6

The diversity of proteins

The incredible diversity shown by the living world ranges from bacteria and viruses to unicellular organisms, eventually culminating with complex multicellular systems of higher plants, including gymnosperms and angiosperms, and animals such as those of the vertebrate kingdom. However, the living world is based around proteins made up of the *same* 20 amino acids. There is no fundamental difference between the amino acids and proteins making up a bacterium such as *Escherichia coli* to those found in higher vertebrates. As a result the principles governing protein structure and function are equally applicable to all living systems. This pattern of similarity is not surprising if one considers that all living systems are related by their evolutionary origin to primitive ancestors that had acquired the basic 20 amino acids to use in the synthesis of proteins. Higher levels of complexity were acquired by evolutionary divergence that led to a subtle alteration in primary sequence and the generation of new or altered functional properties. Although the exact nature of the 'first' ancestral cell is unclear along with details of self-replicating systems advances have been made in understanding the origin of proteins.

Prebiotic synthesis and the origins of proteins

The origin of life represents one of the greatest puzzles facing scientists today. What originally seemed to be an impossible problem has gradually become better understood via experimentation. In order to synthesize proteins it is first necessary to make amino acids containing carbon, hydrogen, oxygen and nitrogen, and occasionally sulfur. The source of all carbon, hydrogen, oxygen and nitrogen would have been the original atmospheric gases of carbon dioxide, nitrogen, water vapour, ammonia, and methane, but not atmospheric oxygen since this was almost certainly lacking in early evolutionary periods. When this series of events is placed in a time span we are describing reactions that occurred more than 3.6 billion years ago.

In a carefully designed experiment Stanley Miller and Harold Urey showed that simulating the primitive conditions present on the Earth around 4 billion years ago could result in the production of biomolecules. This study involved adding inorganic molecules to a closed system under a reducing atmosphere (lacking oxygen). The gaseous mixture of mainly ammonia, hydrogen and methane simulated the early Earth's atmosphere. The whole mixture was refluxed in a closed evacuated system with the water phase representative of the 'oceans' of the early earth analysed at the end of an experiment lasting several days (Table 6.1). Subjecting the system to electrical discharge (lightning) and high amounts of ultraviolet light (Sun) formed biomolecules, including the amino acids glycine, alanine and aspartate. The formation of hydrogen cyanide, aldehydes and other cyano

Proteins: Structure and Function by David Whitford

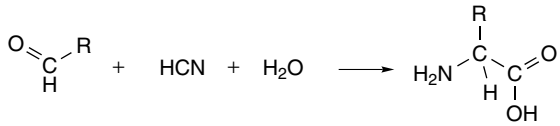

Figure 6.1 Prebiotic synthesis of amino acids from simple organic molecules

Table 6.1 Yields of biomolecules from simulating prebiotic conditions using a mixture of methane, ammonia, water and hydrogen

Biomolecule	Approximate yield (%)
Formic acid	4.0
Glycine	2.1
Glycolic acid	1.9
Alanine	1.7
Lactic acid	1.6
β-Alanine	0.76
Propionic acid	0.66
Acetic acid	0.51
Iminodiacetic acid	0.37
α-Hydroxybutyric acid	0.34
Succinic acid	0.27
Sarcosine	0.25
Iminoaceticpropionic acid	0.13
N-Methylalanine	0.07
Glutamic acid	0.051
N-Methylurea	0.051
Urea	0.034
Aspartic acid	0.024
α-Aminoisobutyric acid	0.007

Shown in red are constituents of proteins (after Miller, S.J. & Orgel, L.E. *The Origins of Life on Earth*. Prentice-Hall, 1975).

compounds was also important since these simple compounds undergo a wide range of further reactions. The general reaction is summarized by a simple addition reaction between aldehydes and hydrogen cyanides in the presence of water (Figure 6.1), although in practice

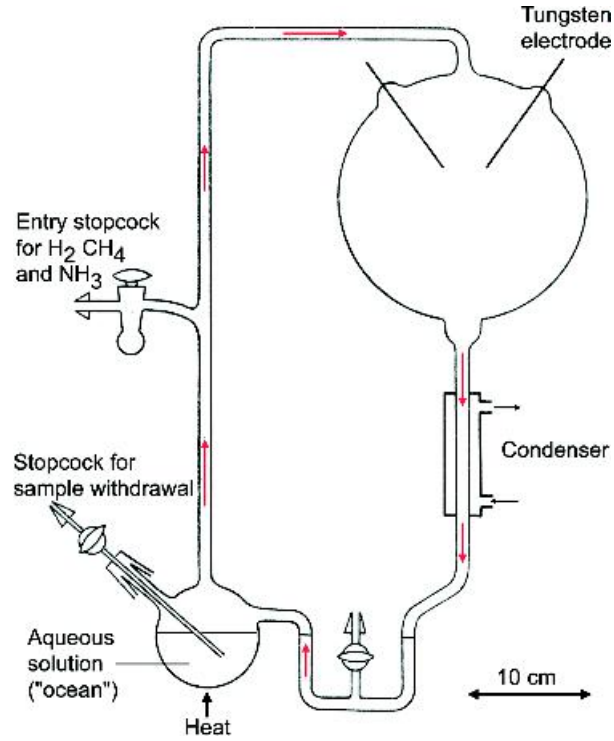

Figure 6.2 An example of the apparatus used by Urey and Miller to demonstrate prebiotic synthesis of organic molecules

the reaction may form a nitrile derivative in the atmosphere followed by hydrolysis in the 'ocean' to yield simple amino acids.

Performing this type of experiment with different mixtures of starting materials yielded additional biomolecules, including adenine. Variations on this basic theme have suggested that although the Earth's early atmosphere lacked oxygen the presence of gases such as CO, CO_2 and H_2S were vital for prebiotic synthesis. The presence of sulfur enhances the number and type of reactions that could occur. More recently, examination of ocean floors has revealed the presence of deep sea vents sometimes called 'smokers' or fumaroles. These vents release hot gases and minerals from the Earth's crust into the ocean and are also prime sites for organic synthesis. This suggests that many potential

sites and sources of energy were available for prebiotic synthesis.

The next barrier to the evolution of life involved the formation of polymers from precursors. The genetic systems found in cells today are specialized polymers. They are able to direct the synthesis of proteins from messenger ribonucleic acid (mRNA), the latter representing the information present in deoxyribonucleic acid (DNA). In addition these polymers are capable of directing their own synthesis in a macromolecular world of DNA/RNA and protein.

$$\text{DNA} \longrightarrow \text{RNA} \longrightarrow \text{protein}$$

These systems are self-replicating and their evolution represents one of the greatest hurdles to be overcome in the development of living systems. Replicating DNA requires proteins to assist in the overall process whilst the scheme above demonstrates that protein synthesis requires DNA and RNA. This creates a paradox often called the 'chicken and egg' puzzle of molecular biology of which came first.

The recent demonstration that RNA molecules have catalytic function analogous to conventional enzymes has revolutionized views of prebiotic synthesis. Catalytic forms of RNA called ribozymes mean that RNA molecules, in theory, at least have the means to direct their own synthesis and to catalyse a limited number of chemical reactions. For this reason a prevalent view of molecular evolution involves a world dominated by RNA molecules that gradually evolved into a system in which proteins carried out catalysis, whilst nucleic acid performed a role of information storage, transfer and control. The details of this transition remain far from complete but supportive lines of evidence include: (i) the existence of different forms of RNA such as rRNA, mRNA, tRNA and genomic RNA; (ii) molecules such as nicotinamide adenine dinucleotide (NAD), adenosine tri- and diphosphate (ATP/ADP), and flavin adenine dinucleotide (FAD) found universally throughout cells are composed of adenine units analogous to those occurring in RNA; (iii) tRNAs have a tertiary structure; (iv) ribosomes represent hybrid RNA–protein systems where catalysis is RNA based; and (v) the enzyme RNaseP from *E. coli* catalyses the degradation of polymeric RNA into smaller nucleotide units in a reaction where RNA is the active component.

A major objection to the view of an 'RNA world' has been the comparative instability of RNA (especially when compared to DNA). RNA is easily degraded and it is difficult to see how stable systems capable of replication and catalysis evolved. Additionally the synthesis of polymers of RNA under conditions similar to those found early in the earth's history has proved remarkably difficult. Despite these problems most researchers view RNA as a likely intermediate between the 'primordial soup' and the systems of replication and catalysis found in modern cell types.

The fossil record evidence shows that bacteria-like organisms were present on earth 3.6 billion years ago. This implies that the systems of replication present today in living cells had already evolved. The 'RNA-directed world' was therefore a comparatively short time interval of ~0.5 billion years!

Having evolved a primitive replication and catalysis system based on RNA the simple amino acids could be used in protein biosynthesis. It is very unlikely that all amino acids were present in the primordial soup since some of the amino acids are relatively unstable especially under acidic conditions, and this includes the side chains of asparagine, glutamine, and histidine. In addition the amino acids found in proteins represent a very small subset of the total number of amino acids known to exist. This might suggest that the present class of amino acids evolved over millions of years to reflect a blend of chemical and physical properties required by proteins although it was essentially complete and intact 3.6 billion years ago.

Evolutionary divergence of organisms and its relationship to protein structure and function

The precise details of the origin of life involving the generation of a self-replicating system and its evolution into the complex multicellular structures found in higher plants and animals are unclear but the fossil record shows that primitive bacteria were present on earth in the pre-Cambrian period nearly 3.6 billion years ago. These fossilized cells resemble a class of bacteria found on present day Earth called

cyanobacteria. Although it is surprising that bacteria leave fossil records cyanobacteria often form a significant cell wall together with layered structures called stromatolites (Figure 6.3). These structures form a mat as cyanobacteria grow trapping sediment and helping the fossilization process. Cyanobacteria are prokaryotic cells, lacking a nucleus and internal membranes, capable of both photosynthetic and respiratory growth and are represented today by genera such as *Nostoc*, *Anacystis* and *Synechococcus*.

Less ambiguously and more importantly the fossil record demonstrates progressive increases in complexity from the simple prokaryotic cell lacking a nucleus to more complicated structures similar to modern eukaryotic cells. Eukaryotic cells became multicellular and evolved by specialization towards specific cellular functions (Figure 6.4). These cells increased in structural complexity by internal compartmentalization with genome organization becoming more complex along with the variety of biochemical reactions catalysed within these cells. Whilst the fossil record aided our understanding of evolution one of the best methods of deciphering evolutionary pathways has come from comparing protein and DNA sequences.

