3

Pairwise Sequence Alignment

Introduction

One of the most basic questions about a gene or protein is whether it is related to any other gene or protein. Relatedness of two proteins at the sequence level suggests that they are homologous. Relatedness also suggests that they may have common functions. By analyzing many DNA and protein sequences, it is possible to identify domains or motifs that are shared among a group of molecules. These analyses of the relatedness of proteins and genes are accomplished by aligning sequences. As we complete the sequencing of many organisms' genomes, the task of finding out how proteins are related within an organism and between organisms becomes increasingly fundamental to our understanding of life.

Two genes (or proteins) are homologous if they have evolved from a common ancestor.

In this chapter we will introduce pairwise sequence alignment. We will adopt an evolutionary perspective in our description of how amino acids (or nucleotides) in two sequences can be aligned and compared. We will then describe algorithms and programs for pairwise alignment.

Protein Alignment: Often More Informative Than DNA Alignment

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is often more informative to compare protein sequences. There are several reasons for this. Many changes in a DNA sequence (particularly at

Bioinformatics and Functional Genomics, Second Edition. By Jonathan Pevsner

the third position of a codon) do not change the amino acid that is specified. Furthermore, many amino acids share related biophysical properties (e.g., lysine and arginine are both basic amino acids). The important relationships between related (but mismatched) amino acids in an alignment can be accounted for using scoring systems (described in this chapter). DNA sequences are less informative in this regard. Protein sequence comparisons can identify homologous sequences from organisms that last shared a common ancestor over 1 billion years ago (BYA) (e.g., glutathione transferases) (Pearson, 1996). In contrast, DNA sequence comparisons typically allow lookback times of up to about 600 million years ago (MYA).

When a nucleotide coding sequence is analyzed, it is often preferable to study its translated protein. In Chapter 4 (on BLAST searching), we will see that we can move easily between the worlds of DNA and protein. For example, the tblastn tool from the National Center for Biotechnology Information (NCBI) BLAST website allows one to search with a protein sequence for related proteins derived from a DNA database (see Chapter 4). This query option is accomplished by translating each DNA sequence into all of the six proteins that it potentially encodes.

Nevertheless, in many cases it is appropriate to compare nucleotide sequences. This comparison can be important in confirming the identity of a DNA sequence in a database search, in searching for polymorphisms, in analyzing the identity of a cloned cDNA fragment, or in many other applications.

Definitions: Homology, Similarity, Identity

Let us consider the globin family of proteins. We will begin with human myoglobin (accession number NP_005359) and beta globin (accession number NP_000509) as two proteins that are distantly but significantly related. The accession numbers are obtained from Entrez Gene (Chapter 2). Myoglobin and the hemoglobin chains (alpha, beta, and other) are thought to have diverged some 600 million years ago, near the time the vertebrate and insect lineages diverged.

ne researchers use the term *logous* to refer to proteins that not homologous, but share e similarity by chance. Such teins are presumed to have not cended from a common estor.

Two sequences are *homologous* if they share a common evolutionary ancestry. There are no degrees of homology; sequences are either homologous or not (Reeck et al., 1987; Tautz, 1998). Homologous proteins almost always share a significantly related three-dimensional structure. Myoglobin and beta globin have very similar structures as determined by x-ray crystallography (Fig. 3.1). When two sequences are homologous, their amino acid or nucleotide sequences usually share significant identity. Thus, while homology is a qualitative inference (sequences are homologous or not), identity and similarity are quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid (or nucleotide) identity. For example, in the globin family, all the members are homologous, but some have sequences that have diverged so greatly that they share no recognizable sequence identity (e.g., human beta globin and human neuroglobin share only 22% amino acid identity). Perutz, Kendrew and others demonstrated that individual globin chains share the same overall shape as myoglobin (see Ingram, 1963), even though the myoglobin and alpha globin proteins share only about 26% amino acid identity. In general, three-dimensional structures diverge much more slowly than amino acid sequence identity between two proteins

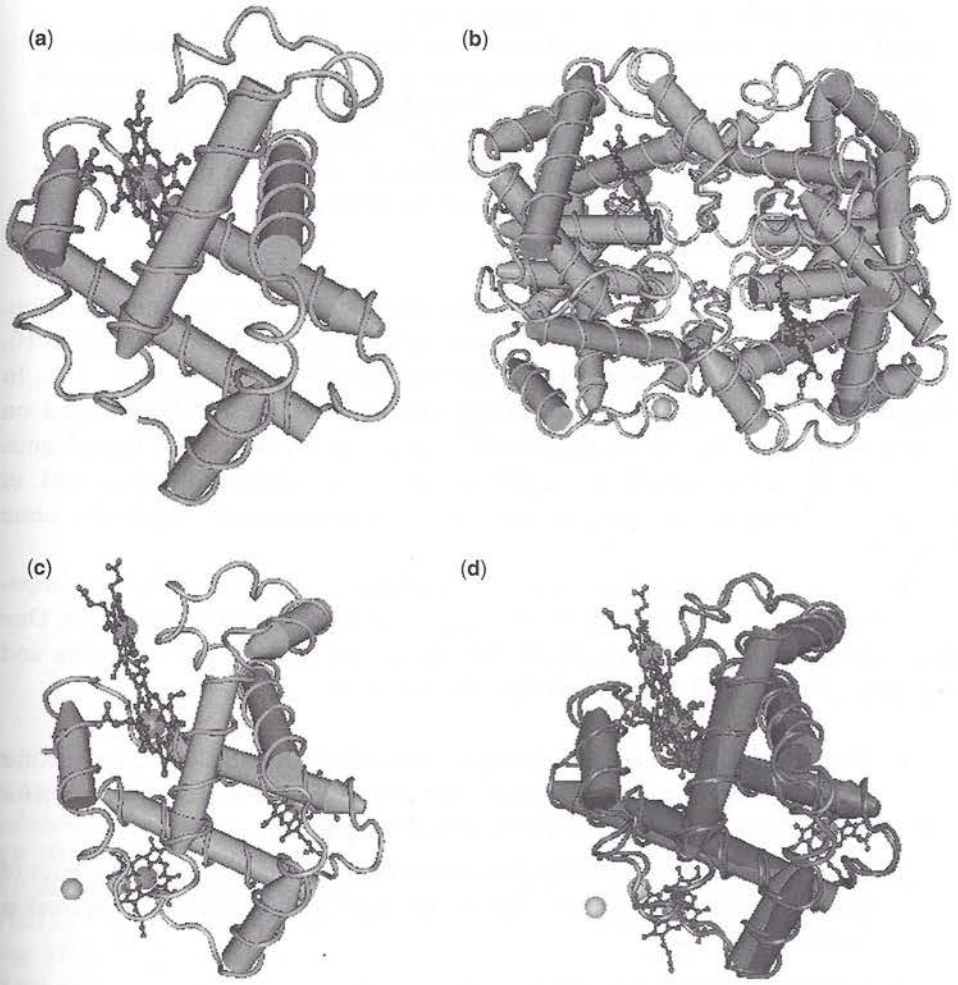


FIGURE 3.1. Three-dimensional structures of (a) myoglobin (accession 2MM1), (b) the tetrameric hemoglobin protein (2H35), (c) the beta globin subunit of hemoglobin, and (d) myoglobin and beta globin superimposed. The images were generated with the program Cn3D (see Chapter 11). These proteins are homologous (descended from a common ancestor), and they share very similar three-dimensional structures. However, pairwise alignment of these proteins' amino acid sequences reveals that the proteins share very limited amino acid identity.

(Chothia and Lesk, 1986). Recognizing this type of homology is an especially challenging bioinformatics problem.

Proteins that are homologous may be orthologous or paralogous. *Orthologs* are homologous sequences in different species that arose from a common ancestral gene during speciation. Figure 3.2 shows a tree of myoglobin orthologs. There is a human myoglobin gene and a rat gene. Humans and rodents diverged about 80 MYA (see Chapter 18), at which time a single ancestral myoglobin gene diverged by speciation. Orthologs are presumed to have similar biological functions; in this example, human and rat myoglobins both transport oxygen in muscle cells. *Paralogs* are homologous sequences that arose by a mechanism such as gene duplication. For example, human alpha-1 globin (NP_000549) is paralogous to alpha-2 globin (NP_000508); indeed, these two proteins share 100% amino acid identity. Human alpha-1 globin and beta globin are also paralogs (as are all the proteins shown in Fig. 3.3). All of the globins have distinct properties, including regional distribution in the body, developmental timing of gene expression, and abundance. They are all thought to have distinct but related functions as oxygen carrier proteins.

You can see the protein sequences used to generate Figs. 3.2 and 3.3 in web documents 3.1 and 3.2 at ► http://www.bioinfbook.org/chapter3.

In general when we consider other paralogous families they are presumed to share common functions. Consider the lipocalins: all are about 20 kilodalton proteins that have a hydrophobic binding pocket that is thought to be used to transport a hydrophobic ligand. Members include retinol binding protein (a retinol transporter), apolipoprotein D (a cholesterol transporter), and odorant-binding protein (an odorant transporter secreted from a nasal gland).

