

**Fig. 1.1** Comparison of the rate of growth of the GenBank sequence (data from Table 1.1) with the rate of growth of the number of transistors in personal computer chips (Moore's law; data from Table 1.2). Dashed lines are fits to an exponential growth law.

**Fig. 1.2** The performance of the world's top supercomputers using the LINPACK benchmark (Gflops). Data from <http://www.top500.org>.

**Fig. 1.3** Cumulative number of scientific articles published from 1981 to the date shown that use specific terms in the title, keywords, or abstract. Data from the Science Citation Index (SCI-EXPANDED) available at <http://wos.mimas.ac.uk/> or <http://isi6.isiknowledge.com>.

**Fig. 2.1** Chemical structure of the RNA backbone showing ribose units linked by phosphate groups.

**Fig. 2.2** The chemical structure of the four bases of DNA showing the formation of hydrogen-bonded AT and GC base pairs. Uracil is also shown.

**Fig. 2.3** Schematic diagram of the DNA double helical structure.

**Fig. 2.4** Secondary structure of tRNA-Ala from *Escherichia coli* showing the anticodon position and the site of amino acid attachment.

**Fig. 2.5** Chemical structure of an amino acid (a) and the protein backbone (b). The peptide bond units (boxed) are planar and inflexible. Flexibility of the backbone comes from rotation about the bonds next to the  $\alpha$  carbons (indicated by arrows).

**Fig. 2.6** Chemical structures of the 20 amino acid side chains.

**Fig. 2.7** Four important mechanisms. (a) Transcription. (b) Structure and processing of prokaryotic mRNA. (c) Translation. (d) DNA replication.

**Fig. 2.8** Plot of amino acid volume against pI – two properties thought to be important in protein folding.

**Fig. 2.9** Schematic illustration of principal component analysis. (a) Original data. (b) Scaled and centered on the origin. (c) Rotated onto principal components.

**Fig. 2.10** Plot of the amino acids on the first two components of the principal component analysis.

**Fig. 2.11** Illustration of the data points as vectors in multidimensional space.

**Fig. 3.1** Part of the alignment of the DNA sequences of the *BRCA1* gene.

**Fig. 3.2** Alignment of the *BRCA1* protein sequences for the same region of the gene as Fig. 3.1.

**Fig. 3.3** Illustration of the coalescence process. Each circle represents one gene copy. Bold lines show the lines of descent of genes in the current generation. Thin lines show lines of descent that do not lead to the current generation. Shaded circles show the inheritance of two different mutations.

**Fig. 3.4** Simulations of the spread of neutral mutations through a population under the influence of random drift.

**Fig. 3.5** Simulations of the spread of advantageous mutations through a population. (a) For selection coefficient  $s = 0.05$  both selection and random drift are important. (b) For  $s = 0.2$  selection dominates random drift. The dashed lines show the predictions of the deterministic theory in Box 3.2.

**Fig. 3.6** Fixation probability in a population of  $N = 200$  as a function of selection coefficient  $s$ , for both advantageous and deleterious mutations. When  $Ns \ll 1$ , both types of mutation behave as nearly neutral mutations.

**Fig. 4.1** The accumulation of substitutions in two sequences descending from a common ancestor.

**Fig. 4.2** Illustration of a single line of descent – used for the derivation in Box 4.1.

**Fig. 4.3** The quantities  $P_{AA}(t)$ ,  $P_{AC}(t)$ ,  $D(t)$ , and  $d(t)$  arising from the solution of the Jukes–Cantor model shown as a function of  $\alpha t$ .

**Fig. 4.4** Evolutionary distances  $d$  as a function of observed fraction of differences  $D$  according to the Jukes–Cantor model with: (i) uniform rate of evolution at all sites; (ii) a fraction  $f = 0.25$  of invariant sites; (iii) with rate variation across sites described by a gamma distribution with  $a = 1$ . These are shown in comparison to the uncorrected distance  $d = D$  (dashed line).