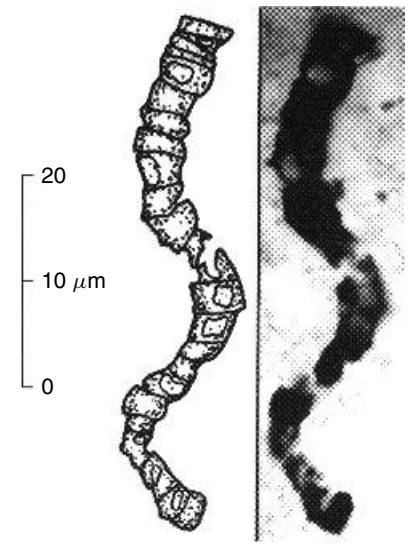

Figure 6.3 Microfossil of filamentous bacterial cells. The fossil shown alongside an interpretive drawing is from Western Australia and rocks dated at ~3.4 × 10^9 years (Reproduced with permission from Voet, D., Voet, J.G. and Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons, Ltd, Chichester, 1999)

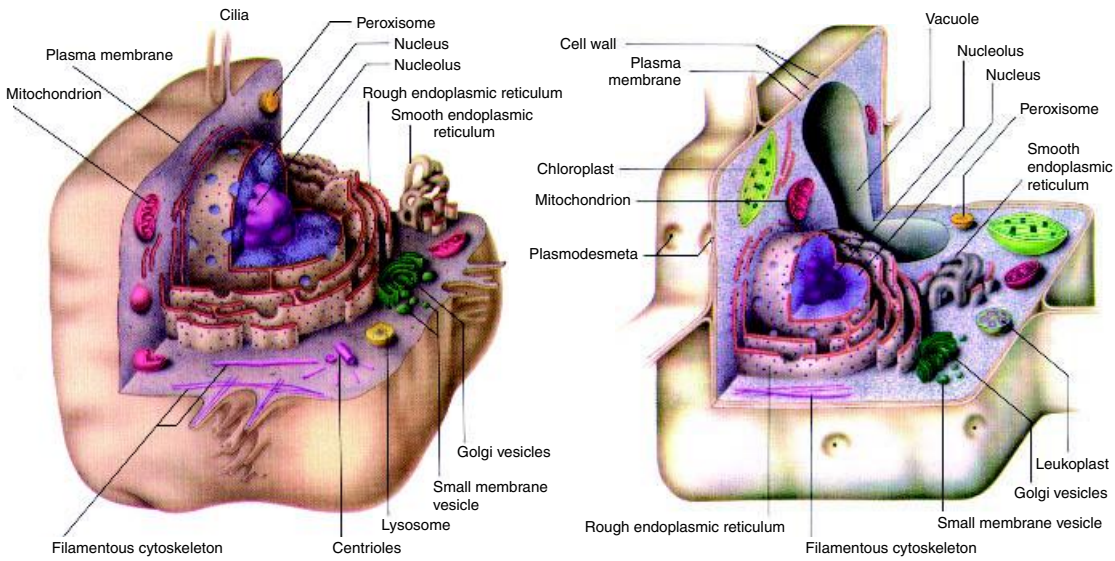

Figure 6.4 Generalized eukaryotic cells (plant and animal) (reproduced courtesy of Darnell, J.E. *et al. Molecular Cell Biology*. Scientific American, 1990)

In a protein of 100 amino acid residues there are 20^{100} unique or possible sequences. It is clear that biological sequences represent only a small fraction of the total number of permutations. Protein sequences often show similarities with these relationships governed by evolutionary lineage. Over millions of years a sequence can change but it cannot cause complete loss of function unless gene duplication has occurred. If mutation provides new enhanced functional activities any selective advantage conferred on the host organism possessing this protein will lead to improved survival and gene perpetuation.

Protein sequence analysis

Protein sequencing

The sequencing of DNA has advanced so rapidly that this method is now by far the most common and effective way of determining the sequence of a protein. By translating the order of nucleotide bases along a DNA sequence one can simply derive the sequence of amino acid residues. However, there are occasions when it becomes important to sequence a protein directly and this might include determining the extent of post-translational processing or arranging peptide fragments in a linear order.

Protein sequencing is an automated technique carried out using sophisticated instruments (sequenators) and based on methods devised by Pehr Edman (it is often called Edman degradation). The unknown polypeptide is reacted under alkaline conditions (pH ~9) with phenylisothiocyanate (PTC) where the free amino group at the N-terminal forms a phenylthiocarbamoyl derivative, which is hydrolysed from the remaining peptide using anhydrous trifluoroacetic acid (Figures 6.5 and 6.6). PTC makes the first peptide bond less stable and easily hydrolysed. Residue rearrangement in aqueous acidic solution yield a phenylthiohydantoin (PTH) derivative of the N-terminal amino acid that is identified using chromatography or mass spectrometry (Figure 6.7). The significance of this series of reactions is that the N-terminal amino acid is 'tagged' by attaching PTC but the remaining polypeptide chain (now containing $n-1$ residues) remains intact and can undergo further reactions with PTC at its new N-terminal residue. The Edman degradation is a repetitive, cyclical, series of reactions, although

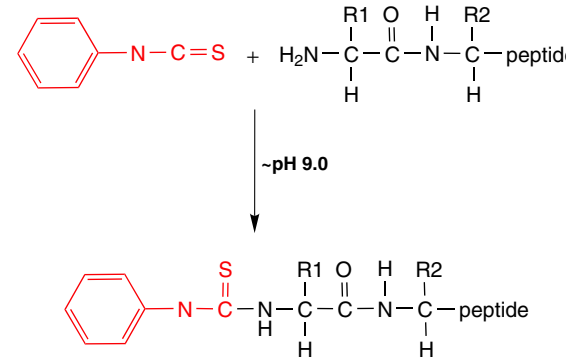

Figure 6.5 The reaction of the N-terminal amino acid residue with phenylisothiocyanate in the first step of the Edman degradation

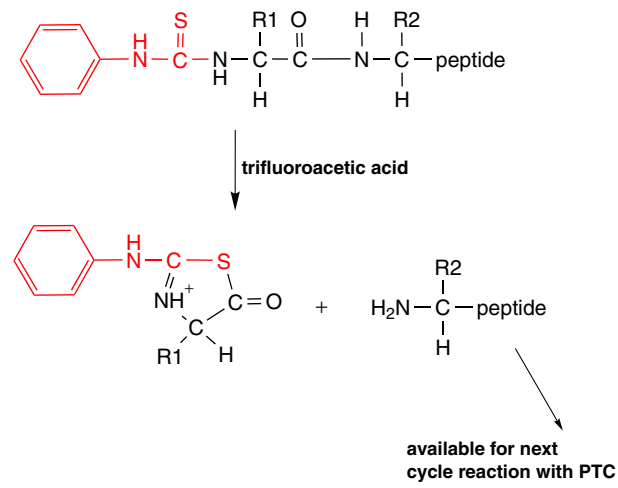

Figure 6.6 Hydrolysis of the phenylthiocarbamoyl derivative of the peptide to yield a protein of $n-1$ residues and a free 'labelled' amino acid

as with most repetitive procedures, errors accumulate and progressively degrade the accuracy of the whole process. Errors include: random breakage of the polypeptide chain producing a second free amino terminal residue; incomplete reaction between PTC and the N-terminal amino acid leading to its appearance in the next reactive cycle; and side reactions that compete with the reaction between PTC and the polypeptide chain. Sequenators are very sensitive

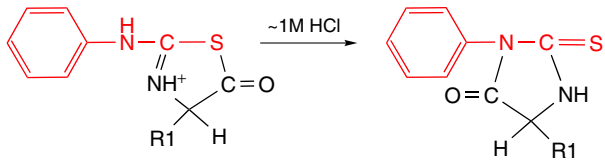

Figure 6.7 Re-arrangement of the PTC-derivative to form a phenylthiohydantoin (PTH) derivative of the N-amino acid

instruments capable of sequencing picomole amounts of polypeptide. Usually an upper limit on the length of the polypeptide chain that can be sequenced directly is about 70 residues. Since most proteins contain far more than 70 residues the sequencing procedures relies on 'chopping' the polypeptide chain into a series of smaller fragments that are each sequenced independently.

Generation of smaller peptide fragments involves using hydrolytic enzymes that cleave the polypeptide at specific sequences or by the use of cyanogen bromide that splits polypeptide chains after methionine residues (Table 6.2).

After purification of the individual fragments the shorter peptides are sequenced, although the major problem is now to deduce the respective order of each

Table 6.2 Enzymes or reagents for generating peptide fragments suitable for sequencing

Enzyme/reagent	Cleavage site
Trypsin	-Arg -↑-Yaa or Lys -↑-Yaa-
Endoprotease Arg-C	-Arg -↑-Yaa
Chymotrypsin	-Phe -↑-Yaa, -Tyr -↑-Yaa, -Trp -↑-Yaa
Clostripain	-Arg -↑-Yaa
Asp-N	-Xaa-↑-Asp
Thermolysin	-Xaa-↑-Leu, -Xaa-↑-Ile, -Xaa-↑-Val, -Xaa-↑-Met,
V-8 protease	-Asp -↑-Yaa, -Glu -↑-Yaa
Cyanogen bromide (CNBr)	-Met -↑-Yaa

In many cases the above enzymes show wider specificity. For example chymotrypsin will cleave other large side chains particularly Leu and care needs to be exercised in interpreting the results of proteolytic cleavage. In other instances the identity of Xaa/Yaa can influence whether cleavage occurs. For example Lys-Pro is not cleaved using trypsin.

Table 6.3 Fragments derived by digestion of unknown protein with Asp-N and trypsin

Digestion with trypsin	
Mass	Peptide sequence
4905.539	ITKPSESIITTIDSNPSWW TNWLIPAISALFVALIYHLYTSEN
2205.928	EQAGGDATENFEDVGHSTDAR
1511.749	FLEEHPGGEEVLR
1412.717	TFIIGELHPDDR
1186.599	YYTLEEIQK
1160.646	STWLILHYK
738.403	VYDLTK
650.299	AEESSK
599.290	HNNSK
476.271	ELSK
317.218	AVK
234.145	SK
Digestion with Asp-N	
Mass	Peptide Sequence
4100.113	AEESSKAVKYYTLEEIQK HNNSKSTWLILHYKVY
3621.805	DSNPSWWTNWLIPAISA LFVALIYHLYTSEN
2411.184	DLTKFLEEHPGGEEVLREQAGG
1825.981	DARELSKTFIIGELHP
1789.007	DRSKITKPSESIITTI
825.326	DATENFE
615.273	DVGHST
134.045	D

peptide. This problem is resolved by repeating the digestion of the intact protein with a second enzyme that reacts at different sites producing fragments whose relationship to the first set is established by sequencing to determine an unambiguous order of residues along a polypeptide chain. The principle of this method is demonstrated with an 'unknown' protein of 133 residues and its digestion with two enzymes; trypsin and Asp-N (Table 6.3 and Figure 6.8). Trypsin shows substrate specificity for lysine and arginine residues cleaving peptide bonds on the C terminal side of these residues whilst Asp-N is a protease that cleaves before aspartate residues.

To verify the primary sequence a total amino acid analysis is usually performed on the unknown protein by complete hydrolysis of the protein into individual amino acid residues. Quite clearly the total amino acid composition must equate with the combined number of amino acids derived from the primary sequence.

Amino acid analysis consists of three steps: (i) hydrolysis of the protein into individual amino acids; (ii) separation of the amino acids in this mixture; and (iii) identification of amino acid type and its quantification. Hydrolysis of the protein is normally complete after dissolving a small amount of the sample in 6 M HCl and heating the sample in a vacuum at 110 °C for 24 hours. The peptide bonds are broken leaving a mixture of individual amino acids. This approach destroys the amino acid tryptophan completely whilst cysteine residues may be oxidized and partially destroyed by these conditions. Similarly, acid hydrolysis of glutamine and asparagine side chains can form aspartate and glutamate and it is not usually possible to distinguish Asn/Asp and Glu/Gln in protein hydrolysates. For this reason protein sequences may be written as Glx or Asx representing the combined number of glutamine/glutamate and asparagine/aspartate residues.

Separation is achieved by cation exchange chromatography (Figure 6.9) using resins, supported within stainless steel or glass columns, containing negatively charge groups such as sulfonated polystyrenes. The negatively charge amino acids such as aspartate and glutamate elute rapidly whilst the flow of positively charged amino acids through the column is retarded due to interaction with the resin. By altering the polarity of the eluting solvent the interaction of hydrophobic amino acids with the column is enhanced. Amino acids with large hydrophobic side chains, for example phenylalanine and isoleucine, elute more slowly than smaller amino acids such as alanine and glycine.

