

# Answers to Exercises

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

- 1.1 (a) 5000 (b)  $5 \times 10^6$  (c) Taking the population of the U.S.A. as  $3 \times 10^8$ , the storage requirements for all DNA sequences of all inhabitants of the U.S.A exceeds the storage requirements for EOS/DIS by a factor of 60.
- 1.2 (a) 20833 (b) 4.6 (c) 1
- 1.3 neurodegenerative disease, polyglutamine tract, huntingtin, anticipation, counselling
- 1.4 (a) Leu-Ala-His-Lys-Tyr  
(b) ... cta ...  
(c) ... ccg ... (Leu → Pro ... )  
(d) ... aag → tag ...  
(e) ... taa → tac
- 1.7 Choose (a)  $E < 10^{-200}$  (b)  $E \sim 0.003$
- 1.8 at least 15 bases long
- 1.9 (a) about 100000 human generations. (b) 100000 bacterial generations = 3.8 years
- 1.10 (a) isoleucine (b) glutamic acid (c) serine (d) arginine (e) lysine (f) phenylalanine
- 1.11 (a) down (b) down
- 1.12 two times
- 1.14 m
- 1.16 75%
- 1.17 change `{[A-Z][a-z]+ [a-z]+}` to `{[A-Z]([a-z]+|\.) [a-z]+}`
- 1.18 change `$species{$1} = 1;` to `$species{$1}++;`
- 1.19 CGCAAAAAGCG or GCGTTTTTCGC
- 2.1 almost 300000
- 2.2 If another enzyme took over the function of the product of the gene knocked out, there might be no effect on phenotype
- 2.3 all of them

2.4 moon landing

2.5 autosomal dominant

2.6 phenotype correlates with allele of retinoblastoma gene (RB1), not with allele of esterase D gene.

2.7 different expression patterns in different tissues

2.8  $7.5 \times 10^5$  bp, 6 genes

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.)

2.10 search human genome sequence

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.

2.12 inversion has occurred in the region around the centromere

2.13 The fluorescent regions would not disappear but appear elsewhere

2.14 average transfer rate:  $5.24 \times 10^{-5}$  ORFs/year,  $5 \times 10^{-5}$  kbp per year

2.15 (a) both, (b) living genome, (c) both, (d) both, (e) both, (f) computer databases.

3.1 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

3.3 `while(<>)`

```
    if (/^\s*DEFINITION/) \{s/^\s*DEFINITION\s*//;print \"$_\";\}
    elsif (/^\s*\d+\s*(\[a-z +\])\s*$/) \{$seq = $1; $seq =~ s/ //g;
    print \"$1\n\";\}
}
```

4.1 4

4.2 6

4.3 agtcc → cgtcc → cgtca → cgtca

4.4 (a) Off main diagonal, expect run of matches on word time and waste, parallel to main diagonal. (b) use PERL program on p. 162, use window = 4, thresh = 2.

4.5 window 2 threshold 2

4.6 PAM250: H↔R more probable, BLOSUM62: W↔F more probable.

4.7 THE .RETORT .COURTEOUS-

THE .REPLY- .CHURLI--SH

4.8 Set weights of all routes into and out of Uppsala to very large negative values.

4.9 Do a dotplot of one sequence against the reverse complement of the other.

4.10 replace line

```
$_ = ~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(.*)\n([A-Za-z\n]*)
\s*\s*\n(.*)\n([A-Za-z\n]*)\s*/;
```

with

```
$_ = ~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(^\\s*\\>.*)\n
n([A-Za-z\n]*)\n(^\\s*\\>.*)\n([A-Za-z\n]*)\s*/;
```

4.11 0.16

4.12 (a) more similar (b) more similar (c) just more similar (d) less similar

4.15 4000000

4.16  $65 = \text{Min}(40 + 25, 45 + 20, 50 + 25)$  where  $40 + 25$  corresponds to the vertical move,  $45 + 20$  to the diagonal move (mismatch) and  $50 + 25$  to the horizontal move. The two arrows appear because two possible predecessors give the same value.

4.18 30, 60, 70, 84

4.19 (a) VDFSAT (b) score of VDFSAT = -1533 score of VDFSAT = -1503

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

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

4.20 (a)

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

4.21 (a) yes, (b) no

4.22 9: a(bc), a(bc), a(cb), ...

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

4.24 distance between ATCC and ATGC = 1 in both original and as sum of edges along shortest path through tree.

distance between ATCC and TTCC = 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.

