CHAPTER 6

Proteomics and systems biology

Chapter contents

DNA microarrays *293*
Analysis of microarray data *295*
Mass spectrometry *301*
Identification of components of a complex mixture *301*
Protein sequencing by mass spectrometry *304*
Genome sequence analysis by mass spectrometry *306*
Systems biology *311*
Networks and graphs *313*
Network structure and dynamics *318*
Protein complexes and aggregates *320*
Properties of protein-protein complexes *321*
Protein interaction networks *324*
Regulatory networks *329*
Structures of regulatory networks *330*
Structural biology of regulatory networks *336*
Recommended reading *339*
Exercises, Problems, and Weblems *339*

Learning goals

1. To understand the goals of proteomics—the measurement of amounts and distributions of proteins within a cell or organism.
2. To be familiar with the data derivable from microarrays and their application to inferring and interpreting similarities and differences in gene expression patterns. To understand the relationship between typical 'raw' microarray data (see for instance Plate VII) and the gene expression table.
3. To be familiar with the data derivable from mass spectrometry, and their application to analysis of mixtures of proteins, to partial protein sequencing, and to high-throughput nucleic acid sequencing and searching for variant genetic sequences.
4. To appreciate a trend towards a new basic point of view: The theme of systems biology is integration.
5. To recognize the distinction between physical and logical networks.
6. To understand the general features of graphs, including the distinction between undirected, directed, and labelled graphs. To understand the representation of networks by graphs.
7. To appreciate the different possible kinds of dynamic states of networks.
8. To understand the characteristics of protein-protein and protein-nucleic acid complexes.
9. To understand the structure and some of the building blocks of regulatory networks.

Proteomics is the study of the distribution and interactions of proteins in time and space in a cell or organism. High-throughput experimental methods of data analysis, including microarray analysis and mass spectrometry, are giving us a large-scale picture of the protein economy in living things, and of gene and protein variation in populations.

The goal of systems biology is the synthesis of proteomic, genomic, and other data into an integrated picture of the structure, dynamics, logistics, and ultimately the logic of living things. For any protein in a cell, a systems biologist will combine study of the protein, its gene, the molecules that control its expression or its activity once expressed, and the set of other proteins with which it interacts. A systems biologist will assemble into a metabolic network the chemical reactions catalysed by the enzymes of a cell, and assemble into control networks the mechanisms that regulate their activities and expression.

DNA microarrays

DNA microarrays analyse (1) the mRNAs in a cell, to reveal the expression patterns of proteins; or (2) genomic DNAs, to reveal absent or mutated genes.

See Box: Applications of DNA microarrays, Chapter 1, page 53

1. **For an integrated characterization of cellular activity, we want to determine what proteins are present, where, and in what amounts.** To find the expression pattern of a cell's genes, we measure the relative amounts of many different mRNAs. Hybridization is an accurate and sensitive way to detect whether any particular nucleic acid sequence is present. The key to high-throughput analysis is to run many hybridization experiments in parallel.
2. **Knowing the human genome sequence can help to identify genes associated with propensities to diseases.** Some diseases, such as cystic fibrosis, arise from mutations in single genes. For these, isolating a region by classical genetic mapping can lead to pinpointing the lesion. Other diseases, such as asthma, depend on interactions among many genes, with environmental factors as complications. To understand the aetiology of multifactorial diseases requires the ability to determine and analyse multiple expression patterns of genes, which may be distributed around different chromosomes.

DNA microarrays, or DNA chips, are devices for checking a sample *simultaneously* for the presence of many sequences.

The basic idea is this: To detect whether one oligonucleotide has a particular known sequence, test whether it can bind to an oligo with the complementary sequence ('one-to-one'). To detect the presence or absence of a query oligo in a mixture, spread the mixture out, and test each component of the mixture for binding to the oligo complementary to the query ('many-to-one'). This is a Northern or Southern blot. To detect the presence or absence of *many* oligos in a mixture, synthesize a set of oligos, one complementary to each sequence of the query list, and test each component of the mixture for binding to each member of the set of complementary oligos ('many-to-many'). Microarrays provide an efficient, high-throughput way of carrying out these tests in parallel. They also permit measuring the expression levels of thousands or even tens of thousands of genes in a single sample.

To achieve parallel hybridization analysis, a large number of DNA oligomers are affixed to known locations on a rigid support, in a regular two-dimensional array. The mixture to be analysed is prepared with fluorescent tags, to permit detection of the hybrids. After exposing the array to the mixture, each element of the array to which some component of the mixture has become attached bears the tag. Because we know the sequence of the oligomeric probe in each spot in the array, measurement of the *positions* of the probes identifies their sequences. This analyses the components present in the sample.

Such a DNA microarray is based on a small wafer of glass or nylon, typically 2 cm square. Oligonucleotides are attached to the chip in a square array, at densities between 10 000 and 250 000 positions per cm^2. The spot size may be as small as ~150 μm in diameter. The grid is typically a few cm across. A *yeast chip* contains over

6000 oligos, covering all known genes of *Saccharomyces cerevisiae*. A DNA array, or DNA chip, may contain 400 000 probe oligomers. Note that this is larger than the total number of genes even in higher organisms (excluding immunoglobulin genes).

To analyse a mixture, expose it to the microarray under conditions that promote hybridization, then wash away any unbound sample. To compare two sets of oligos, tag the samples with differently-coloured fluorophores (Plate VII). Scanning the array collects the data in computer-readable form.

Different types of chips designed for different investigations differ in the types of DNA immobilized.

1. In an **expression chip**, the immobilized oligos are cDNA samples, typically 20–80 base pairs long, derived from mRNAs of known genes. The goal of such an experiment is to determine the expression patterns of genes that correspond to the CDNA samples. This is by far the most common application of microarrays. The target sample might be a mixture of mRNAs from normal or diseased tissue.
2. In **mutation microarray analysis**, one looks for patterns of single-nucleotide polymorphisms (SNPs).
3. In **genomic hybridization**, one looks for gains or losses of genes, or changes in copy number. The probe sequences, fixed on the chip, are large pieces of genomic DNA, from known chromosomal locations, typically 500–5000 base pairs long. The probe mixtures contain genomic DNA from normal or disease states. For instance, some types of cancer arise from chromosome deletions, which can be identified by microarrays.

Microarrays are capable of comparing concentrations of components of the sample. This allows quantitative investigation of responses to changed conditions. However, the precision is low. Moreover, mRNA levels, detected by the array, do not always accurately reflect protein levels. Indeed, usually mRNAs are reverse-transcribed into more stable cDNAs for microarray analysis; the yields in this step may also be nonuniform. Microarray data are therefore semiquantitative, in that a distinction between presence and absence is possible, determination of relative levels of expression in a controlled experiment is more difficult, and measurement of absolute expression levels are beyond the capacity of current microarray techniques.

Web resources: Microarray databases

Microarrays provide another high-throughput stream of data production in bioinformatics. A standard called MIAME (Minimum Information About a Microarray Experiment) describes the contents and format of the information to be recorded in the experiment, and deposited.

Major publicly-available microarray databases include:

The European Bioinformatics Institute hosts a database, ArrayExpress:
http://www.ebi.ac.uk/arrayexpress/

The US National Center for Biotechnology Information hosts the Gene Expression Omnibus database:
http://www.ncbi.nlm.nih.gov/geo/

The Stanford Microarray Database:
http://genome-www5.stanford.edu/MicroArray/SMD/

A listing of microarray databases for plants appears in:
http://www.univ-montp2.fr/~plant_arrays/databases.html

For additional lists, see:

Butte, A. (2002), The use and analysis of microarray data, *Nat. Revs. Drug Discov.*, **1**, 951–960.

Penkett, C. J. & Bähler, J. (2004), Getting the most from public microarray data, *European Pharmaceutical Review*, **1**, 8–17.

Analysis of microarray data

The raw data of a microarray experiment is an image, in which the colour and intensity of the fluorescence reflect the extent of hybridization to alternative probes. The two sets of probes are tagged with red and green fluorophores. If only one probe hybridizes, the spot appears red; if only the other probe hybridizes, the spot appears green. If both hybridize, the colour of the corresponding spot appears red + green = yellow (see Plate VII).

The initial goal of data processing is a **gene expression table.** This is a matrix in which the rows correspond to different genes, and the columns to different samples. Different spots in a microarray pattern such as that shown in Plate VII correspond to different genes. For each gene, results from different sets of samples appear in the red or green channel, respectively (or neither, or both). There is extensive redundancy in the oligos in a microarray—each gene may be represented by several spots, corresponding to different regions of the gene sequence; inclusion of controls with a deliberate mismatch allows data verification. Typically one gene may correspond to ~30–40 spots.

The samples may vary according to experimental conditions and/or physiological states, or they may be extracted from different individuals, different tissues or different developmental stages.

The process of data reduction to produce the gene expression matrix involves many technical details of image processing, checking internal controls, dealing with missing data, selecting reliable measurements, and putting the results of different arrays on consistent scales. The derived gene expression table indicates *relative* expression levels. A change in expression levels of a gene between two samples by a factor $\geq$1.5–2 is generally considered significant.

Extraction of reliable biological information from a gene expression table is not straightforward. Despite extensive internal controls, there is considerable noise in the experimental technique. In many cases, variability is inherent within the samples themselves. Micro-organisms can be cloned; animals can be inbred to a comparable degree of homogeneity. However, experiments using RNA from human sources—for example, a set of patients suffering from a disease and a corresponding set of healthy controls—are at the mercy of the large individual

variations that humans present. Indeed, inbred animals, and even apparently identical eukaryotic tissue-culture samples, show extensive variability.

Another intrinsic disadvantage—and a severe one—in interpreting gene expression data, is the fact that the number of genes is much larger than the number of samples. Computationally, we are trying to understand the relationship of a space of very many variables (the genes) to a space of observations (the phenotype), from only a few measured points (the samples). The sparsity of the observations does not give us anywhere near adequate coverage. Statistical methods bear a heavy burden in the analysis to give us confidence in the significance of our conclusions.

Two general approaches to the analysis of a gene expression matrix involve (a) *comparisons focussed on the genes*; that is, comparing distributions of expression patterns of different genes by comparing rows in the expression matrix, or (b) *comparisons focussed on samples*; that is, comparing expression profiles of different samples by comparing columns of the expression matrix:

(a) Comparisons focussed on genes: How do gene expression patterns vary among the different samples? Suppose a gene is known to be involved in a disease, or to a change in physiological state in response to changed conditions. Other genes coexpressed with the known gene may participate in related processes contributing to the disease or change in state. More generally, if two rows (two genes) of the gene expression matrix show similar expression patterns across the samples, this suggests a common pattern of regulation, and possibly some relationship between their functions, including but not limited to a possible physical interaction.

(b) Comparisons focussed on samples: How do samples differ in their gene expression patterns? A consistent set of differences among the samples may characterize the classes from which the samples originate. If the samples are from different controlled groups (for instance, diseased and healthy animals), do samples from different groups show consistently different expression patterns? If so, given a novel sample, we can assign it to its proper class on the basis of its observed gene expression pattern.

How then do we measure the similarity of different rows or columns? Each row or column of the expression matrix can be considered as a vector, in a space of many dimensions. The row-vectors (a row corresponds to a gene), each entry of which refers to the same gene in different samples, has as many elements as there are samples. The column-vectors (a column corresponds to a sample), each entry of which refers to a different gene in a single sample, has as many elements as there are genes reported. It is possible to calculate the 'angle' between different row-vectors, or between different column-vectors, to provide a measure of their similarities. It is then natural to ask whether subsets of the points form natural clusters—points with high mutual similarity—characterizing either sets of genes or sets of samples.

Depending on the origin of the samples, what is already known about them, and what we want to learn, data analysis can proceed in different ways.

(1) The simplest case is a carefully controlled study, using two different sets of samples *of known characteristics*. For instance, the samples might be taken from

bacteria grown in the presence or absence of a drug, from juvenile or adult fruit flies, or from healthy humans and patients with a disease. We can focus on the question, what differences in gene expression pattern characterize the two states? Can we design a classification rule such that, given another sample, we can assign it to its proper class? This would be applicable in diagnosis of disease. For instance, determination of the subtype of a leukaemia permits more accurate treatment and prognosis. Subject to the availability of adequate data, such an approach can be extended to systems of more than two classes.

Computationally, training such a classification algorithm is called 'supervised learning'. The expression pattern of each sample is given by a vector corresponding to a single column of the matrix. This corresponds to a point in a many-dimensional space—as many dimensions as there are genes. In favourable cases, the points may fall in separated regions of space. Then a scientist, or a computer program, will be able to draw a boundary between them. In other cases, separation of classes may be more difficult. Consider the distribution of football players during a match. At the start of play, a line drawn across the midfield separates the teams; that is, the midfield line divides the field into two regions, each region containing exclusively the players of one of the teams. During play, the teams become commingled, and it is very difficult to divide the field into regions that separate the teams.

(2) In a different experimental situation, we might not be able to *preassign* different samples to different categories. Instead, we hope to extract the classification of samples from the analysis. The goal is to cluster the data to *identify* classes of samples and the differences between the genes that characterize them.

Many clustering algorithms have been applied to microarray data, including those that try to work out simultaneously both the *number* of clusters and the *boundaries* between them. All algorithms must face the difficulty arising from the sparsity of sampling of the very high-dimensional space of the measurement. Sometimes it is possible to simplify the problem by identifying a small number of combinations of genes that account for a large portion of the variability. This is called **reduction of dimensionality** (see Box).

Reduction of dimensionality

The distribution of gene expression data in a space of a large number of dimensions means that (1) coverage of the space with a limited number of samples is sparse and (2) it is difficult to visualize the distribution of sample points. In some cases, the distribution may depend primarily on fewer equivalent variables, and it is very advantageous to find them and transform the data accordingly.

A simple example illustrates the basic idea. Consider a distribution of groups of people picnicking on a beach. Represent the position of each person by the x, y, and z coordinates of the tip of his or her nose. Take the x-axis to be parallel

→

→

Reduction of dimensionality (*continued*)

to the shoreline, the *y*-axis perpendicular to the shoreline, and the *z*-axis vertical. Obviously height is irrelevant: this is really a two-dimensional, not a three-dimensional, distribution. To cluster the people into groups (perhaps families, or surfing clubs) the *x* and *y* coordinates carry all the significant data, and the *z*-coordinate carries only irrelevant information, such as the heights of the people and whether or not they are standing up or sitting on the sand. In this case, to reduce the dimensionality from three to two we need only ignore the *z*-coordinate. (Indeed, if the tide comes in and the beach area becomes narrower, the dimension along the shoreline carries the bulk of the information and the dimensionality could be further reduced from two to one.)

Now suppose the people are not distributed on the beach, but climbing a vertical rock face rising parallel to the shoreline. This also is really a two-dimensional, not a three-dimensional, distribution, but in this case it is the *x* and *z* coordinates that carry the information.

In more complex cases, reduction in dimension requires more than simply picking coordinates to ignore. Suppose the people are distributed on a ski slope. To reduce the distribution from three to two dimensions, we could not simply ignore a coordinate, but would have to *project* the data onto the oblique plane parallel to the slope. (The plane parallel to the sloping ground is oblique to a coordinate system oriented along horizontal and vertical directions.) This idea of *projection* of the data onto a lower-dimensional space, which contains the important components of the variation, is the key to the methods.

Practical problems of data analysis are harder than these simple illustrations. For one thing, the starting dimensions are much higher than three and the reduction in dimensionality is potentially much greater. For another, it is not obvious how to achieve the dimensionality reduction because we don't have the easily visualizable picture of the physical space and the distribution of people on a beach, rock face, or ski slope.

Nevertheless, the questions to be answered remain: Along what directions should we project the data to retain the largest discrimination using the fewest dimensions? Mathematical methods known as **Principal-Component Analysis (PCA)** using the **Singular Value Decomposition (SVD)** can solve this problem. These methods automatically select a new coordinate system that best represents the variability of the data along the fewest axes, and, for each new coordinate axis, the calculation gives a measure of the contribution of that coordinate to accounting for the variability of the data.

