indica* EST sequences leads to 4 contigs (Sept. 2016). Each of the contigs is built by two ESTs. In addition, many singletons are found. These have no similarities to other ESTs and therefore are not assembled into contigs.

Solution 4.5

Analyze the four contigs individually by copying them into the search field at the NCBI BLASTx page. Select the database *nonredundant protein sequences (nr)* and start the BLAST search by clicking on the *BLAST* button. Some of the contigs are very similar to already known genes or proteins, e.g., the WRKY transcription factor 58 of *Manihot esculenta*, the cassava plant. Not all contigs lead to reliable hits. The less reliable hits belong to new, as yet unknown genes. So far, nothing is known about the function of these genes.

Solution 4.6

Go to the database search system Entrez of the NCBI. In the drop-down menu at the top of the page, select *Nucleotide* and enter the AN AI590371

into the text field to the right. To display the sequence in FASTA format, click on *FASTA*. Save the sequence on your computer's hard disk by selecting the option *File* in the *Send to* drop-down menu. You can display the file using any text editor.

Solution 4.7

Go to the home page of the NCBI and run a blastn search in *Basic BLAST*. Cut and paste the EST FASTA sequence you saved earlier into the box *Enter Query Sequence*. Select the database *Nucleotide collection (nr/nt)* and click on the *BLAST* button to start the search. Forty sequences result in clear hits in the nonredundant nucleotide database, 12 of which come from *Homo sapiens*, 26 from other primates (e.g., *Pan troglodytes*, *Gorilla gorilla*, *Macaca mulatta*), and 2 from *Sus scrofa*, the domestic pig (Dec. 2016).

Solution 4.8

The NCB databases contain cross references to other databases. To go to the UniGene database, open the GenBank entry of the sequence by following the hyperlink *NM_080870.3* or the cross-reference link *GenBank*. In both cases the GenBank entry of the sequence is displayed. In the right column you will find a section, *Related Information*, that contains hyperlinks to the UniGene database and to the OMIM database. First, follow the hyperlink for the UniGene database and open the UniGene cluster *Diffuse panbronchiolitis critical region 1, HS.631993*. Even before you open that cluster you can already see in the summary that 41 sequences belong to this cluster. Once you open the cluster, you can find the information that 35 of the sequences are EST sequences, while 6 of the sequences are mRNA sequences (Dec. 2016).

For information on the relatedness to diseases you need to use another database. Go back to the GenBank entry and follow the hyperlink to the OMIM database. If you follow the first link of the resulting page, you will find information on the cloning and expression of that very gene, on its gene structure, on its mapping, and on its nomenclature. For information about related diseases follow the second link. In the section *Description* you will find the information that the gene product is involved in a rare chronic inflammation of the bronchioles. This disease occurs almost exclusively in East Asia. Only a few cases are reported outside this area, mainly in patients of East Asian descent.

Solution 4.9

Go back again to the Unigene entry (Exercise 4.8) and follow the hyperlink *EST Profile* in the section *Gene Expression*. The origin of the ESTs allows for the conclusion that the gene will be expressed in the stomach, the colon, the pancreas, and the adrenal gland. Moreover, the protein can be found in several tumors.

Solution 4.10

Go to the database query system Entrez of the NCBI. Select *Protein* in the drop-down menu at the top of the page and enter the AN P01108

into the text field at the right. Select the *Display* option *FASTA* and save the sequence in FASTA format on your computer's hard disk by selecting the option *File* in the *Send to* drop-down menu. You can display the file using any text editor.

Solution 4.11

Go to the BLAST home page of the NCBI and run a tblastn search under *Basic BLAST*. Cut and paste the saved c-myc FASTA sequence into the box *Enter Query Sequence*. Alternatively, you can run the BLAST analysis by following the hyperlink *Run BLAST* in the section *Analyze this sequence* in the database entry in Exercise 4.10. If you take this approach, it is important to change to the tab *tblastn* on the BLAST page.

Select the database *Expressed sequence tags (est)* and enter mouse (taxid: 10090) into the field *Organism*. Start the analysis by clicking on the *BLAST* button. The graphical analysis *Distribution of the top [number of hits] Blast Hits on the Query Sequence* tells us that more than 100 murine ESTs are similar to the proto-oncogene c-myc. It is striking that the majority of EST sequences show a remarkable identity either to the 5′ or 3′ end of the sequence. The reason for this distribution lies in EST production. ESTs are generated by sequencing the ends of cDNA clones.

Solution 4.12

While the vast majority of excellent hits (alignment score > 200, red bars) show a 100% alignment with the murine c-myc, ESTs with alignment scores of 80–200 (magenta bars) are only identical to 60–80%. This might be due to the fact that these ESTs code for a second, very similar, protein. To prove this hypothesis, we can blast these hits against the UniProtKB database using the BLAST algorithm blastx. The best hit we get is a protein called B-myc that is highly similar to c-myc. This is the proof for our hypothesis and means we identified a similar gene by analyzing ESTs.

Solution 4.13

The NCBI has a very large bookshelf of online textbooks. You can find this bookshelf at the NCBI home page in the section *Literature* or in the section *Popular Resources*. If you want to search all textbooks for a given term at the same time, select *Books* in the Entrez drop-down menu and enter the search text into the text box on the right. For this exercise enter the search term *Genes and disease* and follow the hyperlink to this book on the results page. If you want to limit the search, you can use quotes around the search term. Go to the section *Nutritional and Metabolic Diseases* in the table of contents and find the hyperlink to *Phenylketonuria*. The human phenylalanine hydroxylase is located on chromosome 12. On the right-hand side, in the section *Gene sequence*, you will find a hyperlink to the *Entrez Gene* database, which will give you hyperlinks to other relevant databases. Entrez Gene is always an interesting starting point for database searches.

Solution 4.14

Go to the NCBI database dbSNP and search for the reference cluster rs334 under *Search by IDs*. The SNP rs334 is an SNP in the human genome. Information on the genetic variation can be found in the section *GeneView*. The colored table lists the kinds

and results of the mutation. For SNP rs334 an adenine is exchanged for a thymine in the gene hemoglobin subunit beta, which leads to an amino acid exchange from glutamate to valine. The hyperlink *HBB* will bring you to the database Entrez Gene. Here you can find information about the related disease. Patients with this mutation suffer from a disease called sickle-cell anemia, which is prevalent in areas where malaria is endemic.

Chapter 5

Solution 5.1

Go to the home page of the PDB database [pdb]. The total number of solved structures is indicated on the upper left edge of the page beside the logo. At the time of writing (November 2016), 124,029 structures were stored in the database.

Solution 5.2

Go to the Expasy page [expasy] and follow the hyperlink UniProtKB at Popular Resources, or use the URL directly: ► http://www.uniprot.org/. Enter the AN P07801 or the ID CHER_SALTY into the text entry field at the top of the page. Click the *Search* button. The database record of the *Salmonella typhimurium* protein chemotaxis protein methyltransferase will be shown. Information about the tertiary structure of this protein can be found by following the hyperlinks to the *Structure* section of the PDB database. Change the *Link Destination* to *RCSB PDB* to go to the PDB. The PDB allows you to download the database entry for visualization in an external visualization tool (e.g., Chimera [chimera] or Swiss-PDB Viewer [spdbv]; see also Excercise 5.9) or directly visualize the protein within the browser. For the latter, follow the link *View in 3D* in the *Structure Summary* tab. The structures stored in the PDB database not only represent a single protein but often display complexes such as bound ligands, dimers, and solvent environments. It is therefore often the case that several records exist in the PDB database for a single gene, e.g., CHER.

Solution 5.3

An alternative method to open a PDB entry is to enter the unique PDB ID 1AF7 directly into the search field of the PDB start page. When you do this, you end up directly at the structure summary of PDB entry 1AF7. The structure summary presents an overview of the database record. In addition to the description of the stored structure and the original reference, information regarding the experimental method used to determine the crystal structure is presented (e.g., X-ray diffraction). Furthermore, in the *Annotations* tab, the structure summary gives some references to other databases (e.g., CATH, SCOP, PFAM).

You can follow the link *View in 3D* at *Structure Summary* to display the 3D structure or go directly to the Table *3D View*. You can set up the viewer directly below the displayed protein structure. The structure will usually be displayed in a secondary structure view as a cartoon. Here you will recognize the protein backbone and the spatial arrangement of secondary structural elements. NGL Viewer can be set up using several options on the right.