To enhance detection the amino acids are reacted with a colored reagent such as ninhydrin, fluorescein, dansyl chloride or PTC. If this procedure is performed prior to column separation the absorbance of derivatized amino acids as they elute from the column is readily recorded at ~540 nm or from their fluorescence. Due to the high reproducibility of these

```
                    *         20          *         40          *
Unknown  :  AEESSKAVKYYTLEEIQKHNNSKSTWLILHYKVYDLTKFLEEHPGGEEVL  :   50
Unknown  :  AEESSKAVKYYTLEEIQKHNNSKSTWLILHYKVYDLTKFLEEHPGGEEVL  :   50

                  60          *         80          *         100
Unknown  :  REQAGGDATENFEDVGHSTDARELSKTFIIGELHPDDRSKITKPSESIIT  :  100
Unknown  :  REQAGGDATENFEDVGHSTDARELSKTFIIGELHPDDRSKITKPSESIIT  :  100

                    *         120         *
Unknown  :  TIDSNPSWWTNWLIPAISALFVALIYHLYTSEN  :  133
Unknown  :  TIDSNPSWWTNWLIPAISALFVALIYHLYTSEN  :  133
```

Figure 6.8 The figure shows the unknown protein whose primary sequence can be deduced from sequencing smaller fragments derived by digestion with trypsin (top line) and Asp-N (bottom line). For clarity alternate fragments derived using trypsin are shown in red and purple whilst the Asp-N fragments are shown in magenta and green

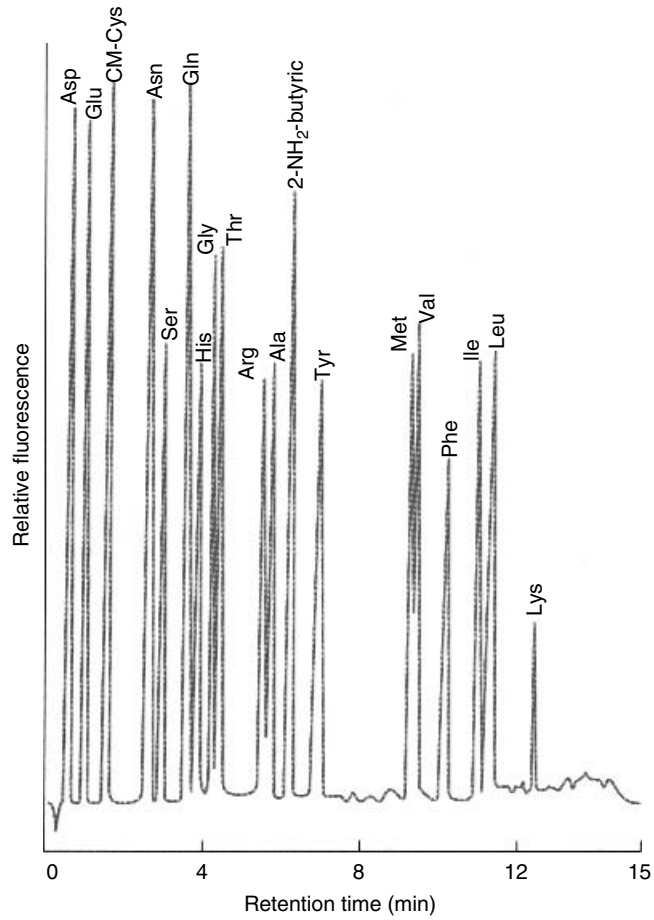

Figure 6.9 Elution of protein hydrolysate from a cation exchange column. The amino acids were derivatized first with a fluorescent tag to aid detection (after Hunkapiller, M.W. *et al. Science* 1984, **226**, 339–344)

profiles the amino acids of each type can be identified and their relative numbers quantified.

DNA sequencing

DNA sequencing methods are now routine and the recent completion of genomic sequencing projects testifies to the efficiency and accuracy of these techniques. The first genome sequencing projects were completed in the early 1980s for viruses and bacteriophages such as ϕX174 as well as organelles such as mitochondria that contain small circular DNA genomes. However in the last decade massive DNA sequencing projects were initiated and have resulted in vast numbers of primary sequences. In 2001 the human genome project was completed as a 'first draft' and the genomes of many other prokaryotic and eukaryotic organisms have been sequenced. This list includes the completion of the genome of the fruit fly *Drosophila melanogaster* as well as the genomes of many bacteria including pathogenic strains such as *Haemophilus influenzae*, *Helicobacter pylori*, *Yersinia pestis*, *Pseudomonas aeruginosa*, *Campylobacter jejuni*, as well as *E. coli* strain K-12. This has been complemented by completion of the genomes of *Saccharomyces pombe* and *cerevisiae* (the fission and baker's yeast,

respectively), the nematode worm *Caenorhabditis elegans* and the plant genome of *Arabidopsis thaliana*. Unfortunately it remains largely true that we have no idea about the structure or function of many of the proteins encoded within these sequenced genomes.

Nowadays it is common to determine the order of bases along a gene and in so doing deduce the primary sequence of a protein. After locating the relevant start codon (ATG) it is straightforward to use the genetic code to translate the remaining triplet of bases into amino acids and thereby determine the primary sequence of the protein. There are many computer programs that will perform these tasks using the primary sequence data to derive additional properties about the protein. These properties can include overall charge, isoelectric point (pI), hydrophobicity and secondary structure elements and are part of the bioinformatics revolution that has accompanied DNA and protein sequencing.

For eukaryotes translation of the order of bases along a gene is complicated by the presence of introns or non coding sequences. The recognition of these sites or the use of cDNA derived from processed mRNA allows protein sequences to be routinely translated. In databases it is normal for DNA sequences of eukaryotic cells to reflect the coding sequence although occasionally sequences containing introns are deposited.

Gene sequencing is an automated technique that involves determining the order of only four components (adenine, cytosine, guanine, and thymine, i.e. A, C, G and T) compared with 20 different amino acids. With fewer components it becomes critical to correctly establish the exact order of bases since a mistake, such as an insertion or deletion, will result in a completely different translated sequence. DNA is amplified to produce many complementary copies of the template using the polymerase chain reaction (PCR) a process that exploits the activity of thermostable DNA polymerases. The PCR technique is divided into three steps: denaturation, annealing and extension (Figure 6.10). Each step is optimized with respect to time and temperature. However, the process is generally performed at three different temperatures of ~96, ~55 and ~72 °C with each phase lasting for about 30, 30 and 120 s, respectively.

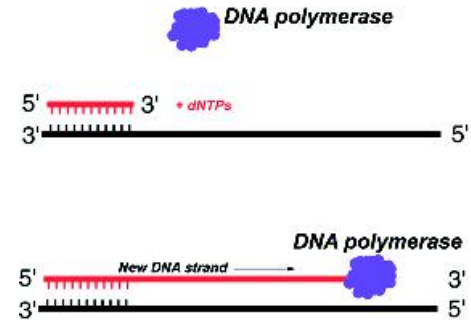

Figure 6.10 General principle of primer-directed DNA synthesis by polymerases. The PCR extends this process in a cyclical series of reactions since the polymerase is thermostable and withstands repeated cycles of high temperature

DNA polymerase activity extends new strands from primers (15–25 bases in length) that are complementary to the two template strands (Figure 6.11). A polymerase commonly used in these studies is that isolated from *Thermus aquaticus* and often referred to as *Taq* polymerase. The double helix is dissociated at high temperatures (~96 °C) and on cooling the suspension primers anneal to each DNA strand (~55 °C). With a suitable supply of nucleotide triphosphates (dNTPs, Figure 6.12) new DNA strand synthesis by the polymerase occurs rapidly at 72 °C forming two complementary strands. After one cycle the PCR doubles the number of copies and the beauty of the process is that it can be repetitively cycled to provide large amounts of identical DNA. It can be seen that starting with one copy of DNA leads after thirty rounds of replication to 2^{30} copies of DNA – all identical to the initial starting material.[1] This procedure is used to amplify DNA fragments as part of a cloning, mutagenesis or forensic study but it can also be used to sequence DNA.

If PCR techniques are carried out with dNTPs and a small amount of dideoxy (ddNTPs) nucleotides then chain termination will result randomly along the new DNA strand. Dideoxynucleotides were first introduced by Frederick Sanger as part of a sequencing strategy based on the absence of an oxygen atom at the

[1]Only with 'proofreading' DNA polymerases. Taq polymerase is not proof reading and exhibits an error rate of ~1 in 400 bases.

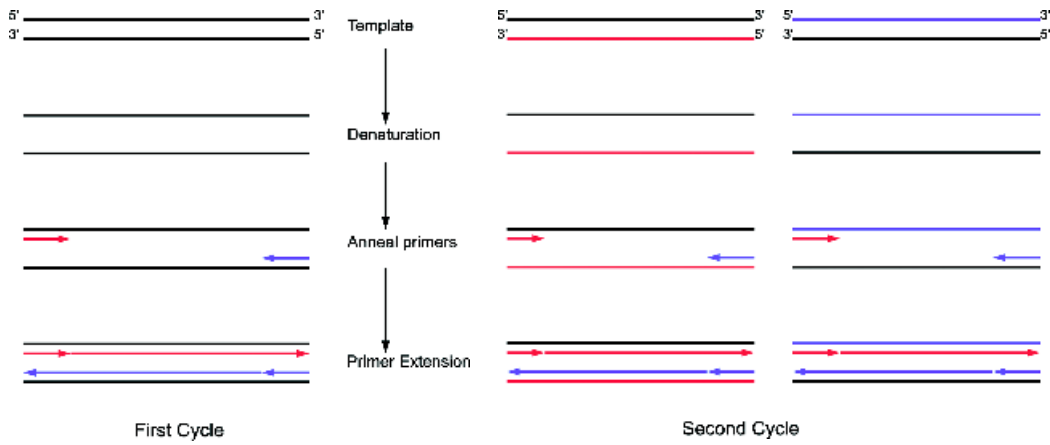

Figure 6.11 The use of PCR-based methods for amplification and replication of template DNA

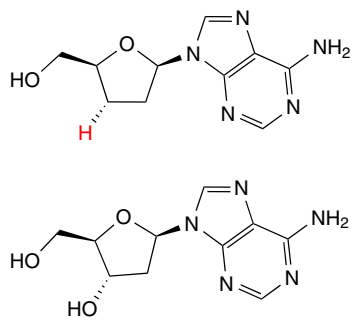

Figure 6.12 The base and sugar components of NTPs. In this case the deoxyadenosine and dideoxyadenosine (note absence of hydroxyl at C3′ position)

C3′ position of the ribose ring. The effect of the missing oxygen is to prevent further elongation of the nucleotide chain in the 5′–3′ direction via an inhibition of phosphodiester bond formation.

When small amounts of the four nucleotides are present as ddNTPs random incorporation will result in chain termination (Figure 6.13). On average the PCR will result in a series of DNA fragments truncated at every nucleotide, each fragment differing from the previous one by just one base in length. Quite clearly if we can establish the identity of the last base we can gradually establish the DNA sequence. The second step of DNA sequencing separates these DNA fragments. To aid identification each ddNTP is also labelled with a fluorescent probe based on the chromophores fluoroscein and rhodamine 6 G. Each of the four dideoxy nucleotides is tagged with a different fluorescent dye that has an emission maxima (λ_{max}) that allows the final base to be discriminated. Dye-labelled DNA fragments are separated according to mass by running through polyacrylamide gels or capillaries. Today the most sophisticated DNA sequencing systems use capillary electrophoresis with the advantages of high separation efficiency, fast separations at high voltages, ease of use with small sample volumes (~1 μl), and high reproducibility. 'Labelled' DNA exits the capillaries and laser-induced fluorescence detected by a charge coupled device (CCD) is interpreted as a fluorescence profile by a computer leading to routine sequencing of over 1000 nucleotides with very low error rates (Figure 6.14).