Although two dimensions may well not contain all important components of the variation, we can always pick the best two-dimensional projection and plot the result on a graph; this has the immense advantage of allowing scientists to stare at the data and think about them. (Three-dimensional distributions can also be represented visually, with somewhat greater difficulty.)

Case Study 6.1: Interpretation of microarray data: Regulation of genes by BRCA1 and implications for the role of BRCA1 dysfunction or silencing in carcinogenesis

The *BRCA1* gene encodes a tumour suppressor. It is mutated in approximately 50% of patients with familial predisposition to breast and ovarian cancer. A single defective *BRCA1* allele is sufficient to increase risk, for in any cell the normal copy of the gene may be lost, or, in a small fraction of cases, rendered inactive by promotor methylation.

BRCA1 is an 1863-residue protein. It has an N-terminal ring finger domain, followed by a predicted helical coiled-coil region, followed by two tandem BRCT domains, that bind other proteins, and also regulate transcription. (BRCT abbreviates BRCA C-terminal domain.)

BRCA1 interacts with many other proteins to form functional complexes and is thereby involved in several different activities, including:

- **Sensing and signalling of lesions in DNA.** *BRCA1* responds to several types of DNA damage—for instance double-strand breaks—and activates repair mechanisms appropriate to each.
- **Preserving chromosome structure.** As chromosome integrity may suffer *as a consequence of* inaccurate repair of DNA damage, these functions are related.
- **Mediating checkpoint tests at points in the cell cycle**, in part at least by regulating transcription of genes encoding proteins involved in checkpoint enforcement.

A unifying idea about *BRCA1* is that the protein encoded mediates responses to DNA damage by eliciting repair mechanisms and, in case repair is unsuccessful, checkpoint mechanisms that stop cells with unrepaired damage from propagating. Loss of BRCA1 function leads to the accumulation of damaged DNA in cells, enhancing the chances of transition to a cancerous state.

The variety and complexity of the processes involving BRCA1 make it difficult to sort out the detailed mechanism of its relation to cancer:

1. Is tumour formation a direct consequence of loss of one or more functions of BRCA1 and its interacting partners? If so, which one(s)?
2. What is the importance of transcriptional regulation—of *BRCA1* by products of other genes, of other genes by BRCA1, or both? To what extent do changing expression patterns involving *BRCA1* lead *indirectly* to tumorigenesis? We shall see that the distinction between direct and indirect effects is not really a hard and fast one: BRCA1 binds directly to some of the proteins the expression of which it regulates.
3. DNA repair mechanisms are common to many types of cells. Why does BRCA1 dysfunction or silencing specifically lead to increased risk of cancers of the breast and ovary (and other epithelial tissues, including pancreas and prostate)?

→

Case Study 6.1. (*continued*)

One function of BRCA1 is control over transcription. In order to investigate the regulatory context of the relationship of *BRCA1* to cancer risk, Welcsh et al. used microarray analysis to compare the expression patterns of genes in cells producing high and low levels of BRCA1, using a cell line in which BRCA1 expression was selectively inducible. (See Plate VIII.) The chip used for detection of the response contained oligonucleotides representing ~6800 human genes. (Note that this is a relatively small fraction of the total human proteome.)

The results implicated 373 genes, differentially expressed by significant and reproducible amounts in response to higher levels of BRCA1 expression. Standing out among these were 57 up-regulated genes and 15 down-regulated genes, for which expression levels changed by factors ≥ 2. These candidates for involvement in functions of BRCA1 relevant to tumourigenesis were checked for differential expression in cancer tissues from patients and normal controls.

Clustering the gene expression matrix shows the clear distinction between up- and down-regulated genes, and gives an impression of the variability among replicates (Plate VIII). Many of the proteins encoded by up-regulated genes are hormone receptors and structural proteins. Many of the proteins encoded by down-regulated genes are involved in DNA replication and translation.

Notable among the genes identified in the study are the following:

Consistent with the tissue-specific appearance of tumours as a result of BRCA1 dysfunction, some of the genes with altered expression patterns are involved in oestrogen-mediated control pathways, suggesting a possible link to the tissue-specificity enigma. The set of proteins implicated includes cyclin D1 and myc, which are up-regulated by lower levels of *BRCA1*. Cyclin D1 and myc are observed to be overexpressed in 20% of breast cancers, consistent with their repression by functional BRCA1. (For comparison with the clinical setting, low levels of BRCA1 expression correspond to patients with reduced or absent BRCA1 function, that is, the high-risk group; and high levels are analogous to normal controls. However, the experiments of Welcsh et al. did not try to reproduce actual endogenous BRCA1 expression levels observed in patients and normal counterparts.)

Conversely, JAK and STAT proteins are down-regulated by decreased levels of BRCA1. These proteins are implicated as growth inhibitors in control pathways that govern proliferation, differentiation, apoptosis and transformation. Loss of BRCA1 activity would be expected to reduce JAK1 and STAT1 levels, promoting cellular proliferation and reducing apoptosis. This is consistent with the observation that *Stat1*-null mice develop tumours more readily than normals.

The relationships detected by Welcsh et al. are part of the cell's control network. However, some of the products of genes regulated by BRCA1 are also known to be involved in formation of functional complexes

→

with BRCA1. For instance, the product of *myc*—a potent oncogene—binds to BRCA1, suggesting a direct inhibition of myc function by BRCA1. Thus reduced BRCA1 levels would have the dual effect of reducing the inhibition of *myc* through binding, and increasing its expression through loss of transcriptional repression.

Thus, *myc* is linked to BRCA1 through both physical and regulatory interactions. We shall see in a later section that the idea of two parallel interaction networks in cells—physical interactions and regulatory interactions—is an attractive distinction. However, it is one that is difficult to maintain in a system such as BRCA1 function in which the two are so closely intertwined.

Mass spectrometry

Mass spectrometry is a physical technique that characterizes molecules by measurements of the masses of their ions. Investigations of large-scale expression patterns of proteins require methods that give high throughput rates as well as fine accuracy and precision. Mass spectrometry achieves this, which has stimulated its development into a mature technology in widespread use. Applications to molecular biology include:

- Rapid identification of the components of a complex mixture of proteins.
- Sequencing of proteins and nucleic acids, including high-throughput genomic sequencing, and surveying populations for genetic variability.
- Analysis of post-translational modifications, or substitutions relative to an expected sequence.

Identification of components of a complex mixture

First the components are separated by electrophoresis, then the isolated proteins digested by trypsin to produce peptide fragments with r.m.m. about 800–4000 (see Fig. 6.1). Trypsin cleaves proteins after Lys and Arg residues. Given a typical amino acid composition, a protein of 500 residues yields about 50 tryptic fragments. The spectrometer measures the masses of the fragments with very high accuracy. The list of fragment masses, called the **peptide mass fingerprint**, characterizes the protein (see Fig. 6.2). Searching a database of fragment masses identifies the unknown sample.

Construction of a database of fragment masses is a simple calculation from the amino acid sequences of known proteins, translations of open reading frames in genomes, or (in a pinch) of segments from EST libraries. The fragments correspond to segments cut by trypsin at lysine and arginine residues, and the masses of the amino acids are known. (Note that trypsin doesn't cleave Lys-Pro peptide bonds, and may also fail to cleave Arg-Pro peptide bonds.)

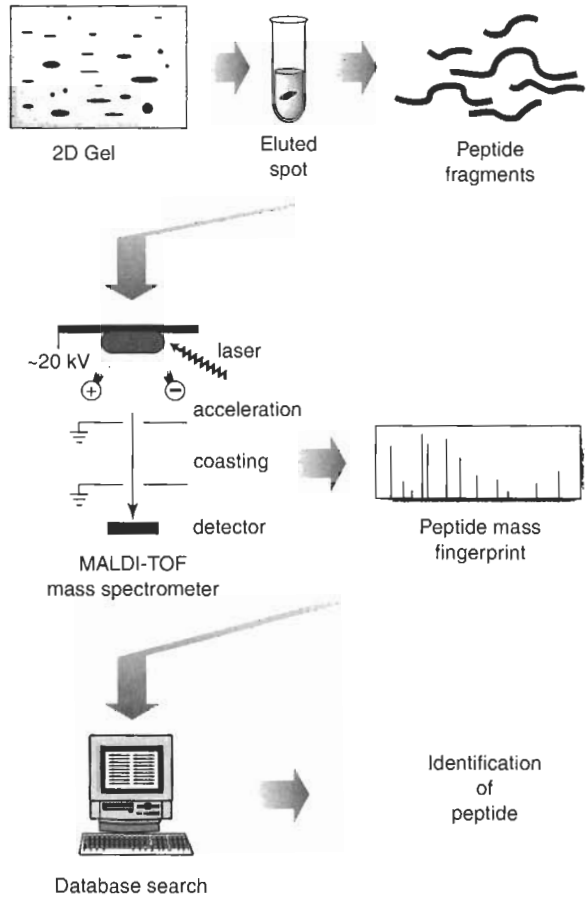


Fig. 6.1 Identification of components of a mixture of proteins by elution of individual spots, digestion and fingerprinting of the peptide fragments by MALDI–TOF (Matrix-Assisted Laser Desorption Ionization–Time of Flight) mass spectrometry, followed by looking up the set of fragment masses in a database.

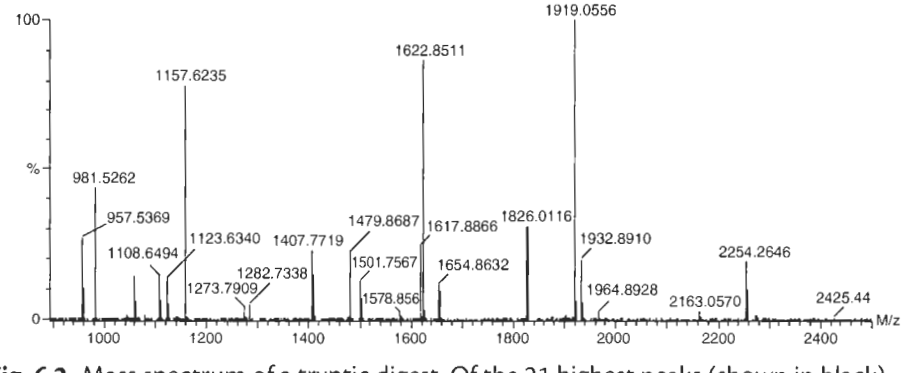


Fig. 6.2 Mass spectrum of a tryptic digest. Of the 21 highest peaks (shown in black), 15 match expected tryptic peptides of the 39 kDa subunit of Cow mitochondrial complex I. This easily suffices for a positive identification. (Figure courtesy of Dr. I. Fearnley.).

Web resources: Identification of proteins from peptide mass fingerprints

Two commonly-used databases compiling predicted peptide mass fingerprints of proteins are:

www.matrixscience.com

prospector.ucsf.edu/ucsf.html3.4/msfit.htm

Mass spectrometry is sensitive and fast. Peptide mass fingerprinting can identify proteins in sub-picomole quantities. Measurement of fragment masses to better than 0.1 mass units is quite good enough to resolve isotopic mixtures. It is a high-throughput method, capable of processing 100 spots/day (though sample preparation time is longer). However, there are limitations. Only proteins of known sequence can be identified from peptide mass fingerprints, because only their predicted fragment masses are included in the databases. (As with other fingerprinting methods, it would be possible to show that two proteins from different samples are likely to be the same, even if no identification is possible.) Also, post-translational modifications interfere with the method by altering the masses of the fragments.

The results shown in Fig. 6.2 are from an experiment in which the molecular masses of the ions were determined from their time-of-flight (TOF) over a known distance, as illustrated in Fig. 6.1. The operation of the spectrometer involves these steps:

1. Production of the sample in an ionized form in the vapour phase.
2. Acceleration of the ions in an electric field. Each ion emerges with a velocity proportional to its charge/mass ratio.
3. Passage of the ions into a field-free region, where they 'coast'.
4. Detection of the times of arrival of the ions. The time-of-flight (TOF) indicates the mass-to-charge ratio of the ions.
5. The result of the measurements is a trace showing the flux as a function of the mass-to-charge ratio of the ions detected.

Because proteins are fairly delicate objects, it has been challenging to vaporize and ionize them without damage. Two 'soft-ionization' methods that solve this problem are:

1. **Matrix-Assisted Laser Desorption Ionization (MALDI)** The sample is introduced into the spectrometer in dry form, mixed with a substrate or matrix that moderates the delivery of energy. A laser pulse, absorbed initially by the matrix, vaporizes and ionizes the protein. The MALDI-TOF combination, that produced the results shown in Fig. 6.2, is a common experimental configuration.
2. **Electrospray Ionization (ESI)** This method starts with the sample in liquid form. Spraying it through a small capillary with an electric field at the tip creates an aerosol of highly-charged droplets. As the solvent evaporates, the

droplets contract, bringing the charges closer together and increasing the repulsive forces between them. Eventually the droplets explode into smaller droplets, each with less total charge. This process repeats, ultimately creating ions, which may be multiply-charged, devoid of solvent. These ions are transferred into the high vacuum region of the mass spectrometer.

Because the sample is initially in liquid form, ESI lends itself to automation in which a mixture of tryptic peptides passes through a high-performance liquid chromatograph (HPLC) into the mass spectrometer directly.

Protein sequencing by mass spectrometry

Fragmentation of a peptide produces a mixture of ions. Conditions under which cleavage occurs primarily at peptide bonds yield series of ions differing by the masses of single amino acids (Fig. 6.3). The amino acid sequence of the peptide is therefore deducible from analysis of the mass spectrum (Fig. 6.4), subject to ambiguities: for instance, Leu and Ile have the same mass and cannot be distinguished in fragments cleaved only at peptide bonds. Discrepancies from the masses of standard amino acids signal post-translational modifications. In practice, the sequence of about 5–10 amino acids can be determined from a peptide of length < 20–30 residues.

In current practice, the fragments are produced *in situ*: First the peptide is vaporized, then it is fragmented by Collision-Induced Dissociation (CID) with Argon gas. This approach requires two mass analysers, operating in tandem in the same instrument (called MS/MS). The vaporized sample first passes through

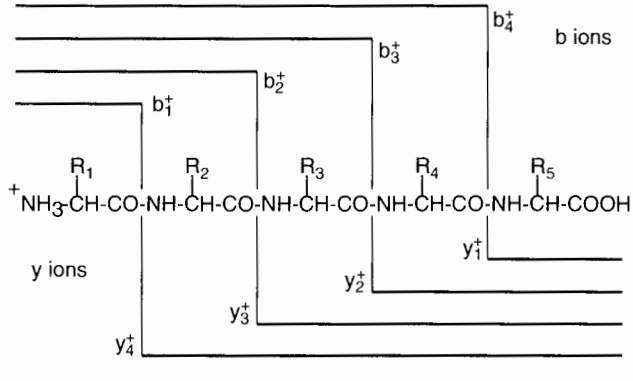


Fig. 6.3 Fragments produced by peptide bond cleavage of a short peptide. b ions contain the N-terminus; y ions contain the C-terminus. The difference in mass between successive b ions or successive y ions is the mass of a single residue, from which the peptide sequence can be determined. Two ambiguities remain: Leu and Ile have the same mass and cannot be distinguished, and Lys and Gln have almost the same mass and usually cannot be distinguished. In Collision-Induced Dissociation, bond breakage can be largely limited to peptide linkages by keeping to low-energy impacts. Higher energy collisions can fragment sidechains, occasionally useful to distinguish Leu/Ile and Lys/Gln.