Solution 5.4

Various protein representation modes are provided by NGL Viewer that can be selected using the field *Style*: a schematic secondary structure representation (*cartoon*), only the protein backbone (*backbone*) or complete side chains (*licorice*), and the protein surface (*surface*). Moreover, the color representation can be adapted using the field *color*.

The ligand SAH (S-Adenosyl-L-homocysteine) has several hydrogen bonding interactions with the protein. After selecting the ligand in the field *Ligand Viewer*, these hydrogen bonds can be analyzed in more detail. The tab *Structure Summary* at *2D Diagram & Interactions* can be used to view the interactions in a schematic 2D diagram.

Solution 5.5

Go to the Swiss-Prot database of the Expasy server and search for the database record of the protein CHER_SALTY, as described in Exercise 5.2. Open the start page of the Expasy server, and look for Jpred or another secondary structure prediction tool at *Categories → Proteomics*. Enter the saved sequence of CHER_SALTY into the input field of the selected server. The input can be performed by copying and pasting, as described in previous exercises. Once the fields have been completed, start the analysis. Some servers will return the results via e-mail, so be sure to enter a valid e-mail address.

The predicted secondary structures agree to a greater or lesser degree with the actual secondary structure depending on the prediction program used. The actual secondary structure is available from the Swiss-Prot database record. Detailed secondary structure information can be found in the section *Structures* after selecting *more details*. The mode of operation of each server affects the quality of the prediction. A distinction is made between methods that align the query sequence with those of known secondary structures and then apply this information to the prediction and methods that perform an *ab initio* calculation. If an alignment can be done with the query sequence, then a significantly better prediction is to be expected.

Solution 5.6

The protein CHER_SALTY is a methyltransferase that is not secreted. Consequently, it is not expected to contain a signal peptide. To verify this, go to the SignalP server [signalp] and enter the sequence into the input field either by copy-paste or by file upload. Then select *Gram negative bacteria* in the section *Organism Group*. The remaining options do not need to be changed. Click *Submit*. A status page will be shown where you can enter an e-mail address at which to receive notification. The analysis takes just a few seconds before the results appear and replace the status page. If the other settings were left unchanged, both the text output and graphical display of the analysis will be displayed. It is obvious that no signal peptide exists.

Solution 5.7

Enter the sequence ABPE_SALTY (AN P41780) into the input field of the SignalP server as described in Exercise 5.6. Because ABPE_SALTY is also a *Salmonella typhimurium* protein, select *Gram negative bacteria* in the section *Organ-*

ism Group and submit the job. SignalP predicts the presence of a signal peptide. The neural network predicts a signal peptide for the first 23 amino acids and that the cleavage site will be between amino acids 23 and 24.

Solution 5.8

Go to the entry page of the Center for Biological Sequence Analysis (► http://www.cbs.dtu.dk/services/) and follow the hyperlink TMHMM at the bottom. Enter the saved amino acid sequence of the Swiss-Prot database record Q99527 via copy-paste or by file upload into the input field of the TMHMM server. In addition, choose one of several output formats. For this exercise select the format *Extensive, with graphics*. Click *Submit*. After a brief status page the results of the analysis are displayed. With the selected settings the results page contains both a text output and a graphical representation. The first few lines of the text output summarize the analysis, and these are followed by lines referring to individual segments of the protein. Each segment is described numerically by the first and last amino acids. In addition, the localization of each segment is given. The keywords *inside*, *outside*, and *transmembrane* indicate that the corresponding segment lies within the cytosol, extracellular matrix or as a transmembrane helix within the lipid bilayer, respectively. These are also displayed in the graphical overview of the results.

The TMHMM server identifies seven transmembrane helices for the protein CML2-HUMAN. Seven transmembrane helices are typical for G protein-coupled receptors. Depending on the program used for secondary structure prediction, the seven transmembrane helices also coincide with the predicted secondary structure.

Solution 5.9

The Swiss-Prot sequence can be retrieved using UniProt [uniprot]. The sequence can be downloaded at *Sequence*. Enter the sequence into the provided file of the SWISS-MODEL server or upload it. Start building the homology model using *Build Model*. A summary page with an integrated visualizer will be shown after successful model generation. Here, you can also download the final model for further analysis and visualization using software like Chimera [chimera]. Chimera is freely available and can be downloaded for Microsoft Windows, Mac OS, and Linux. Several tutorials are available at ► https://www.cgl.ucsf.edu/chimera/docindex.html as a starting point and introduction on how to use chimera.

Chapter 6

Solution 6.1

Go directly to the GEO database (► https://www.ncbi.nlm.nih.gov/geo/), or select GEO data sets at the NCBI start page (► https://www.ncbi.nlm.nih.gov/) and enter GDS1399 in the search field. In the latter case, select the data set GDS1399 with the title *DNA adenine methyltransferase and mismatch repair mutants [Escherichia coli]*.

1. Click on *Experiment design and value distribution* and then click on *click for details*. The newly opened window shows the number of wild-type and mutant replications (in each case 3).

2. After selecting *Compare 2 sets of samples*, select *Value means difference*, *2+ fold, lower* and *higher*, respectively, in the pull-down menu *Step 1*. In *Step2* the selection must be *Group A* for all wild-type and *Group B* for DAM mutants. The result is 3349 or 3129 genes that are up- or down-regulated, respectively, in the DAM mutant.
3. In Exercise 6.1-2, the mean values from the corresponding three replicates of the wild-type and DAM mutant are compared. The variation within the three replicates is not considered at this point. Therefore, this kind of calculation is statistically not supported. To obtain statistically significant results, a *t*-test is used by default for the analysis of microarray data. This well-known test asks the question whether the observed differences in the mean values of the wild-type and DAM mutant are due to the mutation or just chance. In such a case, there is no difference in the expression of the gene between the wild-type and DAM mutant.

   In this data set there are 581 genes that are significantly up- or downregulated in the DAM mutant as compared to the wild type with a significance level of 0.05. On the right side of the results page, all expression data in all replicates for each gene are displayed. Clicking on the expression profiles will show a detailed view. Check randomly some of the results.

Solution 6.2

The World session can be activated in the field *researcher login* on the start page. Click on *activate a world session* in the first section. Afterward you can find the standard basic search in the field search. Select *Publications* and *Plasmodium falciparum* as organism. The result is the desired publication. The abstract is shown through *Display Data*, the available data by a second click on *Display Data*. The picture of the microarray, for example, can be shown using *Clickable Image*. You can get information about a target gene of a colored spot by clicking on it.

Solution 6.3

The program GenePattern offers many functions for analysis and visualization that allow a comprehensive analysis of microarray experiments. GenePattern offers many individual software modules and has a clear and easy-to-use interface.

Solution 6.4

1. On the left side, click on *[description, ID or gene]* and enter HSP60 in the field *Enter search keyword*. Select *CH60_HUMAN* in the list of results and click on the 2D–PAGE figure of the entry *HEPG2_HUMAN*.

   The 2D gel of the HepG2 cells shows five spots that represent HSP60. All these spots have the same molecular weight (approx. 60 kDa) but differ in their pI values. The differences are probably due to posttranslational modifications, e.g., phosphorylation, which affect the pI value. The phosphate group changes the charge of the protein and, thus, also the pI

value. HSP60 can be phosphorylated at several sites simultaneously, which explains why there are several spots for HSP60.

2. The 2D gel of liver tissue reveals only three spots for HSP60, unlike HepG2 cells. Thus, there seems to be less posttranslational modification of HSP60 in the liver compared to HepG2 cells.
3. Click on *Maps* in the protein list. Select *HEPG2_HUMAN* and *HEPG2SP_HUMAN*, respectively, as a reference map (at *Choose a map*). Then choose *Execute Query*.

   In the 2D gel of secreted proteins from HepG2 cells, there are no spots for HSP60 (HEPG2SP_HUMAN). This indicates that the protein is not secreted.
4. Three methods were used to identify the proteins.

   {Gm} - *Gel matching*: Here, existing 2D gels are compared. If spots with the same molecular weight or pI value are found, and the proteins are already known, then it is assumed that these proteins are in fact identical.

   {Im} - *Immunodetection*: Immunodetection uses specific antibodies for unequivocal identification. A protein is unambiguously recognized if it is recognized by an antibody.