**4.25** three nodes, all pairs connected

**4.26** sheep

**4.28** (a) as  $n^2$  (b) as  $n$ .

**5.1** 25.5 kJ/hydrogen bond

**5.2** (a) orthologue (b) paralogue (c) paralogue (d) paralogue (e) paralogue (f) neither orthologue nor paralogue

**5.5** b

**5.6** (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 appears 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 connectivity 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.

**5.8** 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.

**5.9** 24

**5.10** 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  $y = 1$ .)

**5.11** Sketch looks like top left diagram on p. 245. Geometric interpretation: selects points below and to the right of the line  $y = x$ .

**6.1** (a) yes (b) yes (c) no

**6.3** (a) RVVWLL (b) y5

**6.4** (a) 22 (b)  $A \leftrightarrow C$  (c) yes (d) 23 (e)  $A \leftrightarrow C$  (f) yes

**6.5** about 400 distinct sequences if recombination occurred with equal probability everywhere. 0.75%

6.6 (a) phage display (b) surface plasmon resonance

6.7 (a)  $V_1-V_3$

(b)  $V_1-V_3$

(c)  $V_5-V_6$

(d)  $V_1-V_4$

(e)  $V_4-V_3$

6.8 (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

6.9 (a) 10 (b) 17 (c) 178.5 kJ

6.10 1/21

6.11 (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

6.12 (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).

6.15 (a) closed (b)  $\beta$ -sheet

6.16 0.201

# Glossary

**B-factor** a measure of the precision to which an atom position is determined in X-ray crystallography

**binomial** identifier of a taxonomic unit in terms of the genus and species

**bootstrapping** a protocol for validating a method in which data sets are sampled randomly from a corpus of data, allowing multiple copies of the same data set

**character string** a linear array of text, each item chosen from a fixed alphabet

**connected graph** a graph in which there is at least one path between every two nodes

**contig map** a chart of the order of appearance in a full DNA molecule of small sequenced fragments

**controlled vocabulary** fixed and carefully-defined set of terms, to which feature tables in database annotations are to be restricted

**convergent evolution** appearance of similar features not shared by a common parent, arising independently rather than by descent from a common ancestor

**differential genomics and proteomics** comparisons of sequences or expression patterns between individuals; for instance, healthy and diseased humans or animals, or drug-susceptible and resistant infectious microorganisms

**docking** prediction of protein (or nucleic acid)-ligand interactions

**domain** a compact structural subunit of a protein

**drug target** a protein, the function of which would be useful to modify to alleviate a symptom of disease

**ecosystem** a coherent localized grouping of living things interacting with one another and with their environment

**edge length** a number assigned to each edge of a graph signifying in some sense the distance between the nodes connected by the edge

**Electrospray Ionization (ESI)** a non-destructive method for vaporizing and ionizing proteins in which the sample is sprayed through a small capillary with an electric field at its tip

**eukaryote** an organism whose cells contain a nucleus

**exon** an expressed region of a split eukaryotic gene

**expression chip** a microarray containing oligonucleotides derived from cDNAs

**fluorescent *in situ* hybridization** identification of the position of a sequence on a chromosome by tagging a complementary oligonucleotide with a fluorescent marker

**fold recognition** attempt to decide which (if any) of a given library of folding patterns is adopted by a polypeptide sequence

**genetic drift** change in allelic frequencies within a population by mechanisms other than natural selection

**gene duplication** a type of mutation in which two or more copies of a gene appear corresponding to a single copy in the parent

**gene flow** redistribution of allele frequencies within or between populations

**genomic hybridization** microarray technique to look for gains or losses of genes or of copy number

**genotype** the hereditary information contained in the nucleotide sequence of the genetic material in the cells of an organism

**graph** an abstract structure containing nodes (points) and edges (lines connecting points)

**Hamming distance** a measure of dissimilarity between two character strings of the same length, equal to the number of mismatches in characters at corresponding positions

**homologous** descended from a common ancestor

**homology modelling** prediction of the three-dimensional structure adopted by a polypeptide chain on the basis of the known structure adopted by a similar sequence

**hydrophobicity profile** graph of hydrophobicity of successive residues in a protein, giving clues to positions of turns, or buried/exposed regions

**hydrophobic effect** unfavourable interaction of non-polar solutes with water accounting for low solubility

**intron** an intervening, non-expressed, region of a split eukaryotic gene

**jackknifing** a protocol for validating a method in which data sets are sampled randomly from a corpus of data

**Levenstein distance** a measure of dissimilarity between two character strings, equal to the minimum number of edit operations (substitution, insertion, deletion) required to convert one sequence into the other