Walter M. Fitch (1970, p. 113) defined these terms. He wrote:
there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example, α and β hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact).

We thus define homologous genes within the same organism as paralogous. But consider further the case of globins. Human α-globin and β-globin are paralogs, as are mouse α-globin and mouse β-globin. Human α-globin and mouse α-globin are orthologs. What is the relation of human α-globin to mouse β-globin? These could be considered paralogs, because α-globin and β-globin originate from a gene duplication event rather than from a speciation event. However, they are not paralogs because they do not occur in the same species. It may thus be most appropriate to simply call them "homologs," reflecting their descent from a common ancestor. Fitch (1970, p. 113) notes that phylogenies require the study of orthologs (see also Chapter 7).

Richard Owen (1804–1892) was one of the first biologists to use the term homology. He defined homology as "the same organ in different animals under every variety of form and function" (Owen, 1843, p. 379). Charles Darwin (1809–1882) also discussed homology in the sixth edition of *The Origin of Species by means of Natural Selection or, The Preservation of Favoured Races in the Struggle for Life* (1872). He wrote: "That relation between parts which results from their development from corresponding embryonic parts, either in different animals, as in the case of the arm of man, the foreleg of a quadruped, and the wing of a bird; or in the same individual, as in the case of the fore and hind legs in quadrupeds, and the segments or rings and their appendages of which the body of a worm, a centipede, &c., is composed. The latter is called serial homology. The parts which stand in such a relation to each other are said to be homologous, and one such part or organ is called the homologue of the other. In different plants the parts of the flower are homologous, and in general these parts are regarded as homologous with leaves."

Notably, orthologs and paralogs do not necessarily have the same function. We will provide various definitions of gene and protein function in Chapter 10. Later we will explore genomes across the tree of life (Chapters 13 to 19). In all genome sequencing projects, orthologs and paralogs are identified based on database searches. Two DNA (or protein) sequences are defined as homologous based on achieving significant alignment scores, as discussed below and in Chapter 4. However, homologous proteins do not necessarily share the same function.

We can assess the relatedness of any two proteins by performing a *pairwise alignment*. In this procedure, we place the two sequences directly next to each other. One practical way to do this is through the NCBI pairwise BLAST tool (Tatusova and Madden, 1999) (Fig. 3.4). Perform the following steps:

1. Choose the protein BLAST program and select "BLAST 2 sequences" for our comparison of two proteins. An alternative is to select blastn (for "BLAST nucleotides") for DNA–DNA comparison.
2. Enter the sequences or their accession numbers. Here we use the sequence of human beta globin in the fasta format, and for myoglobin we use the accession number (Fig. 3.4).
3. Select any optional parameters.
   - You can choose from five scoring matrices: BLOSUM62, BLOSUM45, BLOSUM80, PAM70, and PAM30. Select PAM250.
   - You can change the gap creation penalty and gap extension penalty.
   - For blastn searches you can change reward and penalty values.
   - There are other parameters you can change, such as word size, expect value, filtering, and dropoff values. We will discuss these more in Chapter 4.
4. Click "BLAST." The output includes a pairwise alignment using the single-letter amino acid code (Fig. 3.5a).

Note that the fasta format uses the single-letter amino acid code; those abbreviations are shown in Box 3.1.

It is extremely difficult to align proteins by visual inspection. Also, if we allow gaps in the alignment to account for deletions or insertions in the two sequences, the number of possible alignments rises exponentially. Clearly, we will need a computer algorithm to perform an alignment (see Box 3.2). In the pairwise alignments shown in Fig. 3.5a, beta globin is on top (on the line labeled query) and myoglobin is below (on the subject line). An intermediate row indicates the presence of *identical* amino acids in the alignment. For example, notice that near the beginning of the alignment the residues WGKV are identical between the two proteins. We can count the total number of identical residues; in this case, the two proteins share

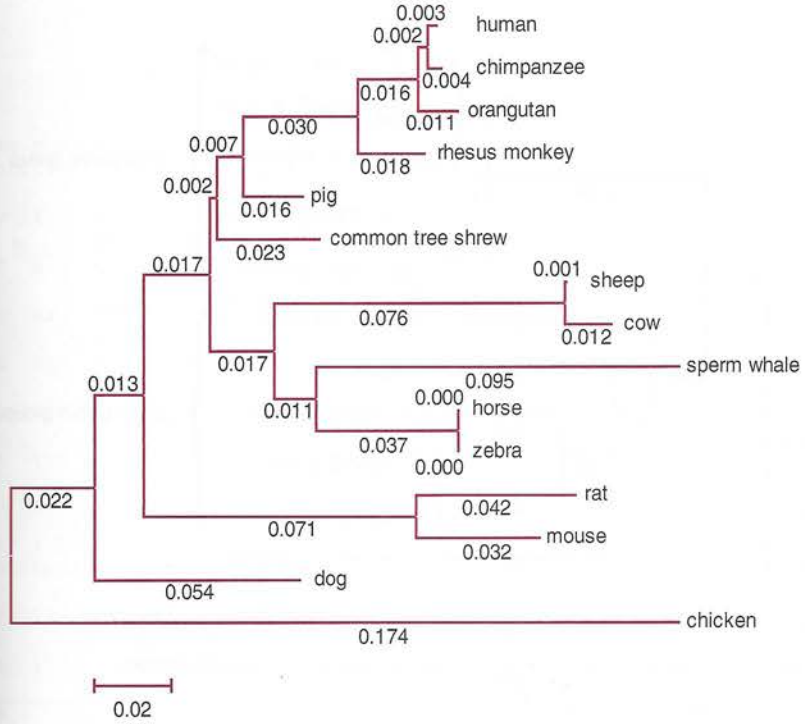


FIGURE 3.2. A group of myoglobin orthologs, visualized by multiply aligning the sequences (Chapter 6) then creating a phylogenetic tree by neighbor-joining (Chapter 7). The accession numbers and species names are as follows: human, NP_005359 (Homo sapiens); chimpanzee, XP_001156591 (Pan troglodytes); orangutan, P02148 (Pongo pygmaeus); rhesus monkey, XP_001082347 (Macaca mulatta); pig, NP_999401 (Sus scrofa); common tree shrew, P02165 (Tupaia glis); horse, P68082 (Equus caballus); zebra, P68083 (Equus burchellii); dog, XP_850735 (Canis familiaris); sperm whale, P02185 (Physeter catodon); sheep, P02190 (Ovis aries); rat, NP_067599 (Rattus norvegicus); mouse, NP_038621 (Mus musculus); cow, NP_776306 (Bos taurus); chicken_XP_416292 (Gallus gallus). The sequences are shown in web document 3.1 (► http://www.bioinfbook.org/chapter3). In this tree, sequences that are more closely related to each other are grouped closer together. Note that as entire genomes continue to be sequenced (Chapters 13 to 19), the number of known orthologs will grow rapidly for most families of orthologous proteins.

25% identity (37 of 145 aligned residues). Identity is the extent to which two amino acid (or nucleotide) sequences are invariant. Note that this particular alignment is called *local* because only a subset of the two proteins is aligned: the first and last few amino acid residues of each protein are not displayed. A global pairwise alignment includes all residues of both sequences.

Another aspect of this pairwise alignment is that some of the aligned residues are similar but not identical; they are related to each other because they share similar biochemical properties. *Similar* pairs of residues are structurally or functionally related. For example, on the first row of the alignment we can find threonine and serine (T and S connected by a + sign in Fig. 3.5a); nearby we can see a leucine and a valine residue that are aligned. These are *conservative substitutions*. Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydroxylated amino acids (S, T), and hydrophobic amino acids (W, F, Y, L, I, V, M, A). Later in this chapter we will see how scores are assigned to aligned amino acid residues.

You can access the pairwise BLAST program at the NCBI blast site, ► http://www.ncbi.nlm.nih.gov/BLAST/. We discuss various options for using the Basic Local Alignment Search Tool (BLAST) in Chapter 4. We discuss global and local alignments below.