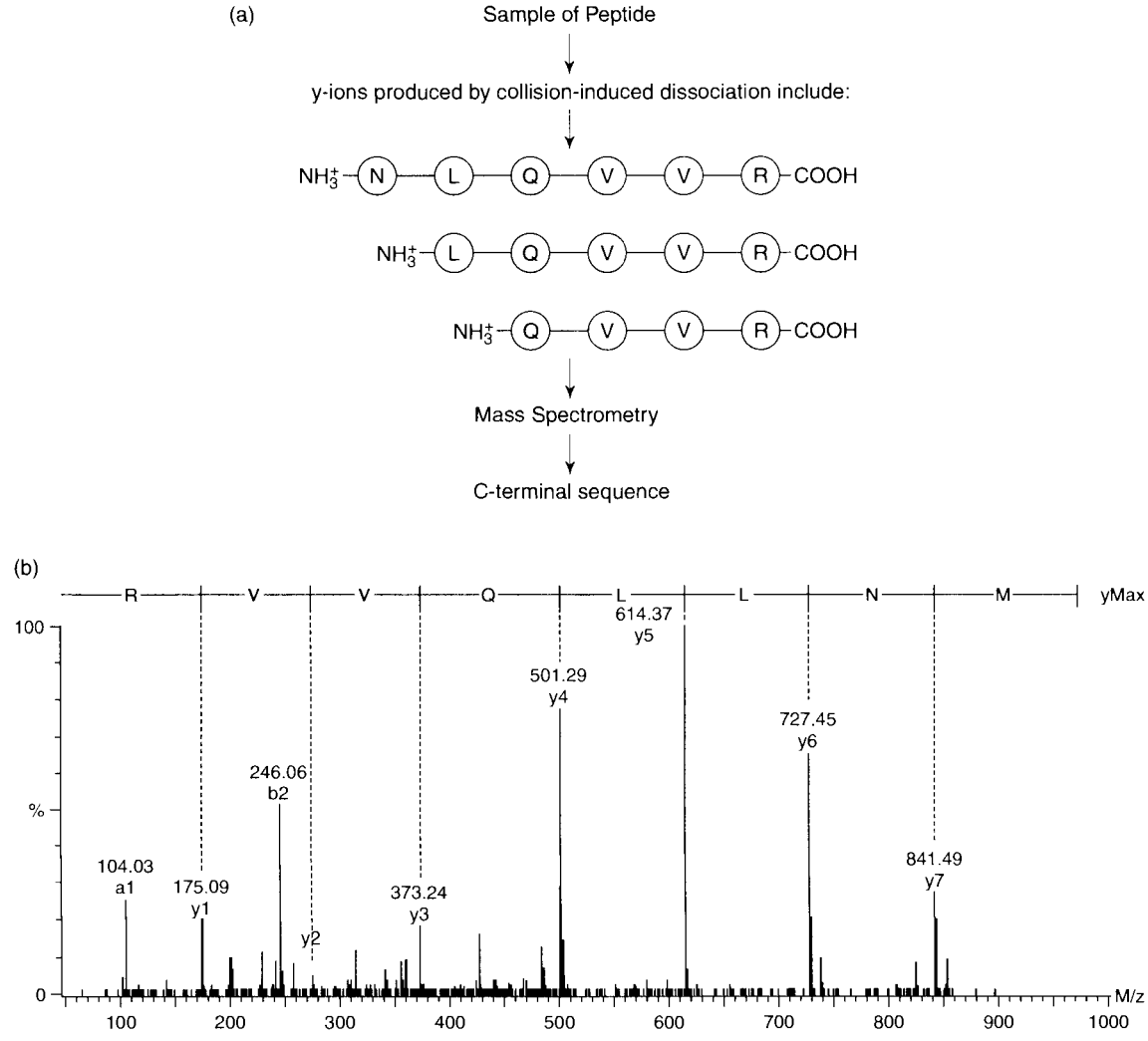


Fig. 6.4 Peptide sequencing by mass spectrometry. Collision-Induced Dissociation (CID) produces a mixture of ions. (a) The mixture contains a series of ions, differing by the masses of successive amino acids in the sequence. In CID the ions are not *produced* in sequence as suggested by this list, but the mass-spectral measurement automatically sorts them in order of their mass/charge ratio. (b) Mass spectrum of fragments suitable for C-terminal sequence determination. The greater stability of y ions over b ions in fragments produced from tryptic digests simplifies the interpretation of the spectrum. The mass differences between successive y-ion peaks are equal to the individual residue masses of successive amino acids in the sequence. Because y ions contain the C-terminus, the y-ion peak of smallest mass contains the C-terminal residue, etc. and therefore the sequence comes out 'in reverse'. The two leucine residues in this sequence could not be distinguished from isoleucine in this experiment. From Carroll, J., Fearnley, I. M., Shannon, R. J., Hirst, J. & Walker, J. E. (2003), Analysis of the subunit composition of complex I from bovine heart mitochondria, *Mol. Cell. Proteomics*, **2**, 117–126 (Supplementary figure S138).

one mass analyser, to separate an ion of interest. The selected ion passes into the collision cell where impact with Argon atoms excite and fragment it. By keeping the energy of impact low, the fragmentation can be limited largely to peptide bond breakage (Fig. 6.3). The second mass analyser determines the masses of the fragments.

Masses of amino acid residues, standard isotopes

Gly	57.02146	Ala	71.03711	Ser	87.03203
Pro	97.05276	Val	99.06881	Thr	101.04768
Cys	103.00919	Leu	113.08406	Ile	113.08406
Asn	114.04293	Asp	115.02694	Gln	128.05858
Lys	128.09496	Glu	129.04259	Met	131.04049
His	137.05891	Phe	147.06841	Arg	156.10111
Tyr	163.06333	Trp	186.07931		

Genome sequence analysis by mass spectrometry

Mass spectrometry of nucleic acids provides a very precise and high-throughput technique for quantitative analysis of DNA and RNA sequences in individuals and in populations. Its advantages include:

- **High precision** The standard deviation of typical mass spectral concentration measurement replicates is ~3% compared with ~200% for microarray measurements.
- **More data per sample** A mass spectrum contains many peaks rather than a single value. This allows analysis of mixtures; and permits 'multiplexing', or simultaneous analysis of several features of a set of mixed samples.
- **High specificity and sensitivity** Very small sample sizes are required. PCR amplification can be pushed to very high gain, as there is little risk of mistaking a contaminant for a true sample amplicon. In fact, it is possible to determine sequences from individual cells or even single DNA strands.

To prepare for the measurement, samples are treated by gene-specific PCR amplification by allele-specific primer extension, to produce single-stranded oligonucleotides. Products are purified and embedded in a matrix suitable for MALDI vaporization and mass analysis. No hybridization step is required for detection. Assembly of many subjects on an array allows for automation of data collection. (Throughput rates can reach 10^5 spectra per instrument per day.)

The typical relative molecular mass of an oligonucleotide measured is ~6000, corresponding to about 20 bases. Under conditions where the amplified products of different alleles contain different numbers of bases, the mass difference is $\geq$300, a very large difference relative to the accuracy of mass spectrometry. In fact,

it is feasible to pick up a single-base substitution in oligos of the same length, or even the methylation of a gPc site. For nucleotide substitutions, the mass differences between bases range from ±9 for t↔a to ±40 for c↔g.

Applications include:

(1) **Measurement of allele frequencies in populations, or detection of alleles in individuals, by identification of single nucleotide polymorphisms (SNPs)** For population studies, samples from several individuals in the selected groups can be pooled, and genotype frequencies measured to about 3% accuracy. Several SNPs can be determined from a single spectrum. Such studies have impact on a wide variety of fields, including anthropology, agriculture, and forensics, but medical applications are the major driving force. For example, controlled comparisons between healthy populations and those predisposed to a disease can identify genetic factors of clinical importance.

(2) **Characterization of individual genotypes** A selection of 100 000 SNPs offers about 3 polymorphisms per gene, enough for fairly thorough characterization of the protein-coding portion of an individual person's genome. Determination of one individual's SNP profile is achievable using one instrument for one day. Clinical applications include: (a) diagnosis, based on systematic differences, between healthy individuals and those with a disease, previously established from controlled population studies, and (b) pharmacogenomics, to distinguish patients who will benefit from treatment with a drug from those who will not benefit or even risk severe side effects.

(3) **Measurement of individual haplotypes** Haplotypes are local combinations of genetic polymorphisms that tend to be co-inherited. (See Box, Haplotype distributions.) Haplotypes simplify the search for phenotype-genotype correlations, because they reduce the number of variables with which to characterize the genotype. Mass-spectrometric methods based on amplifying regions around SNPs in a sample containing a single DNA molecule provide an accurate and high-throughput method of individual haplotype determination.

Haplotype distributions

Our individual genomes are characterized by a distribution of genetic markers. Single-nucleotide polymorphisms are convenient features to observe, and to study within and across populations. Although the overall density of SNPs in our genomes is ~1 SNP/5 kbp, many 100 kbp regions show only a few (typically 2–4) of the possible combinations of SNPs, suggesting that recombination is rare within the region. These segments, which remain intact, are separated by intervals in which recombination is more frequent.

The few discrete combinations of SNPs define the **haplotype** of an individual. The International HapMap project collects and curates haplotype distributions from several human populations.

→

Haplotype distributions (*continued*)

Haplotypes are difficult to measure, because it is essential to determine which SNPs appear in the *same* DNA strand. Clearly, study of mixed samples from several individuals can determine the frequencies of individual SNPs but not their correlation into haplotypes. Even a sample containing both chromosomes from a single diploid cell mixes the contributions of both copies of the region. However, mass spectral studies of amplified single-copy DNA molecules, produced by dilution, can identify the *combination* of SNPs appearing together on the same chromosome, allowing unambiguous haplotyping.

(4) **Measurement of gene expression levels on an absolute scale, with a precision of ~3%** This is achieved by spiking the RNA extracted from a sample with a known amount of a related oligoribonucleotide, and measuring the relative amounts of signal from the calibrating oligo and the natural ones.

(5) **Noninvasive prenatal diagnosis based on the small amount of foetal DNA that leaks into maternal blood** Because of the 95–99% maternal DNA background, only paternal contributions to the foetus can be identified. However, the technique is sensitive enough to detect the *SRY* gene, demonstrating that the foetus is male, or other paternal alleles that may be useful in diagnosing genetic abnormalities. It should be emphasized that the use of only a *maternal blood sample* avoids the significant risks of an invasive procedure to sample amniotic fluid.

(6) **Genomic sequencing** Mass spectrometry has the potential to compete in accuracy and throughput with gel-based methods for large-scale DNA sequence determination.

Case Study 6.2: Application of combined genomic, proteomic, and structural methods to antibiotic resistance in tuberculosis

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis*. Despite development of vaccines and drugs, it remains a potent killer. Tuberculosis, and AIDS, are the most common cause of death from infectious disease, claiming about three million victims per year. The World Health Organization estimates that there are eight million new cases per year. HIV infection exacerbates the mortality of tuberculosis infection by lowering resistance.

Our front-line defences against most bacterial infections include macrophages, cells of the immune system that engulf bacteria and attack them with a variety of chemical and biochemical agents. *M. tuberculosis*, exceptionally, is adapted to survive *within* the macrophage. Part of its adaptation is structural: cells of *M. tuberculosis* and close relatives surround themselves with a waxy coat. The low permeability of the coat shields them from the inhospitable environment within the macrophage, including low

pH and oxidative stress. The bacteria also make substantial changes to gene expression patterns, to adapt their physiological state to these surroundings.

After several decades of decline following the development of effective drugs, the incidence of tuberculosis began to increase in the mid-1980s. One reason is emergence of resistant strains.

A primary drug used in prevention and therapy of tuberculosis is isoniazid (isonicotinic acid hydrazide). Isoniazid attacks *M. tuberculosis* by interfering with the synthesis of its cell wall, without which the bacterium cannot survive. Targets of isoniazid include an NADH-dependent enoyl acyl carrier protein reductase (InhA), and a β-keto-acyl ACP synthase (KasA). These enzymes participate in synthesis of mycolic acids, major components of the cell wall.

Isoniazid must be converted to an active form after absorption by the bacterial cell. The enzyme that effects the conversion, KatG, is a natural suspect for involvement in resistance. Its natural function is to detoxify peroxides.

Several methods have been applied to elucidate the adaptations responsible for isoniazid resistance:

(1) Changes in gene expression patterns were detected using microarrays.

(2) Genes that change expression were sequenced in susceptible and resistant strains, and mutations observed.

(3) The crystal structure of isoniazid bound to InhA has been determined.

(1) Changes in gene expression patterns

Wilson and colleagues* examined susceptible and resistant strains of *M. tuberculosis* at times up to 8 hours of exposure to isoniazid. (Plate IX shows the results after 4 hours exposure.) The array included almost all ORFs identified in the *M. tuberculosis* genome. Although biochemical studies had already implicated some proteins in resistance, a general screen was carried out in order to identify as many potential drug targets as possible.

The genome of *M. tuberculosis* is about 4.4 Mbp long and contains about 4000 genes.

Exposure to isoniazid greatly enhanced the transcription of two classes of genes. One set is involved in cell-wall synthesis, including an operon-like cluster encoding components of a fatty-acid synthase complex (FAS-II). Additional genes, including a subunit of alkyl hydroxyperoxide reductase (AhpC), that handles oxidative stress, were also up-regulated. The logic of the experiment is that the treated cells are recognizing the effects of the drug, and feedback mechanisms are acting to try to compensate for reduced activities, by enhanced expression.

(2) Mutations conferring resistance to isoniazid

On the basis of the changed expression profiles, Ramaswamy et al. (2003) sequenced a total of 2.6 Mbp from 124 *M. tuberculosis* isolates.† These include

* Wilson, M., DeRisi, J., Kristensen, H. H., Imboden, P., Rane, S., Brown, P. O. & Schoolnik, G. K. (1999), Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization, *Proc. Natl. Acad. Sci. USA*, **96**, 12833–12838.

† Ramaswamy, S. V., Reich, R., Dou, S. J., Jasperse, L., Pan, X., Wanger, A., Quitugua, T. & Graviss, E. A. (2003), Single nucleotide polymorphisms in genes associated with isoniazid, *Antimicrob. Agents Chemother.*, **47**, 1241–1250.

Case Study 6.2. (*continued*)

mutations in KatG that impede activation of isoniazid, and mutations in InhA to escape inhibition by the activated form.

Note that because oxidative stress is part of the host's natural defence to infection, simple knockout of KatG could be a dangerous strategy for the bacterium. Ideally, to achieve resistance, the bacterium would reduce the activity of the enzyme in isoniazid activation but retain activity against small peroxides. In this way it would reduce susceptibility to the drug while maintaining its general fitness in the environment within the macrophage. Precisely this balance is attained by the most common KatG mutation in resistant strains, S315T.

The most common mutation in InhA is S94A. The inhibitory effectiveness of activated isoniazid is reduced in this modified protein.

(3) Crystallography

Rozwarski et al. (1998) solved the structure of the complex between the activated form of isoniazid and InhA (Fig. 6.5). The drug is covalently attached to the nicotinamide ring of NAD, bound to the active site of InhA. The sidechain of S94 is also shown. In the inhibitory complex, the protein binds the NAD-activated isoniazid adduct. The coupling of these molecules can occur *only* on the enzyme (in solution activated isoniazid and NADH do not react).

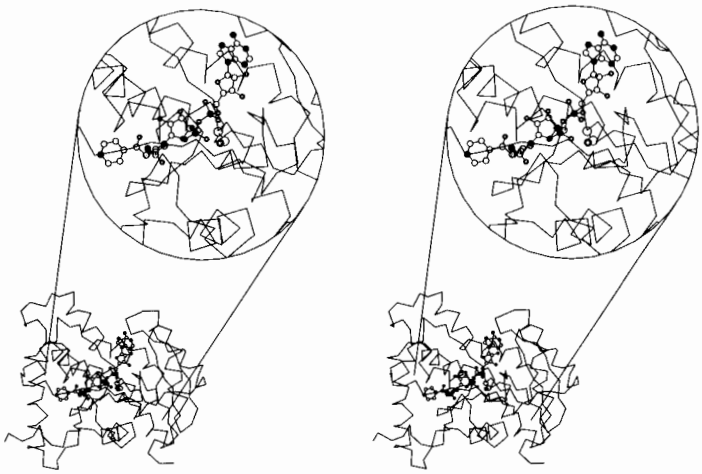


Fig. 6.5 Structure of long fatty acid chain enoyl-acp reductase (InhA) in complex with inhibitor [1ZID]. The ligand is an adduct of activated isoniazid and NADH. Shown in red are the isoniazid moiety of the inhibitor (at left in blown-up circle), and the sidechain of Ser94 (centre of blown-up circle). The mutation S94A contributes to isoniazid resistance. (See Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs, W. R., Jr. & Sacchettini, J. C. (1998), Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*, *Science*, **279**, 98–102.)