Sequence homology

Advances in both DNA and protein sequencing have generated enormous amounts of data. This data represents either the order of amino acids along a polypeptide chain or the order of bases along a nucleotide chain and contains within this 'code' significant supplementary information on proteins. With the introduction of powerful computers sequences can be analysed and compared with each other. In particular computational analysis has

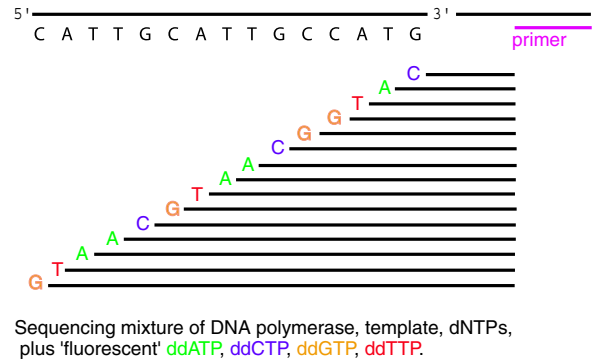

Figure 6.13 Chain termination of an extending DNA sequence by incorporation of dideoxynucleotide triphosphates (ddNTPs)

allowed the recognition of sequence similarity. For any sequence there is a massive number of permutations and similarity does not arise by chance. Instead sequence similarity may indicate evolutionary links and in this context the term homology is used reservedly. Protein sequences can be similar without needing to invoke an evolutionary link but the term homology implies evolutionary lineage from a common ancestor. Both DNA and protein sequences can show homology. In order to establish that protein sequences are homologous we have to establish rules governing this potential similarity. Consider the partial sequences

-D-E-A-L-V-S-V-A-F-T-S-I-V-G-G-

-D-E-A-F-T-S-I-V-G-G-M-D-D-P-G-

This represents a small section of the polypeptide chain (15 residues) in each of two sequences. Are they

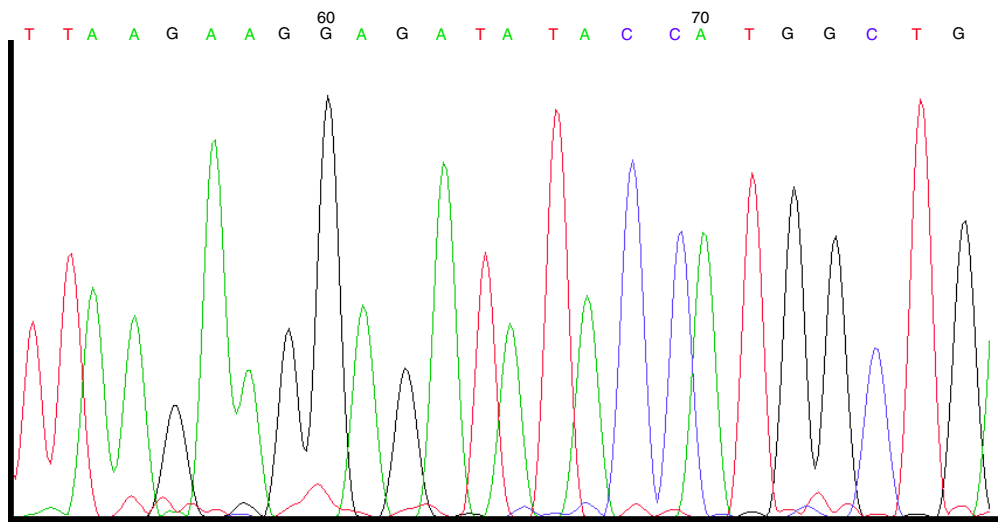

Figure 6.14 A small section of the computer-interpreted fluorescence profile of a DNA sequence. A is denoted by the green trace, C in blue, G in black and T by the red trace. Sequence anomalies arise when 'clumps' of bases are found together, for example, a block of T residues and may require user intervention

similar and if so how similar or are they different? An initial inspection 'by eye' might fail to establish any relationship. However, closer examination would reveal that the two sequences could be 'aligned' by introducing a gap of seven residues into the second sequence. When the sequences are re-written as

D-E-A-L-V-S-V-A-F-T-S-I-V-G-G-
D-E-*-*-*-*-*-*-*-T-S-I-V-G-G-M-D-D-P-G

a good alignment exists for some of the residues and significantly these residues are identical. Clearly to some observers this is significant similarity (8 out of 15 residues are identical) whilst to others it could be viewed as a major difference (7 residues are missing and information is lacking about the similarity of the last 5 residues in the second sequence!). What is needed is a way of quantifying this type of problem.

Alignment of protein sequences is the first step towards quantifying similarity between one or more sequences. As a result of point mutations or larger mutational events sequences change giving proteins containing different residues. This obscures relationships between proteins and one reason for comparing and aligning sequences is to deduce these relationships. For newly determined sequences this allows identification to previously characterized proteins and highlights a shared common origin.

In Chapter 3 the tertiary structures of haemoglobin α and β chains were superimposed to reveal little difference in respective fold (Figure 3.43). This suggested an evolutionary relationship as a result of ancestral gene duplication. Structural homology, supported by significant sequence homology between α and β chains, reveals a level of identity between the two chains of over 40 percent (62/146) (Figure 6.15).

Domains are key features of modular or mosaic proteins. Sequence alignments reveal that gene duplication leads to a proliferation of related domains in different proteins. As a result, proteins are related by the presence of similar domains – one example is the occurrence of the SH3 domain in proteins that share little else in common except the presence of this motif. SH3 (or Src Homology 3) domains are small, non-catalytic, modules of 50–70 residues that mediate protein–protein interactions by binding to proline-rich peptide sequences. The domain was discovered in tyrosine kinase as one module together with SH1 (tyrosine kinase) and SH2 (phosphotyrosine binding). The domain is found in kinases, lipases, GTPases, structural proteins and viral regulatory proteins.

Alignment methods offer a way of pictorially representing similarity between one or more 'test' sequences and a library of 'known' sequences derived from databases. In addition alignment methods offer a route towards quantifying the extent of this similarity by incorporating 'scoring' schemes. A number of different approaches exist for aligning sequences. A prevalent approach is called a 'pairwise similarity' and involves comparing each sequence in the database (library) with a 'test' sequence (Figure 6.16). The observation of 'matches' indicates sequence similarity. A second level of comparison involves comparing families of sequences with libraries to establish relationships (Figure 6.17). This approach establishes 'profiles' for the initial family and then attempts to fit this profile to other members of the database.

A third approach is to use known motifs found within proteins. These motifs are invariant or highly conserved blocks of residues characteristic of a protein family. These motifs are used to search databases for other sequences bearing the corresponding motif.

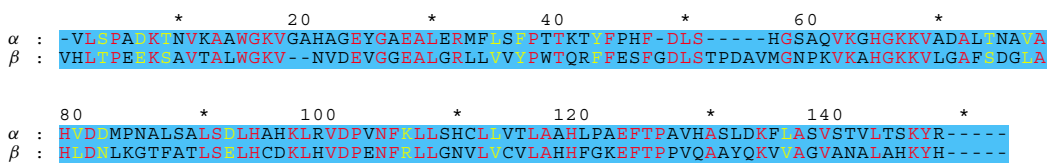

Figure 6.15 The sequence of the α and β chains of haemoglobin. Identical residues are shown in red whilst conserved residues are shown in yellow

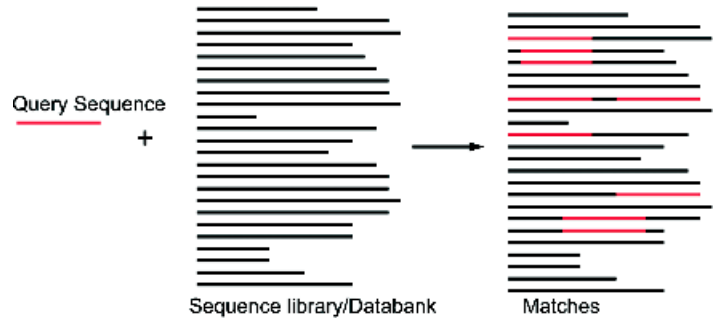

Figure 6.16 Pairwise similarity search of the databank using single 'query' sequence

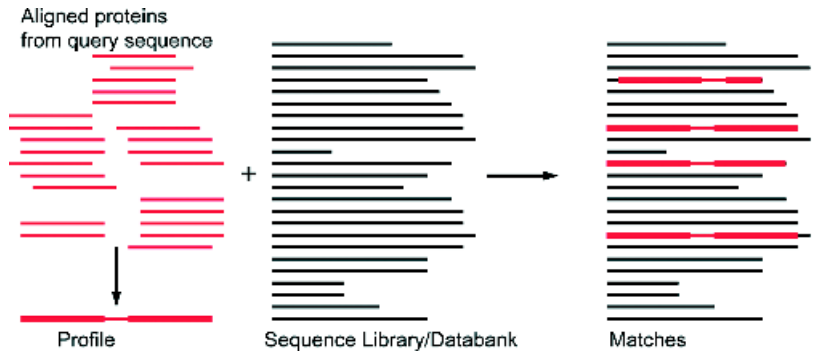

Figure 6.17 The use of a profile established from aligned proteins to improve quality of matches

Table 6.4 A selection of the characteristic motifs used to identify proteins

Motif sequence	Protein family	Example
CX_2CH	Class I soluble c type cytochromes	Bacterial cytochrome c551
$F(Y)L(IVMK)X_2HPG(A)G$	Cytochrome b_5 family	Nitrate reductase
$CX_7L(FY)X_6F(YW)XR(K)X_8CXCX_6C$	Ribosomal proteins	L3 protein
$A(G)X_4GKS(T).$	ATPases	ATP synthetase
$CX_2CX_3L(IVMFYWC)X_8HX_3H$	Zn finger-DNA binding proteins	TFIIIA
LKEAExRAE	Tropomyosin family	Tropomyosin

Motifs are simply short sequences of amino acid residues within a polypeptide chain that facilitate the identification of related proteins. The residues in parenthesis are alternative residues that make up the consensus motif. The X indicates the number of intervening residues lacking any form of consensus. The number of intervening residues can show minor variations in length.

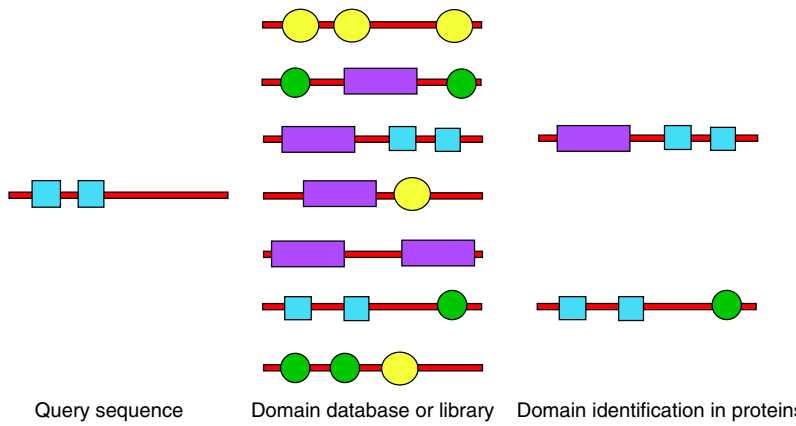

Figure 6.18 A query sequence is used with a domain-based database to identify similar proteins containing recognized domains

Diagnostic motifs occur regularly in proteins with whole databases devoted to their identification (Table 6.4). Many protein families have conserved sequence motifs and it is often sufficient to 'query' the database only with this motif to identify related proteins (Figure 6.18).

The best forms of sequence alignment are derived via computational methods known as dynamic programming that detect optimal pairwise alignment between two or more proteins. The details of computational programming are beyond the scope of this book but the procedure compares two or more sequences by looking for individual characters, or a series of character patterns, that are in the same order in each sequence. Identical or similar characters are placed in the same column, and non-identical characters can be placed in the same column either as a mismatch or opposite a 'gap' in the other sequences. Irrespective of the approach used to derive alignments a query sequence is compared with all sequences within a selected database to yield a 'score' that indicates a level of similarity. An optimal alignment will result in the arrangement of non-identical characters and gaps minimized to yield identical or similar characters as a vertical register. The art of these methods is to create a feasible scoring scheme, and the identity matrix or matrices based on physicochemical properties often fail to establish relationships, even for proteins known to be related.

The most important improvement in scoring schemes came from methods based on the observed changes in amino acid sequence in homologous proteins. This allowed mutation rates to be derived from their evolutionary distances and was pioneered by Margaret Dayhoff in the 1970s. The study measured the frequencies with which residues are changed as a result of mutation during evolution and involved carefully aligning (by eye) all proteins recognized to be within a single family. The process was repeated for different families and used to construct phylogenetic trees for groups of proteins. This approach yielded a table of relative frequencies describing the rate of residue replacement by each of the nineteen other residues over an evolutionary period. By combining this table with the relative frequency of occurrence of residues in proteins a family of scoring matrices were computed known as point accepted mutation (PAM) matrices. The PAM matrices are based on estimated mutation rates from closely related proteins and are effective at 'scoring' similarities between sequences that diverged with evolutionary time. In the PAM 250 matrix the data reflects aligned protein sequences extrapolated to a level of 250 amino acid replacements per 100 residues per 100 million years. A score above 0 indicates that amino acids replace each other more frequently than expected from their distribution in proteins and this usually means that such residues are functionally equivalent. As expected mutations involving the substitution of D > E or D > N are relatively common whilst transitions such as D > R or D > L are rare.