   {Mi} - *Microsequencing*: Protein spots are excised from the gel, eluted from the gel slices, trypsin-digested, and then sequenced.
5. Microsequencing identified the sequence LVKKQTYHI.
6. The protein is human S100-A4. This is an abbreviation for S100 calcium-binding protein A4. The protein has a molecular weight of 14.4 kDa and two alternative names, CAPL and MTS1, that can be accessed by selecting the UniProtKB entry *AN P26447*.

Solution 6.5

Change in the tab *Settings* the meaning of network edges to *molecular action* and minimum required interaction score to *highest confidence*. TrxC, trx-2, and MRA_3953 are the only interactions with two connections to trxB. In fact, TrxC and trx-2 are the natural substrates of thioredoxin reductase from *Mycobacterium tuberculosis*. Clicking on the connection reveals that this information was taken from curated databases. MRA_3953 is similar to TrxC, which was retrieved by analyzing homolog proteins of other species.

Solution 6.6

Enter the accession number P12931 in the search field, select the enzyme Trypsin, and choose 1000 under *Display the peptide with a mass bigger than*. After clicking *Perform*, 21 peptides with a mass > 1.000 Da are shown. They are all the results of a tryptic digest of the human protein kinase src. The largest peptide has a mass of 5072 Da.

Solution 6.7

Enter 1-Methlyxanthine in the search field. 1-methylxanthine is the main metabolite of caffeine. It was assigned to the origin *Drug Metabolites* and *Endogenous*, although caffeine is a food product. You can find a link to the KEGG database and to caffeine metabolism by following the links at *Biological Properties*. There, you can see that caffeine is degraded to 1-methylxanthine via theophylline.

Chapter 7

Solution 7.1

Go to the GOLD Genomes OnLine Database (► https://gold.jgi.doe.gov/). At the time of writing (December 2016), the first table lists 121,393 genome sequencing projects; 9092 genomes are completed and 66,684 genomes are in permanent draft status. The hyperlinks in the table fields lead to listings of the corresponding genome sequencing projects that contain further information regarding the individual projects. Using the hyperlink *Download Excel Data File* the data can be downloaded to your computer's hard disk for a quick and easy statistical analysis.

Solution 7.2

Go to the KEGG home page (► http://www.kegg.jp/) and follow the hyperlink *PATHWAY* to the PATHWAY database. TGglycolysis/gluconeogenesis metabolism is a part of carbohydrate metabolism, so the corresponding metabolic chart is found in the section *Carbohydrate Metabolism*. Select the hyperlink *Glycolysis/Gluconeogenesis* to display the metabolic map. Alternatively, the map can be found by following the hyperlink *KEGG Atlas* and then selecting *Metabolic Pathways*. Then open the *Glycolysis/Gluconeogenesis* pathway by clicking the colored area of the corresponding pathway in the graphical map.

Solution 7.3

The entry *Pyruvate* is in the lower third of the metabolic map and *L-Lactate* is to the right of it. A double arrow connects the two entries. An enzyme (EC 1.1.1.27) is written on this arrow that catalyzes the conversion of L-lactate to pyruvate. By clicking on the EC number the corresponding enzyme entry can be found. EC 1.1.1.27 is an oxidoreductase (L-lactate dehydrogenase).

To see whether this conversion takes place in humans, select the organism, in our case *Homo sapiens* (human), in the drop-down menu above the pathway map and click the *Go* button. In the reloaded pathway map, the actually used enzymes are marked in green, including enzyme EC 1.1.1.27. This conversion takes place in the human body.

For comparison with *S. cerevisiae* you can follow the same procedure. The organism list is quite long, so you can just enter the first letters of the organism name. Once the search term is unambiguous, the full name is automatically entered and you can click on the *Go* button. EC 1.1.1.27 is not marked green, i.e., *S. cerevisiae* does not have a gene coding for this protein and, therefore, does not use this metabolic pathway.

Solution 7.4

Follow the hyperlink to EC 1.1.1.27 in the metabolic map from Exercise 7.3 (glycolysis/gluconeogenesis metabolism in humans). The entries for LDHA, LDHB, LDHC, LDHAL6A, and LDHAL6B from the GENES database are shown. This means that in species-specific metabolic charts, the hyperlinks of these enzymes lead to individual records in the GENES database. In the reference map, however, the enzyme hyperlinks lead to entries in the ORTHOLOGY database.

Solution 7.5

Go to the KEGG home page (► http://www.kegg.jp/) and open the chart for human glycolysis/gluconeogenesis metabolism as described in Exercise 7.3. In a second browser window, display the species-specific metabolic pathway of *Helicobacter pylori* 26,695. A direct comparison of the two pathways indicates that enzymes EC2.7.1.11 and EC2.7.1.40 are missing within *H. pylori*. Both proteins are kinases, i.e., they are phosphate group–transferring enzymes. Information about the function of both can be found by retrieving the entries for the corresponding EC number in the ENZYME database. To do this, go back to the KEGG home page and enter the two EC numbers one by one in the text field at the top of the page. Press the *Search* button. The reaction can be found in the section *Reaction (IUBMB)*. The hyperlink in this field leads to a graphical representation of the structural formula of the reactants. Phosphofructokinase (EC 2.7.1.11) catalyzes the conversion of fructose-6-phosphate to fructose-1.6-bisphosphate via an irreversible reaction. In a subsequent irreversible reaction, pyruvate kinase (EC 2.7.1.40) catalyzes the conversion of phosphoenolpyruvate to pyruvate – the last step of glycolysis. Comparison of both metabolic maps leads to the conclusion that *H. pylori* lacks two important enzymes of glycolysis. Consequently, *H. pylori* has an incomplete glycolysis pathway. This is not difficult to understand when one considers that the natural habitat of *H. pylori* is the acidic environment of the stomach of mammals. Therefore, if pyruvate were to be produced, a further acid burden would result. Consequently, the bacterium does not utilize this metabolic step.

Solution 7.6

Go to the BLAST home page of the NCBI (► https://blast.ncbi.nlm.nih.gov/Blast.cgi) and follow the hyperlink *Microbes* beneath the text field *BLAST Genomes*. The resulting special BLAST page can be used to BLAST against microbial genomes. Because a BLASTP is desired, choose the tab *blastp*. Enter the accession number Q9ZK41 in the text field. Go to the organism selection and enter the names of the desired organisms in the text field. Additional text fields can be generated by clicking on the button *PLUS (+)*. Start the analysis by clicking the *BLAST* button at the bottom of the page.

Relevant database hits for the following organisms are found for *Helicobacter pylorii*. Obviously, the sequence with the accession number Q9ZK41 is the glucose/galactose transporter of *H. pylori* that is encoded by the gene gluP. No homologous proteins were found in the genera *Staphylococcus* and *Streptococcus*.

Solution 7.7

Go to the home page of the eggNOG database (► http://eggnog.embl.de/) and click on the *Search* button. Enter the search term cyclin-dependent kinase 1 in the text field. As soon as the entry is unambiguous, it is automatically completed and you need enter only a few letters and select the correct entry. In the next step, click on the yellow hyperlink *2 species* and select the organism *Homo*

sapiens. Enter Apicomplexans in the next text field. This entry is also checked and completed in real time. Start the query by clicking the *Explore and Download Orthologous Groups* button. Right below the orthologous group's ID *KOG0594* you can find the taxonomic level, which in this case is *Eukaryotes*.

Solution 7.8

Either go back to the home page and repeat the search as described in Exercise 7.7 or delete the target organisms on the results page of Exercise 7.7 by clicking on the small cross icon in the upper right in the field *taxa*. Then enter the search term Marsupials in the text field *Add target taxa…* Again, the search term is checked and completed in real time. Right below the orthologous group's ID *ENOG410URJI* you can find the taxonomic level, which in this case is *Mammals*. To see which PFAM domains have been found, follow the hyperlink *Functional Profile* at the end of the description and select the tab *Domains*. The PFAM domain Pkinase was found in 32 sequences, which is a frequency of occurrence of 97% (December 2016).

Follow the hyperlink *Phylogenetic Tree*. The blue markup at the branching between the Tasmanian devil (Sarcophilus harrisii) and the Brazilian opossum (*Monodelphis domestica*) denotes orthologs. A red markup further up in the tree at a branching denotes a paralog sequence in the proteome of the gray mouse lemur (*Microcebus murinus*). This part of the phylogenetic tree is drawn in light gray because the organisms were not included in either the query or the target organisms.