How does the S94A mutant achieve resistance? In the absence of inhibitor, the enzyme can bind either substrate first and then cofactor, or cofactor first and then substrate. Because the substrate occupies the same site on the enzyme as the inhibitor, only if cofactor is bound first can an inhibitory complex form. Two pathways are possible, the first leading exclusively to product, the other producing an inhibitory complex (E = enzyme, S = substrate, C = cofactor and I = inhibitor):

If substrate binds first:

(1a) $E + S \rightarrow ES + C \not\rightarrow ESC \rightarrow E + C + P$

(1b) $E + S \rightarrow ES + I \rightarrow$

If cofactor binds first:

(2a) $E + C \rightarrow EC + S \rightarrow ESC \rightarrow E + C + P$

(2b) $E + C \rightarrow EC + I \rightarrow ESI = \underset{\text{complex}}{\text{inhibitory}}$

If substrate binds first (1a and 1b), the inhibitory complex cannot form. If cofactor binds first (2a and 2b), a stable inhibitory complex may form, taking the enzyme out of the game permanently.

The S94A mutation reduces the affinity of the enzyme for NADH. This enhances the substrate-bound-first pathway (1a and 1b), lowering the amount of inhibitory complex produced (2b), and also enhancing the dissociation rate of the inhibitory complex.

It is also possible that S94A and other mutations reduce the affinity of the adduct.

Research and development of anti-tuberculosis drugs is a continuing challenge. This example shows the effectiveness of coordinated application of many different techniques.

Systems biology

Systems biology focusses on the integration and control of gene and protein activity.

Proteins are social animals, and life depends on their interactions. Because individual proteins have specialized functions, control mechanisms are required to integrate their activities. The right amount of the right protein must function in the right place at the right time. Failure of control mechanisms can lead to disease and even death.

Under unchanging environmental conditions, an organism's biochemical systems must be stable. Under changing conditions, the system must be robust, accommodating both neutral and stressful perturbations. Over longer periods of time, processes must have their rates altered, or even be switched on and off. This regulation includes short-term adjustments, for instance in the stages of the cell cycle; or responses to external stimuli such as changes in the composition or levels of nutrients or of oxygen. Longer-term regulatory activities include control over developmental stages during the entire lifetime of an organism.

Metabolism is the flow of molecules and energy through pathways of chemical reactions. Of course many metabolic reactions involve proteins and nucleic

acids as well as small compounds such as amino acids and sugars. The full panoply of metabolic reactions forms complex traffic patterns. Some patterns are linear pathways, such as the multistep synthesis of tryptophan from chorismate. Others form closed loops, such as the tricarboxylic acid (Krebs) cycle. Moreover, the pathways interlock densely. The structure of the totality of metabolic pathways—its connectivity or topology—and its activity patterns, can be analysed in terms of a mathematical apparatus dealing with graphs and flows and throughputs.

To control metabolic flow patterns, **regulatory pathways** connect proteins and metabolite concentrations. The structure and dynamics of the regulatory pathways are different from those of the metabolic pathways. Corresponding to the succession of enzymatic transformations in metabolism, a regulatory pathway is an assembly of signalling cascades.

Systems biology describes metabolic and regulatory interactions in terms of **interaction networks.**

Two parallel networks: physical and logical

In cells, the two interaction networks operate in parallel: (1) a **physical network** of protein-protein and protein-nucleic acid complexes, and (2) a **logical network** of control cascades. Metabolic pathways partake of both: many but not all metabolic pathways are mediated by physical protein-protein interactions and regulated by logical interactions.

Examples of purely physical interactions include the assembly of photosynthetic reaction centres, complexes of proteins and cofactors that convert light to chemical energy; and assemblies of collagen in connective tissue. Examples of logical interactions, *not* mediated entirely by direct physical interaction between proteins, include feedback loops in which the increase in concentration of a product of a metabolic pathway inhibits an enzyme catalysing one of the early steps in the pathway, or the secretion of a small molecule as a signal to other cells ('fire and forget' mode). In these cases the logical interaction is transmitted by a diffusing small molecule. Many other examples appear in the very common theme of regulation of gene expression. A transcription factor, binding to DNA, may never interact physically with the proteins the expression of which it controls.

See Box: Cell-cell communication in micro-organisms: Quorum sensing

The allosteric change in haemoglobin is an example of simultaneous physical and logical interaction: The subunits of haemoglobin respond to changes in oxygen levels by a conformational change that alters oxygen affinity. Another example is the transmission of a signal from the surface of a cell across the membrane to the interior by dimerization of a receptor. This can be the initial trigger of a cascade that ultimately affects gene expression. Not all links of this process need involve protein-protein interactions; some may be mediated by diffusion of small molecules such as cyclic AMP.

Even though certain protein-protein and protein-nucleic acid complexes participate in both physical and logical networks, the two networks remain distinct in terms of their logic and their biological function, and it is useful to keep the distinction in mind, *especially* when considering proteins that participate in both.

Cell-cell communication in micro-organisms: Quorum sensing

Control mechanisms *not* involving direct protein-protein interactions mediate intercellular signalling in micro-organisms. *Vibrio fischeri* is a marine bacterium that can adopt alternative physiological states in which bioluminescence is active or inactive. (Literally a 'light switch'.) The organism can live free in seawater, or colonize the light organs of certain species of fish or squid. It is bioluminescent only when growing within the animal.

The bacteria respond to the local density of cells; a form of communication called **quorum sensing.** In *V. fischeri*, quorum sensing is mediated by secretion and detection of a small signalling molecule, N-(3-oxohexanoyl)-homoserine lactone. Related species use other N-acyl homoserine lactones, abbreviated to AHL. AHL can diffuse freely out of the cells in which it is synthesized. Within the light organs, culture densities can reach 10^{10}–10^{11} cells/ml, and the AHL concentration can exceed the threshold of about 5–10 nM for triggering the physiological switch.

Bacterial genes *LuxI* and *LuxR* govern the regulation. The product of *LuxI* is involved in the synthesis of AHL. The *LuxR* gene product contains a membrane-bound domain, that detects the AHL signal; and a transcriptional activator domain. *LuxR* activates an operon that includes (1) genes for synthesis of luciferase (the enzyme responsible for the bioluminescence), and (2) *LuxI*, expression of which synthesizes additional AHL, amplifying the signal and sharpening the transition.

The host also senses the bacteria: the light organs of squid grown in sterile salt water do not develop properly. This appears to be a reaction to the luminescence, rather than to the AHL. For the animal, the luminescence contributes to camouflage: disguise from predators living at lower depths, by blending with illumination from the sky. The masking of shadows is a natural form of 'make-up'. (The bioluminescence also regularly surprises diners in seafood restaurants, who jump to the conclusion that their glowing dinner is of extraterrestrial origin. However, most bioluminescent bacteria are harmless, although some strains of the related *Vibrio* species *V. cholerae*, the causative agent of cholera, are weakly bioluminescent. In fact the virulence of *V. cholerae* is also under the control of quorum sensing, by a related mechanism.)

Networks and graphs

In the abstract, networks have the form of graphs.

The routes between cities on the map of Sweden in Fig. 4.3 (page 177) is a network represented by a graph, similar to those appearing in systems biology. Each city is a node. The thick lines joining them indicate roads. Other examples

See Box: The idea of a graph (page 315).

familiar to many readers are the map of the London Underground,‡ and maps of the subway systems of other cities. Each station is a node of the graph, and edges correspond to tracks connecting the stations. The modern London Underground map shows the *topology* of the network; it does not quantitatively represent the geography of the area. An early map, from 1925, did maintain geographic accuracy.§ This was possible when the system was simpler than it is now. Some of the maps now posted in the Paris Métro are fairly accurate geographically. *Considered as networks, a geographically accurate map and a simplified map with the same topology correspond to the same graph.*

The London Underground network is fully connected, in that there is a path between any two stations. Many questions familiar to commuters are shared in the analysis of biological networks; for example: What are the paths connecting station A and station B? Regarding different lines as subnetworks, how easy is it to transfer from one to another; that is, what is the nature of the patterns of connectivity? In case of failure of one or more links, is the network still robust (that is, does it remain fully connected?).

Biological systems need to be robust, both for survival of individuals under stress and for the plasticity required for evolution. In yeast, for example, single gene knockouts of over 80% of the ~6200 open reading frames are survivable injuries.

In principle, networks can achieve robustness through redundancy. The most direct mechanism is simple **substitutional redundancy**: if two proteins are each capable of doing a job, knock out one and the other takes over. In the London Underground this would correspond to a second line running over the same route. For instance, when the Circle Line is not running, passengers travelling between Paddington and King's Cross stations can travel by the Hammersmith & City line running on the same tracks.

In cells, some genes have closely-related homologues resulting from gene duplication, and some of these contribute to substitutional redundancy. For example, in investigating mouse models for diabetes it appeared that mice and rats (but not humans) have two similar but non-allelic insulin genes. However, substitutional redundancy requires equivalence not only of function but of control of expression. In the mouse, knocking out either insulin gene leads to compensatory increased expression of the other and normal phenotype. Equivalent expression patterns are more probable among duplicated genes than among unrelated ones. For example, *E. coli* contains two fructose-1,6-bisphosphate aldolases. One, expressed only in the presence of special nutrients, is nonessential under normal growth conditions. However, the other is essential. In this case functional redundancy does *not* provide robustness. These two enzymes are probably homologous, but they are distant relatives, not the product of a recent gene duplication. One is a member of a family of fructose-1,6-bisphosphate aldolases typical of bacteria and eukaryotes, and the other is a member of another family that occurs in archaea. *E. coli* is unusual in containing both.

‡ http://www.transportforlondon.gov.uk/tfl/tube_map.shtml or http://www.afn.org/~alplatt/tube.html Exercises 6.11 and 6.12, Problem 6.5 and Weblem 6.4 also make use of this map.

§ http://www.ltmuseum.co.uk/collections/posters_b.html

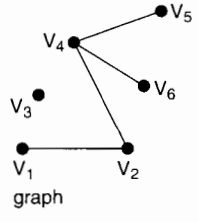
An alternative mechanism of network robustness is **distributed redundancy:** the same effect achieved through different routes. In normal *E. coli* approximately two-thirds of the NADPH produced in metabolism arises via the pentose phosphate shunt, which requires the enzyme glucose-6-phosphate dehydrogenase. Knocking out the gene for this enzyme leads to metabolic shifts, after which increased NADH produced by the tricarboxylic acid cycle is converted to NADPH by a transhydrogenase reaction. The growth rate of the knockout strain is comparable to that of the parent.

The idea of a graph

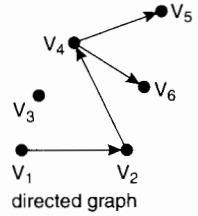
- Mathematically, a graph consists of a set of vertices V and a set of edges E.
- Each edge is specified by a pair of vertices.
- In a **directed graph** the edges are **ordered** pairs of vertices.
- In a **labelled graph** there is a value associated with each edge. (A directed graph is a special case of a labelled graph: consider the arrowheads as labels.)

An undirected unlabelled graph specifies the connectivity of a network but not the distances between vertices (the topology but not the geometry, as in the modern London Underground map). Labels on the edges can indicate distances. For example, some phylogenetic trees indicate only the topology of the ancestry. Others indicate quantitatively the amount of divergence between species. Phylogenetic trees are often drawn with the lengths of the branches indicating the time since the last common ancestor. This is a pictorial device for labelling the edges.

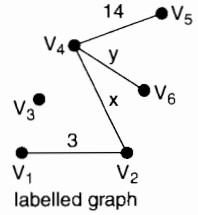
Some graphs do not correspond to physical structures, and in any event edge labels need not reflect geometry in the usual sense. For example, the links in a network of metabolic pathways might be labelled to reflect flow patterns.



graph



directed graph



labelled graph

Examples of graphs

- Sets of people who have met each other
- Electricity distribution systems
- Phylogenetic trees
- Metabolic pathways
- Chemical bonding patterns in molecules
- Citation patterns in the scientific literature
- The World Wide Web

A fundamental property of a network is its **connectivity.**

If V_A and V_Z are vertices in a graph, a **path** from V_A to V_Z is a series of vertices: V_A, V_B, V_C, ... V_Z, such that an edge in the graph connects each successive pair of vertices. The number of vertices in the chain is called the **length** of the path. A **cycle** is a path of length >2 in a nondirected graph for which the initial and final endpoints are the same, but in which no intermediate link is repeated.

A graph that contains a path between any two vertices is called **connected.** Alternatively, a graph may split into several connected components, called **cliques.** The graph in the margin on page 315 contains two cliques, one containing five vertices and one containing only one vertex. (In the extreme, a graph could contain many vertices but no edges at all.) It is often useful to determine the *shortest path* between any two nodes, and to characterize a network by the distribution of path lengths. The well-known assertion that any pair of people are connected by 'six degrees of separation' means that: if the people in the world are vertices of a graph and the graph contains an edge whenever two people know each other, then the graph is fully connected, and there is a path between any two vertices with length ≤6.

A **tree** is a special form of graph. A tree is a connected graph containing only one path between each pair of vertices. A hierarchy is a tree: examples include military chains of command, and Linnean taxonomy. Note that some family trees are not trees in the mathematical sense; examples are plentiful in the royal families of Europe. A tree cannot contain a cycle: if it did, there would be two paths from the initial point (= the final point) to each intermediate point. In the graph in the margin on page 315, the subgraph consisting of vertices V_1, V_2, V_4, V_5, and V_6, is a tree. Adding an edge from V_1 to V_5 would create an alternative path from V_1 to V_5, and the cycle $V_1 \rightarrow V_2 \rightarrow V_4 \rightarrow V_5 \rightarrow V_1$; the graph would no longer be a tree.

The **density of connections**, that is, the mean number of edges per vertex, characterizes the structure of a graph. A fully-connected graph of N vertices has $N - 1$ connections per vertex; a graph with no edges has 0. Nervous systems of higher animals achieve their power not only by containing large numbers of neurons but also by high connectivities.

In some systems there are limits on numbers of connections: For many human societies, in the graph in which individuals are the vertices, and edges link people married to each other, each node has connectivity 0 or 1. In hydrocarbons, the graphs in which carbon and hydrogen atoms are the vertices and edges link atoms bonded to each other, each node has ≤4 connections. In other networks, connectivities follow observable regularities. (See Box: 'Small-world' networks.) For instance, the World Wide Web can be considered as a directed graph. Individual documents are the nodes, and hyperlinks are the edges. It is observed that the distribution of incoming and outgoing links follow power laws: $P(k)$ = probability of k edges is proportional to k^{-q}, where $q = 2.1$ for incoming links, and $q = 2.45$ for outgoing links.

'Small-world' networks

Many observed networks, including biological networks, the World Wide Web, and electric power distribution grids, have the characteristics of high clustering and short path lengths. They include relatively few nodes with very large numbers of connections, called 'hubs', and many that contain few connections. These combine to produce short path lengths between all nodes. From this feature they are called 'small-world' networks. Such networks tend to be fairly robust—staying connected after failure of random nodes. Failure of a hub would be disastrous but is unlikely, because there are few hubs.

Many networks, notably the World Wide Web, are continuously adding nodes. The connectivity distribution tends to remain fairly constant as the network grows. These are called **'scale-free' networks**.

The density of connections is very important in defining the properties of a network. For instance, the interactions that spread disease among humans and/or animals form a network. Whether a disease will cause an epidemic depends not only on the ease of transmission in any particular interaction, but on the density of connections. As the density of connections—the rate of interactions—increases, the system can exhibit a *qualitative* change in behaviour, analogous to a phase change in physical chemistry, from a situation in which the disease remains under control to an epidemic spreading through an entire population. The classic approach of 'quarantine'—isolating people for forty days—works by cutting down the degree of connectivity of the disease-transmission network. Note that a carrier who shows no symptoms—'Typhoid Mary[¶]' was a classic case—serves as a hub of the disease transmission network.