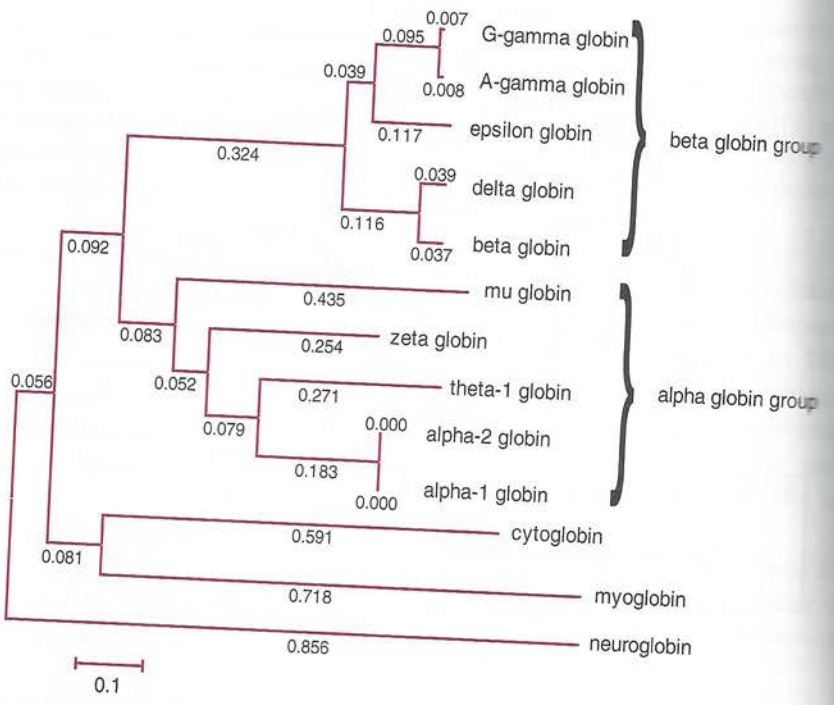


FIGURE 3.3. Paralogous human globins: Each of these proteins is human, and each is a member of the globin family. This unrooted tree was generated using the neighbor-joining algorithm in MEGA (see Chapter 7). The proteins and their RefSeq accession numbers (also shown in web document 3.2) are delta globin (NP_000510), G-gamma globin (NP_000175), beta globin (NP_000509), A-gamma globin (NP_000550), epsilon globin (NP_005321), zeta globin (NP_005323), alpha-1 globin (NP_000549), alpha-2 globin (NP_000508), theta-1 globin (NP_005322), hemoglobin mu chain (NP_001003938), cytoglobin (NP_599030), myoglobin (NP_005359), and neuroglobin (NP_067080). A Poisson correction model was used (see Chapter 7).

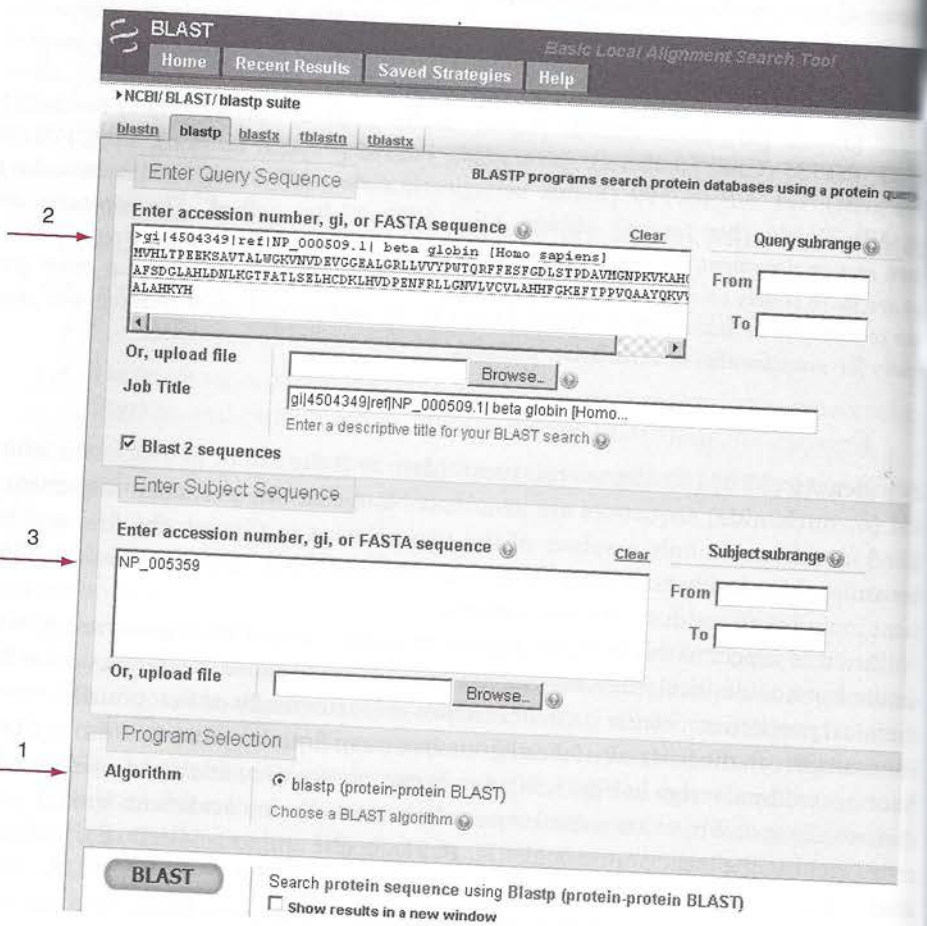


FIGURE 3.4. The BLAST program at the NCBI website allows the comparison of two DNA or protein sequences. Here the program is set to blastp for the comparison of two proteins (arrow 1). Human beta globin (NP_000509) is input in the fasta format (arrow 2), while human myoglobin (NP_005359) is input as an accession number (arrow 3).

(a)

```
 Score = 43.9 bits (102),  Expect = 1e-09, Method: Composition-based stats.
 Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

Query  4    LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D    G E L RL   +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A    L    + +     L++ H  K  +  +    +  VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

(b)