Solution 7.9

Go to the home page of the MBGD (► http://mbgd.genome.ad.jp/), click on the blue button marked *Taxonomy Browser*, and select the desired organisms. To do so, you first must delete the preselection by clicking on the button *Clear All* at the top of the page; then you can select the organisms. If you click on the button *Expand All* and use your browser's search functionality, it is easier to find the organisms. *Staphylococcus aureus* RF122 can be found under *Firmicutes-Bacilli-Bacillales-Staphylococcaceae-Staphylococcus-Staphylococcus aureus*, *Escherichia coli* 536 can be found under *Proteobacteria-Gammaproteobacteria-Enterobacteriales-Entereobacteriaceae-Escherichia-Escherichia coli*, and *Saccharomyces cerevisiae* S288C can be found under *Eukaryota-Ascomycota-Saccharomycetes-Saccharomycetales-Saccharomycetaceae-Saccharomyces-Saccharomyces cerevisiae*. Click the button *Choose checked taxa*. On the next page, you can either click the button *Create/View Cluster Table* or alter the homology parameters beforehand by clicking the button *Change Homology Parameters*. The calculation of the cluster can take a few minutes. While the calculation is being done, a self-refreshing HTML page is displayed. Once the calculation is finished, the cluster table is displayed.

Solution 7.10

The phylogenetic profiles of the selected organisms are shown on the page *Occurrence Pattern* of the cluster table of Exercise 7.9. The columns of the table (occurrence patterns) correspond to the organisms, the rows correspond to the individual profiles. If one organism has proteins in a

cluster, a green block is drawn at the corresponding position in the table. Thus, the phylogenetic pattern we are looking for is a green bar that spans a full row because all organisms have proteins in the cluster. To this phylogenetic profile (December 2016) correspond 476 clusters. To display an individual cluster, click on the colored bar to the right of the profile. Which cluster is displayed depends on the area of the colored bar you clicked. The colors correspond to the functional categories. To display the first cluster, click on the purple section of the bar. The purple color coding shows that this cluster contains proteins that belong to the functional category amino acid biosynthesis. The legend of the color code can be found at the home page following the hyperlink *Function Categories*.

Solution 7.11

Go to the start page of the MBGD (► http://mbgd.genome.ad.jp/). If the selected organisms are not highlighted in the *Organism selection* window, click *Reload/Refresh*. Then enter the search term fructokinase in the text entry field to the left of the organism overview and click the *Go* button. Three entries are found in the cluster table.

Glossary

@ In 1972, the engineer Ray Tomlinson wrote the first e-mail program (Bolt Beranek and Newman, Inc.). He needed a character that would separate the first part of an e-mail address from the host's designation or domain. In addition, the required character should not appear in any name. Tomlinson decided to use the @ character on the keyboard of his teletypewriter Model 33. The character had been known from writings and prints of the baroque period (seventeenth century) and represented the Latin *ad*. Today @ is read as *at* and is an essential component of every e-mail address

Accession Number A unique identification number for a database record in a sequence database. Accession numbers are static, i.e., they do not change even after database updates

Affinity chromatography A technique for the purification of proteins that makes use of the affinity of a protein for a distinct substrate or ligand (e.g., antibodies for antigens)

Algorithm Derived from al-Chwarizmi (Abu Dscha'far Muhammad ibn Musa al-Chwarizmi, Arab mathematician, A.D. 825). A logical sequence of steps for solving mathematical problems

Alignment Adjustment of two (pairwise alignment) or several (multiple alignment) sequences so that similar or identical amino acids or nucleotides are arranged vertically to produce matches

Alpha (α) helix An ordered folding pattern of the secondary structure of proteins. The α-helix displays a pitch of 0.54 nm with 3.6 amino acid residues per turn

Alternative splicing Generation of different mRNA transcripts from one pre-RNA using different splice sites

Amino acids The building blocks of proteins. Proteins are built from the 20 naturally occurring amino acids

Analogy A classification according to common features of the structure or the function considered to be essential (e.g., proteins that have similar folds or functional centers yet cannot be grouped by a common ancestral protein; e.g., head and mouthparts of arthropods, such as insects, compared to those of vertebrates). See also Homology, Character, Relationship, Phylogeny

Annotation Information on possible relationships and the derivation of possible biological functions

Antigen Compound that activates the immune system to generate antibodies. An antigen, for example, is a surface protein of a bacterium

Antibody A protein (also referred to as immunoglobulin) that binds to an antigen and consequently allows cells of the immune system to neutralize the antigen

Applet A small computer program that is downloaded from a server and executed on one's own computer. Applets are usually written in the programming language JAVA

Array See Microarray

Array express A database at the European Bioinformatics Institute where the results of microarray experiments can be stored and are accessible at any time

ASCII American Standard Code for Information Interchange. Code table for the encoding of 128 accent-free characters (a–z, A–Z, 0–9, as well as special and control characters). ASCII files are often referred to as *plain text* or *flat file*

Assembly See Sequence assembly

Base Basic building block of DNA and RNA. A sequence of bases (nucleotide sequence) forms the blueprint for a gene product

Base pair Pairing between two bases on opposite nucleotide strands of DNA or RNA. In DNA, adenine pairs with thymine, and in RNA, it pairs with uracil; cytosine always pairs with guanine

Beta (β) sheet Regular secondary structure element as building block of overall folding

pattern of proteins. β-sheets are built of different, extended parts of the polypeptide chain, the β-strands. These strands can be orientated either in the same or opposite directions, leading to parallel or antiparallel β-sheets. Successive amino acid residues are on opposite sides of the plane of the β-sheet with a repetition unit of two residues and a distance of 0.7 nm

Binary file A file that includes nonreadable text, such as, for example, executable programs, videos, and sound files

Biochip See Oligonucleotide array

Bioinformatics (applied) Application of informatic and mathematical concepts to large sets of biological data in order to accelerate and improve biological research. Applied bioinformatics is important in the fields of molecular biology, biochemistry, chemistry, and medicine

Bioinformatics (theoretical) The development of computer-based databases, algorithms, and programs to accelerate and improve biological research. Theoretical bioinformatics is important in the field of computer science

Biomarker Characteristic biological markers that can be used for personalized medicine. These are, for example, metabolites or specific gene expressions.

BLAST Basic Local Alignment Search Tool. Heuristic algorithm to search for sequences in sequence databases

BLOSUM *BLO*cks *SU*bstitution *M*atrix, substitution matrix for the alignment of protein sequences. BLOSUM matrices were introduced by Henikoff and Henikoff in 1992 and are well suited to the alignment of remotely related protein sequences. BLOSUM matrices are characterized by an affixed number that indicates the sequence identity of the sequences used to derive the matrix. Accordingly, the BLOSUM62 matrix is based on the observed substitution patterns of sequences that share 62% identity and is well suited for the alignment of sequence with a similar identity

Broad-spectrum antibiotic A substance that kills bacteria or stops bacteria from reproducing and for which the mode of action is based on a ubiquitous target

Browser Software to access the World Wide Web (e.g., Firefox, Internet Explorer, Opera, Chrome)

CAP3 Sequence assembly program based on the Smith–Waterman algorithm

CATH Structural protein database that hierarchically classifies protein domains into four groups: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H)

cDNA Complementary DNA; a DNA that is produced from mRNA as a template with the help of the viral enzyme reverse transcriptase. Like mRNA, cDNA does not have introns

cDNA array DNA microarray consisting of in vitro–amplified cDNA that is spotted onto a support material

cDNA library A cDNA library contains all cDNA transcripts of a cell, tissue, or whole organism. Unlike a genomic library, it contains only coding DNA

CDS See Coding sequence

Central dogma of molecular biology DNA is transcribed in the process of transcription to mRNA, which is then translated into proteins during translation (Francis Crick, 1957)

CERN Conceil Européen pour la Recherche Nucléaire, or Organisation Européenne pour la Recherche Nucléaire. European organization for nuclear research based in Geneva and with a research station in nearby Meyrin. The development of the World Wide Web started at CERN to organize research data in such a way as to make it available to researchers in other countries

Character A property of a protein or species (e.g., motif, structure, function, morphology, physiological process) that distinguishes it from other proteins or species. Phylogenetic research always deals with either character pairs or character strings that can be resolved into character pairs. Such character pairs can be differentiated into relatively ancestral (plesiomorph) or relatively derived (apomorph) character partners. See also Analogy, Homology, Relationship, and Phylogeny

Cheminformatics or Chemoinformatics Analogous to the term bioinformatics, cheminformatics includes all scientific disciplines that

apply concepts from mathematics and informatics to large data sets facilitating chemical research. It mainly deals with the processing of molecular structures and huge chemical databases in chemical and pharmaceutical research. In the wide sense, it also includes all computer-based methods of molecular design.