**Fig. 4.5** A short protein sequence alignment and the phylogenetic tree obtained for these sequences using the parsimony method. Internal nodes 1–5 are labeled with the deduced amino acid sequence at each point. Amino acid substitutions are labeled on the branch where they occur. Trees like this are the first stage of derivation of the PAM model.

**Fig. 4.6** Above the diagonal – numbers of observed substitutions,  $A_{ij}$ , between each pair of amino acids in the data of Jones, Taylor, and Thornton (1992). On and below the diagonal – log-odds scoring matrix corresponding to PAM250 calculated from these data. This is calculated as  $S_{ij} = 10 \log_{10} R_{ij}$  and rounded to the nearest integer. Cells shaded gray have positive scores, meaning that these amino acids are more likely to interchange than would be expected by chance. Values written in white on a black box correspond to amino acid substitutions that are possible via a single nucleotide substitution at one position in the codon.

**Fig. 4.7** PAM1 matrix calculated by Jones, Taylor, and Thornton (1992). Values are multiplied by  $10^5$  for convenience.  $M_{ij}$  is the probability that the amino acid in row  $i$  changes to the amino acid in column  $j$  in a small time corresponding to 1 PAM unit. The two highest non-diagonal elements in each row are highlighted in black. These are the two most rapid substitution rates for

each amino acid. Frequencies  $\pi_i$  and relative mutabilities  $m_i$  of each amino acid are shown at the bottom of the figure.

**Fig. 4.8** The relationship between the evolutionary distance  $d$  and the fraction of sites that differ  $D$  according to the PAM model of evolution. Solid line, the empirical relationship calculated from Dayhoff *et al.* (1978) data (Eq. (4.31)); dashed line, the Kimura distance formula (Eq. (4.32)). Data points are calculated by the Phylip package for a particular set of aligned proteins using the same evolutionary model.

**Fig. 5.1** Example GenBank entry for the human prion protein, illustrating the use of keywords, sub-keywords, and the Feature Table. For convenience, the nucleotide sequence has been abbreviated (. . .).

**Fig. 5.2** The tripartite International Nucleotide Sequence Database (INSD), comprising EMBL (Europe), GenBank (USA), and DDBJ (Japan).

**Fig. 5.3** Excerpt from the Swiss-Prot entry for the human prion protein, illustrating the EMBL-like structured format, with extensive annotations and database cross-references (note: dotted lines denote points at which, for convenience, material has been excised). Compare the GenBank entry for human prion protein illustrated in Fig. 5.1. The characteristic tandem octapeptide-repeat region thought to be associated with various prion diseases has been highlighted in bold.

**Fig. 5.4** Excerpt from the TrEMBL entry for the canine prion protein, illustrating the EMBL-like structured format, with automatically generated annotations and database cross-references – there are no annotation-rich CC or FT fields (note: dotted lines denote points at which several DR lines have been excised). Compare the Swiss-Prot entry for human prion protein illustrated in Fig. 5.3.

**Fig. 5.5** PIR-NRL3D entry for the hamster prion protein, showing the sequence, together with relevant bibliographic references and structural annotations. Note that the sequence is truncated by comparison with those in the GenBank, Swiss-Prot, and TrEMBL entries illustrated in Figs. 5.1, 5.3, and 5.4.

**Fig. 5.6** Illustration of the three principal methods for building family databases, based on the use of single motifs, multiple motifs, and full-domain alignments.

**Fig. 5.7** Example PROSITE entry, showing one of two data files for the prion protein family.

**Fig. 5.8** Example PROSITE entry, showing the documentation file for the prion protein family.

**Fig. 5.9** Example PRINTS entry, showing the fingerprint for the prion protein family. For convenience, only the first motif is depicted. The two-letter code in the left-hand margin separates the information into specific fields (relating to text, references, motifs, etc.), which allows indexing of the data for rapid querying.

