

## Answers to Exercises

Answers are provided to all exercises except those requiring annotation or drawing of diagrams

**Ex. 1-1.** (a) 5000 (b)  $5 \times 10^6$  (c) complete independent storage of all genome sequences of the U.S. population would amount to approximately 70 times as much data as EOS/DIS

**Ex. 1-2.** (a) about 4 CDs. (b) 1.

**Ex. 1-3.** neurodegenerative disorder, New England colonial period, age of onset, discovery of gene for Huntington disease, polyglutamine block, huntingtin, trinucleotide repeat, anticipation, counseled

**Ex. 1-4.** (a) Leu-Ala-His-Lys-Tyr-His-STOP (b) change initial c  $\rightarrow$  t (c) change initial c  $\rightarrow$  a (d) change fourth codon aag  $\rightarrow$  tag (e) change final a  $\rightarrow$  t

**Ex. 1-5.**

PAX6_human	-----MQNSHSGVNLGGVFNGRPLPDSTRQ	27
eyeless	MFTLQPTPTAIGTVVPPWSAGTLIERLPSLEDMAHKGHSGVNLGGVFNGRPLPDSTRQ	60
PAX6_human	KIVELAHSGARPCDISRILQVSNCGVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKI	87
eyeless	KIVELAHSGARPCDISRILQVSNCGVSKILGRYYETGSIRPRAIGGSKPRVATAEIVVSKI	120
PAX6_human	AQYKRECPSIFAWEIRDRLLESEGVCTNDNIPSVSSINRVLRLNLAASEKQQ-----	136
eyeless	SQYKRECPSIFAWEIRDRLLENVCTNDNIPSVSSINRVLRLNLAQKEQQSTGSGSSSTS	180
PAX6_human	-----MG-----ADG	141
eyeless	AGNSISAKVSVSIGNVSNVAGSGRGLSSSTDLMQTATPLNSSES GGATNSGEGSEGEA	240
PAX6_human	MYDKLRMLNGQTGS-----WGTRP-----	160
eyeless	IYEKLRLNLTQHAAGPGPLEPARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSW	300
PAX6_human	-----GWYPG-----TSVP-----GQP-----	172
eyeless	PPRHYSGSWYPTSLSEIPISSAPNIASVTAYASGPSLAHSLSPNDIKSLASIGHQRNCP	360
PAX6_human	-----TQDGCQQEGG---GENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQ	219
eyeless	VATEDIHLKKELDGHQSDETGSGEGENSNGASNIGNTEDDQARLILKRKLQRNRTSFTN	420

**Ex. 1-6.**

PAX6\_human A QYKRECPSIFAW EIRDRL LSEGVC **TNDNIPSVSSINRVLRNLASEKQQ**----- 136

PAX6\_human -----**MG**-----**ADG** 141

PAX6\_human **MYDKLRMLNGQTGS**-----**WGTRP**----- 160

PAX6\_human -----**GWYPG**-----**TSVP**-----**GQP**----- 172

PAX6\_human -----**TQDGCQQQEGG**---**GENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQ** 219

**Ex. 1-7.** Choose (a)  $E < 10^{-200}$  (b)  $E \sim 0.003$

**Ex. 1-8.** at least 15 bases long

**Ex. 1-9.** (a) about 10000 human generations (b) not quite a year

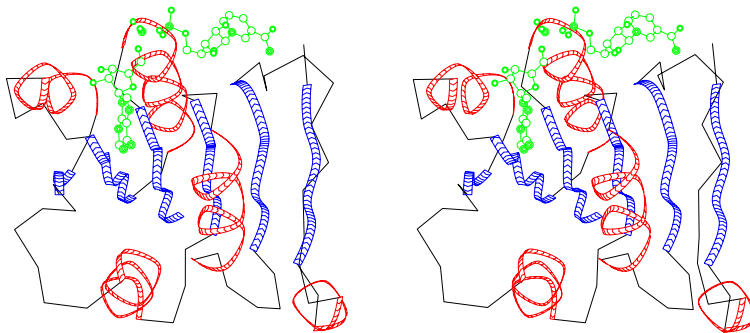
**Ex. 1-10.** (a) Ile. (b) Glu (c) Ser (d) Arg (e) Gly (f) Arg