Chromatography A method for the separation of substance mixtures involving stationary and mobile phases. The term chromatography was coined by the Russian botanist Mikhail S. Tsvet (1872–1919), who used the method to isolate pigments from plant extracts

CIB Center for Information Biology. Japanese bioinformatics institute that manages the nucleotide database DDBJ

Clade A branch, monophylum, monophyletic group, or closed descent community. A systematic unit that includes a common ancestor and all descendants

Classical proteomics Classical proteomics deals with the identification and quantification of proteins in cell lysates

Client Computer program that communicates with a server. Browsers are classical clients that communicate with Web servers

Clone A population of genetically identical organisms, cells, or bacteria that have a common origin. For example, a bacterial clone in a cDNA library consists of several thousand bacteria that all possess the same cloned DNA sequence on a plasmid. Another meaning for the term clone refers to a group of recombinant DNA molecules that are descended from an initial molecule (DNA clone)

Cloning A specific DNA sequence is inserted into plasmids that serve as vectors. The DNA sequence, as part of the plasmid, is then propagated by transformation into bacteria

Cloning vector See Vector

Cluster A group that contains similar objects. Examples are expressed sequence tag sequences that are clustered owing to sequence similarities or genes that are assigned to a cluster owing to similar expression profiles

Clustering Process of grouping together objects into single clusters due to concurrences

Coding sequence Part of the DNA that is transcribed into mRNA during transcription and then translated into protein

Codon Set of three nucleotides (base triplet) of DNA or RNA that code for one of the 20 natural amino acids

Codon usage Species-specific use of the different possible codons that encode amino acids

Comparative genomics Simultaneous comparison of two or more genomes with the aim of identifying similarities and differences between those genomes

Compiling Assembly of a new complete database from a number of individual databases

Computer model A mathematical model to simulate a biological system that allows the prediction of certain properties (e.g., the concentration of metabolites at a given time) and, because of its complexity, can only be solved with the aid of a computer

Consensus sequence A single common DNA or protein sequence derived from a multiple alignment. Each position of the consensus sequence comprises that nucleotide or amino acid that occurs most often in the sequence alignment

Conserved sequence Part of a DNA or protein sequence that has remained constant during evolution

Contig Contiguous segment of a genome that was generated by joining overlapping sequences

CORBA Common Object Request Broker Architecture. Industry standard that allows the connection of different objects and programs regardless of the programming language, machine architecture, or locations of the computers

Database Collection of data organized to allow easy access to its content

dbEST Publicly accessible database at NCBI that stores expressed sequence tags (ESTs)

dbGSS Publicly accessible database at NCBI that stores Genome Survey Sequences (GSSs)

dbSNP Publicly accessible database at NCBI that stores short genetic variations such as single nucleotide polymorphisms (SNPs)

DDBJ DNA Data Bank of Japan. Together with the databases EMBL and GenBank, DDBJ forms the International Nucleotide Sequence Database

Deletion Mutation in a nucleotide sequence where single nucleotides or whole regions are missing compared to the original sequence

DNA Deoxyribonucleic acid. DNA carries genetic information. It consists of a pair of nucleotide strands that wind around a common axis to form a double helix. The pairing of the nucleotide strands occurs via hydrogen bonds between specific base pairs

DNA denaturation Conversion of double-stranded nucleotide sequences into single-stranded sequences. The hydrogen bonds between the single strands can be destroyed by strong heating, for example. The generation of single-stranded nucleotide sequences is a prerequisite to hybridization with complementary single-stranded sequences, e.g., in the assembly of a DNA microarray

DNA microarray Miniaturized technology based on the method of nucleic acid hybridization. With DNA microarrays, gene expression profiles of cells can be analyzed, for example. One differentiates between oligonucleotide and cDNA microarrays

DNA sequence Sequence of base pairs in a DNA fragment, gene, chromosome, or complete genome

DNA sequencing Method to determine the nucleotide sequence of a DNA molecule. A common method is the dideoxy chain-termination method published by Frederick Sanger in 1977

Docking Computer-assisted fitting of a ligand into the binding pocket of a protein

Domain Delimited functional unit of a protein with its own discrete folding. The complete functionality of a protein results from the combination of different domains

Dynamic methods Breakdown of a problem into subproblems and reuse of the solutions for subproblems. To solve a problem of size n, all subproblems of size 1, 2,…, $n-1$ are solved. The solutions are saved in a table from which the solution for n is derived. Dynamic methods are usually very exact; however, they can be very slow (e.g., the Smith–Waterman algorithm)

EBI European Bioinformatics Institute that is part of EMBL and is located in Hinxton near Cambridge, UK

E-cell project International research project with the aim of simulating biological phenomena on the computer and developing tools, technologies, and programs for the computational simulation of a complete cell

Edman degradation Method to determine the sequence of polypeptides

EMBL European Molecular Biology Laboratory. It was founded in 1974 and is funded by 16 European countries and Israel. Its headquarters are in Heidelberg, Germany. Other sites are in Hamburg (D), Grenoble (F), Hinxton (GB), and Monterotondo (I)

ENTREZ A general query system for all available databases at NCBI

Enzyme A protein that works as a catalyst, i.e., to reduce the activation energy of a reaction and thereby influence reaction rate. Catalysts do not change the direction of a reaction

Epitope Part of a protein bound by antibody

ESI Electrospray ionization; in mass spectrometry, a method to generate ions. Because of the gentle ionization of the analyte molecule, the method is particularly suitable for the analysis of biomolecules

EST Expressed sequence tag. Partial sequence of a cDNA clone

Eukaryotes Organisms in which cells have a nucleus and other subcellular compartments, such as mitochondria. All organisms are eukaryotic with the exception of viruses, bacteria, cyanobacteria, and archaebacteria

European Nucleotide Archive A data of nucleotide sequences, located at the European Bioinformatics Institute

Exon Coding region of a eukaryotic gene. Exons may be separated from one another by noncoding introns

ExPASY Expert Protein Analysis System. A WWW server of the Swiss Institute of Bioinformatics to analyze protein sequences. The Expasy server hosts the Swissprot database, among others

Expression profiling Determination of gene expression pattern of a cell or tissue with the aid of DNA microarrays

FASTA Heuristic algorithm to search for sequences in databases

FASTA format Simple database format to store sequence data. The FASTA format consists of a single header line that starts with the character >. It is directly followed (without a space) by an identifier and, optionally (separated by a space), a short description. Subsequent lines contain the sequence information

Fingerprint A number of sequence motifs that were derived from multiple alignments and form a characteristic signature for members of a protein family

Flat file Contains data that do not have any structural relationship to one other. Most biological databases consist of flat files

Frameshift Deletion or insertion in a DNA sequence that leads to a shift in the reading frame of all subsequent codons. In nature, frameshifts can arise by accidental mutations. In DNA sequences, frameshifts are frequently observed owing to reading errors by sequencing machines

Functional genomics Parallel analysis of genes of a given organism to identify the function of gene products. Methods used to identify gene function are, for example, DNA microarrays, serial analysis of gene expression (SAGE), and proteomics

Functional proteomics The aim of functional proteomics is to identify the functions of proteins. An important aspect of functional proteomics is the identification of protein–protein interactions

Fusion protein Product of a hybrid gene. Such hybrid genes are frequently produced experimentally so that the resulting fusion proteins can be purified or detected

Gap Gap in a sequence alignment that arises from insertions or deletions

GCG Genetics Computer Group. A number of bioinformatics programs to analyze DNA and protein sequences. GCG was founded in 1982 as a service of the University of Wisconsin and is, therefore, also known under the name Wisconsin Package. GCG became a commercial software in 1990 and is distributed worldwide by Accelrys Inc.