**Fig. 5.10** Example Blocks entry, showing the fifth block used to characterize the prion protein family.

**Fig. 5.11** Example Blocks-format PRINTS entry, showing the third block used to characterize the prion protein family.

**Fig. 5.12** (*opposite*) Excerpt from a PROSITE profile entry, illustrating part of the profile used to characterize WD-40 repeats.

**Fig. 5.13** Excerpt from a Pfam entry, illustrating some of the technical parameters stored for the prion protein family.

**Fig. 5.14** Excerpt from an InterPro entry, illustrating some of the annotation stored for the prion protein family.

**Fig. 5.15** Example graphical output from InterPro, showing matches to the PRINTS (black), PROSITE (spotted), Pfam (striped), and SMART (checked) prion protein signatures.

**Fig. 6.1** (a) Sorting a list of numbers into ascending order. (b) The traveling salesman problem.

**Fig. 6.2** Gap cost functions  $W(l)$  – linear, affine, and general gap functions are shown.

**Fig. 6.3** Pairwise alignment of SHAKE and SPEARE. (a) Pairwise amino acid scores taken from the PAM250 matrix. (b) Alignment scores  $H(i, j)$  using algorithm 1 and  $g = 6$ . (c) Alignment scores  $H(i, j)$  using algorithm 2 and  $g = 6$ . The pathways through the matrix corresponding to the optimal alignments in (b) and (c) are indicated by the thick arrows.

**Fig. 6.4** Global pairwise alignments of hexokinase proteins from human and *Schistosoma mansoni* using an affine gap penalty function. The three parameters used for the three alignments differ only in the value of the gap opening parameter. Regions of alignments (b) and (c) that differ from alignment (a) are written in bold.

**Fig. 6.5** Phylogenetic tree of hexokinase sequences from human, rat, *Schistosoma mansoni*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Plasmodium falciparum*. This tree is produced by Clustal and used as a guide tree during progressive multiple alignment.

**Fig. 6.6** Multiple alignment of hexokinase sequences constructed by Clustal using the guide tree in Fig. 6.5. Bold sections illustrate points discussed in the text.

**Fig. 7.1** (a) Selected parts of the output from MPsrch using the Swiss-Prot database and the sequence PTP1\_YEAST as the query with scoring matrix PAM 300,  $g_{open} = 12$ ,  $g_{ext} = 2$ .

**Fig. 7.1** (b) Selected parts of the output from MPsrch using the Swiss-Prot database and the sequence PTP1\_YEAST as the query with scoring matrix PAM 300,  $g_{open} = 40$ ,  $g_{ext} = 7$ .

**Fig. 7.2** Query of PTP1\_YEAST against Swiss-Prot using BLASTP. The top 10 hits are shown, plus the hits to other yeast

sequences and a few other sequences discussed in the text.

**Fig. 7.3** Probability distributions for match scores using the Simplest Possible Case example of a database search tool.  $P(m)$  is a binomial distribution (Eq. (7.4)).  $F(m_{max})$  is an extreme value distribution. The solid curve is calculated from simulated data, and the dashed line is obtained by fitting Eq. (7.6) to the simulated data.

**Fig. 7.4** Probability distributions for matching word lengths in the word-matching example. Solid curves show simulated data for the distributions of  $l$  and  $l_{max}$ . Both are extreme value distributions. Dashed curves are obtained by fitting Eq. (7.6) to the simulated data.

**Fig. 7.5** BLAST results using PTP1\_YEAST as query sequence. (a) Search made only against other yeast sequences in Swiss-Prot. The complete list of hits is shown. (b) Search made against all proteins in the non-redundant database. Only hits against yeast sequences are shown.

**Fig. 8.1** Rooted trees with a time axis. Tree (a) can be converted to tree (b) by swinging around the horizontal branches like mobiles. Hence (a) and (b) are equivalent to one another.