Two historical epidemics associated with wars demonstrate the distinction between topology and geometry in network connectivity. (1) In the early years of the Peloponnesian War, Athens suffered a severe epidemic. (Despite Thucydides' detailed description of the symptoms, the disease has not been definitively identified, but was probably bubonic plague.) A factor contributing to its transmission was the crowding of people into the city from the more vulnerable surrounding countryside. (2) After the First World War, an epidemic of influenza killed an estimated 20 million people, more than died in the war itself. Long-distance travel by soldiers returning from the war helped spread the disease. Any epidemic needs an infectious agent, and a high density of routes of transmission. The controlling factor is the density of the *connections* and not the density of the people.

A change in behaviour analogous to the transition to an epidemic appears in nuclear fission. In a sample of Uranium-235, decaying nuclei produce neutrons that can trigger fission of other nuclei. If the sample is small, so many secondary neutrons are lost through the surface that the sample remains stable. Above a critical mass, enough neutrons are captured within the sample to create a chain

[¶] Mary Mallon (1869–1938) presented the following unfortunate combination of features: (1) she was infected with typhoid, (2) she did not show symptoms, and (3) she worked for many families as a cook.

reaction. If the atoms are vertices of a graph, and the edges are the trajectories of neutrons from one atom to another, the change in behaviour can be seen as the effect of increasing the connectivity of a network. (The background of Michael Frayn's recent popular play, Copenhagen, involves the attempts, before and during the Second World War, to estimate the size of the critical mass, in order to determine whether nuclear weapons would be feasible.)

Network structure and dynamics

An unlabelled, undirected graph gives a *static* structure of the topology of a network. For our molecular interaction networks, this may be an adequate description of many of the physical interactions.

For some networks, such as metabolic pathways or patterns of traffic in cities, the *dynamics* of the system depend on the transmission capacities of the individual links. These capacities can be indicated as labels on the edges of the graph. This allows modelling of patterns of flow through the network. Examples include route planning, in travel or deliveries. Note that the shortest path may well not give optimal throughput. Taxi drivers are exquisitely sensitive—and, in some cities, insensitively voluble—about optimal traffic paths.

In molecular biology, metabolic pathways and signal transduction cascades are networks that lend themselves to pathway and flow analysis. Even optimal sequence alignment by dynamic programming (see Chapter 4) involves determining the optimal path through an edit graph.

Although much is known about the mechanisms of individual elements of control and signalling pathways, understanding their integration is a subject of current research. For instance, the idea that healthy cells and organisms are in stable states is certainly no more than an approximation (and in most cases an idealization). The description of the actual dynamic state of the metabolic and regulatory networks is a very delicate problem. Understanding *how* cells achieve even an apparent approximation to stability is also quite tricky. It is likely that great redundancy of control processes lies at its basis. Regulation is based on the resultant of many individual control mechanisms—here a short feedback loop, there a multistep cascade. Somehow the independent actions of all the individual signals combine to achieve an overall, integrated result. It is like the operation of the 'invisible hand' that, according to Adam Smith, coordinates individual behaviour into the regulation of national economies.

Several types of dynamic states of a network are possible (see Box):

- Equilibrium
- Steady-state
- States that vary periodically
- Unfolding of developmental programs
- Chaotic states
- Runaway or Divergence
- Shutdown

States of a network of processes

- At **equilibrium** one or more forward and reverse processes occur at compensating rates, to leave the amounts of different substances unchanging:

$$A \rightleftarrows B$$

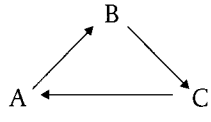
  Chemical equilibria are generally self-adjusting upon changes in conditions, or in concentrations of reactants or products.

- A **steady state** will exist if the total rate of processes that produce a substance is the same as the total rate of processes that consume it. For instance, the two-step conversion:

$$A \rightarrow B \rightarrow C$$

  could maintain the amount of B constant, provided that the rate of production of B (the process A → B) is the same as the rate of its consumption (the process B → C). The net effect would be to convert A to C.

  A cyclic process could maintain a steady state in all its components:



  A steady state in such a cyclic process with all reactions proceeding in one direction is very different from an equilibrium state. Nevertheless, in some cases, it is still true that altering external conditions produces a shift to another, neighbouring, steady state.

- **States that vary periodically** appear in the regulation of the cell cycle, e.g. circadian rhythms, and seasonal changes such as annual patterns of breeding in animals and flowering in plants. Circadian and seasonal cycles have their origins in the regular progressions of the day and year, but have evolved a certain degree of internalization.

- Many equilibrium and some steady-state conditions are **stable**, in the sense that concentrations of most metabolites are changing slowly if at all, and the system is robust to small changes in external conditions. The alternative is a **chaotic state**, in which small changes in conditions can cause very large responses. Weather is a chaotic system: the meteorologist Lorenz asked, 'Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?' In a carefully regulated system, chaos is usually well worth avoiding, and it is likely that life has evolved to damp down the responses to the kinds of fluctuations that might give rise to it. Chaotic dynamics does sometimes produce the approximations to stable states—these are called **strange attractors**. Understanding stability in dynamical systems subject to changing environmental stimuli is an important topic, but beyond the scope of this book.

→

States of a network of processes (*continued*)

- **Unfolding of developmental programs** occurs over the course of the lifetime of the cell or organism. Many developmental events are relatively independent of external conditions, and are controlled primarily by regulation of gene expression patterns.
- **Runaway** or **Divergence**. Breakdown in control over cellular proliferation leads to unconstrained growth, in cancer.
- **Shutdown** is part of the picture. Apoptosis is the programmed death of a cell, as part of normal developmental processes, or in response to damage that could threaten the organism, such as DNA strand breaks. Breakdown of mechanisms of apoptosis—for instance, mutations in protein p53—is an important cause of cancer.

Protein complexes and aggregates

The basis of our understanding of how life within a cell is organized and regulated is the set of protein-protein and protein-nucleic acid interactions. The development of high-throughput methods for detecting interactions has been a focus of recent interest.

Interacting proteins and nucleic acids span a range of structures and functions:

- Simple dimers or oligomers in which the monomers appear to function independently.
- Oligomers with functional 'cross-talk', including ligand-induced dimerization of receptors, and allosteric proteins such as haemoglobin, phosphofructokinase and asparate carbamoyltransferase.
- Large fibrous proteins such as actin or keratin.
- Non-fibrous structural aggregates such as viral capsids.
- Large aggregates with dynamic properties such as F1-ATPase, pyruvate kinase, the GroEL-GroES chaperonin, and the proteasome.
- Protein-nucleic acid complexes, including ribosomes, nucleosomes, transcription regulation complexes, splicing and repair particles, and viruses. In many cases initial binding is followed by recruitment of additional proteins to form large complexes.
- Many proteins, whether monomeric or oligomeric, function by interacting with other proteins. These include all enzymes with protein substrates, and many antibodies, inhibitors, and regulatory proteins.
- Protein interactions are frequently associated with disease, as misfolded or mutant proteins are prone to aggregation. (See Box: Diseases associated with protein aggregates.)

Diseases associated with protein aggregates

Disease	Aggregating protein	Comment
Sickle-cell anaemia	Deoxyhaemoglobin–S	Mutation creates hydrophobic patch on surface
Classical amyloidoses	Immunoglobulin light chains, transthyretin, and many others	Extracellular fibrillar deposits
Emphysema associated with Z-antitrypsin	Mutant α_1-antitrypsin	Destabilization of structure facilitates aggregation
Huntington	Altered huntingtin	One of several polyglutamine repeat diseases
Parkinson	α-synuclein	Found in Lewy bodies
Alzheimer	Aβ,τ	Aβ = 40–42 residue fragment
Spongiform encephalopathies	Prion proteins	Infectious, despite containing no nucleic acid

Properties of protein-protein complexes

Stoichiometry—what is the composition of the complex?

Stable oligomeric proteins may contain many copies of one protein, or combine different ones. Among aggregates of a single protein, complexes containing *odd* numbers of molecules are less common than those containing even numbers. Oligomers (complexes containing a few copies of the same protein—dimers, trimers, . . .) usually show symmetry. For instance, insulin is a hexamer with three-fold and two-fold axes.

Some prokaryotic proteins containing identical subunits are homologous to eukaryotic proteins containing related but nonidentical subunits, arising by gene duplication and divergence. The proteasome is an example. Some viruses achieve diversity *without* duplication, by combining proteins with the same sequence but different conformations.

Protein complexes vary widely in the numbers and variety of molecules they contain. Some complexes contain only a few proteins, but others are very large: for example, pyruvate dehydrogenase contains hundreds of subunits, and some viral capsids contain thousands.

Many very large aggregrates have clinical importance, including Bovine spongiform encephalopathy (BSE, or 'mad cow disease'), Alzheimer and Huntington disease. Amyloidoses are diseases characterized by extracellular fibrillar deposits, usually with a common crossed-β-sheet structure. They arise from a variety of causes, including destabilizing mutations, overproduction of a protein, and inadequate clearance in renal failure. Misfolded proteins are more prone to aggregate, and mutated proteins are more prone to misfold. Large local concentrations, such as can occur in myelomas that overproduce immunoglobulin light chains, also heighten the threat of aggregation.

See Box, Diseases associated with protein aggregation

Affinity—how stable is the complex?

A common index of the affinity of a complex is the **dissociation constant,** K_D, the equilibrium constant for the *reverse* of the binding reaction:

$$\text{Protein} \cdot \text{Ligand} = \text{Protein} + \text{Ligand} \qquad K_D = \frac{[p][l]}{[pl]}$$

The Michaelis constant of an enzyme is the dissociation constant of the Enzyme-Substrate complex.

[P], [L] and [PL] denote the numerical values of the concentrations of Protein, Ligand, and Protein-Ligand complex, respectively, expressed in mol ℓ^{-1}. The lower the K_D, the tighter the binding. K_D corresponds to the concentration of free ligand at which half the proteins bind ligand and half are free: [P] = [PL].

The K_D is related to the Gibbs free energy change of dissociation by the relationship:

$$\text{PL} = \text{P} + \text{L} \qquad \Delta G^{\ominus} = \Delta H^{\ominus} - T\Delta S^{\ominus} = -RT \ln K_D$$

Dissociation constants of protein-ligand complexes span a wide range:

Biological context	Ligand	Typical K_D
Allosteric activator	Monovalent ion	$10^{-4} - 10^{-2}$
Coenzyme binding	NAD, for instance	$10^{-7} - 10^{-4}$
Antigen-antibody complexes	Various	$10^{-4} - 10^{-11}$
Thrombin inhibitor	Hirudin	5×10^{-14}
Trypsin inhibitor	Bovine pancreatic trypsin inhibitor	10^{-14}
Streptavidin	Biotin	10^{-15}

Structural studies have elucidated several important features of the interactions between soluble proteins, that contribute to affinity:

- **What holds the proteins together?** Burial of hydrophobic surface, hydrogen bonds and salt bridges.
- **Do proteins change conformation upon formation of complexes?** In some cases they do. In these cases the interaction energy has to 'pay for' the conformational change, and the interface tends to be larger.
- **What determines specificity?** Complementarity of the occluding surfaces—in shape, hydrogen-bonding potential, and charge distribution. (See Box: Features of protein-protein interfaces.) Prediction of protein complexes from the structures of the partners is the **docking problem.** Reliable solution of this problem, together with progress in structural genomics, would permit *in silicio* screening of proteomes for interacting partners.

Kinetics of formation and breakup, average lifetime

The dissociation constant of a complex indicates the fraction of time that the components spend in the bound state and the fraction of time in which they are

unbound. But the **average lifetime** of the bound state can vary without affecting K_D. Defining individual rate constants for association and dissociation, k_{on} and k_{off}, the dissociation constant is equal to their ratio:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} [PL] \quad K_D = k_{off}/k_{on}$$

A short average lifetime, corresponding to large values of both k_{off} and k_{on}; or a long average lifetime, corresponding to small values of both k_{off} and k_{on}, can produce the same K_D. Lifetimes are important: if you want to purify a complex, it is important that its average lifetime be longer than the duration of the isolation procedure! Conversely, if a protein-protein complex is to mediate transmission of

Features of protein-protein interfaces

- **Burial of protein surface.** The accessible surface area (ASA) of a protein is calculated by rolling a probe sphere the size of a water molecule (radius 1.4 Å) over the protein surface. The surface buried by formation of a complex is the difference between the ASA of the complex and the sum of the ASAs of the components separately.

  A typical protein-protein interface might involve 22 residues, and 90 atoms of which 20% would be mainchain atoms, and an occasional water molecule. A histogram of surface area buried in binary protein complexes shows a peak centred at 1600 Å^2.

  The minimum buried surface for stability of a protein-protein complex is about 1000 Å^2. Complexes that bury > 2000 Å^2 tend to involve conformational changes upon complex formation.

  Each Å^2 of hydrophobic surface buried contributes about 105 J mol^{-1} to the free energy of stabilization.

- **The composition of the interface.** The chemical character of protein-protein interfaces is intermediate between that of the surfaces and interiors of monomeric globular proteins. Interfaces are enriched in neutral polar atoms at the expense of charged atoms. The amino acid compositions of interfaces are enriched in aromatic residues—His, Phe, Tyr, Trp—relative to remaining exposed surface. There is a lesser degree of enrichment in aliphatic sidechains—Leu, Ile, Val, Met—and Arg (but surprisingly, not Lys).

- **Complementarity** of interfaces is responsible for specificity. Complementarity involves both good packing at the occluding surfaces and proper juxtaposition of hydrogen-bonded and charged atoms. Typically there is one hydrogen bond per 170 Å^2 of interface area. Isolated water molecules occupy sites in many interfaces. Typically there is one fixed water molecule per 100 Å^2 of interface.

a signal, a short lifetime provides a natural 'reset mechanism' to preclude the signal's being locked 'on' for too long.

The 'on rate' is limited by diffusion rates. Under ordinary conditions $k_{on} \leq 10^{-9}$ M s^{-1}. If a conformational change is required for binding k_{on} may be considerably smaller. Typical k_{on} rates are 10^{-6}–10^{-7} M^{-1} s^{-1}, and typical lifetimes ~1 s.

How are complexes organized in three dimensions?

When two proteins form a complex, each leaves a 'footprint' on the surface of the other, defining the portion of the surface involved in the interaction. If two proteins interact using the *same* surface on both, the complex is **closed.** If two proteins interact through *different* surfaces, the complex is **open.** The significance is that a closed complex does not allow additional proteins to bind with the same interaction. An open complex, in which the surface of potential interaction is not occluded, can grow by accretion of additional subunits. Thus, open but not closed complexes are compatible with formation of aggregates by replication of the interaction.

Do proteins change conformation upon complex formation?

Some protein complexes form by the coming together of rigid subunits. The subunits in these complexes have the same structure in the complex that they have separately. Other protein complexes involve structural changes upon complex formation. These include complexes of subunits that are not independently stable separately.

Protein interaction networks

The units from which interaction networks are assembled are:

- For physical networks, a protein-protein or protein-nucleic acid complex.
- For logical networks, a dynamic connection in which the activity of a process is affected by a change in external conditions, or by the activity of another process.

Most experiments reveal only pairwise interactions. The challenges are to integrate pairwise interactions into a network and then to study the structure and dynamics of the system.

Many techniques detect physical interactions directly. These include:

- **X-ray and NMR structure determinations** cannot only identify the components of the complex, but reveal how they interact, and whether conformational changes occur upon binding.
- **Two-hybrid screening systems.** Transcriptional activators such as Gal4 contain a DNA-binding domain and an activation domain. Suppose these two domains are separated, and one test protein is fused to the DNA-binding domain and a

second test protein is fused to the activation domain. Then a reporter protein will be expressed only if the components of the activator are brought together by formation of a complex between two test proteins. High-throughput methods allow parallel screening of a 'bait' protein for interaction with a large number of potential 'prey' proteins. (See Box: Protein interaction networks determined by two-hybrid screening systems.)