An alternative approach is the BLOCKS database based on ungapped multiple sequence alignments that correspond to conserved regions of proteins. These 'blocks' were constructed from databases of families of related proteins (such as Pfam, ProDom, InterPro or Prosite). This has yielded approximately 9000 blocks representing nearly 2000 protein families. The BLOCKS database is the basis for the BLOSUM substitution matrices that form a key component of common alignment programs such as BLAST, FASTA, etc. (BLOSUM is an acronym of BLOcks SUbstitution Matrices.) These substitution matrices are widely used for scoring protein sequence alignments and are based on the observed amino acid substitutions in a large set of approximately 2000 conserved amino acid patterns, called blocks. The blocks act as identifiers of protein families and are based around a greater dataset than that used in the PAM matrices. The BLOSUM matrices detect distant relationships and produce alignment that agrees well with subsequent determination of tertiary structure. In general if a test sequence shares 25–30 percent identity with sequences in the database it is likely to represent a homologous protein. Unfortunately, when the level of similarity falls below this level of identity it proves difficult to draw firm conclusions about homology.

Structural homology arising from sequence similarity

With large numbers of potential sequences for proteins the number of different folded conformations might also be expected to be large. However, the tertiary or folded conformations of proteins are less diverse than would be expected from the total number of sequences. Protein folds have been conserved during evolution with conservation of structure occurring despite changes in primary sequence.

In most cases structural similarities arise as a result of sequence homology. However, in a few instances structural homology has been observed where there is no obvious evolutionary link. One family of proteins that clearly indicates both sequence and structural homology is the cytochrome c family. Vertebrate cytochrome c isolated from mitochondria, such as those from horse and tuna, share very similar primary sequences (only 17 out of 104 residues are different) and this is emphasized by comparable tertiary structures (Figure 6.20). The conservation of structure is expected as both horse and tuna proteins occupy similar functional roles. Yeast cytochrome c also exhibits homology and a similar tertiary structure but a reduced level of sequence identity to either horse or tuna cytochrome c (~59 out of 104 residues are identical) reflects a more distant evolutionary lineage (hundreds of millions of years). In bacteria a wide range of *c* type cytochromes are known all containing the heme group covalently ligated to the polypeptide via two thioether bridges derived from cysteine residues. The proteins are soluble, contain between 80 and 130 residues, and function as redox carriers.

If the sequences of cytochrome c_2 from *Rhodobacter rubrum*, cytochrome c-550 from *Paracoccus denitrificans* and the mitochondrial cytochromes c from yeast, tuna and horse are compared only 18 residues remain invariant (Figure 6.19). These residues include His

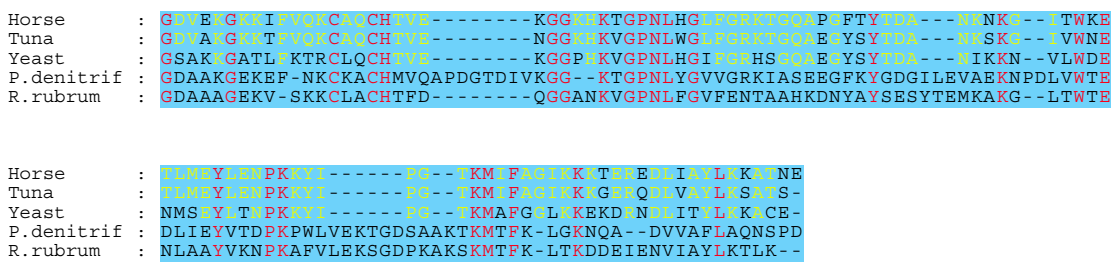

Figure 6.19 Sequence homology between the cytochromes c of horse, tuna, yeast, *R. rubrum* and *P. denitrificans*. Shown in red are identical residues between the sequences whilst yellow residues highlight those residues that are conserved within the horse, tuna and yeast eukaryotic sequences. Only 18 out of 104 residues show identity

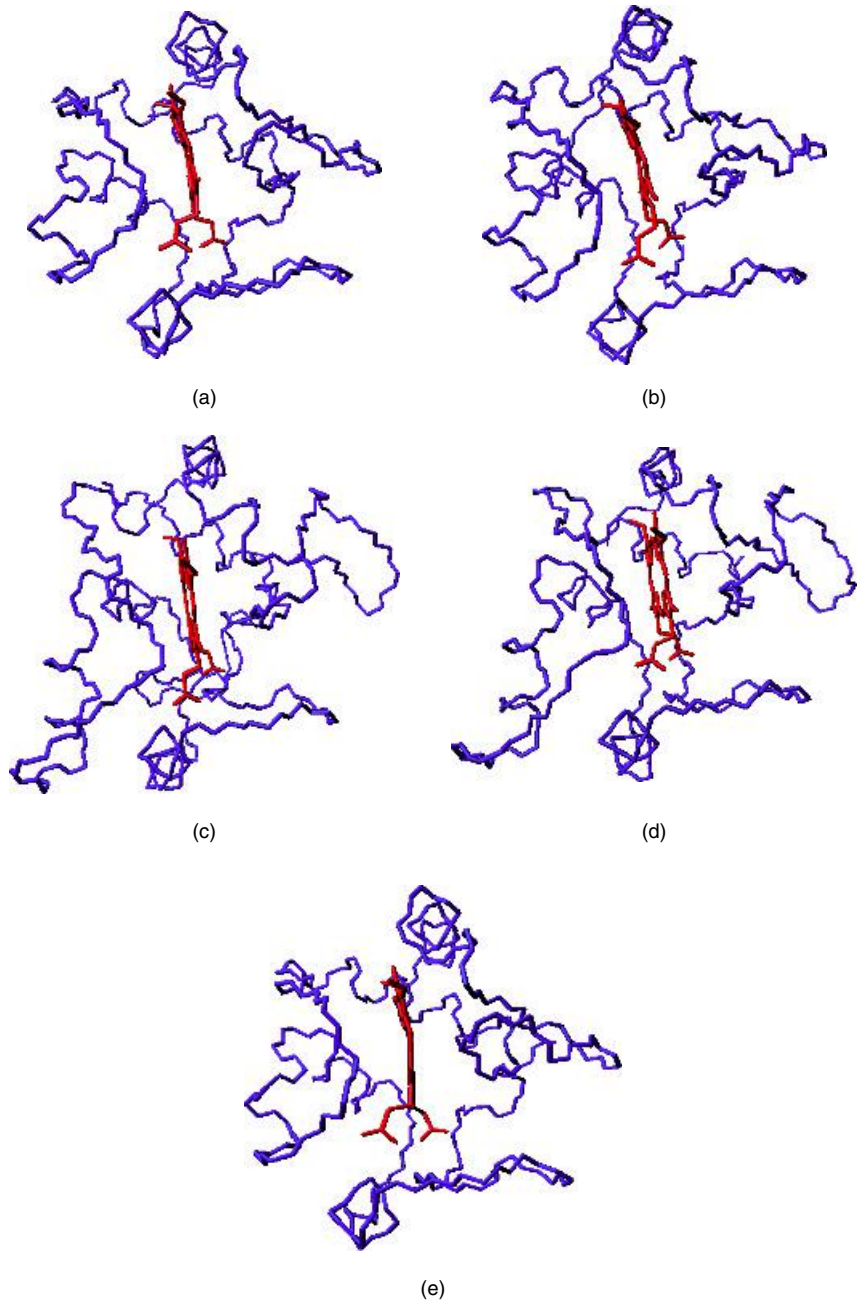

Figure 6.20 The structures of cytochrome c from different source organisms. The structures shown for five cytochrome c were obtained from (a) tuna (PDB: 3CYT), (b) yeast – the iso-1 form (PDB:1YCC), (c) *P. denitrificans* (PDB:155C), (d) *R. rubrum* cytochrome c_2 (PDB:1C2R), and (e) horse (PDB:1CRC). The insertions of residues in the bacterial sequences can be seen in the additional backbone structure shown in the bottom left region of the molecule

18, Met80, Cys14 and Cys17 and are critical to the functional role of cytochrome c in electron transfer. Consequently their conservation is expected. There is very little sequence similarity between cytochrome c_2 and horse cytochrome c but an evolutionary lineage is defined by tracking progressive changes in sequence through micro-organisms, animal and plants. In this manner it is clear that cytochrome c represents the systematic evolution of a protein designed for biological electron transfer from a common ancestral protein. A more thorough analysis of the sequences reveals that although changes in residue occur at many positions the majority of transitions involve closely related amino acid residues. For example, near the C terminus of each cytochrome c a highly conserved Phe residue is usually found but in the sequence of *P. denitrificans* this residue is substituted with Tyr.

Irrespective of the changes in primary sequence for these cytochromes c considerable structural homology exists between all of these proteins. This homology extends from the protein found in the lowliest prokaryote to that found in man. It is only by comparing intermediate sequences between cytochrome c_2 and horse cytochrome c that evolutionary links are established but structural homology is a strong indicator that the proteins are related. As a caveat although structural homology is normally a good indicator of relationships it is not *always* true (see below) and must be supported by sequence analysis.

The serine proteases are another example of structural homology within an evolutionarily related group of proteins. This family of enzymes includes familiar proteins from higher organisms such as trypsin, chymotrypsin, elastase and thrombin and they have the common function of proteolysis. If the sequences of trypsin, chymotrypsin and elastase are compared (Figure 6.21) they exhibit an identity of ~40 percent and this is comparable to the identity shown between the haemoglobin α and β chains. The three-dimensional structure of all of these enzymes are known and they share similar folded conformations with invariant His and Ser residues equivalent to positions 57 and 195 in chymotrypsin located in the active sites along with the third important residue of Asp 102. Together these residues make up the catalytic triad and from their relative positions other members of the family of serine proteases have been identified. The similarity in sequence and structure between chymotrypsin, elastase and trypsin indicates that these proteins arose from gene duplication of an ancestral protease gene with subsequent evolution accounting for individual differences.

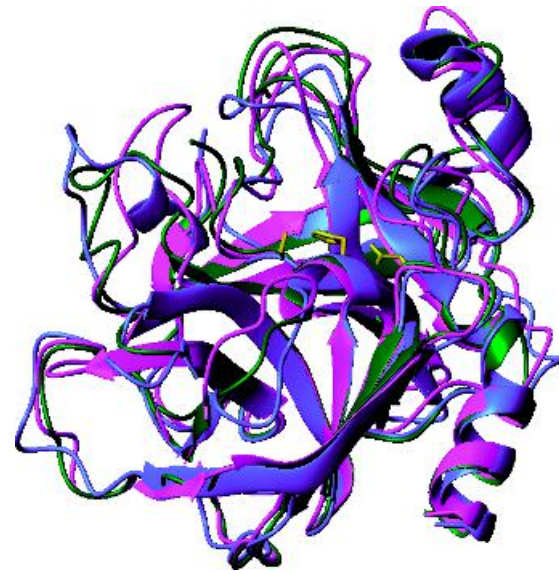

Figure 6.21 The structures of chymotrypsin, trypsin and elastase shown with elements of secondary structure superimposed. Chymotrypsin (blue, PDB: 2CGA), elastase (magenta, PDB: 1QNJ) and trypsin (green, PDB: 1TGN). Shown in yellow in a cleft or active site are the catalytic triad of Ser, His and Asp (from left to right) that are a feature of all serine protease enzymes

By following changes in sequence for different proteins in a family an average mutation rate is calculated and reveals that 'house-keeping' proteins such as histones, enzymes catalysing essential metabolic pathways, and proteins of the cytoskeleton evolve at very slow rates. This generally means the sequences incorporate between 1 and 10 mutations per 100 residues per 100 million years. Consequently to obscure all evolutionary information, which generally requires ~250–350 substitutions per 100 residues, takes a considerable length of time. As a result of this slow

Table 6.5 Rate of evolution for different proteins (adapted from Wilson, A.C. *Ann. Rev. Biochem.* 1977, 46, 573–639)

Protein	Accepted point mutations/100 residues/10^8 years	Protein	Accepted point mutations/100 residues/10^8 years
Histone H2	0.25	Insulin	7
Collagen α_1	2.8	Glucagon	2.3
Cytochrome c	6.7	Triose phosphate isomerase	5.3
Cytochrome b_5	9.1	Lactate dehydrogenase M chain	7.7
Lysozyme	40	α-lactalbumin	43
Ribonuclease A	43	Immunoglobulin V region	125
Myoglobin	17	Haemoglobin α	27
Histone H1	12	Haemoglobin β	30

rate of evolution 'house-keeping' proteins are excellent tools with which to trace evolutionary relationships over hundreds of millions of years. Higher rates of evolution are seen in proteins occupying less critical roles.