**Fig. 8.2** A rooted tree with branches scaled according to the amount of evolutionary change.

**Fig. 8.3** The unrooted tree in (a) can be converted to the rooted trees in (b) and (c) and to the tree in Fig. 8.2 by placing the root in different positions.

**Fig. 8.4** Part of the alignment of the mitochondrial small subunit rRNA gene from primates, tree shrews, and rodents.

**Fig. 8.5** The JC distance matrix for some of the rRNA sequences in Fig. 8.4. The matrix has been divided into sections in order to indicate the most important split within the Catarrhini group: that between the old-world monkeys and the apes.

**Fig. 8.6** Tree obtained using the UPGMA method with the matrix in Fig. 8.5. The PHYLIP package was used for tree construction (Felsenstein 2001) and the Treeview program (Page 2001) was used to prepare the figure.

**Fig. 8.7** Step one of the neighbor-joining method.

**Fig. 8.8** Tree obtained with the neighbor-joining method and JC distances. (a) Unrooted. (b) Rooted with the rodents as outgroup, and with bootstrap percentages added.

**Fig. 8.9** An illustration of resampling columns for the bootstrapping method.

**Fig. 8.10** Examples of changes in tree topology. Trees 1, 2, and 3 all differ from each other by a single nearest neighbor interchange. Tree 4 differs from tree 1 by a subtree pruning and regrafting operation.

**Fig. 8.11** An illustration of calculating the likelihood for a single site on a given tree. Letters A, G, and T label bases on the known sequences. Letters X, Y, Z, and W label unknown bases on internal nodes. The times (or equivalently, branch lengths) on each branch are labeled  $t_1, t_2$ , etc.

**Fig. 8.12** Maximum-likelihood topology using the HKY model with invariant sites, plus six gamma-distributed rate categories.

**Fig. 8.13** The parsimony criterion used with morphological character states shows that tree (a) is preferable to trees (b) and (c).

**Fig. 8.14** The parsimony criterion applied to molecular data shows that tree (a) is preferable to tree (b) according to this informative site. Parsimony does not distinguish between trees (c) and (d) as this is a non-informative site.

**Fig. 8.15** The top four trees for the Platyrrhini group obtained by MCMC using the HKY model with invariant sites, plus six gamma-distributed rate categories.

**Fig. 8.16** Top seven trees for the principal groups obtained by the MCMC method. The consensus tree is also shown.

**Fig. 9.1** Schematic illustration of a sequence alignment, showing how gap insertion brings equivalent parts of the alignment into the correct register, leading to the formation of conserved regions, or motifs (shaded blocks). These provide tell-tale signatures that can be used to diagnose new family members.

**Fig. 9.2** Graphs used to visualize protein fingerprints. The horizontal axis represents the query sequence, the vertical axis the % score of each motif (0–100 per motif), and each block a residue-by-residue match in the sequence, its leading edge marking the first position of the match. Solid blocks appearing in a systematic order along the sequence and above the level of noise indicate matches with the constituent motifs. The graphs depict prion fingerprints of the human prion protein (a) and of its chicken homolog (b). The human prion protein is clearly a true-positive match, containing all eight motifs; the chicken homolog fails to make a complete match, but can still be identified as a family member because of the diagnostic framework provided by the five well-matched motif neighbors.

**Fig. 9.3** Block for the prion protein family, in which sequence segments are clustered and weighted according to their relatedness – the most distant sequence within the block scores 100 (for convenience, part of the block has been deleted, as denoted by . . .).

**Fig. 9.4** Example PROSITE profile, showing position-specific scores for insert and match positions. Penalties within insert positions are highlighted bold: here, the values are more tolerant of indels by comparison with the large overall penalties set by the DEFAULT parameter line.

**Fig. 9.5** BLAST output from a search of Swiss-Prot (release 40.29) with the human urotensin II receptor sequence (UR2R\_HUMAN, Q9UKP6). Note, there is no clear cut-off between the urotensin, somatostatin, and galanin receptor matches.