Protein interactions detected by two-hybrid screening systems*

	H. pylori	*S. cerevisiae*	*C. elegans*	*D. melanogaster*
Total proteome size	1 576	5 585	33 469	13 843
Proteins tested	732	987 / 790	1 415	4 685
Interactions detected	1 465	936 / 800	2 131	4 876

The two sets of numbers for yeast are the results of independent investigations.

* From: Aloy, P. & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, **22**, 1317–1321.

- **Chemical crosslinking** fixes complexes so that they can be isolated. Subsequent proteolytic digestion and mass spectrometry permits identification of the components.
- **Coimmunoprecipitation.** An antibody raised to a 'bait' protein binds the bait together with any other 'prey' proteins that interact with it. The interacting proteins can be purified and analysed, for instance by Western blotting, or mass spectrometry.
- **Chromatin immunoprecipitation** identifies DNA sequences that bind proteins. Treatment with formaldehyde crosslinks proteins and DNA, fixing the complexes that exist within a cell. Then, isolation of the chromatin and breaking the DNA into small fragments allows separation of proteins by binding to specific antibodies, carrying the DNA sequences along with them. Reversal of the crosslink followed by sequencing of the DNA identifies the specific DNA sequence to which each protein binds.
- **Phage display.** Genes for a large number of proteins are individually fused to the gene for a phage coat protein, to create a population of phages each of which carries copies of one of the extra proteins exposed on its surface. Affinity purification against an immobilized 'bait' protein selects phages displaying potential 'prey' proteins. DNA extracted from the interacting phages reveals the amino acid sequences of these proteins.
- **Surface plasmon resonance** analyses the reflection of light from a gold surface to which a protein has been attached. The signal changes if a ligand binds to the immobilized protein. (The method detects localized changes in the

refractive index of the medium adjacent to the gold surface. This is related to the mass being immobilized.)

- **Fluorescence Resonance Energy Transfer.** If two proteins are tagged by different chromophores, transfer of excitation energy can be observed over distances up to about 60 Å.

Other methods provide complementary information:

- **Domain recombination networks.** Many eukaryotic proteins contain multiple domains. A feature of eukaryotic evolution is that a domain may appear in different proteins with different partners. In some cases proteins in a bacterial operon catalysing successive steps in a metabolic pathway are fused into a single multidomain protein in eukaryotes. The domains of the eukaryotic protein are individually homologous to the separate bacterial proteins. (Examples of proteins fused in prokaryotes and separate in eukaryotes are also known.)

  It is possible to create a network by defining an interaction between two protein domains whenever homologues of the two domains appear in the same protein. This is evidence for some functional link between the domains, even in species where the domains appear in separate proteins.

- **Coexpression patterns.** Clustering of microarray data identifies proteins with common expression patterns. They may have the same tissue distribution, or be up- or down-regulated in parallel in different physiological states. This is also suggestive evidence that they share some functional link. In the response of *M. tuberculosis* to isoniazid (page 308), genes for the Fatty Acid Synthesis complex are coordinately up-regulated. They are on an operon-like gene cluster, and in fact these proteins do form a physical complex. On the other hand, alkyl hydroperoxidase (AHPC) is also up-regulated in response to isoniazid. AHPC acts to relieve oxidative stress. There is no evidence that it physically interacts with the Fatty Acid Synthesis complex, or that it mediates a metabolic transformation coupled to fatty acid synthesis. It is a second component of the response to isoniazid.

- **Phylogenetic distribution patterns.** The **phylogenetic profile** of a protein is the set of organisms in which it and its homologues appear. Proteins in a common structural complex or pathway are functionally linked and expected to coevolve. Therefore proteins that share a phylogenetic profile are likely to have a functional link, or at least to have a common subcellular origin. There need be no sequence or structural similarity between the proteins that share a phylogenetic distribution pattern. A welcome feature of this method is that it derives information about the function of a protein from its relationship to *nonhomologous* proteins.

There are many ways to link proteins, including direct physical protein-protein interactions, two-hybrid complementarity, domain recombination, coexpression patterns, and phylogenetic profiles. Each provides a basis for a protein interaction network. The networks formed by combining each set of interactions are different, although they overlap, to a greater or lesser extent. They give different views

of the kinds of relationships between proteins that exist in cells. It is possible to form a more comprehensive network by combining different types of interactions. For instance, the DIP database http://dip.doe-mbi.ucla.edu/ is a curated collection of experimentally-determined protein-protein interactions. It contains data about 44 349 interactions between 17 048 proteins from 107 organisms.

Plate X shows a portion of an interaction network of yeast proteins, based on sets of proteins that have been found together in solved structures.

Web resources: Interaction databases

Intact: An open source molecular interaction database
http://www.ebi.ac.uk/intact/

DIP: Database of Interacting Proteins
http://dip.doe-mbi.ucla.edu/

MIPS Comprehensive Yeast genome database
http://mips.gsf.de/

BIND: Biomolecular Interaction Network Database
http://www.bind.ca/

MINT: A Molecular Interactions database
http://cbm.bio.uniroma2.it/mint/

SPiD: *Subtilis* Protein interaction Database
http://genome.jouy.inra.fr/cgi-bin/spid/index.cgi

GRID: General Repository for Interaction Datasets
http://biodata.mshri.on.ca/grid/servlet/Index

PathCalling: Protein-protein interactions in *S. cerevisiae*
http://portal.curagen.com/extpc/com.curagen.portal.servlet.Yeast

HPID: The Human Protein Interaction Database
http://www.hpid.org/

Case Study 6.3: Components of the primosome assembly in *Bacillus subtilis*

The first step in DNA replication in *Bacillus subtilis* is the binding of initiator proteins to specific DNA sequences that serve as origins of replication. These then recruit a nucleoprotein complex called the primosome. A major component of the primosome is DnaC, a hexameric replicative helicase.‖

It is believed that steps in the process include:

1. Binding of an initiator protein, DnaA or PriA, to an appropriate single-stranded DNA sequence.

‖ Be aware that the nomenclature of these proteins differs between *E. coli* and *B. subtilis*.

→

Case Study 6.3 (*continued*)

2. Other proteins—DnaB, DnaC and DnaI—are recruited. DnaB and DnaI are regulators of DnaC activity.
3. DnaC is loaded onto the single-stranded DNA, forming a hexameric assembly.
4. DnaG is recruited to prime DNA synthesis.

Scientists at the *Institut National de la Récherche Agronomique* maintain a database of the protein interaction network of *B. subtilis*.[**] (See http://genome.jouy.inra.fr/cgi-bin/spid/index.cgi)

Figure 6.6 shows a small fragment of the network, limited to immediate neighbours of DnaC.

The web site is active: Clicking on a node either *adds* the interaction partners of the node to the graph, or *replaces* the graph with another centred on the selected protein. By adding partners, one can look at more extended neighbourhoods of DnaC. By replacing the graph, one can walk through the network. (See Weblem 6.5.)

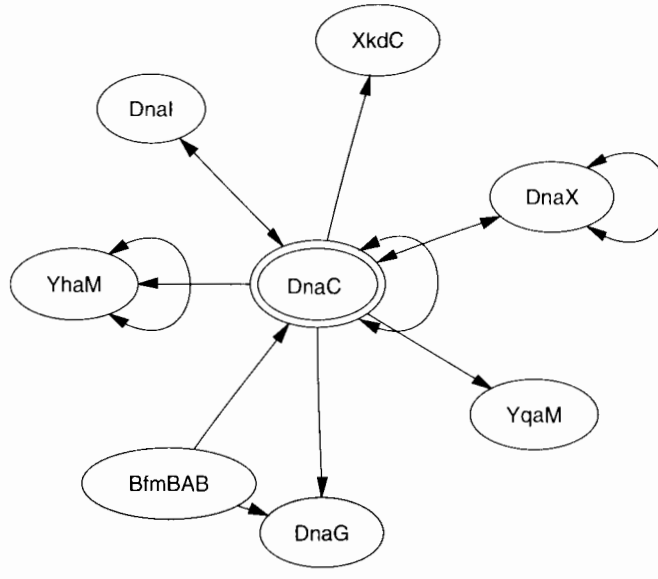


Fig. 6.6 DnaC and proteins that interact directly with it. Arrows linking partners point from 'bait' to 'prey'; bidirectional arrows indicate cases where the interaction was detected in reciprocal experiments. In the original web site, the arrows are colour coded according to the nature of the evidence for the interaction. Reproduced by permission.

[**] Hoebeke, M., Chiapello, H., Noirot, P. & Bessiéres, P. (2001), SPiD: a subtilis protein interaction database, *Bioinformatics*, **17**, 1209–1212; Noirot-Gros, M. F., Dervyn, E. Wu, L. J., Mervelet, P., Errington, J., Erlich, S. D. & Noirot, P. (2002), An expanded view of bacterial DNA replication, *Proc. Natl. Acad. Sci. USA*, **99**, 8342–8347.

Regulatory networks

Individual control interactions are organized into linear signal transduction cascades, and reticulated into control networks. Regulatory networks pervade living processes.

Any regulatory action requires (1) a stimulus, (2) transmission of a signal to a target, (3) a response, and (4) a 'reset' mechanism to restore the resting state (see Fig. 6.7). Many regulatory actions are mediated by protein-protein complexes. Transient complexes are common in regulation, as dissociation provides a natural reset mechanism.

Some stimuli arise from genetic programs. Some regulatory events are responses to current internal metabolite concentrations. Others originate outside the cell; the signal is detected by surface receptors, and transmitted across the membrane to an intracellular target.

Two components of regulatory networks are (1) the **signal transduction network**, and (2) the **transcriptional control network**. The signal transduction network exerts control '*in the field*', by a variety of mechanisms such as: inhibitors; dimerization, ligand-induced conformational changes including but not limited to allosteric effects; GDP-GTP exchange or kinase-phosphorylase switches; and differential turnover rates. This component acts fast, on sub-second timescales. The transcriptional regulatory network exerts control '*at headquarters*', through control over gene expression. This component is slower, acting on a timescale of minutes.

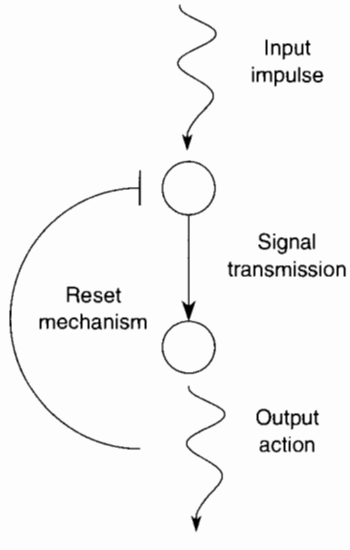


Fig. 6.7 The elementary step in a regulatory network. An input impulse is received by a node, which transmits a signal to a downstream node, causing an output action. This is followed by reset of the upstream node to its inactive state. Combination of such elementary diagrams gives rise to the complex regulatory networks in biology.

General characteristics of all control pathways include:

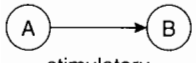
- a single signal can trigger a single response or many responses
- a single response can be controlled by a single signal or influenced by many signals
- each response may be stimulatory—increasing an activity—or inhibitory—decreasing an activity
- transmission of signals may damp out stimuli or amplify them

There are ample opportunities for complexity, and cells have taken extensive advantage of these.

Structures of regulatory networks

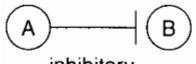
Think of control, or regulatory, networks as assemblies of *activities*. Although mediated in part by physical assemblies of macromolecules—protein-protein and protein-nucleic acid complexes—regulatory networks:

(1) **Tend to be unidirectional**. A transcription activator may stimulate the expression of a metabolic enzyme, but the enzyme may not be involved directly in regulating the expression of the transcription factor. (See page 82 for a discussion of control of the tryptophan synthase pathway in *E. coli*.)



stimulatory interaction

(2) **Have a logical component**. It is not enough to describe the connectivity of a regulatory network. Any regulatory action may stimulate or repress the activity of its target. If two interactions combine to activate a target, activation may require *both* stimuli (logical AND), or *either* stimulus may suffice (logical OR).

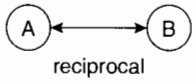


inhibitory interaction

(3) **Produce dynamic patterns**. Signals may produce combinations of effects with specified time courses. Cell-cycle regulation is a classic example.

autoregulatory interaction

A regulatory network can be described by a graph, in which edges indicate steps in pathways of control. Regulatory networks are directed graphs (see page 315): the influence of vertex A on vertex B is expressed by a directed edge connecting A and B. An edge directed from vertex A to vertex B is called an **outgoing connection** from A and an **incoming connection** to B. Often, an arrow indicates a stimulatory interaction, and a T-symbol indicates an inhibitory interaction. An edge connecting



reciprocal interaction

Case Study 6.4: Architecture and dynamics of the genetic regulatory network of *Saccharomyces cerevisiae*[††]

A recent study of transcription regulation in yeast treated a network containing 3562 genes, corresponding to approximately half the known proteome of *S. cerevisiae*. The genes included 142 that encode transcription regulators, and 3420 that encode target genes exclusive of transcription regulators. There are

[††] Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. & Gerstein, M. B. (2004), Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature*, **431**, 308–312.

→

7074 known regulatory interactions among these genes, including effects of regulators on one another, and of regulators on non-regulatory targets.

Analysis of the overall network architecture reveals that:

- The distribution of incoming connections to target genes has a mean value of 2.1, and is distributed exponentially. Most target genes receive direct input from about two transcription regulators. The probability that a gene is controlled by k transcription regulators, $k = 1, 2, \ldots$, is proportional to $e^{-\alpha k}$, with $\alpha = 0.8$.
- The distribution of outgoing connections has a mean value of 49.8, and obeys a power law. The probability that a given transcription regulator controls k genes is proportional to $k^{-\beta}$, with $\beta = 0.6$. Power-law behaviour is common in networks, and characterizes topologies in which a few nodes—the 'hubs'—have many connections, and many nodes have few. (See page 317.) In regulatory networks, hubs tend to be fairly far upstream, forming important foci of regulation with far-reaching control.
- The average number of intermediate nodes in a minimal path between a transcription regulator and a target gene is 4.7. The maximum number of intermediate nodes in a path between two nodes is 12.
- The clustering coefficient of a node is a measure of the degree of local connectivity within a network. If all neighbours of a node are connected to one another, the clustering coefficient of the node is 1. If no pair of neighbours of a node is connected to each other, the clustering coefficient of the node is 0. The mean clustering coefficient, averaged over all nodes, is a measure of the overall density of the network. For the yeast transcriptional regulatory network, the mean clustering coefficient is 0.11.

Figure 6.8 is a cartoon-like sketch of a fragment of such a network, indicating rather loosely some of the general features. Nodes are divided into **transcription regulators**, shown as circles, and **target genes**, shown as squares. Target genes are distinguished by having no output connections. There is extensive interregulation among the transcription factors, to a much higher density of interconnections than can intelligibly be shown in this diagram. Think of a seething broth of transcription factors, within the shaded area, sending out signals to target genes. The shaded area indicates only the *logical* clustering of the transcription regulators. There is no suggestion about physical localization; indeed, transcription regulators interact with DNA, and almost never interact physically with the proteins the expression of which they control.