**LINE** long interspersed nuclear element

**macroevolution** large-scale changes in genomes within populations

**MATRIX-ASSISTED Laser Desorption Ionization (MALDI)** a non-destructive method for vaporizing proteins in which the sample is embedded in surroundings (the matrix) that initially absorbs the energy from a laser pulse

**microevolution** small-scale gene changes within a population, some not even affecting phenotype

**modular proteins** proteins composed of a linear concatenation of domains

**multiple sequence alignment** assignment of mutual residue correspondences in three or more related sequences

**RFLP (restriction fragment length polymorphism)** distribution of sizes of fragments produced by cleaving DNA with a restriction enzyme

**secondary structure** interresidue hydrogen-bonded standard structures in a protein or nucleic acid

**secondary structure prediction** an attempt to compute from the amino acid sequence which regions of a polypeptide chain form  $\alpha$ -helices and strands of  $\beta$ -sheet, without knowing the three-dimensional structure; or an analogous prediction of base-pairing regions of single-stranded RNA

**sequence tagged site** short sequenced region of DNA, typically 200–600 base pairs long, appearing in a unique and mapped position in a genome

**similarity** resemblance not necessarily derived from homology

**SINE** short interspersed nuclear element

**SNP (single nucleotide polymorphism)** one or more possible bases appearing at a given position of a genome

**somatic cell hybrids** rodent cells containing one or a few human chromosomes or fragments of human chromosomes

**STRP (short tandem repeat polymorphism) or microsatellite** short repeated unit of only 2–5 base pairs but appearing as many consecutive copies

**substitution matrix** a table specifying the relative expected probability of replacement of any amino acid (or nucleotide) by another

**supersecondary structure** combinations of secondary structure elements that are consecutive in the sequence and interacting in space

**tertiary structure** the spatial arrangement of the polymer chain of a protein or nucleic acid

**tree** a connected graph with exactly one path between every two points

**turns** regions in which a polypeptide (or polynucleotide) chain changes direction in space

**twilight zone** region from 18%–25% sequence identity, in which it is difficult to decide on the basis of sequence alone whether two proteins are related or not

**VNTR (variable number tandem repeat) or minisatellite** short repeated units about 10–100 base pairs long

**window** the size of each successive consecutive region in a sequence considered at one time

**Z-score** a measure of the unexpectedness of a result in a distribution of possible results;  $Z\text{-score} = (\text{score} - \text{mean of distribution}) / \text{standard deviation of distribution}$



# Index

## A

accessible surface area 323  
algorithm 16  
alignment 26, *plate III*  
alignment *plate IV*  
alleles 5  
analgesic drugs 277, 281  
analysis of algorithms 16  
anthropology 103  
apoptosis 320  
archaea 22  
artificial chromosome 77  
aspirin 283

## B

B-factor 133  
bacteria 22  
bacterial artificial chromosome 77  
banding patterns of chromosomes 72  
Berardinelli-Seip syndrome 79  
biological classification 21  
BioMagResBank 131  
BLOCKS database 173  
BLOSUM matrices 173  
bootstrapping 209  
BRCA1 152, 299, *plate VIII*

## C

Cambridge Crystallographic Data Centre 131  
CATH (Class/Architecture/Topology/Homologous superfamily) 135  
cattle domestication 104  
chaos 319  
chemical cross-linking 325  
chemoinformatics 274  
chloroplast DNA 88  
chromatin immunoprecipitation 325  
chromosome banding pattern 89, 114  
cladistic methods 206  
clinical trials 270  
CLUSTAL-W 62, 149  
clustering 199, 205  
    hierarchical 199  
coexpression pattern 326  
coimmunoprecipitation 325  
collision-induced dissociation (CID) 304

comparative genomics 110  
conformational energy calculations 255  
contiguous clone map (contig) 78  
controlled vocabulary 119  
convergent evolution 21  
core of homologous proteins 237  
Critical Assessment of Fully-Automatic Structure Prediction (CAFASP) 242  
Critical Assessment of Structure Prediction (CASP) 52, 242, 287  
cystic fibrosis 75