**Fig. 9.6** Output from a search of InterPro with the human vasopressin 1A receptor sequence (V1AR\_HUMAN). The receptor matches the Pfam HMM (PF00001 – white bar), PRINTS fingerprint (PR00237 – black bars), and PROSITE regular expression (PS00237 – spotted bar) and profile (PS50262 – striped bar) for the rhodopsin-like superfamily of GPCRs. However, only PRINTS gives family- and subtype-level diagnoses, with matches to the vasopressin receptor family (PR00896) and vasopressin V1A receptor subtype (PR00752) fingerprints.

**Fig. 9.7** Output from a search of the human urotensin II receptor sequence against (a) its own fingerprint and (b) the somatostatin receptor family fingerprint using PRINTS' GRAPHScan tool (Scordis, Flower, and Attwood 1999). Within each graph, the horizontal axis represents the sequence, and the vertical axis the percentage score (identity) of each fingerprint element (0–100 per motif). Filled blocks mark the positions of motif matches above a 20% threshold. Here, we see that the receptor matches all nine motifs of its own fingerprint but fails to make any significant matches to the somatostatin receptor fingerprint.

**Fig. 9.8** Schematic diagram representing the endothelial differentiation gene (EDG) family of sphingosine 1-phosphate receptors. Positions of the fingerprint motifs for this family are indicated by rectangles; circles mark the positions of residues known to be important in ligand binding in the EDG1

receptor (black) (Remm and Sonnhammer 2000, Attwood *et al.*

2002) and in the EDG6 receptor (white) (Joost and Methner 2002). Models of these receptors predict that the ligand binds within the TM regions, close to domains 2, 3, 5, and 7.

**Fig. 10.1** (a) Dirichlet prior distributions for a single variable  $\theta$  (see Eq. (10.15)), with mean value  $q = 0.05$  and different values of  $A$ . Larger values of  $A$  give distributions more closely peaked about the mean. (b) The effect of increasing the amount of data  $N$  on the posterior probability distribution. The prior has  $q = 0.05$  and  $A = 20$ . The data are chosen with frequency 0.5. As  $N$  increases, the posterior distribution moves towards a peak centered at  $\theta = 0.5$ .

**Fig. 10.2** A simple HMM with two hidden states for distinguishing helical and loop regions of proteins.

**Fig. 10.3** (a) The helical wheel representation of a two-helix coiled-coil domain, showing sites  $a$  to  $g$ . (b) The model contains

nine groups of states, plus an initiation/end group 0. (c) Each of the nine groups contains seven states representing the seven possible positions in the helix. States in one group are linked to the state at the following helix position in the next group. Redrawn from Delorenzi and Speed (2002).

**Fig. 10.4** A detailed HMM for prediction of the positions of TM helices in membrane proteins. The model contains specialized states, representing the helix core, the cap regions at the end of the helices, loop regions connecting helices, and globular regions both inside and outside the membrane. From Krogh *et al.* (2001). Copyright (2001) with permission from Elsevier.

**Fig. 10.5** A profile HMM representation of a protein sequence alignment, showing match (M), insert (I), and delete (D) states.

**Fig. 10.6** Architecture of a typical feed-forward NN with an input layer, a hidden layer, and an output layer.

**Fig. 10.7** A single artificial neuron, with several inputs and one output. The output is a function of the total input, and can either be a step function (solid line) or a sigmoidal function (dotted line), as in Eq. (10.38).

**Fig. 10.8** Illustration of the perceptron problem defined by Eq. (10.40). The two black points must be mapped to output 1, and the two white points must be mapped to output 0. Finding a solution consists of choosing a line that separates the black and white points.

**Fig. 11.1** (a) Alignment of tRNA-Leu genes from mitochondrial genomes, with conserved secondary structure illustrated using bracket notation. Gray-scale shading illustrates sequence conservation. (b) Alignment of the two halves of the aminoacyl acceptor stem of the tRNA, with shading added to illustrate compensatory substitutions.