```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12   VTALWGKVNVD--EVGGEALGRLL  33
            V  +WGKV  D    G E L RL
Sbjct  11   VLNVWGKVEADIPGHGQEVLIRLF  34
```

match	4 11 5 6 4	6	6 5 4	5 4		sum of matches: +60
mismatch	-1 1 -2 0 0	-2 -3	-2 0	-4	0	sum of mismatches: -13
gap open		-11				sum of gap penalties: -12
gap extend		-1				
						total raw score: 60 - 13 - 12 = 35

FIGURE 3.5. Pairwise alignment of human beta globin (the "query") and myoglobin (the "subject"). Panel (a) shows the alignment from the search shown in Fig. 3.4. Note that this alignment is local (i.e., the entire lengths of each protein are not compared), and there are many positions of identity between the two sequences (indicated with amino acids intervening between the query and subject lines). The alignment contains an internal gap (indicated by two dashes). Panel (b) illustrates how raw scores are calculated, using the result of a separate search with just amino acids 10–34 of HBB (corresponding to the region between the arrowheads in panel a). The raw score is 35; this represents the sum of the match scores (from a BLOSUM62 matrix in this case), the mismatch scores, the gap opening penalty (set to −11 for this search), and the gap extension penalty (set to −1).

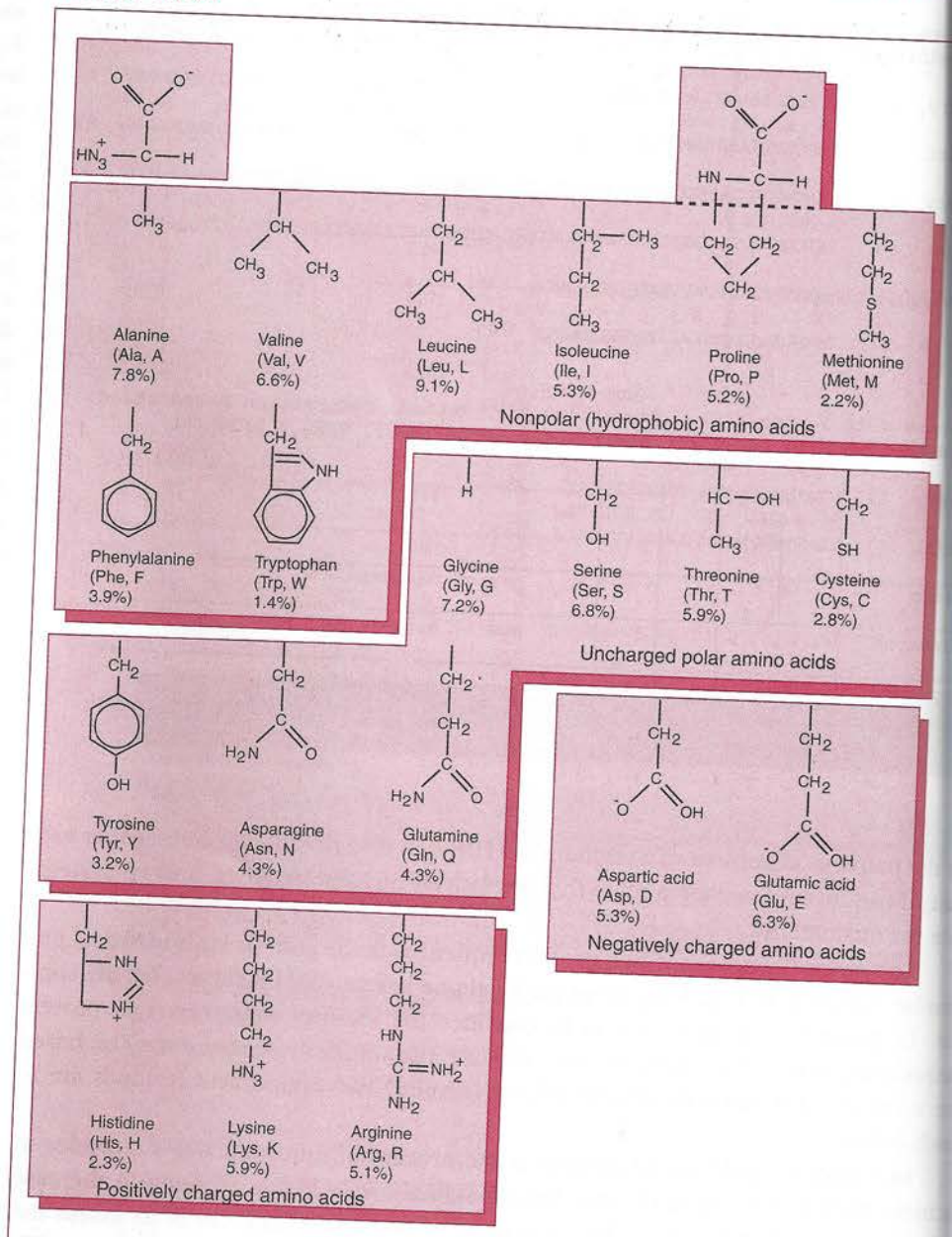
In the pairwise alignment of a segment of HBB and myoglobin, you can see that each pair of residues is assigned a score that is relatively high for matches, and often negative for mismatches.

The *percent similarity* of two protein sequences is the sum of both identical and similar matches. In Fig. 3.5a, there are 57 aligned amino acid residues that are similar. In general, it is more useful to consider the identity shared by two protein sequences, rather than the similarity, because the similarity measure may be based on a variety of definitions of how related (similar) two amino acid residues are to each other.

In summary, pairwise alignment is the process of lining up two sequences to achieve maximal levels of identity (and maximal levels of conservation in the case of amino acid alignments). The purpose of a pairwise alignment is to assess the degree of similarity and the possibility of homology between two molecules. We may say that two proteins share, for example, 25% amino acid identity and 39% similarity. If the amount of sequence identity is sufficient, then the two sequences are probably homologous. It is never correct to say that two proteins share a certain percent homology, because they are either homologous or not. Similarly, it is not appropriate to describe two sequences as "highly homologous"; instead one can say that they share a high degree of similiarity. We will discuss the statistical significance of sequence alignments below, including the use of expect values to assess whether an alignment of two sequences is likely to have occurred by chance (Chapter 4).

Two proteins could have similar structures due to convergent evolution. Molecular evolutionary studies are essential (based on sequence analyses) to assess this possibility.

Box 3.1
Structures and One- and Three-Letter Abbreviations of Twenty Com Amino Acids



It is very helpful to memorize these abbreviations and to become familiar with the physical properties of the amino acids. The percentages refer to the relative abundance of each amino acid in proteins.

Such analyses provide evidence to assess the hypothesis that two proteins are homologous. Ultimately the strongest evidence to determine whether two proteins are homologous comes from structural studies in combination with evolutionary analyses.

Box 3.2
Algorithms and Programs

An *algorithm* is a procedure that is structured in a computer program (Sedgewick, 1988). For example, there are many algorithms used for pairwise alignment. A computer *program* is a set of instructions that uses an algorithm (or multiple algorithms) to solve a task. For example, the BLAST program (Chapters 3 to 5) uses a set of algorithms to perform sequence alignments. Other programs that we introduce in Chapter 7 use algorithms to generate phylogenetic trees.

Computer programs are essential to solve a variety of bioinformatics problems because millions of operations may need to be performed. The algorithm used by a program provides the means by which the operations of the program are automated. Throughout this book, note how many hundreds of programs have been developed using many hundreds of different algorithms. Each program and algorithm is designed to solve a specific task. An algorithm that is useful to compare one protein sequence to another may not work in a comparison of one sequence to a database of 10 million protein sequences.

Why is it that an algorithm that is useful for comparing two sequences cannot be used to compare millions of sequences? Some problems are so inherently complex that an exhaustive analysis would require a computer with enormous memory or the problem would take an unacceptably long time to complete. A *heuristic algorithm* is one that makes approximations of the best solution without exhaustively considering every possible outcome. The 13 proteins in Fig. 3.2 can be arranged in a tree over a billion distinct ways (see Chapter 7)—and finding the optimal tree is a problem that a heuristic algorithm can solve in a second.

Gaps

Pairwise alignment is useful as a way to identify mutations that have occurred during evolution and have caused divergence of the sequences of the two proteins we are studying. The most common mutations are *substitutions*, *insertions*, and *deletions*. In protein sequences, substitutions occur when a mutation results in the codon for one amino acid being changed into that for another. This results in the alignment of two nonidentical amino acids, such as serine and threonine. Insertions and deletions occur when residues are added or removed and are typically represented by dashes that are added to one or the other sequence. Insertions or deletions (even those just one character long) are referred to as *gaps* in the alignment.

In our alignment of human beta globin and myoglobin there is one gap (Fig. 3.5a, between the D and E residues of the query). Gaps can occur at the ends of the proteins or in the middle. Note that one of the effects of adding gaps is to make the overall length of each alignment exactly the same. The addition of gaps can help to create an alignment that models evolutionary changes that have occurred. In a typical scoring scheme there are two gap penalties: one for creating a gap (−11 in the example of Fig. 3.5b) and one for each additional residue that a gap extends (−1 in Fig. 3.5b).

Pairwise Alignment, Homology, and Evolution of Life

If two proteins are homologous, they share a common ancestor. Generally, we observe the sequence of proteins (and genes) from organisms that are extant. We

It is possible to infer the sequence of the common ancestor (see Chapter 7).

can compare myoglobins from species such as human, horse, and chicken, and see that the sequences are homologous (Fig. 3.2). This implies that an ancestral organism had a myoglobin gene and lived sometime before the divergences of the lineages that gave rise to human and chicken (over 300 MYA; see Chapter 18). Descendants of that ancestral organism include many vertebrate species. The study of homologous protein (or DNA) sequences by pairwise alignment involves an investigation of the evolutionary history of that protein (or gene).

For a brief overview of the time scale of life on Earth, see Fig. 3.6 (refer to Chapter 13 for a more detailed discussion). The divergence of different species is established through the use of many sources of data, especially the fossil record. Fossils of prokaryotes have been discovered in rocks 3.5 billion years old or even older (Schopf, 2002). Fossils of methane-producing archaea, representative of a second domain of life, are found in rocks over 3 billion years old. The other main domain of life, the eukaryotes, emerged soon after. In the case of globins, in addition to the vertebrate proteins represented in Fig. 3.2, there are plant globins that must have shared a common ancestor with the metazoan (animal) globins some 1.5 billion years ago. There are also many bacterial and archaeal globins suggesting that the globin family arose earlier than two billion years ago.

Databases such as Pfam (Chapter 6) and COGS (Chapter 15) summarize the phylogenetic distribution of gene/protein families across the tree of life.

As we examine a variety of homologous protein sequences, we can observe a wide range of conservation between family members. Some are very ancient and well conserved, such as the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). A multiple sequence alignment, which is essentially a series of pairwise alignments between a group of proteins, reveals that GAPDH orthologs are extraordinarily well conserved (Fig. 3.7). Such highly conserved proteins may have any degree of representation across the tree of life, from being present in most known species to only a select few.

The GAPDH sequences used to generate Fig. 3.7 and the kappa casein sequences used to generate fig. 3.8 are shown in web documents 3.3 and 3.4 at ► http://www.bioinfbook.org/chapter3.

Orthologous kappa caseins from various species provide an example of a less well-conserved family (Fig. 3.8). Some columns of residues in this alignment are perfectly conserved among the selected species, but most are not, and many gaps needed to be introduced. Several positions at which four or even five different residues occur in an aligned column are indicated.

We can see from the preceding examples that pairwise sequence alignment between any two proteins can exhibit widely varying amounts of conservation. We will next examine how the information in such alignments can be used to decide how to quantitate the relatedness of any two proteins.

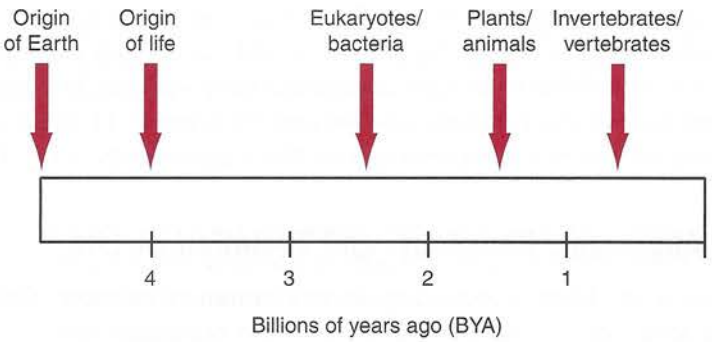


FIGURE 3.6. Overview of the history of life on Earth. See Chapter 13 for details. Gene/protein sequences are analyzed in the context of evolution: Which organisms have orthologous genes? When did these organisms evolve? How related are human and bacterial globins?