Mutations arising in DNA are not always converted into changes in protein primary sequence. Some mutations are silent due to the degeneracy inherent in the genetic code. A nonsense mutation results from the insertion of a stop codon within the open reading frame of mRNA and gives a truncated polypeptide chains. The generation of a stop codon close to the start codon will invariably lead to a loss of protein activity whilst a stop codon located relatively near to the original 'stop' sight may well be tolerated by the protein. Missense mutations change the identity of a residue by altering bases within the triplet coding for each amino acid. From the standpoint of evolutionary analysis it is these mutations that are detected *via* changes in primary sequences.

Proteins with different structures and functions evolve at significantly different rates. This is seen most clearly by comparing proteins found in *Homo sapiens* and *Rhesus* monkeys. For cytochrome c the respective primary sequences differ by less than 1 percent of their residues but for the α and β chains of haemoglobin these differences are at a level of 3–5 percent whilst for fibrinopeptides involved in blood clotting the differences are ~30 percent of residues. This emphasizes the need to use families of proteins to extrapolate rates of protein evolution. Systematic studies have determined rates of evolution for a wide range of proteins of different structure and function (Table 6.5) with mutation rates expressed as the number of point mutations per 100 residues per 10^8 years.

Conotoxins are a family of small peptides derived from the *Conus* genus of predatory snails. These snails have a proboscis containing a harpoon-like organ that is capable of injecting venom into fish, molluscs and other invertebrates causing rapid paralysis and death. The venom contains over 75 small toxic peptides that are between 13 and 35 residues in length and disulfide rich. The toxins have been purified and exhibit varying degrees of toxicity on the acetylcholine receptor of vertebrate neurones. The peptides represent a molecular arms race whereby rapid evolution has allowed *Conus* to develop 'weapons' in the form of toxins of different sequence and functional activity. As potential prey adapt to the toxins *Conus* species evolve new toxins based around the same pattern. Whereas proteins such as histones show evolutionary rates of change of ~0.25 point mutations/100 residues/10^8 years the conotoxins have much higher rates of change estimated at ~60–180 point mutations/100 residues/10^8 years. This pattern is supported by analysis of other toxins such as those from snake venom where typical mutation rates are ~100 point mutations/100 residues/10^8 years.

The rapid development of molecular cloning techniques, DNA sequencing methods, sequence comparison algorithms and powerful yet affordable computer workstations has revolutionized the importance of protein sequence data. Thirty years ago protein sequence determination was often one of the final steps in the characterization of a protein, whilst today one premise of the human genome mapping project is that sequencing all of the genes found in man will uncover their function via data analysis. There is no doubt that this premise is beginning to yield rich rewards with the identification of new homologues as well as new open reading frames (ORFs), but in many cases sequence data has remained impervious to analysis.

The above examples highlight structural homology that has persisted from a common ancestral protein despite the slow and gradual divergence of protein sequences by mutation. Occasionally structural homology is detected where there is no discernable relationship between proteins. This is called convergent evolution and arises from the use of similar structural motifs in the absence of sequence homology. In subtilisin and serine carboxypeptidase II a catalytic triad of Ser-Asp-His is observed (Figure 6.22) that might imply a serine protease but the arrangement of residues within their primary sequences are different to chymotrypsin and these proteins differ in overall structure. It is therefore very unlikely that these three proteins could have arisen from a common ancestor of chymotrypsin or a related serine protease since the order

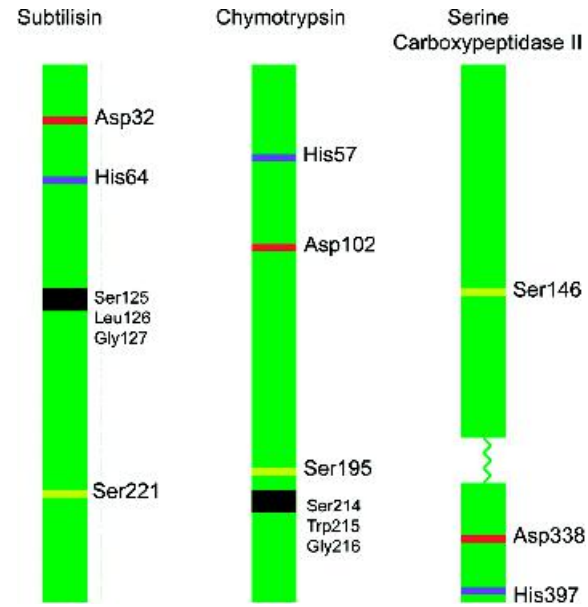

Figure 6.22 The positions of the catalytic triad together with hydrophobic residues involved in substrate binding in unrelated serine proteases. Subtilisin, chymotrypsin and serine carboxypeptidase II

Table 6.6 Potential examples of convergent evolution in proteins

Protein 1	Protein 2	Functional motif
Chymotrypsin	Subtilisin Carboxypeptidase II	Catalytic triad of Ser/Asp/His residues in the proteolysis of peptides
Triose phosphate isomerase	As many as 17 different groups of enzymes possess β barrel. Luciferase, pyruvate kinase and ribulose 1,5 bisphosphate carboxylase-oxygenase (rubisco)	TIM barrel fold of eight parallel β strands forming a cylinder connected by eight helices arranged in an outer layer of the protein
Carbonic anhydrase (α)	Carbonic anhydrase (β) form plants.	Zn ion ligated to protein and involved in catalytic conversion of CO_2 to HCO_3^-
Thermolysin	Carboxypeptidase A	Zn ion ligated to two imidazole side chains and the carboxyl side chain of Glu. Also coordinated by water molecule playing a crucial role in proteolytic activity.

of residues in the catalytic triad differs as do other structural features present in each of the enzymes. This phenomenon is generally viewed as an example of convergent evolution where nature has discovered the same catalytic mechanism on more than one occasion (see Table 6.6).

The term convergent evolution has been applied to the Rossmann fold found in nucleotide binding domains consisting of three β strands interspersed with two α helices with a common role of binding ligands such as NAD^+ or ADP. In many proteins these structural elements are identified, but little sequence homology exists between nucleotide-binding proteins. The sequences reflect the production of comparable topological structures for nucleotide-binding domains by different permutations of residues. Are these domains diverged from a common ancestor or do they result from convergent evolution? The large number of domains found in proteins capable of binding ATP/GTP or NAD/NADP suggests that it is unlikely that proteins have frequently and independently evolved a nucleotide-binding motif domain. It is now thought more likely that the Rossmann fold evolved with numerous variations arising as a result of divergent evolution from a common and very distant ancestral protein. The β barrel exemplified by triose phosphate isomerase is also widely found in proteins and may also represent a similar phenomenon.

Protein databases

A number of important databases are based on the sequence and structure information deposited and archived in the Protein Data Bank. These databases attempt to order the available structural information in a hierarchical arrangement that is valuable for an analysis of evolutionary relationships as well as enabling functional comparisons. In the SCOP (Structural Classification of Proteins) database all of the deposited protein structures are sorted according to their pattern of folding. The folds are evaluated on the basis of their arrangement and in particular the composition and distribution of secondary structural elements such as helices and strands. Frequently construction of this hierarchical system is based on manual classification of protein folds and as such may be viewed as subjective. Currently, non-subjective methods are being actively pursued in other databases in an attempt to remove any bias in classification of protein folds. The levels of organization are hierarchical and involve the classes of folds, superfamilies, families and domains (the individual proteins).

Domains within a family are homologous and have a common ancestor from which they have diverged. The homology is established from either sequence and/or functional similarity. Proteins within a superfamily have the same fold and a related function and therefore also probably have a common ancestor. However, within a superfamily the protein sequence composition or function may be substantially different leading to difficulty in reaching a conclusive decision about evolutionary relationships. At the next level, the fold, the proteins have the same topology, but there is no evidence for an evolutionary relationship except a limited structural similarity.

The CATH database attempts to classify protein folds according to four major hierarchical divisions; Class, Architecture, Topology and Homology (see Figure 6.23 for an example). It also utilizes algorithms to establish definitions of each hierarchical division. Class is determined according to the secondary structure composition and packing within a protein structure. It is assigned automatically for most structure (>90 percent) with manual inspection used for 'difficult' proteins. Four major classes are recognized; mainly α, mainly β and α–β protein domains together with domains that have a very low secondary structure content. The mixed α–β class can be further divided into domains with alternating α/β structures and domains with distinct α rich and β rich regions (α + β). The architecture of proteins (A-level hierarchy) describes the overall shape of the domain and is determined by the orientations of the individual secondary structural elements. It ignores the connectivity between these secondary structures and is assigned manually using descriptions of secondary structure such as β barrel or β–α–β sandwich. Several well-known architectures have been described in this book and include the β propellor, the four-helix bundle, and the helix-turn-helix motif. Structures are grouped into fold families or topologies at the next

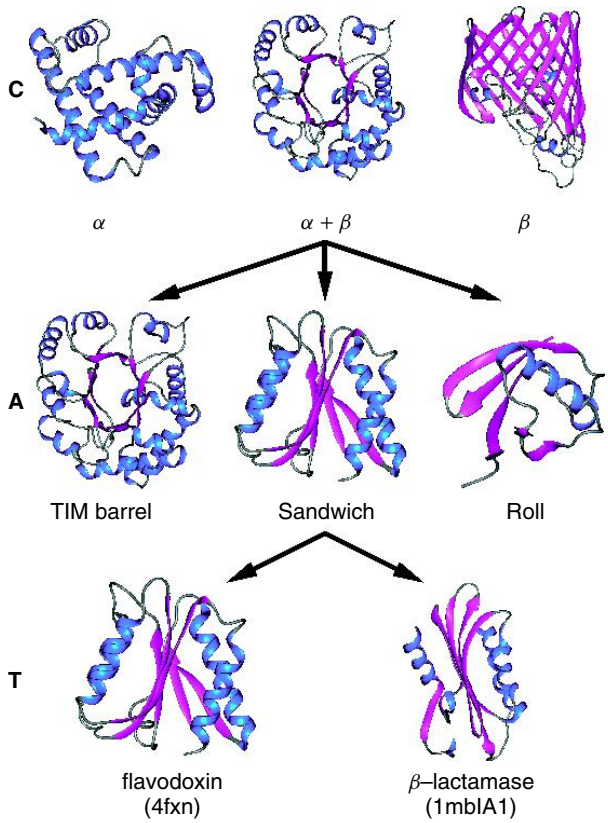

Figure 6.23 The distinguishing of flavodoxin and β lactamase; two α/β proteins based around a β sandwich structure

level of organization. Assignment of topology depends on the overall shape and connectivity of the secondary structures and has started to be automated via algorithm development. A number of topologies or fold families have been identified, with some, such as the β two-layer sandwich architecture and the α–β three-layer sandwich structures, being relatively common. The H level of classification represents the homologous superfamily and brings together protein domains that share a common ancestor. Similarities are identified first by sequence comparisons and subsequently by structural comparisons. Hierarchical levels of organization are shown at the C, A and T states for α/β containing proteins.