**Fig. 11.2** Secondary structure of a short RNA molecule illustrating the different types of loop structure.

**Fig. 11.3** Three possible relative positions of two RNA base pairs  $i - j$  and  $k - l$ . Diagrams (a) and (b) show compatible pairs.

Diagram (c) is a type of pseudoknot, and is usually excluded in secondary-structure prediction programs.

**Fig. 11.4** Illustration of the dynamic programming algorithm for maximizing the number of base pairs in an RNA secondary structure (see Eq. (11.1)).

**Fig. 11.5** The probability,  $P_{ii}(p)$ , that a base pair in state  $i$  in one sequence is also in state  $i$  in a related sequence, shown as a function of the percentage,  $p$ , of base-pair changes. Data points are measured in a large set of SSU rRNA genes from bacteria (more details in Higgs 2000). Curves are the best-fit lines for the general reversible seven-state model.

**Fig. 11.6** Distributions,  $P(\delta)$ , for the likelihood ratio tests described in Section 11.2.1.

**Fig. 11.7** Phylogeny of the mammalian orders obtained using the complete set of rRNA and tRNA genes from mitochondrial genomes (Hudelot *et al.* 2003). Two models of evolution were combined in this analysis: one for the paired regions of the secondary structure, and one for the unpaired regions. Posterior probabilities given on the nodes are obtained using the MCMC

method. Where no percentage is given, the node has 100% support.

**Fig. 11.8** Phylogeny of the metazoan taxa. Reproduced from Adoutte *et al.* (2000). Copyright 2000 National Academy of Sciences, USA. (a) Traditional phylogeny based on morphology and embryology. (b) Molecular phylogeny based on rRNA.

**Fig. 12.1** Correlation between the total genome length and the estimated number of genes on bacterial genomes. Each point corresponds to one of the species of Proteobacteria listed in Table 12.1. The correlation coefficient for the linear regression is 0.98, i.e., there is an extremely good straight-line fit.

**Fig. 12.2** Phylogenetic tree of Proteobacteria obtained using a set of 26 concatenated tRNA genes that are present in every genome. The bars indicate the genome length (scale bar shows 2,000 kb). Color indicates percentage of G+C bases.

**Fig. 12.3** A dot-plot illustrating regions of colinearity between the circular chromosomes of *A. tumefaciens* and *S. meliloti* (reproduced from Wood *et al.* 2001. Copyright 2001 AAAS). Each dot represents a bidirectional best hit between protein sequences using BLASTP. A and B indicate putative origin and terminus of replication. C indicates another sizeable region lacking colinearity.

**Fig. 12.4** Comparison of the gene content in three strains of *E. coli* (reproduced from Welch *et al.* 2002. Copyright 2002 National Academy of Sciences USA). The regions of the diagram illustrate the numbers of predicted proteins present in one, two, or all three strains. Remarkably few are shared between all three.

**Fig. 12.5** Phylogeny based on shared gene content of completely sequenced genomes, calculated using the default options of the SHOT program (Korbel *et al.* 2002).

**Fig. 12.6** The pattern of presence and absence of a gene in the four genomes shown here can be explained either by horizontal transfer from B to C, or vice versa, or by two independent losses of the gene from species A and D.

**Fig. 12.7** Comparison of a bacteria-like mitochondrial genome from *Reclinomonas americana* with a mitochondria-like bacterial genome from *Rickettsia prowazekii* (reproduced from Andersson *et al.* 1998, with permission of Nature Publishing Group). (a) Several regions of conserved gene order are illustrated. S10, *spc*, and  $\alpha$  are each operons composed of several consecutive genes. (b) The phylogeny, constructed using ribosomal protein genes from bacteria, mitochondria, and chloroplasts, demonstrates that the mitochondria are related to *Rickettsia* and the chloroplasts to cyanobacteria.