```
NP_002037.2          164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
XP_001162057.1       164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
NP_001003142.1       162   IHDHFGIVEGLMTTVHAITATQKTVDGPSGKMWRDGRGAAQNII   205
XP_893121.1          168   IHDNFGIMEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   211
XP_576394.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_058704.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001070653.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001062726.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_989636.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_525091.1          161   INDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
XP_318655.2          161   INDNFGILEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
NP_508535.1          170   INDNFGIIEGLMTTVHAVTATQKTVDGPSGKLWRDGRGAGQNII   213
NP_595236.1          164   INDTFGIEEGLMTTVHATTATQKTVDGPSKKDWRGGRGASANII   207
NP_011708.1          162   INDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRTASGNII   205
XP_456022.1          161   INDEFGIDEALMTTVHSITATQKTVDGPSHKDWRGGRTASGNII   204
NP_001060897.1       166   IHDNFGIIEGLMTTVHAITATQKTVDGPSSKDWRGGRAASFNII   209
```

FIGURE 3.7. Multiple sequence alignment of a portion of the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein from thirteen organisms: Homo sapiens *(human),* Pan troglodytes *(chimpanzee),* Canis lupus *(dog),* Mus musculus *(mouse),* Rattus norvegicus *(rat; four variants),* Gallus gallus *(chicken),* Drosophila melanogaster *(fruit fly),* Anopheles gambiae *(mosquito),* Caenorhabditis elegans *(worm),* Schizosaccharomyces pombe *(fission yeast),* Saccharomyces cerevisiae *(baker's yeast),* Kluyveromyces lactis *(a fungus), and* Oryza sativa *(rice). Columns in the alignment having even a single amino acid change are indicated with arrowheads. The accession numbers are given in the figure. The alignment was created by searching HomoloGene at NCBI with the term gapdh. The full alignment is given in Web Document 3.3 at ► http://www.bioinfbook.org/chapter3.*