Gene fusion and duplication

The preceding sections highlight that some proteins share similar sequences reflected in domains of similar structure. The α + β fold of cytochrome b_5 first evolved as an electron transfer protein and then became integrated as a module found in larger multi-domain proteins. The result was that ligated to a hydrophobic tail the protein became a membrane-bound component of the endoplasmic reticulum fatty acid desaturase pathway. When this domain was joined to a Mo-containing domain an enzyme capable of converting sulfite into sulfate was formed, sulfite oxidase. Joining the cytochrome to a flavin-containing domain formed the enzyme, nitrate reductase. Proteins sharing homologous domains arose as a result of gene duplication and a second copy of the gene. This event is advantageous to an organism since large genetic variation can occur in this second copy without impairing the original gene.

The globin family of proteins has arisen by gene duplication. Haemoglobin contains 2α and 2β chains and each α and β chain shows homology and is similar to myoglobin. This sequence homology reflects evolutionary origin with a primitive globin functioning simply as an oxygen storage protein like myoglobin. Subsequent duplication and evolution allowed the subtle properties of allostery as well as the differences between the α and β subunits. During embryogenesis other globin chains are observed (ξ and ε chains) and fetal haemoglobin contains a tetramer made up of $\alpha_2\gamma_2$ subunits that persists in adult primates at a level of ~1 percent of the total haemoglobin. The δ chain is homologous to the β subunit and leads to the evolution of the globin chains in higher mammals as a genealogical tree, with each branch point representing gene duplication that gives rise to further globin chains (Figure 6.24).

Secondary structure prediction

One of the earliest applications of using sequence information involved predicting elements of secondary structure. If sufficient numbers of residues can be placed in secondary structure then, in theory at least, it is possible to generate a folded structure based on

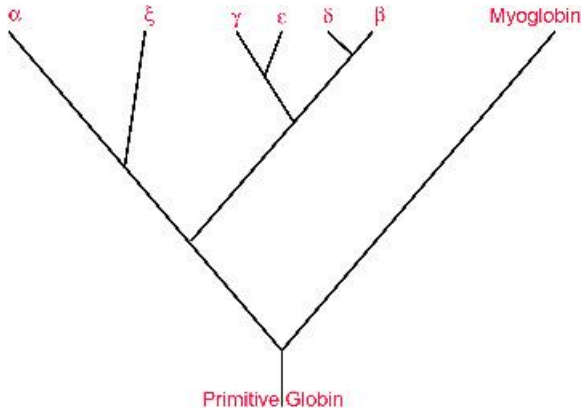

Figure 6.24 Evolution of the globin chains in primates. Each branch point represents a gene duplication event with myoglobin evolving early in the evolutionary history of globin subunit

Table 6.7 Propensity for a given amino acid residue to be found in helix, strand or turn

Residue	P_α	P_β	P_t	Residue	P_α	P_β	P_t
Ala	1.41	0.72	0.82	Leu	1.34	1.22	0.57
Arg	1.21	0.84	0.90	Lys	1.23	0.84	0.90
Asn	0.76	0.48	1.34	Met	1.30	1.14	0.52
Asp	0.99	0.39	1.24	Phe	1.16	1.33	0.65
Cys	0.66	1.40	0.54	Pro	0.34	0.31	1.34
Gln	1.27	0.98	0.84	Ser	0.57	0.96	1.22
Glu	1.59	0.52	1.01	Thr	0.76	1.17	0.90
Gly	0.43	0.58	1.77	Trp	1.02	1.35	0.65
His	1.05	0.80	0.81	Tyr	0.74	1.45	0.76
Ile	1.09	1.67	0.47	Val	0.90	1.87	0.41

(After Chou, P.Y. & Fasman, G.D. *Ann. Rev. Biochem* 1978, **47**, 251–276; and Wilmot, C.M. & Thornton, J.M. *J. Mol. Biol.* 1988, **203**, 221–232).
A value of 1.41 observed for alanine in helices implies that alanine occurs 41 percent more frequently than expected for a random distribution

the packing together of helices, turns and strands. Secondary structure prediction methods are often based on the preference of amino acid residues for certain conformational states.

One algorithm for secondary structure prediction was devised by P.Y. Chou and G.D. Fasman, and although superior methods and refinements exist today it proved useful in defining helices, strands and turns in the absence of structural data. It was derived from databases of known protein structures by estimating the frequency with which a particular amino acid was found in a secondary structural element divided by the frequency for all other residues. A value of 1 indicates a random distribution for a particular amino acid whilst a value greater than unity suggests a propensity for finding this residue in a particular element of secondary structure. A series of amino acid preferences were established (Table 6.7).

The numbers in the first three columns, P_α, P_β P_t, are equivalent to preference parameters for the 20 amino acids for helices, strands and reverse turns respectively. From this list one can assign the residues on the basis of the preferences; Ala, Arg, Glu, Gln, His, Leu, Lys, and Met are more likely to be found in helices; Cys, Ile, Phe, Thr, Trp, Tyr and Val are likely to be found in strands while Asn, Asp, Gly, Pro and Ser are more frequently located in turns. The algorithm of Chou and Fasman pre-dated the introduction of inexpensive computers but contained a few simple steps that could easily be calculated (Table 6.8). In practice the Chou–Fasman algorithm has a success rate of ~50–60 percent, although later developments have succeeded in reaching success rates above 70 percent. A major failing of the Chou–Fasman algorithm is that it considers only local interactions and neglects long-range order known to influence the stability of secondary structure. In addition the algorithm makes no distinction between types of helices, types of turns or orientation of β strands.

The availability of large families of homologous sequences (constructed using the algorithms described above) has revolutionized methods of secondary structure prediction. Traditional methods, when applied to a family of proteins rather than a single sequence have proved much more accurate in identifying secondary structure elements. Today the combination of sequence data with sophisticated computing techniques such as neural networks has given accuracy levels in excess of 70 percent. Besides the use of advanced computational techniques recent approaches to the problem of secondary structure prediction have focussed on the inclusion of additional constraints and parameters to assist precision. For example, the regular periodicity of

Table 6.8 Algorithm of Chou and Fasman

Step	Procedure
1	Assign all of the residues in the peptide the appropriate set of parameters
2	Scan through the peptide and identify regions where four out of 6 consecutive residues have $P_\alpha > 1.0$ This region is declared to be an α helix
3	Extend the helix in both directions until a set of four residues yield an average $P_\alpha < 1.00$. This represents the end of the helix
4	If the segment defined in step 3 is longer than 5 residues and the average value of $P_\alpha > P_\beta$ for this segment it can also be assigned as a helix
5	Repeat this procedure to locate all helical regions
6	Identify a region of the sequence where 3 out of 5 residues have a value of $P_\beta > 1.00$. These residues are in a β strand
7	Extend the sheet in both directions until a set of four contiguous residues yielding an average $P_\beta < 1.00$ is reached. This represents the end of the β strands
8	Any segment of the region located by this procedure is assigned as a β-strand if the average $P_\beta > 1.05$ and the average value of $P_\beta > P_\alpha$ for the same region
9	Any region containing overlapping helical and strand assignments are taken to be helical if the average $P_\alpha > P_\beta$ for that region. If the average $P_\beta > P_\alpha$ for that region then it is declared a β strand

the α helix of 3.6 residues per turn means that in proteins many regular helices are amphipathic with polar residues on the solvent side and non-polar residues facing the inside of the protein. Recognition of a periodicity of i, i + 3, i + 4, i + 7, etc. in hydrophobic residues has proved particularly effective in predicting helical regions in membrane proteins.

Genomics and proteomics

The completion of the human genome sequencing project has provided a large amount of data concerning the number and distribution of proteins. It seems likely that the human genome contains over 25 000 different polypeptide chains, and most of these are currently of unknown structure and function. Genomics reflects the wish to understand more about the complexity of living systems through an understanding of gene organization and function. Advances in genomics have provided information on the number, size and composition of proteins encoded by the genome. It has highlighted the organization of genes within chromosomes, their homology to other genes, the presence of introns together with the mechanism and sites of gene splicing and the involvement of specific genes in known human disease states. The latter discovery is heralding major advances in understanding the molecular mechanisms of disease particularly the complex interplay of genetic and environmental factors. Genomics has stimulated the discovery and development of healthcare products by revealing thousands of potential biological targets for new drugs or therapeutic agents. It has also initiated the design of new drugs, vaccines and diagnostic DNA kits. So, although genomic based therapeutic agents include traditional 'chemical' drugs, we are now seeing the introduction of protein-based drugs as well as the very exciting and potentially beneficial approach of gene therapy.

However, as the era of genomics reaches maturity it has expanded from a simple definition referring to the mapping, sequencing, and analysis of genomes to include an emphasis on genome function. To reflect this shift, genome analysis can now be divided into 'structural genomics' and 'functional genomics'. Structural genomics is the initial phase of genome analysis with the end point represented by the high-resolution genetic maps of an organism embodied

by a complete DNA sequence. Functional genomics refers to the analysis of gene expression and the use of information provided by genome mapping projects to study the products of gene expression. In many instances this involves the study of proteins and a major branch of functional genomics is the new and expanding area of proteomics.

Proteomics is literally the study of the proteome via the systematic and global analysis of *all* proteins encoded within the genome. The global analysis of proteins includes specifically an understanding of structure and function. In view of the potentially large number of polypeptides within genomes this requires the development and application of large-scale, high throughput, experimental methodologies to examine not only vast numbers of proteins but also their interaction with other proteins and nucleic acids. Proteomics is still in its infancy and current scientific research is struggling with developing methods to deal with the vast amounts of information provided by the genomic revolution. One approach that will be used in both genomic and proteomic study is statistical and computational analysis usually called bioinformatics.

Bioinformatics

Bioinformatics involves the fusion of biology with computational sciences. Almost all of the areas described in the previous sections represent part of the expanding field of bioinformatics. This new and rapidly advancing subject allows the study of biological information at a gene or protein level. In general, bioinformatics deals with methods for storing, retrieving and analysing biological data. Most frequently this involves DNA, RNA or protein sequences, but bioinformatics is also applied to structure, functional properties, metabolic pathways and biological interactions. With a wide range of application and the huge amount of 'raw' data derived from sequencing projects and deposited in databanks (Table 6.9) it is clear that bioinformatics as a major field of study will become increasingly important over the next few decades.

Bioinformatics uses computer software tools for database creation, data management, data storage, data mining and data transfer or communication. Advances in information technology, particularly the use of the Internet allows rapid access to increasing amounts of biological information. For example, the sequenced genomes of many organisms are widely available at multiple web sites throughout the world. This allows researchers to download sequences, to manipulate them or to compare sequences in a large number of different ways.

Central to the development of bioinformatics has been the explosion in the size, content and popularity

Table 6.9 A selection of databases related to protein structure and function; some have been used frequently throughout this book to source data

Database/repository/resource	URL (web address)
The Protein DataBank.	http://www.rcsb.org/pdb
Expert Protein Analysis System (EXPASY)	http://www.expasy.ch/
European Bioinformatics Institute	http://www2.ebi.ac.uk
Protein structure classification (CATH)	http://www.biochem.ucl.ac.uk/bsm/cath
Structural classification of proteins (SCOP)	http://scop.mrc-lmb.cam.ac.uk/scop
Atlas of proteins side chain interactions.	http://www.biochem.ucl.ac.uk/bsm/sidechains
Human genome project (a tour!)	http://www.ncbi.nlm.nih.gov/Tour
An online database of inherited diseases	http://www.ncbi.nlm.nih.gov/Omim
Restriction enzyme database	http://rebase.neb.com
American Chemical Society	http://pubs.acs.org

of the world wide web, the most recognized component of the internet.