**Fig. 13.1** Steps involved in DNA microarray experiments for (a) cDNA arrays and (b) oligonucleotide arrays. The top illustrates preparation of the array and the bottom illustrates preparation of the sample (reproduced from Schulze and Downward, 2001, with permission of Nature Publishing Group).

**Fig. 13.2** Intensity-ratio data from a mouse cDNA array (reproduced from Quackenbush 2002, with permission of Nature Publishing Group): (a) raw data before normalization, and (b) after normalization with the LOWESS method. R-I stands for Ratio-Intensity, which we have called *M* and *A*.

**Fig. 13.3** Spot-position dependent bias in a yeast microarray study calculated by fitting a smooth quadratic function of spot coordinates to the intensity ratios of spots in each of the 16 print-tip regions (reproduced from Fang *et al.*, 2003, with permission of Oxford University Press).

**Fig. 13.4** Hierarchical clustering of gene-expression profile data for a series of seven time points. The data are artificial and intended to illustrate the sensitivity of the clustering procedure to the details of the measure of similarity between gene profiles (reproduced from Leung and Cavalieri 2003. Copyright 2003, with permission from Elsevier).

**Fig. 13.5** Analysis of expression profiles of 534 genes that vary during the metamorphosis of *Drosophila* using (a) hierarchical clustering and (b) a self-organizing map (reproduced from White *et al.* 1999. Copyright 1999, AAAS). PF stands for puparium formation. Time points are labeled by the number of hours before or after this point (BPF and APF). The PF stage is used as the reference for the other samples; hence the solid black stripe in the third column of data. The SOMs figure shows the mean expression profile of genes in each cluster, together with lines indicating one standard deviation above and below the mean. The number of genes contributing to each cluster is shown above each profile. Cluster c15 corresponds to the co-regulated set of genes that is the uppermost block of expanded gene profiles illustrated in (a).

**Fig. 13.6** 2D gel electrophoresis map of nuclear proteins from *Arabidopsis* (reproduced from Bae *et al.* 2003). The vertical scale is protein molecular weight (in kDa). The horizontal scale is pI. Two gels were prepared from the same sample using different ranges of pH in the isoelectric focusing stage. Numbers indicate protein spots that were successfully identified from their mass fingerprint. Copyright 2003 Blackwell Publishing Ltd.

**Fig. 13.7** An example of the TAP technique applied to the proteins in the polyadenylation machinery complex (Gavin *et al.* 2002, with permission of Nature Publishing Group), which adds a poly-A tail onto mRNAs. A schematic diagram of the complex is shown in (b). In (a), band patterns are shown from SDS-PAGE, with bands labeled according to the corresponding protein. The proteins labeled at the top of each lane were used as the target. The band from the target protein is marked with an arrowhead. This shows that the same bands are observed in each case, confirming that these proteins really do interact in a complex.

**Fig. 13.8** Schema diagram for PEDRo (Taylor *et al.* 2003, with permission of Nature Publishing Group), a proposed Proteomics Experiment Data Repository.

**Fig. M1** Differentiation measures the slope of a curve.

**Fig. M2** Integration measures the area under a curve.

**Fig. M3** A binomial distribution can sometimes be approximated by a normal distribution with the same mean and variance. The dashed line is a binomial distribution with mean 4 and variance 2, and the solid line is the normal distribution approximation.

**Fig. M4** A binomial distribution tends to a Poisson distribution when  $n \gg 1$  but the mean value remains constant. Solid line: Poisson distribution with mean  $\lambda = 5$ . Dashed line: a binomial distribution with  $n = 50$  and  $a = 0.1$ . Dot-dashed line: a binomial distribution with  $n = 10$  and  $a = 0.5$ .

**Fig. M5** Gamma distributions  $f(r, a)$ . Labels indicate the value of  $a$  on each curve.