**Ex. 1-11.** (a) down. (b) down.

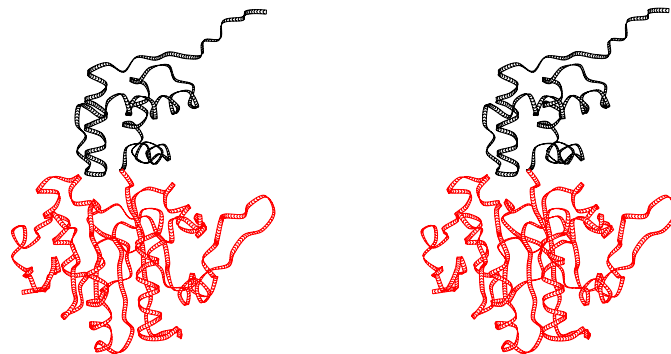
**Ex. 1-12.** twice.

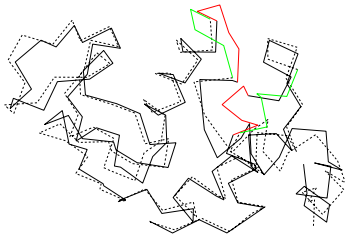
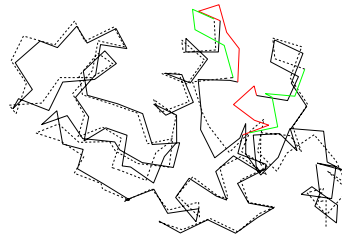
**Ex. 1-13.**

(k)



(m)



**Ex. 1-14.**lysozyme /  $\alpha$ -lactalbuminlysozyme /  $\alpha$ -lactalbumin**Ex. 1-15.** m

**Ex. 1-16.** about 1/10 (omitting the genetic code table, about )

**Ex. 1-17.** Change line 14 of the program to:

```
if (/([A-Z][a-z]+\s+[a-z]+)/ || /([A-Z]\.\s+[a-z]+)/) {
```

**Ex. 1-18.** Change line 16 of the program to:

```
$species{$1}++;
```

and change line 20 of the program to:

```
print "\"$_ $species{$_}\n\"";
```

**Ex. 1-19.** CGCAAAAAGCG or GCGTTTTTTCGC

**Ex. 2-1.** almost 300000

**Ex. 2-2.** If another enzyme provided the function of the product of the gene knocked out, there might be little or no effect on the phenotype.

**Ex. 2-3.** all of them.

**Ex. 2-4.** moon landing

**Ex. 2-5.** autosomal dominant

**Ex. 2-6.** phenotype correlates with allele of retinoblastoma gene (RB1), not with allele of esterase D gene.

**Ex. 2-7.** different expression patterns in different tissues

**Ex. 2-8.** On average  $7.5 \times 10^5$  bp, 6 genes.

**Ex. 2-9.** risk no higher than if father normal (but see Schwartz, M. & Vissing, J. (2002). Paternal inheritance of mitochondrial DNA. *New Eng. J. Med.* 347, 576-80.)

**Ex. 2-10.** search for amino acid sequence in all-six-frame translation of human genome.

**Ex. 2-11.** exceptions: trpE and trpD, and trpB and trpA, are two pairs of genes contributing to one enzyme; trpD encodes [parts of] two enzymes.

**Ex. 2-12.** inversion has occurred in the region around the centromere.

**Ex. 2-13.** the fluorescent regions would not disappear but appear elsewhere.

**Ex. 2-14.** average transfer rate:  $5.24 \times 10^{-5}$  ORFs/year,  $5 \times 10^{-5}$  kbp/year

**Ex. 2-15.** (a) both, (b) living genome, (c) both, (d) both, (e) both, (f) computer databases.

**Ex. 3-1.** (a) the librarian could allocate resources according to likelihood of use. (b) probably not.

**Ex. 3-2.** (a) convert both the speech database and the query to text and use standard text-searching algorithms. (b) characterize each successive instant of the speeches and the query by a frequency spectrum and adapt text-searching algorithms to search for patterns in the time-ordered set of frequency spectra.

(1/8000 seconds is an appropriate time interval).