Initially expected to be of use only to scientists the world wide web is a vast resource allowing data transfer in the form of pages containing text, images, audio and video content. Pages, linked by pointers, allow a computer on one side of the world to access information anywhere in the world via series of connected networks. The pointers refer to URL's or 'uniform resource locators' and are the basic 'sites' of information. So, for example, the protein databank widely referred to in this book has a URL of http://www.rcsb.org. The 'http' part of a web address simply refers to the method of transferring data and stands for hypertext transfer protocol. Hypertext is the language of web pages and all web pages are written in a specially coded set of instructions that governs the appearance and delivery of pages known as hypertext markup language (HTML). Finally, the web pages are made comprehensible (interpreted) by 'browsers' – software that reads HTML and displays the content. Browsers include Internet Explorer, Netscape and Opera and all can be used to view 'online content' connected with this book. In 10 years the web has become a familiar resource.

The development of computers has also been rapid and has seen the introduction of faster 'chips', increased memory and expanded disk sizes to provide a level of computational power that has enabled the development of bioinformatics. In a trite formulation of a law first commented upon by Gordon Moore, and hence often called Moore's law, the clock speed (in MHz) available on a computer will double every 18 months. So in 2003 I am typing this book on a 1.8 GHz-based computer. Although computer speeds will undoubtedly increase there are grounds for believing rates of increase will slow. Searching and manipulating large databases, coupled with the use of complex software tools to analyse or model data, places huge demands on computer resources. It is likely that in the future single powerful computer workstations with a very high clock speed, enormous amounts of memory and plentiful storage devices will become insufficient to perform these tasks. Instead bioinformatics will have to harness the power of many computers working either in parallel or in grid-like arrays. These computers cumulatively will allow the solution to larger and more complex problems. However, today bioinformatics is an important subject in its own right that directs biological studies in directions likely to result in favourable outcomes.

One of the results of genome sequencing projects is that software tools can be developed to compare and contrast genetic information. One particular avenue that is being pursued actively at the moment is the prediction of protein structure from only sequence information. The determination of three dimensional protein structures is an expensive, time consuming and formidable task usually involving X-ray crystallography or NMR spectroscopy. Consequently any method that allows a 'by-pass' of this stage is extremely attractive especially to pharmaceutical companies where the prime objective is often product development. One approach is to compare a protein sequence to other proteins since sequential homology (identity >25 percent) will always be accompanied by similar topology. If the structure of a sequentially homologous protein is known then the topology of the new protein can be deduced with considerable certainty. A more likely scenario, however, is that the sequence may show relatively low levels of identity and one would like to know how similar, or different, the three dimensional structures might be to each other. This is a far more difficult problem. Comparative modelling will work when sequentially homologous proteins are compared but may fail when the levels of identity fall below a benchmark of $\sim$25 percent. Additionally some proteins show very similar structures with very low levels of sequence homology.

A second approach is the technique of 'threading'. This approach attempts to compare 'target sequences' against libraries of known structural templates. A comparison produces a series of scores that are ranked and the fold with the best score is assumed to be the one adopted by the unknown sequence. This approach will fail when a new 'fold' or tertiary structure is discovered and it relies on a representative database of structures and sequences.

The *ab initio* approach (Figure 6.25) ignores sequence homology and attempts to predict the folded state from fundamental energetics or physicochemical properties associated with the constituent residues. This involves modelling physicochemical parameters in terms of force fields that direct the folding of the

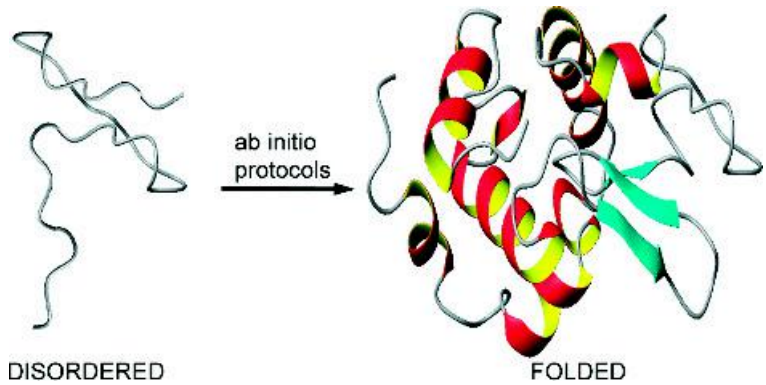

Figure 6.25 The *ab initio* approach to fold prediction

primary sequence from an initial randomized structure to one satisfying all constraints. These constraints will reflect the energetics associated with charge, hydrophobicity and polarity with the aim being to find a single structure of low energy. The resulting structures should have very few violations with respect to bond angles and length and can be checked for consistency against the Ramachandran plot. *Ab initio* protocols do *not* utilize experimental constraints but depend on the generation of structure from fundamental parameters *in silico*.

This approach is based on the thermodynamic argument that the native structure of a protein is the global minimum in the free energy profile. *Ab initio* methods place great demands on computational power but have the advantage of not using peripheral information. Despite considerable technical difficulties success is being achieved in this area as a result of regular 'contests' held to judge the success of *ab initio* protocols. These proceedings go under the more formal name of critical assessment of techniques for protein structure prediction (CASP). As well as involving the comparative homology based methods CASP involves the use of *ab initio* methods to predict tertiary structures for 'test' sequences (Figure 6.26). The emphasis is on *prediction* as opposed to "postdiction" and involves a community wide attempt to determine structures for proteins that have been assessed independently (but are not available in public domain databases). Independent structural knowledge allows an accurate assessment of the eventual success of the different *ab initio* approaches. Generally the results are expressed as rmsd (root mean square deviations), reflecting the difference in positions between corresponding atoms in the experimental and calculated (predicted) structures. In successful predictions rmsd values below 0.5 nm were seen for small proteins (<100 residues) using *ab initio* approaches. Although the agreement between predicted and experimentally determined structures is still relatively poor the resolution allows in favourable

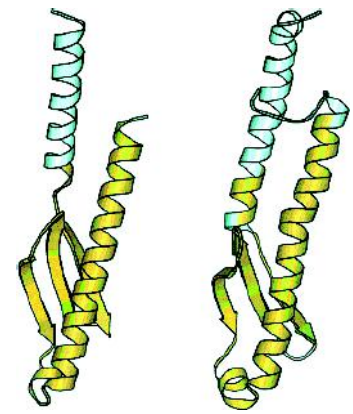

Figure 6.26 One good model predicted for target protein (T0091) in CASP4. The native structure of T0091 is shown on the left and the predicted model is shown on the right. Structurally equivalent residues are marked in yellow (reproduced with permission Sippl, M.J. *et al. Proteins: Structure, Function, and Genetics*, Suppl. 5, 55–67. John Wiley & Sons, Chichester, 2001)

circumstances the backbone of 'target proteins' to be defined with reasonable accuracy and for overall fold to be identified. For larger proteins the refinement of structure is worse but *ab initio* methods are becoming steadily better and offer the possibility of true protein structure prediction in the future.

Summary

Despite the incredible diversity of living cells all organisms are made up of the same 20 amino acid residues linked together to form proteins. This arises from the origin of the amino acid alphabet very early in evolution before the first true cells. All subsequent forms of life evolved using this basic alphabet.

Prebiotic synthesis is the term applied to non-cellular based methods of amino acid synthesis that existed over 3.6×10^9 years ago. The famous experiment of Urey and Miller demonstrated formation of organic molecules such as adenine, alanine and glycine that are the precursors today of nucleic acids and protein systems.

A major development in molecular evolution was the origin of self-replicating systems. Today this role is reserved for DNA but the first replicating systems were based on RNA a molecule now known to have catalytic function. An early prebiotic system involving RNA molecules closely associated with amino acids is thought to be most likely.

The fossil record shows primitive prokaryotic cells resembling blue-green cyanobacteria evolved 3.6 billion years ago. Evolution of these cells through compartmentalization, symbiosis and specialization yielded single cell and metazoan eukaryotes.

Protein sequencing is performed using the Edman procedure and involves the labelling and identification of the N-terminal residue with phenylisothiocyanate in a cyclic process. The procedure can be repeated ~50–80 times before cumulative errors restrict the accuracy of sequencing.

Nucleic acid sequencing is based on dideoxy chain termination procedures. The result of efficient DNA sequencing methods is the completion of genome sequencing projects and the prevalence of enormous amounts of bio-information within databases.

Databases represent the 'core' of the new area of bioinformatics – a subject merging the disciplines of biochemistry, computer science and information technology together to allow the interpretation of protein and nucleic acid sequence data.

One of the first uses of sequence data was to establish homology between proteins. Sequence homology arises from a link between proteins as a result of evolution from a common ancestor. Serine proteases show extensive sequence homology and this is accompanied by structural homology. Chymotrypsin, trypsin and elastase share homologous sequences and structure.

Structural homology will also result when sequences show low levels of sequence identity. The c type cytochromes from bacteria and mitochondria exhibit remarkably similar folds achieved with low overall sequence identity. The results emphasize that proteins evolve with the retention of the folded structure and the preservation of functional activity.

The bioinformatics revolution allows analysis of protein sequences at many different levels. Common applications include secondary structure prediction, conserved motif recognition, identification of signal sequences and transmembrane regions, determination of sequence homology, and structural prediction *ab initio*.

In the future bioinformatics is likely to guide the directions pursued by biochemical research by allowing the formation of new hypotheses to be tested *via* experimental methods.

Problems

1. Use the internet to locate some or all of the web pages/databases listed in Table 6.9. Find any web page(s) describing this book.
2. A peptide containing 41 residues is treated with cyanogen bromide to liberate three smaller peptides whose sequences are

(i) Phe-Leu-Asn-Ser-Val-Thr-Val-Ala-Ala-Tyr-Gly-Gly-Pro-Ala-Lys-Pro-Ala-Val-Glu-Asp-Gly-Ala-Met

(ii) Ala-Ser-Ser-Glu-Glu-Lys-Gly-Met and

(iii) Val-Ser-Thr-Asn-Glu-Lys-Ala-Ala-Val-Phe

Trypsin digestion of the same 41 residue peptide yields the sequences:

(i) Ala-Ala-Val-Phe

(ii) Gly-Met-Phe-Leu-Asn-Ser-Val-Thr-Val-Ala-Ala-Tyr-Gly-Gly-Pro-Ala-Lys-Pro-Ala-Val-Glu-Asp-Gly-Ala-Met-Val-Ser-Thr-Asn-Glu-Lys and

(iii) Ala-Ser-Ser-Glu-Glu-Lys

Can the sequence be established unambiguously?

3. Obtain the amino acid sequences of human α-lactalbumin and lysozyme from any suitable database. Use the BLAST or FASTA tool to compare the amino acid sequences of human α-lactalbumin and lysozyme. Are the sequences sufficiently similar to suggest homology?

4. Identify the unknown protein of Table 6.3.

5. Download the following coordinate files from the protein databank 1CYO and 1NU4. In each case use your molecular graphic software to highlight the number of Phe residues in each protein, any co-factors present in these proteins and the number of charged residues present in each protein.

6. What is the EC number associated with the enzyme β-galactosidase. Locate the enzyme derived from *E. coli* in a protein or sequence database. How many domains does this protein possess? Are these domains related? Describe the structure of each domain. Estimate the residues encompassing each domain. Identify the active site and any conserved residues.

7. Find a program on the internet that allows the prediction of secondary structure. Use this program to predict the secondary structure content of myoglobin. Does this value agree with the known secondary structure content from X-ray crystallography? Repeat this trial with your 'favourite' protein? Now repeat the analysis with a different secondary structure prediction algorithm. Do you get the same results?

8. Identify proteins that are sequentially homologous to (a) α chain of human haemoglobin, (b) subtilisin E from *Bacillus subtilis*, (c) HIV protease. Comment on your results and some of the implications?

9. The β barrel structure is a common topology found in many proteins. An example is triose phosphate isomerase. Use databases to find other proteins with this structural motif. Do these sequences exhibit homology? Comment on the evolutionary relationship of β barrels.

10. Assuming a point accepted DNA mutation rate of 20 (20 PAMs/100 residues/10^8 years) establish the degree of DNA homology for two proteins each of 200 residues that diverged 500 million years ago. What is the maximal level of amino acid residue homology between proteins and what is the lowest level?