Each transcription regulator directly influences approximately 50 genes on average, although, as with other 'small-world' networks following power-law distributions of connectivities, the distribution is very skewed—some 'hubs' have very many output connections, but most nodes have very few. A few of the interregulatory connections between transcription factors are shown in red. In about 10% of the cases, two neighbours of the same transcription factor

Case Study 6.4 (*continued*)

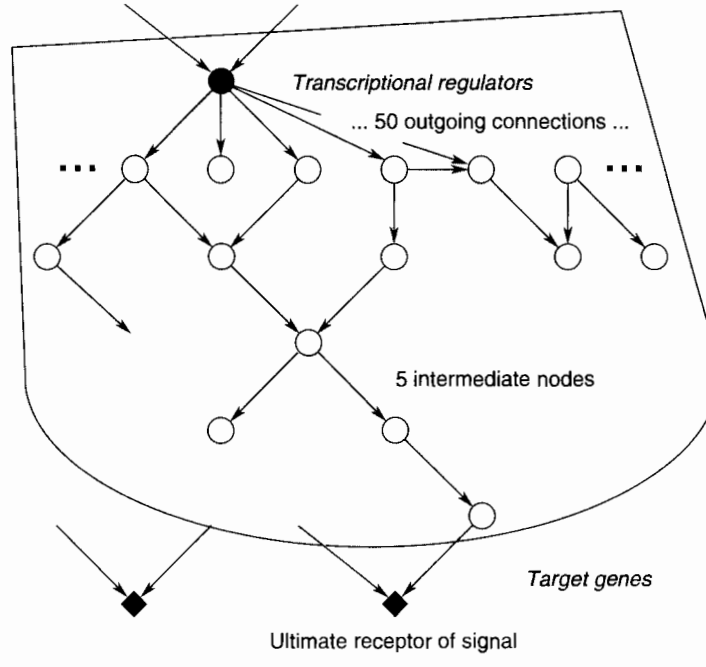


Fig. 6.8 Simplified sketch illustrating some features of an 'average' segment of the pathways in the yeast interaction network. Transcription regulators appear as circles. Target genes appear as squares. A transcription regulator typically has direct influence over about 50 genes, indicated by multiple connections from the filled black circle to the circles on the line below it. Roughly one in ten of the neighbours of any node is connected to another neighbour, indicated by the horizontal arrow on the second row. The ultimate receptor of the signal lies at the end of a pathway typically containing about five intermediate nodes (shown in black). This ultimate target gene receives on average about two inputs. This diagram shows only a small fragment of a network that is in fact quite dense.

interact with each other. A pathway from one regulator (filled black circle) to one ultimate receptor (filled black square), through five intermediate nodes, is shown in black. The intermediate nodes are other transcription regulators, connected both within the path drawn in black, and off this path. Even the transcription factor used as the origin of the path receives input connections. Although it is possible to identify target genes from the absence of outgoing connections, it is more difficult to identify ultimate initiators of signal cascades.

The ultimate receptor is a target gene that receives regulatory input but itself has no output links. This target is expected to receive (on average) a second control input. The black target node receives input via a black arrow, along the selected path, and via a red arrow suggesting the second input. Of course the

second input may arrive via a path that shares common nodes with the black path, including other routes from the filled black circle.

The dense forest of additional pathways, from which this fragment is extracted, is not shown. Some 'back-of-the-envelope' calculations: There are ~3500 nodes, each receiving on average 2 input connections. There are ~140 transcription factors, making an average of 50 output connections. The number of input connections must equal the number of output connections, and indeed $3500 \times 2 = 140 \times 50 = 7000$.

Given the complexity, it is difficult to illustrate larger segments of the network in more detail than the simplified version appearing in Fig. 6.8. However, dissections of yeast and other regulatory networks have defined certain recurrent motifs that serve as building blocks. These might be considered the 'secondary structures' of network architectures. (See Box, page 335: Common motifs in biological control networks.)

The high ratio of interactions to transcription regulators implies that we cannot expect to associate individual regulatory molecules with single, dedicated, activities (as we can, for the most part, with metabolic enzymes). Instead, the activity of the network involves the coordinated activities of many individual regulatory molecules.

The network achieves versatility and responsiveness by reconfiguring its activities. This is seen by comparing the changes in the activities of networks controlling yeast gene expression patterns in different physiological regimes of the organism: cell cycle, sporulation, diauxic shift (the change from anaerobic fermentative metabolism to aerobic respiration as O_2 levels increase), DNA damage, and stress response. Cell cycling and sporulation involve the unfolding of endogenous gene expression programs; the others are responses to environmental changes.

Different states are characterized both by similarities and differences in gene expression patterns, and by the components of the regulatory network that are active. There is considerable shift in expression of *target genes*. About a quarter of the target genes are specialized to individual physiological states. That is, of the total of 3420 target genes, the expression of almost half (1514) do not show major changes in the different states. Of the 1906 that show altered expression levels in different states, almost half (803) are specialized to a single physiological state.

In contrast, different states show much more overlap in the usage of *transcription regulators*. For instance, for cell-cycle control, 280 target genes (8%) are differentially regulated by 70 (49%) of the transcription regulators. Clearly there is a much lower degree of specialization than of the target genes. In general, half the transcription factors are active in at least three out of the five physiological regimes. However, contrasting with the high overlap of usage of the transcription regulators (the nodes), the overlap of the activities within the network (the connections) is relatively low. Different components

Case Study 6.4 (*continued*)

of the interaction network organize the different gene expression patterns in different states.

Whereas different physiological states are characterized by substitutions of different sets of synthesized proteins, the regulatory network uses much of the same structure but reconfigures the pattern of activity. Think of the transcription factors as 'hardware' and the connections as reprogrammable 'software'. The molecules do not change but the interactions do: in different states, many transcription regulators change most, or a substantial part, of their interactions. In particular, the set of transcription regulators that form the hubs of the network—those with many outgoing nodes that form foci of control—are not a constant feature of the system. Some hubs are common to all states, but others step forward to take control in different physiological regimes. The result of the reconfiguration of activity is that over half of the regulatory interactions are *unique* to the different states.

The effect of the changes in the active interaction patterns is to alter the topological characteristics of the network in different states. For instance, under panic conditions—DNA damage and stress—the average number of genes under control of individual transcription regulators increases, the average minimal path length between regulator and target decreases, and the clustering becomes less dense (that is, there is less interregulation among transcription factors). This can be understood in terms of a need for fast and general mobilization—the equivalent of shouting 'Go! Go! Go!' over the radio. Normal circumstances—cell-cycle control for instance—allow for a more dignified and precise regulatory state, which permits finer control over the temporal course of expression patterns. In cell-cycle control and sporulation, there is a much denser interregulation among transcription factors, and longer minimal path lengths between transcription regulators and target genes.

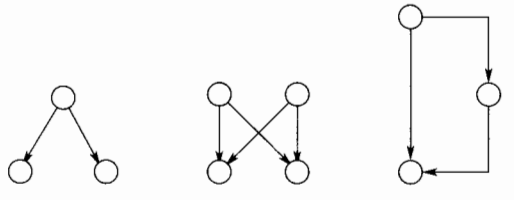
Different physiological states also differ in their usage of the common motifs—fork, scatter, and 'one-two punch'. (See Box: Common motifs in biological control networks). Forks are more used in conditions of stress, diauxic shift, and DNA damage. They are appropriate to the need for quick action. Requirements for buildup of intermediates would delay the response. Conversely, the 'one-two punch' motif is more common in cell-cycle control. This is consistent with the need for a signal from one stage to be stabilized before the cell enters the next stage.

Much of evolution proceeds towards greater specialization. The human eye is a classic example. It is an intricate and fine-tuned structure, (features that were once adduced as evidence *against* Darwin's theory). Many evolutionary pathways show a trade-off between specialized adaptation and generalized adaptability.

Regulatory networks are an exception. Evolution has produced structures that are both specialized *and* versatile. The reconfigurability of regulatory networks allows them to respond robustly to changes in conditions, by creating many different structures, all specialized to the conditions that elicit them.

Common motifs in biological control networks

Within the high complexity of typical regulatory networks, certain common patterns appear frequently. In the architecture of networks, these form building blocks which contribute to higher levels of organization. Shen-Orr, Milo, Mangan and Alon* have described examples including: the *fork*, the *scatter*, and the '*one-two punch*' (a phrase from the boxing ring):



The **fork**, also called the single-input motif, transmits a single incoming signal to two outputs. Successive forks, or forks with higher branching degrees, are an effective way to activate large sets of genes from a single impulse. Generalizations of the binary fork include more downstream genes under common control (more tines to the fork), and autoregulation of the control node. Forks can achieve general mobilization. Moreover, if the regulated genes have different thresholds for activation, the dynamics of building up the signal can produce a temporal pattern of successive initiation of the expression of different genes.

The **scatter** configuration, also called the multiple input motif, can function as a logical OR operation: both downstream targets become active if *either* of the input impulses is active. Generalizations of the square scatter pattern shown may contain different numbers of nodes on both layers. Note that scatter patterns are superpositions of forks.

The '**one-two punch**', also called the 'feed-forward loop', affects the output both directly through the vertical link; and indirectly and subsequently, through the intermediate link. This motif can show interesting temporal behaviour if activation of the target requires simultaneous input from both direct and indirect paths (logical AND). Because buildup of the intermediate requires time, the direct signal will arrive before the indirect one. Therefore a short pulsed input to the complex will not activate the output—by the time the intermediate signal builds up, the direct signal is no longer active. The system can thereby filter out transient stimuli in noisy inputs. Conversely, the active state of the system can shut down quickly upon withdrawal of the external trigger.

A driver's action at a traffic light is an example of this control mechanism: The response to an amber light followed by a green light should be a cautious acceleration. The response to a red light should be an immediate stop. Control over drinking unfortunately *fails* to show this behaviour: the first drink itself up-regulates additional drinking, and there is no quick way to sober up!

* Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002), Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nat. Genet.* **31**, 64–68.

a vertex to itself indicates autoregulation. A double-headed arrow indicates reciprocal stimulation of two nodes; note that this is *not* the same as an undirected edge.

Structural biology of regulatory networks

Any regulatory interaction involves one or more proteins and nucleic acids. Examples of regulatory mechanisms include a protein binding a ligand, undergoing chemical modification such as phosphorylation/dephosphorylation, changing conformation, or all of the above. X-ray crystallography and NMR spectroscopy have elucidated some of the general mechanisms underlying control processes.

Many molecules involved in regulation are multidomain proteins. A domain is a segment of a protein that has independent stability and can appear in conjunction with different partners through evolutionary recombination. Most multidomain proteins contain a linear sequence of domains each of which is relatively free to interact with other molecules. Assembly of a protein from domains therefore permits the joining into one molecule of a set of functions. 'Mixing and matching' of domains gives evolution access to a wide variety of functional combinations. (See Fig. 2.3, page 111.)

One important feature of regulatory proteins is recognition. An **interaction domain** is a part of a protein that confers specificity in ligation of a partner. Regulatory proteins contain a limited number of types of interaction domains, which have diverged to form large families with different individual specificities. For instance, the human genome contains 115 SH2 domains, and 253 SH3 domains. (*S*rc-*H*omology domains SH2 and SH3 are named for their homologies to domains of the src family of cytoplasmic tyrosine kinases.) Many individual interaction domains even interact with different partners as they participate in successive steps of a control cascade. Initial interactions may also trigger recruitment of additional proteins to form large regulatory complexes.

Many interaction domains are sensitive to the state of post-translational modification of their ligands, for instance binding preferentially to states of a ligand in which specific tyrosines, serines, or threonines are phosphorylated. These and other post-translational modifications function as switches, turning on or interrupting/resetting a signalling cascade.

Protein-protein complex formation allows a cell to detect a signal molecule in the external medium, and report its arrival to the cell interior, without the signal molecule itself ever needing to enter the cell. Many receptors use an ingenious dimerization mechanism: The receptor has external, transmembrane, and internal segments. An external ligand binds to *two* molecules of receptor. The juxtaposition of the external portions brings the internal portions together also, because they are tethered to the external regions by the transmembrane segments. Interaction between the interior segments triggers a conformational change that activates a process such as phosphorylation of a protein. This may initiate a signal transduction cascade that can amplify the original stimulus.

Figure 6.9 shows types of interaction domain complexes with ligands, including binding of peptides (which may be attached to proteins), protein-protein complexes, extracellular dimer formation upon binding a hormone, and a protein-nucleic acid complex.

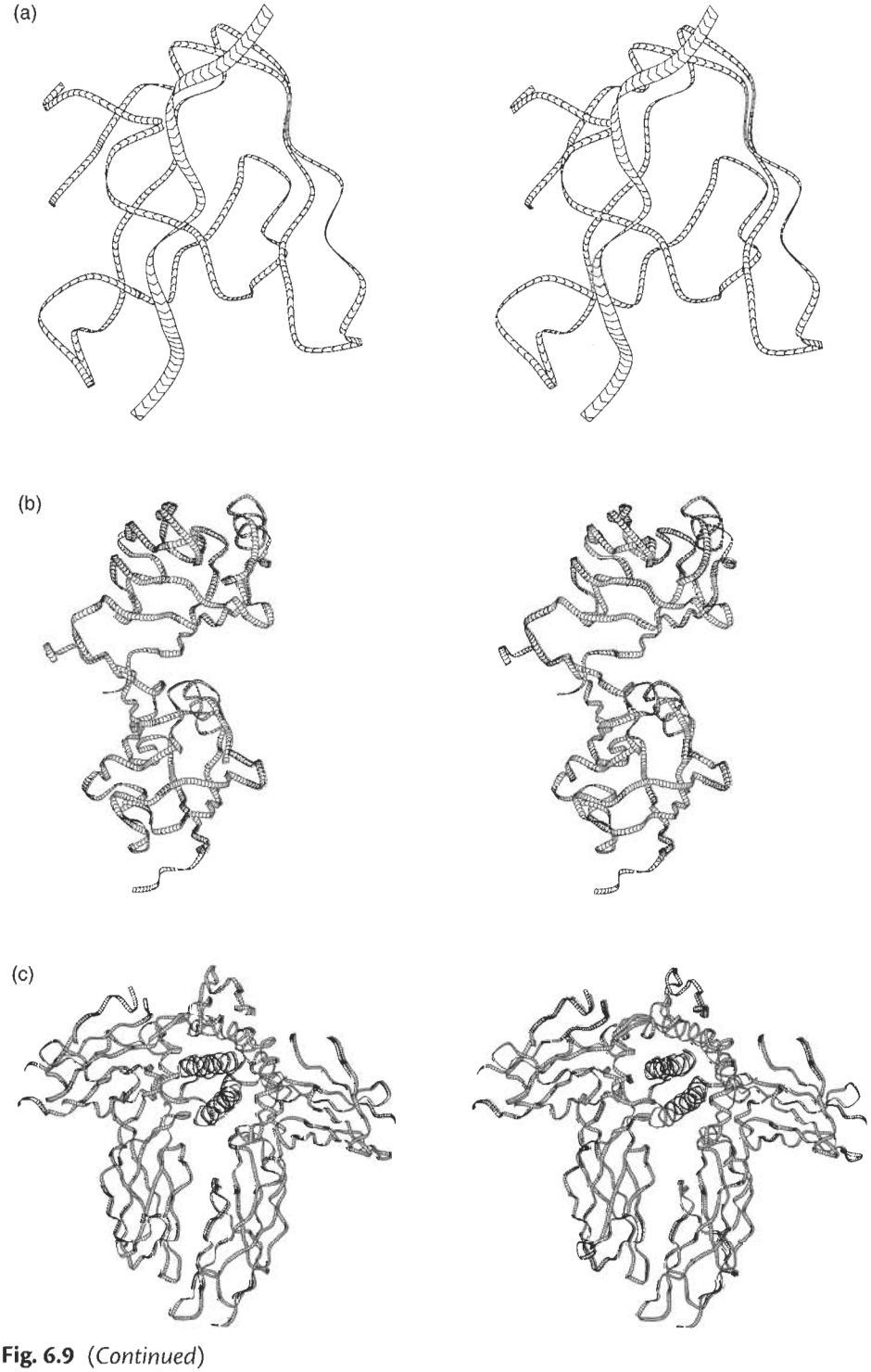


Fig. 6.9 (*Continued*)

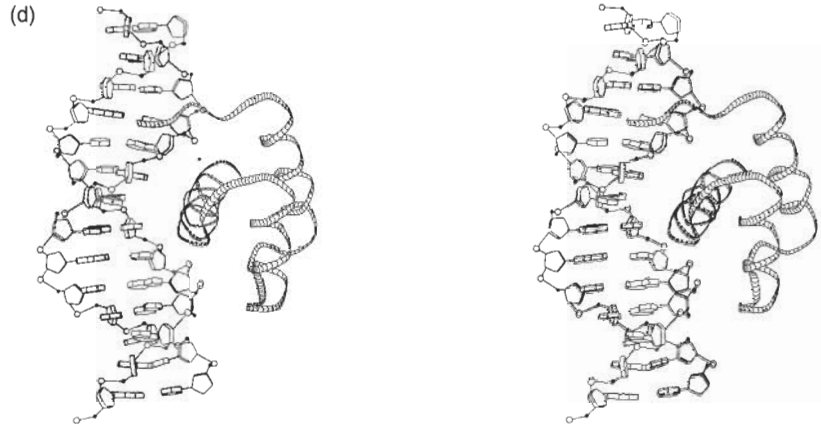


Fig. 6.9 Types of interactions involved in regulatory signalling. (a) Binding of a peptide by an SH3 domain [1CKA]. SH3 domains are common constituents of regulatory proteins. Functions of SH3 domains include signal transduction, protein and vesicle trafficking, cytoskeletal organization, cell polarization, and organelle biosynthesis. (b) domain-domain interaction: PDZ domains in syntrophin (black) and neuronal nitric oxide synthase (red) [1QAV]. (c) Binding of a molecule of human growth hormone (red) to two molecules of the external segment of the human growth hormone receptor (black). (d) The homeodomain antennapedia-DNA complex [9ANT]. Homeodomains are highly-conserved eukaryotic proteins, active in control of animal development. They regulate homeotic genes; that is, genes that specify locations of body parts. Antennapedia is a *Drosophila* protein responsible for initiating leg development. The earliest mutations found in antennapedia produced ectopic legs at the positions of, and instead of, antennae. Loss-of-function mutations convert legs into antennae. As with many DNA-binding proteins, an α-helix binds in the major groove of DNA.