**Ex. 3-3.** Cys.

**Ex. 3-4.** (a) virtual reality (computer graphics). (b) chemicals. (c) bioinformatics sequences. (d) log reports of web servers

**Ex. 3-5.**

```
<mammals>
  <genus>Homo
    <species>sapiens</species>
    <common_name>human</common_name>
    <species>neanderthalis</species>
    <common_name>neanderthal man</common_name>
  </genus>
</mammals>
```

**Ex. 3-6.** (a) (1) Time appears to pass very quickly. (2) Measure the speed of flies the same way that you would measure the speed of an arrow. (3) A certain type of flies (time flies) are fond of arrows. (b) (3) because there are no time flies, probably (2) because it doesn't make much sense.

**Ex. 3-7.** <association of> ... <protein name> ... <with—and> ... <protein name>

**Ex. 3-8.** Examples: base, briefcase, chase, erase, phrase, purchase, staircase; many geological terms such as plagioclase

**Ex. 4-1.** bicycle, tricycle: human propulsion

bicycle, motorcycle: number of wheels = 2

bicycle, car: (number of wheels = 2 AND human propulsion) OR (number of wheels = 4 AND engine propulsion)

tricycle, motorcycle: (number of wheels = 3 AND human propulsion) OR (number of wheels = 2 AND engine propulsion)

tricycle, car: number of wheels > 2 motorcycle, car: engine propulsion

**Ex. 4-2.** (a) Database housekeeping: material described as or introduced by terms LOCUS, DEFINITION, ACCESSION, VERSION, DBSOURCE. (b) Peripheral data: material described as or introduced by terms KEYWORDS, SOURCE ORGANISM, REFERENCE. (c) results of experimental measurements: material described as or introduced by terms COMMENT, ORIGIN. (d) information inferred from experimental measurements: material described as or introduced by terms FEATURES. (e) links to other databases: material described as or introduced by terms [WEB RESOURCE], within COMMENT field.

**Ex. 4-3.** The PDB reports entries that contain the term ELASTASE in descriptive material (for instance, a molecule that is described as binding to elastase), even if the entry does not actually contain a molecule of elastase.

**Ex. 4-4.**

```
#extract amino acid sequence from EMBLBANK entry

$sequence = "";                # initialize sequence empty
while (<>) {                    # read successive input lines
  if (/^DE/) {s/DE\s*/>/; print "$_";} # use DE line as title
  if (/^FT.*\/translation="/    # first sequence line?
    || length($sequence) > 0) { # subsequent sequence line?
    s/^FT\s*\/; s\/translation="\//; # kill extraneous text
    /([A-Z"]+)/;                 # extract sequence
    $sequence .= $1;            # append to current sequence
    if ($sequence =~ s\/"\/) { last; } # check for end of sequence
  }
}

$sequence =~ s/(.{1,60})/$1\n/g; #break out 60-character blocks
print "$sequence";             #print sequence, 60 char/line
```

**Ex. 4-5.** Common: material described as or introduced by terms DEFINITION, KEYWORD serine protease, SOURCE ORGANISM, REFERENCE, /translation in nucleotide entry corresponds to sequence in protein entry.

**Ex. 4-6.** Common ancestor of human and aardvark was a primitive placental mammal, probably a small shrew-like animal.

**Ex. 5-1.** 4

**Ex. 5-2.** 6

**Ex. 5-3.** agtcc → cgtcc → cgtca → cgetca

**Ex. 5-4.** off main diagonal, expect run of matches on words time and waste, appearing parallel to main diagonal. (b) use PERL program on page 248, use window = 4, thresh = 2.