```
mouse     AIPNPSFLAMPTNENQDNTAIPTIDPITPIVST--PVPTM------ESIVNTVANPEAST
rabbit    S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT------EPVVSTEVIAEASP
sheep     PHPHLSFMAIPPKKDQDKTEIPAINTIASAEPTVHSTPTT------EAVVNAVDNPEASS
cattle    PHPHLSFMAIPPKKNQDKTEIPTINTIASGEPT--STPTT------EAVESTVATLEDSP
pig       PRPHASFIAIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
human     PNLHPSFIAIPPKKIQDKIIIPTINTIATVEPT--PAPAT------EPTVDSVVTPEAFS
horse     PCPHPSFIAIPPKKLQEITVIPKINTIATVEPT--PIPTP------EPTVNNAVIPDASS
          .  :  *:*: .:: *:    *  :.*:.  .*    *:       *.  .     :  .
```

FIGURE 3.8. Multiple sequence alignment of seven kappa caseins, representing a protein family that is relatively poorly conserved. Only a portion of the entire alignment is shown. Note that just eight columns of residues are perfectly conserved (indicated with asterisks), and gaps of varying length form part of the alignment. In several columns, there are four different aligned amino acids (arrowheads); in two instances there are five different residues (double arrowheads). The sequences were aligned with MUSCLE 3.6 (see Chapter 6) and were human (NP_005203), equine (Equus caballus; NP_001075353), pig (Sus scrofa NP_001004026), ovine (Ovis aries NP_001009378), rabbit (Oryctolagus cuniculus P33618), bovine (Bos taurus NP_776719), and mouse (Mus musculus NP_031812). The full alignment is available as web document 3.3 at ► http://www.bioinfbook.org/chapter3.

Scoring Matrices

When two proteins are aligned, what scores should they be assigned? For the alignment of beta globin and myoglobin in Fig. 3.5a there were specific scores for matches and mismatches; how were they derived? Margaret Dayhoff (1978) provided a model of the rules by which evolutionary change occurs in proteins. We will now examine the Dayhoff model, which provides the basis of a quantitative scoring system for pairwise alignments. This system accounts for scores between any proteins, whether they are closely or distantly related. We will then describe the BLOSUM matrices of

The Dayhoff (1978) reference is to the *Atlas of Protein Sequence and Structure*, a book with 25 chapters (and various coauthors) describing protein families. The 1966 version of the *Atlas* described the sequences of just several dozen proteins (cytochromes c, other respiratory proteins, globins, some enzymes such as lysozyme and ribonucleases, virus coat proteins, peptide hormones, kinins, and fibrinopeptides). The 1978 edition included about 800 protein sequences.

Dayhoff et al. (1972) focused on proteins sharing 85% or more identity. Thus, they could construct their alignments with a high degree of confidence. Later in this chapter, we will see how the Needleman and Wunsch algorithm (described in 1970) permits the optimal alignment of protein sequences.

Steven Henikoff and Jorja G. Henikoff (1992). Next, we will discuss the two main kinds of pairwise sequence algorithms, global and local. Many database searching methods such as BLAST (Chapters 4 and 5) depend in some form on the evolutionary insights of the Dayhoff model.

Dayhoff Model: Accepted Point Mutations

Dayhoff and colleagues considered the problem of how to assign scores to aligned amino acid residues. Their approach was to catalog thousands of proteins and compare the sequences of closely related proteins in many families. They considered the question of which specific amino acid substitutions are observed to occur when two homologous protein sequences are aligned. They defined an *accepted point mutation* as a replacement of one amino acid in a protein by another residue that has been accepted by natural selection. Accepted point mutation is abbreviated *PAM* (which is easier to pronounce than APM). An amino acid change that is accepted by natural selection occurs when (1) a gene undergoes a DNA mutation such that it encodes a different amino acid and (2) the entire species adopts that change as the predominant form of the protein.

Which point mutations are accepted in protein evolution? Intuitively, conservative replacements such as serine for threonine would be most readily accepted. In order to determine all possible changes, Dayhoff and colleagues examined 1572 changes in 71 groups of closely related proteins (Box 3.3). Thus, their definition of "accepted" mutations was based on empirically observed amino acid substitutions. Their approach involved a phylogenetic analysis: rather than comparing two amino acid residues directly, they compared them to the inferred common ancestor of those sequences (Fig. 3.9 and Box 3.4).

For the PAM1 matrix, the proteins have undergone 1% change (that is, 1 accepted point mutation per 100 amino acid residues). The results are shown in Fig. 3.10, which describes the frequency with which any amino acid pairs i, j are aligned. Inspection of this table reveals which substitutions are unlikely to occur (for example, cysteine and tryptophan have noticeably few substitutions), while others such as asparagine and serine tolerate replacements quite commonly. Today, we could generate a table like this with vastly more data (refer to Fig. 2.1 and the explosive growth of GenBank). Several groups have produced updated versions of the PAM matrices (Gonnet et al., 1992; Jones et al., 1992). Nonetheless the findings from 1978 are essentially correct.

The main goal of Dayhoff's approach was to define a set of scores for the comparison of aligned amino acid residues. By comparing two aligned proteins, one can then tabulate an overall score, taking into account identities as well as mismatches, and also applying appropriate penalties for gaps. A scoring matrix defines scores for the interchange of residues i and j. It is given by the probability $q_{i,j}$ of aligning original amino acid residue j with replacement residue i relative to the likelihood of observing residues i by chance (p_i). The scoring matrix further incorporates a logarithm to generate log-odds scores. For the Dayhoff matrices, this takes the following form:

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_i}\right) \tag{3.1}$$

Here the score $s_{i,j}$ refers to the score for aligning any two residues (including an amino acid with itself) along the length of a pairwise alignment. The probability $q_{i,j}$ is the

Box 3.3
Dayhoff's Protein Superfamilies

Dayhoff (1978, p. 3) and colleagues studied 34 protein "superfamilies" grouped into 71 phylogenetic trees. These proteins ranged from some that are very well conserved (e.g., histones and glutamate dehydrogenase; see Fig. 3.7) to others that have a high rate of mutation acceptance (e.g., immunoglobulin [Ig] chains and kappa casein; see Fig. 3.8). Protein families were aligned (compare Fig. 3.7); then they counted how often any one amino acid in the alignment was replaced by another. Here is a partial list of the proteins they studied, including the rates of mutation acceptance. For a more detailed list, see Table 11.1. There is a range of almost 400-fold between the families that evolve fastest and slowest, but within a given family the rate of evolution (measured in PAMs per unit time) varies only two- to threefold between species. Used with permission.

Protein	PAMs per 100 million years
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome *c*	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

observed frequency of substitution for each pair of amino acids. The values for q_{ij} are called the "target frequencies," and they are estimated in reference to a particular amount of evolutionary change. For example, in a comparison of human beta globin versus the closely related chimpanzee beta globin, the likelihood of any particular residue matching another in a pairwise alignment is extremely high, while in a comparison of human beta globin and a bacterial globin the likelihood of a match is low. If in a particular comparison of closely related proteins an aligned serine were to change to a threonine 5% of the time, then that target frequency $q_{S,T}$ would be 0.05. If in a different comparison of differently related proteins serine were to change to threonine more often, say 40% of the time, then that target frequency $q_{S,T}$ would be 0.4.

Equation 3.1 describes an odds ratio (Box 3.5). For the numerator, Dayhoff et al. (1972) considered an entire spectrum of models for evolutionary change in determining target frequencies. We begin with the PAM1 matrix, which describes substitutions that occur in very closely related proteins. For the denominator of Equation 3.1, $p_i p_j$, is the probability of amino acid residues i and j occurring by chance. We will

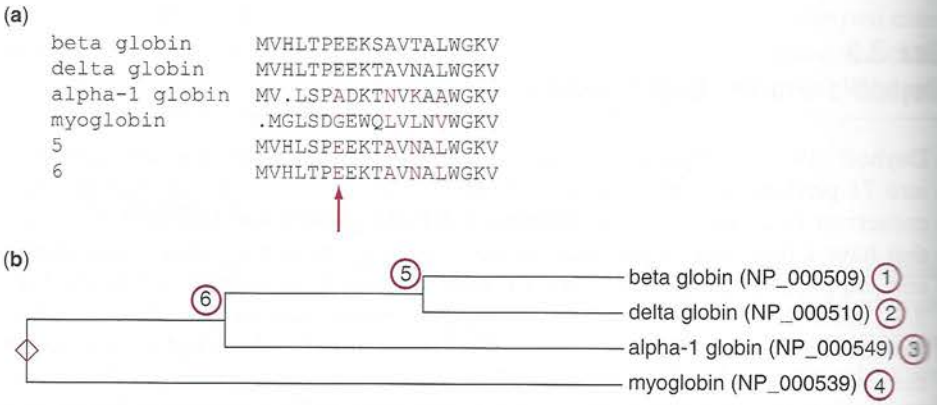


FIGURE 3.9. Dayhoff's approach to determining amino acid substitutions. Panel (a) shows a partial multiple sequence alignment of human alpha-1 globin, beta globin, delta globin, and myoglobin. Four columns in which alpha-1 globin and myoglobin have different amino acid residues are indicated in red. For example, A is aligned with G (arrow). Panel (b) shows a phylogenetic tree that shows the four extant sequences (labeled 1 to 4), as well as two internal nodes that represent the ancestral sequences (labeled 5 and 6). The inferred ancestral sequences were identified by maximum parsimony analysis using the software PAUP (Chapter 7), and are displayed in panel (a). From this analysis it is apparent that at each of the columns labeled in red, there was not a direct interchange of two amino acids between alpha-1 globin and myoglobin. Instead, an ancestral residue diverged. For example, the arrow in panel (a) indicates an ancestral glutamate that evolved to become alanine or glycine, but it would not be correct to suggest that alanine had been converted directly to glycine.

Box 3.4
A Phylogenetic Approach to Aligning Amino Acids

Dayhoff and colleagues did not compare the probability of one residue mutating directly into another. Instead, they constructed phylogenetic trees using parsimony analysis (see Chapter 7). Then, they described the probability that two aligned residues derived from a common ancestral residue. With this approach, they could minimize the confounding effects of multiple substitutions occurring in an aligned pair of residues. As an example, consider an alignment of the four human proteins alpha-1 globin, beta globin, delta globin, and myoglobin. A direct comparison of alpha-1 globin to myoglobin would suggest several amino acid replacements, such as ala ↔ gly, asn ↔ leu, lys ↔ leu, and ala ↔ val (Fig. 3.9a, residues highlighted in red). However, a phylogenetic analysis of these four proteins results in the estimation of internal nodes that represent ancestral sequences. In Fig. 3.9b the external nodes (corresponding to the four existing proteins) are labeled, as are internal nodes 5 and 6, which correspond to inferred ancestral sequences. In one of the four cases that are highlighted in Fig. 3.9a, the ancestral sequences suggest that a glu residue changed to ala and gly in alpha-1 globin and myoglobin, but ala and gly never directly interchanged (Fig. 3.9a, arrow). Thus, the Dayhoff approach was more accurate by taking an evolutionary perspective.

In a further effort to avoid the complicating factor of multiple substitutions occurring in alignments of protein families, Dayhoff et al. also focused on using multiple sequence alignments of closely related proteins. Thus, for example, their analysis of globins considered the alpha globins and beta globins separately.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

FIGURE 3.10. Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences. This figure is modified from Dayhoff (1978, p. 346). Amino acids are presented alphabetically according to the three-letter code. Notice that some substitutions are very commonly accepted (such as V and I or S and T). Other amino acids, such as C and W, are rarely substituted by any other residue. Used with permission.