Understanding the mechanism of regulation will require the structures of large protein and protein-nucleic acid complexes. The sizes of many of the large complexes challenges the limits of NMR spectroscopy. X-ray diffraction has had some major successes, but is at the mercy of being able to grow adequate crystals. Cryo-electron microscopy is another approach to structure determination of larger assemblies.

Electron microscopy of specimens at liquid nitrogen temperatures has revealed structures in the range r.m.m. = 5×10^5 to 4×10^8, 100–1500 Å in diameter. These results do not achieve atomic resolution. However, if the structures of individual components of a complex are known to high resolution from X-ray diffraction or NMR spectroscopy, the component structures can be fitted into the low-resolution structure determined by electron microscopy, to produce a detailed model of the entire assembly.

A limitation that remains is the difficulty of determining structures of transient complexes, or of systems showing substantial conformational changes upon assembly. The situation is shared with much of current molecular biology: we are coming to grips with static structures, but awaiting the development of methods for treating the dynamics.

Recommended reading

Albert, R. & Barabàsi, A. -L. (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.*, **74**, 47–97.
Barabàsi, A. -L., *Linked: How Everything Is Connected to Everything Else and What It means* (New York: Plume Books, 2003)
Ideker. T. (2004), A systems approach to discovering signaling and regulatory pathways—or, how to digest large interaction networks into relevant pieces, *Adv. Exp. Med. Biol.*, **547**, 21–30.
Babu, M. M., Luscombe, N. M, Aravind, L., Gcrstein, M. & Teichmann, S. A., (2004), Structure and evolution of transcriptional regulatory networks, *Curr. Opin. Struct. Biol.*, **14**, 283–291.
Tyers M. & Mann M., From genomics to proteomics, *Nature*, **422**, 193–197.

Exercises, Problems, and Weblems

Exercises

6.1 Hen egg white lysozyme has a relative molecular mass of about 14 300. If mass spectroscopy can measure mass to within 0.01%, could the following be confidently distinguished from the unmodified protein: (a) N-terminal acetylation? (b) phosphorylation of a single serine residue? (c) a single Lys→Gln substitution?

6.2 On photocopies of Fig. 6.4b, indicate the positions of the peaks if the sequence were: (a) MNLVQVR, (b) GNLQVVR, (c) MNLQVVG.

6.3 (a) What is the sequence of the fragment y_6 in Fig. 6.4b? (b) To which peak in Fig. 6.4b does the fragment NH_3^+-LQVVR-COOH correspond?

6.4 Oligonucleotide samples may vary by the binding of a Na^+ or K^+ ion to a phosphate, instead of a proton.

(a) What is the difference in mass between an oligonucleotide binding a proton or a Na^+ ion at a single site?

(b) What base change has the closest mass difference to the H^+-Na^+ mass difference?

(c) Would measuring mass to within 1 D be sufficiently accurate to distinguish this base change from the binding of a Na^+ ion instead of a proton, at a single site?

(d) In a mass spectrum of an oligonucleotide, what is the difference in mass between an oligonucleotide with a proton or a Mg^{2+} ion at a single site?

(e) What base change has the closest mass difference to the H^+-Mg^{2+} mass difference?

(f) Would measuring mass to within 1 D be sufficient accuracy to distinguish this base change from the binding of a Mg^{2+} ion instead of a proton, at a single site?

6.5 Assuming a typical single-nucleotide polymorphism (SNP) density of 1 SNP/5 kbp in a human genome, and only two possible bases observed at the position of any SNP, how many sequences could you expect to find throughout a population, within a 100 kbp region, if recombination were common at every position in the region? If only three of the possible combinations of SNPs—that is, three haplotypes—are observed, what fraction of possible sequences does this represent?

6.6 For which of the methods for determining interacting proteins (pages 325–326) (a) must one of the proteins be purified before testing for the interaction? (b) both of the proteins be purified before testing for the interaction?

6.7 In the top graph in the margin on page 315,

(a) Name two vertices such that if you add an edge between them at least one vertex has exactly two neighbours. (Note that two edges may cross without making a new vertex at their point of intersection.)

(b) Name two vertices such that if you add an edge between them to the original graph, the graph remains a tree.

(c) Name two vertices (neither of them V_1) such that if you add an edge between them to the original graph, the graph does not remain a tree.

(d) Name two vertices such that if you add an edge between them to the original graph, there are alternative paths, of lengths 3 and 4, between V_1 and V_5, with no vertices repeated. (In determining the length of a path, you have to count the initial and final vertices. A path of length 3 between V_1 and V_5 contains one intermediate vertex.)

(e) Name two vertices such that if you add an edge between them to the original graph, there is exactly one path between V_1 and V_3, with no vertices repeated, and it has length 4.

6.8 Of the examples of graphs at the bottom of page 315, (a) Which are directed graphs? (b) Which are labelled graphs? (c) In each example, what is the set of nodes? (d) In each example, what is the set of edges?

6.9 In a typical protein-protein interface of area 1700 Å^2: (a) How many intermolecular hydrogen bonds would you expect to be formed? (b) How many fixed water molecules would you expect to find in the interface? (c) If the entire buried area were hydrophobic, what contribution to the free energy of stabilization would you estimate it to make?

6.10 From the fragment of the *B. subtilis* protein interaction network shown in Fig. 6.6, what is the clustering coefficient of DnaC? The clustering coefficient of a node in a graph is defined as follows: Suppose the node has k neighbours. Then the total possible connections between the neighbours is $k(k-1)/2$. The clustering coefficient is the observed number of connections between neighbours divided by this maximum potential number of connections between neighbours.

6.11 In the London Underground: (a) What is the shortest path between Moorgate and Embankment stations. Note that, considered as a graph, the shortest path between two nodes is the path with the fewest intervening nodes, not the path that would take the minimal time or fewest interchanges. (b) What is the shortest cycle containing King's Cross, Holborn, and Oxford Circus stations? (c) If the neighbours of a station are the other stations that can be reached without passing through any intervening stations, what is the clustering coefficient of the Oxford Circus station? (See Exercise 6.10 for the definition of the clustering coefficient.)

6.12 In the London Underground: (a) What is the maximum path length between any two stations? That is, for which two stations does the shortest trip

between them involve the maximum number of intervening stops? (b) If the District Line were not active, what stations if any would be inaccessible by underground? (c) If the Jubilee line were not active, what stations if any would be inaccessible by underground?

6.13 On a photocopy of the three common combinations of network control motifs (Box, page 335) (a) indicate which nodes are controlled by only *one* upstream node; (b) indicate which node exerts control over only *one* downstream node.

6.14 On a photocopy of the simplified fragment of the yeast regulatory network (Fig. 6.8) indicate examples of the network control motifs (a) fork (b) 'one-two punch'. (c) Add one arrow to create a scatter motif.

6.15 In the dimer between syntrophin and neuronal nitric oxide synthase (Fig. 6.9b), (a) is the dimer structure open or closed? (b) What secondary structure element is shared between the two domains?

6.16 In the overall yeast transcriptional regulatory network the number of incoming connections to target genes follows an exponential distribution. That is, the probability that a gene is controlled by k transcription regulators is proportional to $e^{-\alpha k}$, with $\alpha = 0.8$, $k = 1, 2, \ldots$. What is the ratio of the number of target genes receiving four input connections to the number receiving two input connections?

Problems

6.1 (a) How many positions in all are there in the microarray in Plate VII? (b) How many are complementary to RNAs from liver? (c) How many are complementary to RNAs from brain? (d) How many are complementary to RNAs from liver and brain? (e) How many are complementary to neither?

6.2 For avidin-biotin, $K_D = 10^{-15}$. Suppose k_{on} were as fast as the diffusion limit, $\sim 10^{-9}$ M s^{-1}. (a) What is the value of k_{off}? (b) What would be the half-life of the avidin-biotin complex? (c) Suppose k_{on} for avidin-biotin were 10^{-7} M^{-1} s^{-1}. What would be the half-life of the complex?

6.3 The anti-tuberculosis drug isoniazid requires activation by the *M. tuberculosis* enzyme KatG (a catalase-peroxidase), but the related drug ethionamide does not require activation. Suppose expression profiles were measured for the following:

(a) a strain with active KatG, not exposed to either drug,

(b) a strain with active KatG, exposed to isoniazid,

(c) a strain without active KatG, exposed to isoniazid,

(d) a strain with active KatG, exposed to ethionamide,

(e) a strain without active KatG, exposed to ethionamide.

Which two of (b), (c), (d), and (e) would show a similar pattern of enhancement of gene expression relative to (a)? Why whould you expect the enhancement pattern to be similar in: (b), (d) and (e) but not (c)?

6.4 J. Foote and G. Winter compared the dissociation constants of a natural mouse antilysozyme antibody (D1.3), an engineered 'humanized' antibody in which the antigen-binding site was grafted onto a human framework (Human-original) and several mutants of the 'humanized form', including Human-mutated. The antigen was hen egg white lysozyme.

Antibody	Number of sequence differences to D1.3	k_{on} $M^{-1}s^{-1}$	K_D
D1.3	0	1.4×10^{-6}	3.7×10^{-9}
Human-original	48	0.7×10^{-6}	260×10^{-9}
Human-mutated	44	1.3×10^{-6}	14×10^{-9}

(a) Calculate the 'off-rate' k_{off} for each antibody. (b) Which has the major effect on the dissociation constant: differences in 'on-rate' or differences in 'off-rate'?

6.5 Analyse the map of the London Underground by counting the number of connections made from each station in zone 1 (the central portion). Count connections *to* stations inside and outside zone 1, as long as they originate within zone 1. Count only one connection if two stations are connected by more than one line; in other words, for each station, the question is: How many other stations can be reached without passing through any intermediate stops?

(a) What is the maximum number of connections of any station?

(b) For each integer k from 1 to this maximum number, how many stations have k connections?

(c) Plot these data on a log-log plot. Does the relationship appear reasonably linear?

(d) If so, fit a straight line to the log-log plot and determine the exponent.

Results of network analysis of this sort are more significant if the data cover several orders of magnitude, but this is not possible for this example.

6.6 In the overall yeast transcriptional regulatory network the number of incoming connections to nodes follows an exponential distribution. That is, the probability P_k that a gene is controlled by k transcription regulators is given by $P_k = Ce^{-\alpha k}$, $k = 1, 2, \ldots$, with $\alpha = 0.8$.

(a) Determine the constant of proportionality C in terms of α, by summing the series $\sum_{k=1}^{\infty} Ce^{-\alpha k} = 1$.

(b) If $\alpha = 0.8$, what is the maximum value of k for which at least 1% of the nodes would be expected to have at least k incoming connections?

(c) If $\alpha = 0.8$, plot the expected histogram for $1 \leq k \leq 7$.

(d) Determine the mean value of k in terms of α. (Hint: in the solution of (a) you expressed $\sum_{k=1}^{\infty} e^{-\alpha k}$ as a function $f(\alpha)$. Differentiate this relationship with

respect to α to produce the equation: $\sum_{k=1}^{\infty} -ke^{-\alpha k} = f'(\alpha)$. Then the mean value of k is given by $-f'(\alpha)/f(\alpha)$.)

(e) What is the mean value $<k>$ corresponding to $\alpha = 0.8$?

(f) What is the median value of k? This is the value κ such that half the nodes have $\leq\kappa$ incoming connections, and half the nodes have $\geq\kappa$ incoming connections. Find κ in terms of α. (Hint: if $\sum_{k=1}^{\infty} Ce^{-\alpha k} = 1$, then $\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k} = \frac{1}{2}$. But $\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k} = e^{-\alpha\kappa}\sum_{k=\kappa+1}^{\infty} Ce^{-\alpha k}$. In general, this approach will provide a non-integral estimate of κ, just round this result to the nearest integer.)

(g) If $\alpha = 0.8$, what is the median value κ? How does it compare with the average value $<k>$? Are the two values approximately equal?

6.7 Indicate how to connect a selection of the three common network control motifs so that a single input node can influence three output nodes.

Weblems

6.1 Define the following terms: (a) interactome (b) metabolome (c) signalome. (d) More difficult: can you think of, and define, a reasonable 'ome' that has not yet been proposed?

6.2 The catalase-peroxidase KatG of *M. tuberculosis* activates the drug isoniazid. The most common mutation in KatG associated with resistant strains is S315T. From a model of *M. tuberculosis* KatG based on the structure of the homologue from *Burkholderia pseudomallei* (PDB entry |1mwv|), suggest a mechanism for the reduced activity of the mutant to activate isoniazid while retaining activity against smaller substrates.

6.3 Find a fragment of the genealogy of the royal families of Europe containing a family tree that is not a graph.

6.4 One model for the growth of a scale-free network suggests that if new nodes and edges are added according to the rule that the probability of adding a new edge is proportional to the number of edges that a node already has, the network will remain scale-free and retain the same exponent. Using historical maps of the London Underground, check whether earlier networks are scale-free. Test whether addition of edges to the network has followed the rule that edges have been added preferentially to more highly-connected nodes.

6.5 From the *B. subtilis* protein interaction database SPiD (see Fig. 6.6), (a) what type of experimental evidence links DnaC to DnaG? (b) What type of experimental evidence links DnaC to DnaI?

6.6 From the *B. subtilis* protein interaction database SPiD (see Fig. 6.6), print a graph showing not only the immediate neighbours of DnaC, but the second neighbours (the neighbours of the neighbours).

Conclusions

How can we extrapolate from the current state of play to the bioinformatics of the future? Clearly, data collection will proceed and continue to accelerate. New high-throughput techniques will provide additional types of data, including information about the integration and control of life processes. Computing facilities of increasing power will be applied to the storage, distribution and analysis of the results. New databases will appear on the Web, and links between databases will become more effective. Improved algorithms will be devised to analyse and interpret the information given us and to transmute it from data to knowledge to wisdom.

One threshold will be reached when our knowledge of sequences and structures becomes more nearly complete, in the sense that a fairly dense subset of the available data from contemporary living forms has been collected. (Of course there is no question of being able to know everything.) This will be recognized operationally when a random dip into the pot of a genome, or the determination of a new protein structure, is far more likely to turn up something already known, rather than to uncover something new. Nature is, after all, a system of unlimited possibilities but finite choices.

Applications will become more feasible, and mature ever more quickly from 'blue-sky' research to standard industrial and clinical practice. Some of the higher levels of biological information transfer—such as the programmes of genetic development during the lifetime of individuals, and the activities of the human mind—will come to be included in the processes we can describe quantitatively and analyse at the level of molecules and their interactions.

In Michaelangelo's frescos on the ceiling of the Sistine Chapel, the serpent offering Eve the fruit of the tree of knowledge is represented with its legs coiled around the tree in the form of a double helix. We can hope that our new temptation to knowledge embodied in another double helix will have more fortunate consequences.