**Ex. 5-5.** window = 2, threshold = 2.

**Ex. 5-6.** PAM250: H↔R more probable, BLOSUM62: W↔F more probable.

**Ex. 5.7** THE.RETORT.COURTEOUS-

THE.REPLY-.CHURLI--SH

**Ex. 5-8.** set the weights of all route segments into and out of Uppsals to very large negative values.

**Ex. 5-9.** Do a dotplot of one sequence against the reverse complement of the other.

**Ex. 5-10.** Change lines 6–10 of the program to:

```
$/ = "";                # read entire paragraph
$_ = <DATA>;           # read input, assume no blank lines
$_ =~ s/#([\n]*)\n/\n/g; # kill comments after # on each line
s/^(([\n]+)\n\s*(\d+)\s+(\d+)/; # extract job title,
$title = $1; $nwind = $2; $thresh = $3; # window, threshold

# What remains is two sequences, break at FASTA signal >

@seqs = split(>/, $_);

$seqs[1] =~ s/^(([\n]+)\n//; # split out title of first sequence
$seqt1 = $1;                # record title of first sequence

$seq1 = $seqs[1];          # record first sequence
$seq1 =~ s/\n//g;         # kill end-of-line characters
```

```

$seqs[2] =~ s/^([\n]+\n)//;    # split out title of second sequence
$seqt2 = $1;                    # record title of second sequence

$seq2 = $seqs[2];              # record second sequence
$seq2 =~ s/\n//g;             # kill end-of-line characters

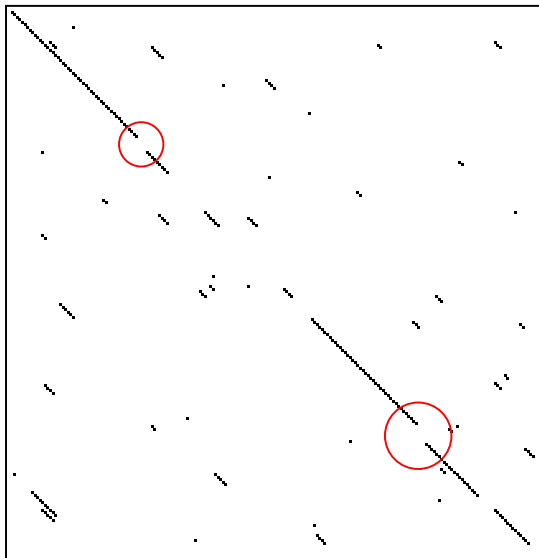
```

**Ex. 5-11.** 0.16.

**Ex. 5-12.** (a) more similar, (b) more similar, (c) just more similar, (d) less similar

**Ex. 5-13.**

PAPA\_CARPA / ACTN\_ACTCH



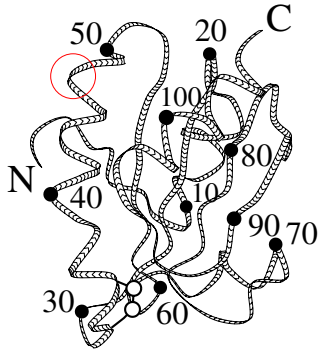
**Ex. 5-14.** Break the sequence into dipeptide or tripeptide sequences, and reassemble them in random order.

**Ex. 5-15.** 4000000

**Ex. 5-16.**  $65 = \text{Min}(40 + 25, 45 + 20, 50 + 25)$  where  $40 + 25$  corresponds to the vertical move,  $45 + 20$  to the diagonal move (mismatch) and  $50 + 25$  corresponds to the horizontal move. The two arrows appear because two possible predecessors give the same value.

**Ex. 5-17.**

(a)



(b) about residue 48

**Ex. 5-18.** (a) VDFSAT. (b) score of VDFSAT = -1533, score of VDFSAT = -1503.

**Ex. 5-19.** (a)

Scores computed by adding up six terms, the first of which is:

inventory score of V  $\times \sum_{i=1}^{20} BLOSUM62(V, \text{amino acid } i)$

Residue number	number of																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
90					8			2	2	1								3		
91	1			1				2						1		7		4		
92						16														
93	15	1																		
94			2						2	2		6				1		3		
95				4					5				3			2		1		

(b) 32. (c) 49.

**Ex. 5-20.** (a) yes. (b) no.

**Ex. 5-21.** 9: a(bc), a(bc), a(cb), ...

**Ex. 5-22.** The reduced distance matrix after combining ATCC and ATGC is shown in the text. The next step is to combine TTCG and TCGG. In the further-reduced distance matrix following that combination, the distance from {ATCC, ATGC} to {TTCG, TCGG} = 3. Therefore the distances from the root to {ATCC, ATGC} or to {TTCG, TCGG} =  $\frac{1}{2} \times 3 = 1.5$ .