Box 3.5
Statistical Concept: The Odds Ratio

Dayhoff et al. (1972) developed their scoring matrix by using odds ratios. The mutation probability matrix has elements M_{ij} that give the probability that amino acid j changes to amino acid i in a given evolutionary interval. The normalized frequency f_i gives the probability that amino acid i will occur at that given amino acid position by chance. The relatedness odds matrix in Equation 3.1 may also be expressed as follows:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

Here, R_{ij} is the relatedness odds ratio. Equation 3.1 may also be represented:

$$\text{Probability of an authentic alignment} = \frac{p(\text{aligned} \mid \text{authentic})}{p(\text{aligned} \mid \text{random})}$$

The right side of this equation can be read, "the probability of an alignment given that it is authentic (i.e. the substitution of amino acid j with amino acid i) divided by the probability that the alignment occurs given that it happened by chance. An *odds ratio* can be any positive ratio. The *probability* that an event will occur is the fraction of times it is expected to be observed over many trials; probabilities have values ranging from 0 to 1. Odds and probability are closely related concepts. A probability of 0 corresponds to an odds of 0; a probability of 0.5 corresponds to an odds of 1.0; a probability of 0.75 corresponds to odds of 75:25 or 3. Odds and probabilities may be converted as follows:

$$odds = \frac{probability}{1 - probability} \quad and \quad probability = \frac{odds}{1 + odds}$$

next explain how they calculated these values, resulting in the creation of an entire series of scoring matrices.

Dayhoff et al. calculated the relative mutabilities of the amino acids (Table 3.1). This simply describes how often each amino acid is likely to change over a short evolutionary period. (We note that the evolutionary period in question is short because this analysis involves protein sequences that are closely related to each other.) To calculate relative mutability, they divided the number of times each amino acid was observed to mutate by the overall frequency of occurrence of that amino acid. Table 3.2 shows the frequency with which each amino acid is found.

Why are some amino acids more mutable than others? The less mutable residues probably have important structural or functional roles in proteins, such that the consequence of replacing them with any other residue could be harmful to the organism. (We will see in Chapter 20 that many human diseases, from cystic fibrosis to the autism-related Rett syndrome to hemoglobinopathies, can be caused by a single amino acid substitution in a protein.) Conversely, the most mutable amino acids—asparagine, serine, aspartic acid, and glutamic acid—have functions in proteins that are easily assumed by other residues. The most common substitutions seen in Fig. 3.10 are glutamic acid for aspartic acid (both are acidic), serine for

You can look up a recent estimate of the frequency of occurrence of each amino acid at the SwissProt website ► http://www.expasy.ch/sprot/relnotes/relstat.html. From the UniProtKB/Swiss-Prot protein knowledgebase (release 51.7), the amino acid composition of all proteins is shown in web document 3.5 (► http://www.bioinfbook.org/chapter3).

TABLE 3-1 Relative Mutabilities of Amino Acids

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

The value of alanine is arbitrarily set to 100.
Source: From Dayhoff (1978). Used with permission.

TABLE 3-2 Normalized Frequencies of Amino Acid

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence
Source: From Dayhoff (1978). Used with permission.

alanine, serine for threonine (both are hydroxylated), and isoleucine for valine (both are hydrophobic and of a similar size).

The substitutions that occur in proteins can also be understood with reference to the genetic code (Box 3.6). Observe how common amino acid substitutions tend to require only a single nucleotide change. For example, aspartic acid is encoded by GAU or GAC, and changing the third position to either A or G causes the codon to encode a glutamic acid. Also note that four of the five least mutable amino acids (tryptophan, cysteine, phenylalanine, and tyrosine) are specified by only one or two codons. A mutation of any of the three bases of the W codon is guaranteed to change that amino acid. The low mutability of this amino acid suggests that substitutions are not tolerated by natural selection. Of the eight least mutable amino acids (Table 3.1), only one (leucine) is specified by six codons, and only two (glycine and proline) are specified by four codons. The others are specified by one or two codons.

PAM1 Matrix

Dayhoff and colleagues next used the data on accepted mutations (Fig. 3.10) and the probabilities of occurrence of each amino acid to generate a *mutation probability*

Box 3.6
The Standard Genetic Code

First nucleotide	Second nucleotide: T	Second nucleotide: C	Second nucleotide: A	Second nucleotide: G	Third nucleotide
T	TTT Phe 171	TCT Ser 147	TAT Tyr 124	TGT Cys 99	T
	TTC Phe 203	TCC Ser 172	TAC Tyr 158	TGC Cys 119	C
	TTA Leu 73	TCA Ser 118	TAA Ter 0	TGA Ter 0	A
	TTG Leu 125	TCG Ser 45	TAG Ter 0	TGG Trp 122	G
C	CTT Leu 127	CCT Pro 175	CAT His 104	CGT Arg 47	T
	CTC Leu 187	CCC Pro 197	CAC His 147	CGC Arg 107	C
	CTA Leu 69	CCA Pro 170	CAA Gln 121	CGA Arg 63	A
	CTG Leu 392	CCG Pro 69	CAG Gln 343	CGG Arg 115	G
A	ATT Ile 165	ACT Thr 131	AAT Asn 174	AGT Ser 121	T
	ATC Ile 218	ACC Thr 192	AAC Asn 199	AGC Ser 191	C
	ATA Ile 71	ACA Thr 150	AAA Lys 248	AGA Arg 113	A
	ATG Met 221	ACG Thr 63	AAG Lys 331	AGG Arg 110	G
G	GTT Val 111	GCT Ala 185	GAT Asp 230	GGT Gly 112	T
	GTC Val 146	GCC Ala 282	GAC Asp 262	GGC Gly 230	C
	GTA Val 72	GCA Ala 160	GAA Glu 301	GGA Gly 168	A
	GTG Val 288	GCG Ala 74	GAG Glu 404	GGG Gly 160	G

In this table, the 64 possible codons are depicted along with the frequency of codon utilization and the three-letter code of the amino acid that is specified. There are four bases (A, C, G, U) and three bases per codon, so there are $4^3 = 64$ codons.

Several features of the genetic code should be noted. Amino acids may be specified by one codon (M, W), two codons (C, D, E, F, H, K, N, Q, Y), three codons (I), four codons (A, G, P, T, V), or six codons (L, R, S). UGA is rarely read as a selenocysteine (abbreviated sec, and the assigned single-letter abbreviations is U).

For each block of four codons that are grouped together, one is often used dramatically less frequently. For example, for F, L, I, M, and V (i.e., codons with a U in the middle, occupying the first column of the genetic code), adenine is used relatively infrequently in the third-codon position. For codons with a cytosine in the middle position, guanine is strongly underrepresented in the third position.

Also note that in many cases mutations cause a conservative change (or no change at all) in the amino acid. Consider threonine (ACX). Any mutation in the third position causes no change in the specified amino acid, because of "wobble." If the first nucleotide of any threonine codon is mutated from A to U, the conservative replacement to a serine occurs. If the second nucleotide C is mutated to a G, a serine replacement occurs. Similar patterns of conservative substitution can be seen along the entire first column of the genetic code, where all of the residues are hydrophobic, and for the charged residues D, E and K, R as well.

Codon usage varies between organisms and between genes within organisms. Note also that while this is the standard genetic code, some organisms use

alternate genetic codes. A group of two dozen alternate genetic codes are listed at the NCBI Taxonomy website, ► http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/. As an example of a nonstandard code, vertebrate mitochondrial genomes use AGA and AGG to specify termination (rather than arg in the standard code), ATA to specify met (rather than ile), and TGA to specify trp (rather then termination).

Source: Adapted from the International Human Genome Sequencing Consortium (2001), fig. 34. Used with permission.

matrix M (Fig. 3.11). Each element of the matrix M_{ij} shows the probability that an original amino acid j (see the columns) will be replaced by another amino acid i (see the rows) over a defined evolutionary interval. In the case of Fig. 3.11 the interval is one PAM, which is defined as the unit of evolutionary divergence in which 1% of the amino acids have been changed between the two protein sequences. Note that the evolutionary interval of this PAM matrix is defined in terms of percent amino acid divergence and not in units of years. One percent divergence of protein sequence may occur over vastly different time frames for protein families that undergo substitutions at different rates.

Examination of Fig. 3.11 reveals several important features. The highest scores are distributed in a diagonal from top left to bottom right. The values in each column sum to 100%. The value 98.67 at the top left indicates that when the original sequence consists of an alanine there is a 98.67% chance that the replacement amino acid will also be an alanine over an evolutionary distance of one PAM. There is a 0.28% chance that it will be changed to serine. The most mutable amino acid (from Table 3.1), asparagine, has only a 98.22% chance of remaining unchanged; the least mutable amino acid, tryptophan, has a 99.76% chance of remaining the same.

For each original amino acid, it is easy to observe the amino acids that are most likely to replace it if a change should occur. These data are very relevant to pairwise sequence alignment because they will form the basis of a scoring system (described below) in which reasonable amino acid substitutions in an alignment are rewarded while unlikely substitutions are penalized. These concepts are also relevant to database searching algorithms such as BLAST (Chapters 4 and 5) which depend on rules to score the relatedness of molecular sequences.

Almost all molecular sequence data are obtained from extant organisms. We can infer ancestral sequences, as described in Box 3.4 and Chapter 7. But in general, for an aligned pair of residues i, j we do not know which one mutated into the other. Dayhoff and colleagues used the assumption that accepted amino acid mutations are undirected, that is, they are equally likely in either direction. In the PAM1 matrix, the close relationship of the proteins makes it unlikely that the ancestral residue is entirely different than both of the observed, aligned residues.

PAM250 and Other PAM Matrices

The PAM1 matrix was based on the alignment of closely related protein sequences, all of which were at least 85% identical within a protein family. We are often interested in exploring the relationships of proteins that share far less than 85% amino acid identity. We can accomplish this by constructing probability matrices for proteins that share any degree of amino acid identity. Consider closely related proteins, such as the GAPDH proteins shown in Fig. 3.7. A mutation from one residue to another

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	98.67	0.02	0.09	0.10	0.03	0.08	0.17	0.21	0.02	0.06	0.04	0.02	0.06	0.02	0.22	0.35	0.32	0.00	0.02	0.18
R	0.01	99.13	0.01	0.00	0.01	0.10	0.00	0.00	0.10	0.03	0.01	0.19	0.04	0.01	0.04	0.06	0.01	0.08	0.00	0.01
N	0.04	0.01	98.22	0.36	0.00	0.04	0.06	0.06	0.21	0.03	0.01	0.13	0.00	0.01	0.02	0.20	0.09	0.01	0.04	0.01
D	0.06	0.00	0.42	98.59	0.00	0.06	0.53	0.06	0.04	0.01	0.00	0.03	0.00	0.00	0.01	0.05	0.03	0.00	0.00	0.01
C	0.01	0.01	0.00	0.00	99.73	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.03	0.02