## D

DALI (Distance-matrix ALIGNment) 135, 235  
data structures 16  
databanks 9  
database annotation 13  
database indexing 118  
database searching 11  
differential genomics 57, 273  
divergence 320  
divergence of function 266  
divergence of sequence 237  
divergence of structure 237  
DNA methylation 70  
DNA sequence databases 122  
DNA structure 6, 8, *plate I*  
docking 276  
domain 42, 111, *plate VI*  
domain recombination network 326  
dotplot 160  
    and alignment 165, 210  
drug discovery 269  
dynamic programming 176

## E

E-value 183  
edge (of a graph) 204  
edge length 204  
elastase *plate IV*  
electron microscopy 338  
electronic publication 15  
Electrospray ionization (ESI) 303  
Ensembl 151  
ENTREZ 124, 141  
Enzyme Commission 51  
ENZYME DB 124

equilibrium 319  
 eukarya 22  
 EVA 249  
 'Eve', mitochondrial 60  
 evolution 4  
 Expert Protein Analysis System (ExPASy) 150  
 expressed sequence tag (EST) 136, 301  
 expression profile 136, 293, 309, *plates VII, VIII, IX*  
 eyeless gene, in *Drosophila* 34

## F

FASTA format 25  
 feature table 123  
 FISH (Fluorescent *in situ* hybridization) 77,  
   113, *plate II*  
 flavodoxin *plate V*  
 fluorescence resonance energy transfer 326  
 fold recognition 51, 242, 252  
 fork 335

## G

gap weighting 174, 181  
 GenBank 1  
 gene duplication 5, 91, 111  
 gene expression table 295  
 gene flow 5  
 gene maps 73  
 gene therapy 57  
 genetic code 7  
 genetic drift 5  
 genome 68  
   databases 124  
   evolution 105  
   sizes 67, 68  
 genotype 4  
 global alignment 26, 175  
 graph 204, 315

## H

haemoglobin gene cluster 91  
 Hamming distance 171, 210  
 haplotype 307  
 helical wheel 231  
 helix 42, 129, 224  
 heteroplasmy 104  
 heterozygote 5  
 Hidden Markov Model (HMM) 188, 193  
 HIV protease database 135  
 homology 21, 29, 197  
 homology modelling 51, 242, 250  
 homoplasmy 104  
 homozygote 5  
 horizontal gene transfer 108, 113  
 HOX genes 218  
 Huntington disease 56  
 hydrophobic effect 226  
 hydrophobicity profile 229

## I

information retrieval 16  
 interaction domain 336  
 interaction network 212, *plate X*  
 International Human Genome Sequencing  
   Consortium 97  
 International Immunogenetics database  
   (IGMT) 136  
 isoniazid 309, *plate IX*

## J

jackknifing 209

## K

knockout 268  
 Kyoto Encyclopedia of Genes  
   and Genomes (KEGG) 138

## L

lead compound 271, 274  
 leukaemia *plate II*  
 Levenshtein distance 171, 210  
 LINE (long interspersed nuclear element) 30  
 linkage maps 72  
 LINUS 259  
 local alignment 26, 175  
   Smith-Waterman algorithm 175  
 London Underground 314

## M

Macromolecular Structure Database (MSD) 50  
 mass spectrometry 54, 71, 301  
   nucleic acid sequencing 306  
   protein sequencing 304  
 Matrix-assisted laser desorption ionization-Time of  
   flight (MALDI-TOF), mass spectrometry 303  
 maximum likelihood 207  
 maximum parsimony 206  
 MEDLINE 139  
 metabolic pathway database 138  
 metazoa, phylogenetic tree of 23  
 Michaelis constant 322  
 microarray 53, 293  
 minimal genome 107  
 mitochondrial DNA 88  
 mitochondrial 'Eve' 60  
 MODBASE 241  
 modular protein 44, 111  
 molecular dynamics 255  
 molecular graphics 141, 222  
 Molecular Modelling Database (MMDB) 50  
 molecular modelling, in drug discovery 275  
 Monte Carlo algorithm 261  
 multiple sequence alignment 186  
 mutation 5

**N**

native state of proteins 8  
 natural selection 5  
 networks 313  
   small-world 317  
 neural networks 246  
 nomenclature 21  
 nonsynonymous substitutions (Ka) 105  
 Nuclear Magnetic Resonance (NMR) 134, 324  
 Nucleic Acid Database (NDB) 131  
 nucleic acid sequence databases 122

**O**

OCA 133  
 'one-two punch' 335  
 Online Mendelian Inheritance in Man  
   (OMIM[TM]) 142, 147  
 open reading frame (ORF) 80  
 organelle 88  
 orthologues 266