**Ex. 5-23.** distance between ATCC and ATGC = 1 in both original and as sum of edges along shortest path through tree.

distance between ATCC and TTCG = 2 in original and 4.5 as sum of edges along shortest path through tree.

distance between ATCC and TCGG = 4 in original and 4.5 as sum of edges along shortest path through tree.

distance between ATGC and TTCG = 3 in original and 4.5 as sum of edges along shortest path through tree.

distance between ATGC and TCGG = 3 in original and 4.5 as sum of edges along shortest path through tree.

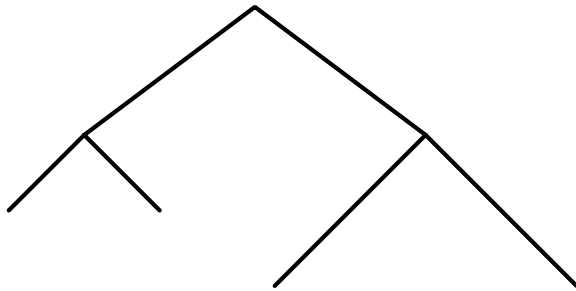
distance between TTCG and TCGG = 2 in both original and as sum of edges along shortest path through tree.

**Ex. 5-24.** three nodes, all pairs connected

**Ex. 5-25.** 2

**Ex. 5-26.** sheep

**Ex. 5-25.** The question should refer to Example 5-6.



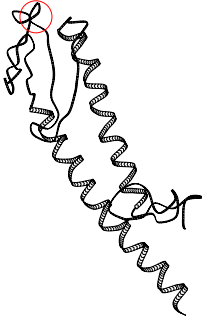
**Ex. 5-28.** (a) as  $n^2$ , (b) as  $n$ .

**Ex. 6-1.** 25.5 kJ/hydrogen bond

**Ex. 6-2.** (a) paralogues, (b) orthologues, (c) paralogues, (d) paralogues, (e) paralogue, (f) neither orthologue nor paralogue

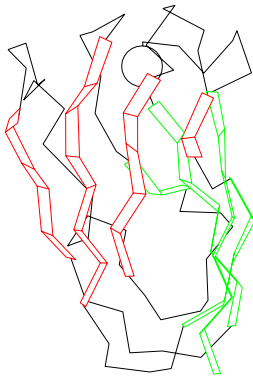
**Ex. 6-3.** point X is in the second strand from the right in Plate VIII. point Y is in the fourth strand from the right in Plate VIII. point X is in the helix appearing in front of the fourth strand from the right in plate VIII.

**Ex. 6-4.**



**Ex. 6-5. b**

**Ex. 6-6.**



**Ex. 6-7.** (a) The third through the sixth strands from the right in the template, and the helices that connect them, have the same topology as the second through the fifth strands from the right in the target. The rightmost strand in the target corresponds to a helix and the second from rightmost strand in the template (which points in the opposite direction from the rightmost strand in the target). The rightmost strand in the template does not correspond to a strand in the target. An extra  $\beta - \alpha - \beta$  unit appear in the target, at the left of the sheet. (b) The six strands in Murzin's prediction have the correct topology as the target; the leftmost (N-terminal) strand in the target is missing from the prediction. (c) The connectivities of the rightmost strand are different; the prediction is like the target structure and the parent is not. The direction of the sixth strand from the right is different between the template and the target.

**Ex. 6-9.** You would search for homologous proteins that had different (and ideally separable) modes of determining specificity. Even better you would look for proteins carrying out functions essential to the pathogen that had NO human homologue.

**Ex. 6-10.** 24

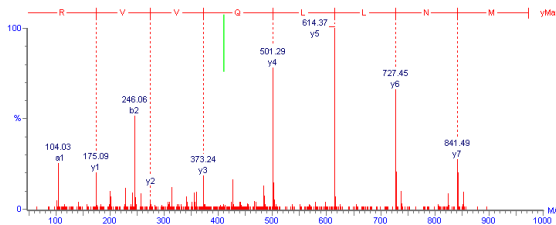
**Ex. 6-11.** It selects points above and to the right of the line  $x + 2y = 2$ . (This line intersects the  $x$ -axis at  $x = 2$  and intersects the  $y$ -axis at  $y = 1$ .)