Gene A DNA segment that contains genetic information encoding protein. A gene comprises several units, including exons and introns and flanking regions that mainly serve in gene regulation. Genes are also described as the functional units of a genome

GenBank A database located at NCBI in which nucleotide sequences are stored

GeneChip See Oligonucleotide array

Genetic code Key for the translation of genetic information into proteins. Three bases (base triplet) encode an amino acid. Different base triplets can code for the same amino acid (degenerate code). With a few exceptions, e.g., in mitochondria or ciliates, the genetic code is universal for all living organisms

Gene expression Process in which the information encoded by a gene is translated into functional structures. Expressed genes are those that are transcribed into RNA and then translated into protein, or those that are only transcribed into RNA (without translation)

Gene family Group of related genes that result in similar protein products

Genome All the genetic information of an organism. The genome represents the sum of all genes, those parts of the DNA that influence the expression of the genetic information and those areas yet to be functionally characterized

Genomics Research field that deals with the analysis of the complete genome of an organism

Genomic library A gene bank that consists of many clones with genomic DNA. Unlike a cDNA library, a genomic library also contains noncoding DNA, such as gene introns, and DNA regions without genes

Genotype Entirety of all genetically determined characteristics of an individual

Genotyping Experimental determination of the genotype of an individual

GEO Gene Expression Omnibus. A database at NCBI that stores a variety of gene expression data and can be queried. This includes the results of DNA microarray and SAGE experiments

Global alignment Alignment over the entire length of two sequences

Glycosylation Posttranslational modification whereby sugar residues (under the release of water) are linked to proteins after translation is completed. Other organic molecules such as lipids can also become glycosylated

GSS Genome Survey Sequence. Like EST sequences, GSSs are generated by single-pass sequencing of the end regions of DNA clones. In contrast to ESTs, to generate GSSs, clones from genomic libraries are sequenced. Therefore, GSSs can also contain regions that lie outside of genes

Heuristic methods Procedures based on a sequence of approximations. Heuristic methods try to find optimal or at least nearly optimal solutions in an exponentially large space of solutions by problem-specific information. Though fast, heuristic methods may not find all possible solutions (e.g., BLAST algorithm)

HGVbase A database at the Karolinska Institute in Sweden that records information regarding variations in the human genome. HGVbase will be developed into a genotype/phenotype database in the near future

Hidden Markov Model The hidden Markov model (HMM) is named after the Russian mathematician A. A. Markov (1856–1922). It is a stochastic process (conjecturing, dependent on randomness) in which parameters that obey the system equations are not directly observable but can only be observed by derived quantities. HMMs consist of states, possible transitions between these states, and the state transition probabilities. In a specific state a result can be generated by taking into consideration all probabilities. The results, not the states, are visible to an external observer, i.e., the states are hidden. HMMs are used for the derivation of profiles from multiple protein alignments to identify new proteins, for example

HomoloGene NCBI database of homologous proteins from different species

Homology A classification based on the phylogenetic origin of structures. Characters that were inherited either unchanged or changed from common ancestors (e.g., specific kinases of mice and humans, or extremities of mice and humans) are considered homologous. See also Analogy, Character, Relationship, and Phylogeny

Homology map Tabular overview of syntenic regions from the chromosomes of two species

Homology modeling Development of a three-dimensional computer model (*in silico*) of a protein structure using as a template the structure of a similar protein that has been solved experimentally by X-ray analysis

Hybridization Pairing of two complementary and single-stranded DNA molecules to generate a double-stranded molecule through the formation of hydrogen bonds between complementary bases. For instance, hybridization is used to isolate complementary sequences in cDNA libraries

Identity Number of identical sequence positions in an alignment

Immobilization Covalent attachment of nucleic acids to solid supports. DNA can be immobilized onto nylon membranes by UV irradiation, for example

In silico In silicon. Silicon is the material computer chips consist of. It means an experiment simulated on a computer

In vitro Latin: with/in glass; outside a living organism. Denotes the location where an experiment is performed or a compound tested, e.g., a drug

In vivo Latin: with/in the living; within (the body of) a living organism. Denotes the location where an experiment is performed or a compound tested, e.g., a drug

Indexing Process describing the contents of databases with the help of descriptors, informative keywords, catchphrases, or text and, thus, allows for the efficient querying of documents within a database

Intergenetic region Noncoding subunit of a DNA sequence between genes

Insertion Incorporation of single nucleotides or whole nucleotide blocks into a DNA strand

Interactome Entirety of all interactions in a cell

Interactomics Bioinformatics discipline that deals with the study of interactomes, i.e., the interaction of all proteins and other molecules in a cell

InterPro Integrated protein motif database at the European Bioinformatics Institute that consists of several individual databases

Intron Noncoding part of a gene in eukaryotes. See also Exon

Isoelectric focusing Electrophoresis technique that separates proteins based on their individual pI values

JAVA Object-oriented, hardware-independent programming language developed by Sun Microsystems, Inc. Java programs or applets can theoretically run on any computer that supports the Java runtime environment (JRE), independently of the respective computer architecture (e.g., PC, MAC, UNIX)

J. Craig Venter Institute Institute for gene analysis. It was funded through a combination of different institutes: Center for the Advancement of Genomics (TCAG), Institute for Genomic Research (TIGR), Institute for Biological Energy Alternatives (IBEA), and J. Craig Venter Institute Joint Technology Center (JTC).

Knockdown Method for elucidating the function of genes or proteins. For example, blocking transcription of a target gene by means of RNAi may result in phenotypic changes that can be analyzed. Because translation may not need to be 100% blocked to achieve the desired effect, the term knockdown applies rather than knockout, where translation is blocked completely

Knockin Method for elucidating the function of genes or proteins. To this end, a transcribable gene is transfected into cells or organisms and the resulting phenotypic changes analyzed. Frequently a knockin is used to reverse the change in phenotype caused by a knockout. If successful, then there is little doubt as to the function of the corresponding gene

Knockout Method for elucidating the function of genes or proteins. With a knockout, the transcription of individual genes is entirely blocked. From the analysis of any resulting phenotype conclusions can be drawn as to the function of the inhibited gene. Frequently, knockout experiments are combined with knockin experiments

Local alignment Alignment of sequences that does not take into account the entire sequence length

Locus Position of a genetic marker or a gene on a chromosome

LocusLink Database at NCBI that contains curated sequence data and descriptive information about genetic loci

Low-complexity region Region of DNA or protein that consists of one or few recurring bases or amino acids

MALDI–TOF Matrix-assisted laser desorption/ionization–time of flight. Mass spectroscopic technique that is frequently used to identify proteins

Mass spectroscopy Spectroscopic technique that is used, for example, to determine the composition of peptides based on the masses of individual amino acids

Metabolite Intermediate of a biochemical metabolic reaction

Metabolome Entirety of all metabolites of an organism

Metabolomics Scientific discipline that deals with the analysis of metabolites, i.e., the metabolic products of the cell

Metagenome Entirety of all genomic information of microorganism community, e.g., of a biotope

Metagenomic Scientific discipline that deals with the analysis of metagenomes

Microarray See DNA microarray

Model organism Organism that is used for the analysis of biological questions relevant also in more complex organisms (e.g., *D. melanogaster, C. elegans, M. musculus, D. rerio, A. thaliana*, *S. cerevisiae*, *E. coli*). However, the functional units being studied must be quite similar in the two organisms

Model system See Model organism

Motif Conserved region within a group of related nucleotide or protein sequences

mRNA Messenger RNA. RNA molecules synthesized during transcription and serve as templates for protein synthesis

Multiple alignment Alignment of at least three sequences. See also Alignment

Mutation Changes in genome due to spontaneous events or triggered by mutagens such as ultraviolet light or chemicals. Leads to permanent loss or exchange of bases in DNA sequence

Narrow-spectrum antibiotic Antibiotic with a mode of action limited to a species-specific target protein found within a small group of bacteria

NCBI National Center for Biotechnology Information. The United States' contribution to the International Database Collaboration, which includes EMBL and CIB. NCBI is part of the U.S. National Library of Medicine, itself a part of the U.S. National Institutes of Health (NIH)

Needleman and Wunsch algorithm Dynamic algorithm to compute a global alignment of two sequences

Nematode Roundworm or threadworm. Example: *Caenorhabditis elegans*

Neural network Computational decision-making process to address complex problems that is analogous to the operation of the brain. A major characteristic of neural networks is their ability to adapt so that newly entered information can be recognized differentially

Next-generation sequencing Different approaches to the sequencing of whole genomes in a short time. It is based on DNA fragmentation that are extended with a known short DNA sequence and subsequently amplified. The amplified DNA strands are amplified

NMR Nuclear magnetic resonance. NMR is a spectroscopic technique to determine protein structures

Nonredundant database Complete database composed of individual databases so that each database record is present only once, even if more than one component database contains the corresponding entry

Normalization Correction of experimentally derived data to ensure accurate comparison between experiments. An example is the normalization of data that is necessary in expression profiling experiments