**P**

P-score 183  
 PAM250 matrix 173  
 paralogues 266  
 path in a graph 204  
 path length 204  
 PAX-6 genes 33  
 percent accepted mutation (PAM) 172  
 periodic variation 319  
 PERL 16  
 personal identification 104  
 phage display 325  
 pharmacogenomics 56  
 pharmacophore 235, 275  
 phenotype 4  
 Phobius 196  
 phylogenetic distribution pattern 326  
 phylogenetic relationship 24, 31  
 phylogenetic tree 203  
 phylogeny 198  
 polyacrylamide gel electrophoresis (PAGE) 71  
 polyglutamine repeats 56  
 population 5  
 position-specific scoring matrix 191  
 positional cloning 76  
 post-translational modification 72  
 prenatal diagnosis 308  
 primary structure of protein 41  
 primosome 327  
 principal component analysis (PCA) 298  
 profiles 188  
   3D 252  
 PROSITE 124  
 protein complexes 320  
 Protein Data Bank (PDB) 1, 50, 128  
   worldwide (wwPDB) 128

protein denaturation 225  
 protein engineering 51  
 protein folding 8, 221, 228  
 protein folding pattern 41  
 protein function prediction 265  
 Protein Identification Resource (PIR) 124, 149  
 protein interaction network 324  
 protein sequence databases 124  
 protein stability 225  
 protein structure 40  
   classification 44, 135, 238  
   evolution 236  
   prediction 51, 240  
 protein-protein interface 323  
 proteome 52, 54, 68, 71, 136  
 pseudogene 90, 100  
 PSI-BLAST 32, 33, 185, 188, 191  
 PubMed 139  
 pyruvate decarboxylase, *plate VI*

**Q**

quantitative structure-activity  
   relationships (QSAR) 272  
 quaternary structure of protein 41  
 quorum sensing 313

**R**

radiation hybrids 77  
 Ramachandran plot (see Sasisekharan-  
   Ramakrishnan-Ramachandran plot) 223  
 recombination 5  
 reduction of dimensionality 297  
 regulatory networks 327  
 resolution, of X-ray structure determination 134  
 retinoblastoma 112  
 reverse genetics 74, 76  
 root, of a tree 201  
 root-mean-square (r.m.s.) deviation 233  
 Rosetta 259  
 runaway 320

**S**

Sasisekharan-Ramakrishnan-  
   Ramachandran plot 223  
 scatter 335  
 SCOP (Structural Classification of Proteins)  
   50, 135, 239, 289, *plate V*  
 scoring schemes, in sequence alignment 171  
 secondary structure 41, 129  
   prediction 51, 242, 244, 287  
 selectivity and sensitivity contrasted 32  
 sequence alignment 158, 26  
   multiple 159  
   pairwise 159  
   scoring schemes in 171  
 sequence databanks 1, 9  
 sequence logo *plate III*

Sequence Retrieval System (SRS) 148  
sequence searching 32  
sequence tagged site (STS) 80  
sheet 42, 129  
short tandem repeat polymorphism (STRP) 78  
sidechains 225  
similarity 29  
simulated annealing 262  
SINE (short interspersed nuclear element) 30  
single-nucleotide polymorphism (SNP) 101  
singular value decomposition 298  
small-world networks 317  
software engineering 16  
somatic cell hybrids 77  
steady state 319  
strand 224  
structural alignment 233  
structural genomics 241  
structural superposition 233  
structure databanks 1  
Structure Explorer 132  
substitution matrix 172  
superposition 233  
supersecondary structure of protein 42  
surface plasmon resonance 325  
SWISS-MODEL 251  
SWISS-PROT 124  
synonymous substitutions (Ks) 105  
systems biology 9, 54, 311

## T

target selection 273  
TATA box 81  
taxonomy 21  
tertiary structure of protein 41  
thioredoxin *plate III*  
threading (method of fold recognition) 254  
time scale of Earth history 199  
transition 172  
transmembrane helix prediction 193, 263

transversion 172  
tree 204  
tree of life 22  
trp operon 82  
tuberculosis 308  
tumour suppressor gene *plate II*  
twilight zone 185  
two-hybrid screening 324  
typhoid Mary 317

## U

UniProt 124  
unweighted pair group method with  
arithmetic mean (UPGMA) 206

## V

variable number tandem repeat (VNTR) 78

## W

wheat 88  
world-wide-web 14  
worldwide Protein Data Bank (wwPDB) 128

## X

X-ray structure determination 324

## Y

yeast artificial chromosome 77

## Z

Z-score 183