**Ex. 6-12.** sketch looks like top left diagram on p. 245. Geometric interpretation: selects points below and to the right of the line  $y = x$ .

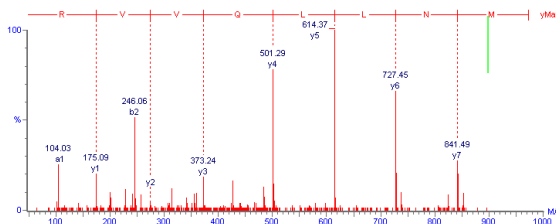
**Ex. 6-13.** a.

**Ex. 7-1** (a) yes (b) yes (c) no

**Ex. 7-2** (a) The  $y_3$  peak would be moved to the position marked in green:



(b) The spectrum of the peptide would look the same. The total molecular weight of the undissociated peptide would appear at the position marked in green in the following figure; but this molecular weight would be measured in a separate stage of the experiment, such as the selection stage of a tandem MS/MS procedure.



(c) The  $y$  peaks would be shifted left by the difference between the mass of the C-terminal R and a C-terminal G. In addition the difference between the  $y_5$  and  $y_6$  peaks would be changed from the mass of L to the mass of N, and the difference between the  $y_6$  peak and the  $y_7$  peak would be shifted by the same amount.

**Ex. 7-3** (a) RVVWLL (b)  $y_5$

**Ex. 7-4** (a) 22 (b)  $A \leftrightarrow C$  (c) yes (d) 23 (e)  $A \leftrightarrow C$  (f) yes

**Ex. 7-5** about 400 distinct sequences if recombination occurred with equal probability everywhere. 0.75%

**Ex. 7-6** (a) phage display. (b) X-ray and NMR structure determinations, Two-hybrid screening systems,

**Ex. 7-7** (a)  $V_1, V_4$  or  $V_3, V_4$ . (b)  $V_1, V_3$ . (c)  $V_5, V_6$ . (d)  $V_1, V_4$  (e)  $V_3, V_4$

**Ex. 7-8** (a) phylogenetic trees, parts of metabolic pathways, citation patterns, the World Wide Web. (b) metabolic pathways, chemical bonding patterns (if single / double / triple bonds are distinguished).

(c,d)

Graph	Nodes	Edges
Sets of people who have met each other	people	relationships
Electricity distribution systems	power stations	cables
Phylogenetic trees	species (or taxa)	lines of descent
Metabolic pathways	metabolites	enzyme-catalyzed reactions
Chemical bonding patterns in molecules	atoms	bonds
Citation patterns in the scientific literature	articles	references
The World Wide Web	sites	references

**Ex. 7-9** (a) 10. (b) 17. (c) 178.5 kJ

**Ex. 7-10** (a) Moorgate → Bank → Waterloo → Embankment

(b) King's Cross → Russell Square → Holborn → Tottenham Court Road → Oxford Circus → Euston Square → King's Cross (c) 2/15

**Ex. 7-11** (a) Amersham – Upminster

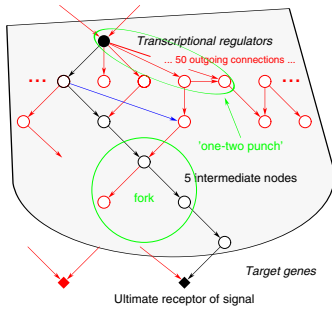
(b) Stations between Upney and Upminster, stations between West Brompton and Wimbledon, Chiswick Park (allowing access by National Rail on Richmond spur).

(c) Stations between Stanmore and Baker Street (except for Wembley Park, West Hampstead – accessible by National Rail – and Finchley Road, Southwark, Bermondsey

**Ex. 7-12** 1/21

**Ex. 7-13** (a) all downstream nodes in fork and scatter motifs. (b) the intermediate node in the 'one-two punch' motif. These answers apply *only* to the motifs in isolation; within the context of a full network there may well be additional connections.

**Ex. 7-14** answers to parts (a) and (b) in green; answer to part (c) in blue



**Ex. 7-15** (a) closed. (b)  $\beta$ -sheet

**Ex. 7-16** 0.201