Northern blot Method to detect mRNA. After electrophoretic separation in an agarose gel, the RNA is transferred onto a nylon or nitrocellulose membrane. On this membrane, individual mRNA transcripts can then be detected by hybridizing with labeled and complementary nucleic acids

Nucleic Acids Research Molecular biological journal of the Oxford University Press. The first issue in January of each year is the database issue. All relevant biological databases are listed in this issue. In July 2003, a software issue was published for the first time that listed and described freely available biological software

Nucleotide Basic building block of DNA and RNA. Nucleotides consist of a base (C, A, T, G in DNA or C, A, U, G in RNA), a phosphate group, and a sugar residue (deoxyribose in DNA, ribose in RNA)

Oligonucleotide array DNA microarray that consists of several thousand single-stranded oligonucleotides. An oligonucleotide array is also called a GeneChip or BioChip

Oligonucleotides Short DNA segments that consist only of a few nucleotides. These can act as starting points for PCR or they can be used in DNA microarrays as gene markers, for example

Open reading frame A region within a DNA sequence that starts with a translation start codon (ATG) and ends with a translation stop codon (e.g., TAA)

Orthologous proteins Homologous proteins that perform the same function in different organisms. Example: A serine protease in the digestive tract of humans and mice

PAGE Polyacrylamide gel electrophoresis. Analytical method to separate proteins based on their individual charges by applying an electric field across a polyacrylamide gel matrix

Palindrome A DNA sequence that is inverse-complementary identical, i.e., where identical bases are present on complementary positions of the sense and antisense strand. For example, the complementary DNA sequence to GAATTC is CTTAAG, and the inverse-complementary to that is again GAATTC. Such palindromes are frequently recognized by restriction enzymes

PAM Matrix Point accepted mutation matrix. A substitution matrix for the alignment of protein sequences. The PAM matrix was developed in 1978 by Margaret Dayhoff and is based on a statistic analysis of sequence differences. The PAM matrix describes the number of accepted mutations between two sequences. A PAM205 matrix represents 80% accepted mutations, which means an identity of 20%

Paralogous proteins Homologous proteins in the same organism that have similar, but non-identical, functions. Example: Two serine proteases in the mouse. See Orthologous proteins

Pathway Metabolic route. Functional network between proteins

Pathway mapping Technique for the identification of multiprotein complexes. These complex proteins belong to a common pathway

PCR See Polymerase chain reaction

PDB Database containing 3D structures of biological macromolecules, such as proteins

Personalized medicine Tailoring of patient treatment to genetic predisposition and the individual metabolic profile

Pfam A protein motif database based on hidden Markov models

Phenotype Appearance of a trait in an organism that is based on both a genetic disposition and environmental influences. Examples of phenotypes are the eye color of humans or the association of certain diseases with families

Pharmacogenetics/genomics Specific field that associates genetic predisposition with the differing reactions individuals might have to drugs

Pharmaco-metabonomics Method that analyzes those factors, e.g., genetics and environment, that influence the effects of drugs

Pharmacophore The whole of steric and electronic properties that are necessary for an optimal interaction with a specific biological target structure. This leads to or blocks a biological response

Pharmacophore model Spatial arrangement of features of one or several molecules essential for that interaction with the protein. This model is normally based on the steric overlap of molecular structures of known drugs or inhibitor molecules and deduction of a pharmacophore from the analysis of congruent molecular properties

Pharmacophore screening Search for molecules in a virtual database with similar spatial feature arrangements to a calculated pharmacophore model

PhenomicDB Multiorganism genotype–phenotype database. PhenomicDB integrates data from a number of different genotype–phenotype databases, thereby allowing cross-organism data comparisons

Phenome Sum of all phenotypes of a cell, tissue, organ, organism, or species

Phenomics Scientific discipline that aims to understand the function of proteins using phenotypes

Phosphorylation Enzymatic process that involves the transfer of a phosphate group to proteins by a protein kinase

Phrap Widely used sequence assembly program

Phylogenetic analysis Analysis of phylogenetic relationship between different organisms and their ancestors. Such analyses can include morphological, physiological, and genetic characters. See also Analogy, Homology, Relationship, Character, and Phylogeny

Phylogenetic tree Graphical representation of phylogenetic relationships between organisms. Among others, phylogenetic trees can be derived from multiple-sequence alignments of DNA or protein

Phylogeny Phylogenetic evolution of living organisms and the origin of species over the course of the Earth's history. See also Analogy, Homology, Relationship, and Character

pI value pH value at which the positive and negative charges of a protein are neutralized and the net charge is zero. The pI value is also called the isoelectric point

PIR Protein Information Resource. A database for protein sequences and their functions at the Georgetown University Medical Center

Plasmid Small ringlike DNA that can replicate independently of the chromosomal DNA of the cell. Plasmids are usually between 5,000 and 40,000 base pairs in length. They contain the information for building proteins, e.g., antibiotic resistance genes. Bacteria can exchange plasmids. Because plasmids replicate quickly and are easily transferable between cells, they are used as vectors in genetic engineering to introduce and propagate genes in bacteria or yeast cells

Polymerase chain reaction PCR. Reaction in which defined DNA fragments are exponentially amplified in vitro with the help of DNA polymerases. PCR was invented by Kary Mullis in 1983, who was awarded the Nobel Prize in Chemistry in 1993

Polymorphism Genetic variation in DNA sequence of individuals within a population

Posttranslational modification Enzymatic modification of a protein upon completion of translation. Examples are the phosphorylation and glycosylation of proteins

Primary database Database that includes biological sequence data (DNA or protein) as well as accompanying annotation data

Primary structure Linear sequence of amino acids in a protein

Profiles Position-specific assessment table to describe sequence information in a complete alignment. For each position in the sequence, profiles describe the appearance of certain amino acids, conserved positions, and deletions or insertions

Prokaryotes Organisms that do not have a defined nucleus or other cellular compartments such as mitochondria. Bacteria belong to the prokaryotes

Promoter A nucleotide sequence preceding a gene that determines where and when the gene is transcribed and to what extent. The enzyme RNA polymerase recognizes the promoter and binds to it, thereby initiating gene transcription

Prosite Protein Database at the European Bioinformatics Institute. It contains information about protein families and domains, together with functional groups and characteristic signatures of proteins

Protease Enzyme that processes or degrades other proteins or peptides. The term peptidase is also used

Protein array Miniaturized technique with many thousands of proteins coupled to a solid support, allowing for their simultaneous functional analysis (e.g., for protein–protein interactions)

Protein families Most proteins can be grouped into a protein family based on sequence similarities. Proteins or protein domains that are part of a protein family have similar functions and can be traced back to a common ancestral protein

Protein kinase Enzyme that transfers phosphate groups onto proteins (phosphorylation). Phosphorylation frequently modulates the activity of target proteins

Protein lysate Protein mixture that arises after the lysis of cells

Protein profiling Experimental technique that allows the understanding of a cell's profile based on the expressed proteins

Protein turnover Time period between the synthesis of a protein and its degradation

Proteins Proteins consist of one or several amino acid chains (polypeptides). Each amino acid is connected to the next by a peptide bond, and a protein's sequence is determined by the nucleotide sequence of the corresponding gene. Proteins have various tasks in a cell (e.g., acting as enzymes, antibodies, hormones)

Proteome Entirety of all proteins of an organism

Proteomics Scientific field that deals with the proteome of an organism by structural and functional analysis of proteins

Proteogenomics Scientific field that deals with the connection between the genome and the proteome

ProtEST Part of the NCBI database UniGene. ProtEST contains the EST sequences of a UniGene cluster that show a hit upon translation into a protein sequence

PSI-BLAST Position-specific iterated BLAST. A program to find new members of a protein family within a protein database. PSI-BLAST also aids the identification of remotely related proteins

PubChem Database at NCBI that contains information of small molecules and their biological activity

Point mutation Single base change in a DNA molecule

Quality score Measure that reflects the quality of each sequenced nucleotide of a DNA sequence as determined by DNA sequencers. Using the quality score, poor-quality DNA regions can be removed from the final sequence

Quaternary structure Association of several protein subunits to form a functional protein

Ramachandran plot Diagram showing torsion angles φ and ψ in a conformation map. Enables the analysis of sterically allowed and disallowed conformation

Reading frame Within a gene, groups of three nucleotides (codons) define an amino acid or a translation start or stop signal. Therefore, during protein translation, the reading frame corresponds to a sequence of consecutive "words" with three "letters" each. If even a single nucleotide (letter) is inserted or lost within the gene, then the reading frame subsequent to the mutation will misalign, resulting in the generation of a premature stop codon and a truncated, nonfunctional protein. On the other hand, the reading frame remains unchanged by the insertion or deletion of three nucleotides, resulting in either the gain or loss of one extra amino acid

Regular expression Formalized description of a set of strings. Regular expressions allow the definition of a number of possible characters for every position in the string. The Prosite database uses regular expressions for the description of the characteristic signatures of protein families

Relationship In a genealogical sense, an abbreviation represents a phylogenetic relationship. Unfortunately, the term is used very differently (e.g., also in terms of related forms = similarity). Two species or types or protein (A and B) are regarded as more closely related compared to a third party C if they are descendants of a common ancestor not shared by C. Therefore, any ancestor that A and B share with C must be older than the common ancestor of A and B. Consequently, the degree of a phylogenetic relationship between species or proteins depends on how close common ancestors are to the present state. See also Analogy, Homology, Character, and Phylogeny

Reporter gene Gene that encodes an easily detectable product. For instance, this can be an enzyme that converts a substrate resulting in a color (change) that can be measured

Restriction enzyme Bacterial enzyme that cuts DNA molecules at specific recognition sequences

Reverse transcriptase Enzyme that catalyzes the transcription of RNA into DNA

RNA Ribonucleic acid. Molecule chemically related to DNA that is central to protein synthesis. DNA is transcribed into mRNA, which in turn is translated into proteins. Besides mRNA, a number of other RNA species exist (e.g., tRNA, rRNA)

RNAi RNA interference. Naturally occurring mechanism in eukaryotic cells that blocks the expression of single genes. See also Knockdown

RT-PCR A version of PCR that amplifies specific sequence regions in RNA. The RNA is first transcribed with the viral enzyme reverse transcriptase into cDNA, and then specific sequences defined by primers are exponentially amplified by DNA polymerases

SAGE Serial analysis of gene expression. Experimental method to analyze gene expression in cells or tissues. SAGE, like DNA microarrays, is adaptable to the high-throughput production of expression data

SBML Systems Biology Markup Language. An XML-based computer-readable format that precisely describes biological networks. Allows an easy data interchange between different programs

SCOP Structural Classification of Proteins. A database that categorizes proteins with a known structure according to structural criteria

Score matrix See Similarity matrix

SDS-PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis. See also PAGE

Secondary database Database that contains information derived from primary database. Fingerprint and motif databases such as Prosite, Blocks, and Pfam are secondary databases

Secondary structure Ordered folding pattern of polypeptide scaffold without consideration of position of amino acid side chains. Example folding patterns are the α-helix, β-sheet, and loops

Sequence Nucleotide or amino acid sequence

Sequence assembly Generation of an alignment from overlapping short sequences of DNA followed by the assembly of a consensus sequence

Sequence retrieval system SRS. Database management and query system to administer flat file databases. Among others, SRS is used on the European Bioinformatics Institute server to query biological databases

Sequencing Determination of nucleotide sequence in DNA or amino acid sequence in proteins. See also DNA sequencing

Server Computer or computer program that transfers information over a network, e.g., the Internet, to a client

SIB Swiss Institute of Bioinformatics

SignalP Computer program to estimate the N-terminal signal peptides of proteins

Signal peptide Short N-terminal amino acid sequence (often between 15 and 30 amino acids) that serves as a signal for cellular transport machinery

Similarity Evaluation of similarity of amino acid sequences. This implies the definition of similarity relationships between the 20 standard amino acids

Similarity matrix Mathematical phrasing of similarity relationships between amino acids on the basis of defined model and the analysis of related amino acid sequences

Significance Significant result that does not occur by chance. The result is, therefore, assumed to be reliable with a high probability. Significance is calculated by a number of statistical tests

Singleton EST sequences that show no overlap with other EST sequences and, therefore, cannot be grouped into contigs

siRNA Small interfering RNA. Small species of RNA (21–28 nucleotides in length) that are important in modulating transcription in eukaryotic cells

Six-frame translation Translation of a DNA fragment into the six possible reading frames. This procedure is necessary when only uncharacterized DNA fragments are available and no details on the direction of the frame exist. See also Reading frame

SMD Stanford Microarray Database. Database that allows the storage and retrieval of both raw data and normalized data from microarray experiments, and pictures of the corresponding arrays

Smith–Waterman algorithm Dynamic algorithm to determine the optimal local alignment of two sequences. The Smith–Waterman algorithm can

also be used to search databases. Though sensitive, the procedure is slow

SNP Single nucleotide polymorphism. Genetic variation caused by a change in a single nucleotide

Splice variants Proteins of different length originating from a process called alternative splicing

Spotting Placing DNA spots onto a cDNA array with the help of a robot

SRS See Sequence retrieval system

Stackpack Computer program developed to cluster EST sequences

Structural genomics Worldwide initiative to automate the experimental analysis of the three-dimensional structures of as many proteins as possible

STS Sequence tagged sites. Short, unique DNA sequences that are used to tag genomes

Substitution matrix See Similarity matrix

Swissprot Curated high-quality protein sequence database of Swiss Institute of Bioinformatics See also Expasy

Synteny Synteny refers to two or more genes lying on the same chromosome of a species

Syntenic regions Chromosomal regions are syntenic if genes of orthologous proteins are in the corresponding chromosomal regions between two species, whereby the gene order is not considered

Systems biology Scientific discipline with the aim of understanding biological organisms in their entirety. It involves the creation of an integrated picture of all regulatory processes from the genome to proteome and metabolome and on up to organelles, and the behavior of the entire organism

TAP Tandem affinity purification. Method to identify multiprotein complexes.

Target Protein that plays a central role in disease and whose activation or inhibition has a direct influence on the course of that disease

Target protein See Target

Target-based approach Modern search for drug targets that is carried out in vitro with a defined target protein

Tertiary structure Spatial organization (including conformation) of an entire protein molecule or other macromolecule consisting of a single chain

TMHMM Computer program to determine the transmembrane domains of proteins using hidden Markov models

Toxicogenomics Scientific field that analyzes the effects of toxic substances on cellular gene expression

Transformation Transfer of nucleic acids into living cells or bacteria (transfection). Also: Transformation of a normal cell into a tumor cell, for example by activation of oncogenes

Transcription Act of producing an RNA copy of DNA using the enzyme RNA polymerase

Transcription factor Protein that positively or negatively influences the transcription of genes, frequently by interacting with RNA polymerase

Transcriptome Entirety of mRNA transcripts of an organism

Transcriptomics Scientific discipline that performs global analyses of gene expression with the help of high-throughput techniques such as DNA microarrays

Translation Synthesis of proteins at ribosomes using mRNA as the template

Transmembrane domain Part of a protein that passes through a cell membrane

Turn Irregular secondary structure element as building block of overall folding pattern of proteins. Turns consist of three to six amino acids and are responsible for the globularity of proteins owing to the conformational space of the polypeptide backbone

Two-dimensional (2D) gel electrophoresis Electrophoretic technique to separate protein mixtures. Proteins are initially separated in the

first dimension according to their individual isoelectric points (pI value) and then in the second dimension according to their molecular weights

UniGene Database at NCBI that contains all nucleotide sequences of a gene and describes them nonredundantly

Uniprot Joint database of EBI, SIB, and Georgetown University that contains all the information of the Swissprot, TrEMBL, and PIR databases and serves as a central repository of protein information

UniSTS Nonredundant NCBI database containing STS markers from different sources

UTR Untranslated region. That part of RNA or cDNA that contains noncoding sequences. One distinguishes between 5′ UTR, which is upstream of the translation start codon and contains important regulatory regions such as the ribosome binding site, from the 3′ UTR, which starts with the translation stop codon and often contains a terminal poly A-sequence

Vector Usually plasmid (DNA ring) or phage (virus that attacks bacteria) to transfer genes between organisms. Vectors can be propagated in cells or bacteria as they include regulatory DNA fragments that are necessary for replication

Virtual screening *In silico*–based searches for putative bioactive molecules in virtual databases. Pharmacophor-based searches and docking are often applied computational methods

Wildcard Character used as placeholder that represents one or more arbitrary characters in file name of a command

X-ray crystallography Technique to determine the three-dimensional structure of proteins based on protein crystals

Yeast two-hybrid system In vivo method to identify protein–protein interactions in yeast cells

Index

F

G

H

I

Q

R

S

T

U

V

W

Y