Q	0.03	0.09	0.04	0.05	0.00	98.76	0.27	0.01	0.23	0.01	0.03	0.06	0.04	0.00	0.06	0.02	0.02	0.00	0.00	0.01
E	0.10	0.00	0.07	0.56	0.00	0.35	98.65	0.04	0.02	0.03	0.01	0.04	0.01	0.00	0.03	0.04	0.02	0.00	0.01	0.02
G	0.21	0.01	0.12	0.11	0.01	0.03	0.07	99.35	0.01	0.00	0.01	0.02	0.01	0.01	0.03	0.21	0.03	0.00	0.00	0.05
H	0.01	0.08	0.18	0.03	0.01	0.20	0.01	0.00	99.12	0.00	0.01	0.01	0.00	0.02	0.03	0.01	0.01	0.01	0.04	0.01
I	0.02	0.02	0.03	0.01	0.02	0.01	0.02	0.00	0.00	98.72	0.09	0.02	0.21	0.07	0.00	0.01	0.07	0.00	0.01	0.33
L	0.03	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.04	0.22	99.47	0.02	0.45	0.13	0.03	0.01	0.03	0.04	0.02	0.15
K	0.02	0.37	0.25	0.06	0.00	0.12	0.07	0.02	0.02	0.04	0.01	99.26	0.20	0.00	0.03	0.08	0.11	0.00	0.01	0.01
M	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.08	0.04	98.74	0.01	0.00	0.01	0.02	0.00	0.00	0.04
F	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.08	0.06	0.00	0.04	99.46	0.00	0.02	0.01	0.03	0.28	0.00
P	0.13	0.05	0.02	0.01	0.01	0.08	0.03	0.02	0.05	0.01	0.02	0.02	0.01	0.01	99.26	0.12	0.04	0.00	0.00	0.02
S	0.28	0.11	0.34	0.07	0.11	0.04	0.06	0.16	0.02	0.02	0.01	0.07	0.04	0.03	0.17	98.40	0.38	0.05	0.02	0.02
T	0.22	0.02	0.13	0.04	0.01	0.03	0.02	0.02	0.01	0.11	0.02	0.08	0.06	0.01	0.05	0.32	98.71	0.00	0.02	0.09
W	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	99.76	0.01	0.00
Y	0.01	0.00	0.03	0.00	0.03	0.00	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.21	0.00	0.01	0.01	0.02	99.45	0.01
V	0.13	0.02	0.01	0.01	0.03	0.02	0.02	0.03	0.03	0.57	0.11	0.01	0.17	0.01	0.03	0.02	0.10	0.00	0.02	99.01

FIGURE 3.11. The PAM1 mutation probability matrix. From Dayhoff (1978, p. 348, fig. 82). The original amino acid j is arranged in columns (across the top), while the replacement amino acid i is arranged in rows. Used with permission.

is a relatively rare event, and a scoring system used to align two such closely related proteins should reflect this. In the PAM1 mutation probability matrix (Fig. 3.11) some substitutions such as tryptophan to threonine are so rare that they are never observed in the data set. But next consider two distantly related proteins, such as the kappa caseins shown in Fig. 3.8. Here, substitutions are likely to be very common. PAM matrices such as PAM100 or PAM250 were generated to reflect the kinds of amino acid substitutions that occur in distantly related proteins.

How are PAM matrices other than PAM1 derived? Dayhoff et al. multiplied the PAM1 matrix by itself, up to hundreds of times, to obtain other PAM matrices (see Box 3.7). Thus they extrapolated from the PAM1 matrix.

To make sense of what different PAM matrices mean, consider the extreme cases. When PAM equals zero, the matrix is a unit diagonal (Fig. 3.12), because no amino acids have changed. PAM can be extremely large (e.g., PAM greater than 2000, or the matrix can even be multiplied against itself an infinite number of times). In the resulting PAM∞ matrix there is an equal likelihood of any amino acid being present and all the values consist of rows of probabilities that approximate the background probability for the frequency of occurrence of each amino acid (Fig. 3.12, lower panel). We described these background frequencies in Table 3.2.

The PAM250 matrix is of particular interest (Fig. 3.13). It is produced when the PAM1 matrix is multiplied against itself 250 times, and it is one of the common matrices used for BLAST searches of databases (Chapter 4). This matrix applies

Box 3.7
Matrix Multiplication

A matrix is an orderly array of numbers. An example of a matrix with rows i and columns j is:

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

In a symmetric matrix, such as the one above, $a_{ij} = a_{ji}$. This means that all the corresponding nondiagonal elements are equal. Matrices may be added, subtracted, or manipulated in a variety of ways. Two matrices can be multiplied together provided that the number of columns in the first matrix M_1 equals the number of rows in the second matrix M_2. Following is an example of how to multiply M_1 by M_2.

Successively multiply each row of M_1 by each column of M_2:

$$M_1 = \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}$$

$$M_{12} = \begin{bmatrix} (3)(5)+(4)(2) & (3)(-2)+(4)(1) \\ (0)(5)+(2)(2) & (0)(-2)+(2)(1) \end{bmatrix} = \begin{bmatrix} 23 & -2 \\ 4 & 2 \end{bmatrix}$$

If you want to try matrix multiplication yourself, enter the PAM1 mutation probability matrix of Fig. 3.11 into a program such as MATLAB® (Mathworks), divide each value by 10,000, and multiply the matrix times itself 250 times. You will get the PAM250 matrix of Fig. 3.13.

FIGURE 3.12. Portion of the matrices for a zero PAM value (PAM0; upper panel) or for an infinite PAM∞ value (lower panel). At PAM∞ (i.e., if the PAM1 matrix is multiplied against itself an infinite number of times), all the entries in each row converge on the normalized frequency of the replacement amino acid (see Table 3.2). A PAM2000 matrix has similar values that tend to converge on these same limits. In a PAM2000 matrix, the proteins being compared are at an extreme of unrelatedness. In constrast, at PAM0, no mutations are tolerated, and the residues of the proteins are perfectly conserved.

original amino acid

replacement amino acid: PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100

original amino acid

replacement amino acid: PAM∞	A	R	N	D	C	Q	E	G
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

to an evolutionary distance where proteins share about 20% amino acid identity. Compare this matrix to the PAM1 matrix (Fig. 3.11) and note that much of the information content is lost. The diagonal from top left to bottom right tends to contain higher values than elsewhere in the matrix, but not in the dramatic fashion of the PAM1 matrix. As an example of how to read the PAM250 matrix, if the original amino acid is an alanine, there is just a 13% chance that the second sequence will also have an alanine. In fact, there is a nearly equal probability (12%) that the alanine will have been replaced by a glycine. For the least mutable amino acids, tryptophan and cysteine, there is more than a 50% probability that those residues will remain unchanged at this evolutionary distance.

FIGURE 3.13. The PAM250 mutation probability matrix. From Dayhoff (1978, p. 350, fig. 83). At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to Fig. 3.11, and the columns sum to 100. Used with permission.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

From a Mutation Probability Matrix to a Log-Odds Scoring Matrix

Our goal in studying PAM matrices is to derive a scoring system so that we can assess the relatedness of two sequences. When we perform BLAST searches (Chapters 4 and 5) or pairwise alignments, we employ a scoring matrix, but it is not in the form we have described so far. The PAM250 mutation probability matrix (Fig. 3.13) is useful because it describes the frequency of amino acid replacements between distantly related proteins. We next need to convert the elements of a PAM mutation probability matrix into a scoring matrix, also called a log-odds matrix or relatedness odds matrix.

The cells in a log-odds matrix consist of scores as defined in Equation 3.1 above. The target frequencies q_{ij} are derived from a mutation probability matrix, such as those shown in Figs. 3.11 (for PAM1) and 3.13 (for PAM250). These values consist of positive numbers that sum to 1. The background frequencies $p_i p_j$ reflect the independent probabilities of each amino acid i, j occurring in this position. Its values were given in Table 3.2.

For this scoring system Dayhoff and colleagues took 10 times the base 10 logarithm of the odds ratio (Equation 3.1). Using the logarithm here is helpful because it allows us to sum the scores of the aligned residues when we perform an overall alignment of two sequences. (If we did not take the logarithm, we would need to multiply the ratios at all the aligned positions, and this is computationally more cumbersome.)

A log-odds matrix for PAM250 is shown in Fig. 3.14. The values have been rounded off to the nearest integer. Try using Equation 3.1 to make sure you understand how the mutation probability matrix (Fig. 3.13) is converted into the log-odds scoring matrix (Fig. 3.14). As an example, to determine the score assigned to two aligned tryptophan residues, the PAM250 mutation probability matrix value is 0.55 (Fig. 3.13), and the normalized frequency of tryptophan is 0.010 (Table 3.2). Thus,

$$S_{(tryptophan,tryptophan)} = 10 \times \log_{10}\left(\frac{0.55}{0.01}\right) = +17.4 \qquad (3.2)$$

Note that this scoring matrix is symmetric, in contrast to the mutation probability matrix in Fig. 3.13. In a comparison of two sequences it does not matter which is given first. In problem [3-6] of this chapter we will calculate the likelihood of changing cys to glu, then of changing glu to cys.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

FIGURE 3.14. Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30.

We state that a score of +17 for tryptophan indicates that the correspondence of two tryptophans in an alignment of homologous proteins is 50 times more likely than a chance alignment of two tryptophan residues. How do we derive the number 50? From Equation 3.1, let $S_{i,j} = +17$ and let the probability of replacement $q_{ij}/p_i = x$. Then $+17 = 10 \log_{10} x$, $+1.7 = \log_{10} x$, and $10^{1.7} = x = 50$.

This value is rounded off to 17 in the PAM250 log-odds matrix (Fig. 3.14). What do the scores in the PAM250 matrix signify? A score of −10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one-tenth as frequent as the chance alignment of these amino acids. This assumes that each was randomly selected from the background amino acid frequency distribution. A score of zero is neutral. A score of +17 for tryptophan indicates that this correspondence is 50 times more frequent than the chance alignment of this residue in a pairwise alignment. A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance. The highest values in this particular matrix are for tryptophan (17 for an identity) and cysteine (12), while the most severe penalties are associated with substitutions for those two residues. When two sequences are aligned and a score is given, that score is simply the sum of the scores for all the aligned residues across the alignment.

It is easy to see how different PAM matrices score amino acid substitutions by comparing the PAM250 matrix (Fig. 3.14) with a PAM10 matrix (Fig. 3.15). In the PAM10 matrix, identical amino acid residue pairs tend to produce a higher score than in the PAM250 matrix; for example, a match of alanine to alanine scores 7 versus 2, respectively. The penalties for mismatches are greater in the PAM10 matrix; for example, a mutation of aspartate to arginine scores −17 (PAM10) versus −1 (PAM250). PAM10 even has negative scores for substitutions (such as glutamate to asparagine, −5) that are scored positively in the PAM250 matrix (+1).

Practical Usefulness of PAM Matrices in Pairwise Alignment

We can demonstrate the usefulness of PAM matrices by performing a series of global pairwise alignments of both closely related proteins and distantly related proteins. For the closely related proteins we will use human beta globin (NP_000509) and beta globin from the chimpanzee *Pan troglodytes* (XP_508242); these proteins share 100% amino acid identity. The bit scores proceed in a fairly linear, decreasing fashion from about 590 bits using the PAM10 matrix to 200 bits using the PAM250 matrix and 100 bits using the PAM500 matrix (Fig. 3.16, black line). In this pairwise alignment there are no mismatches or gaps, and the high bit scores associated with low PAM matrices (such as PAM10) are accounted for by the lower relative entropy (defined below). The PAM10 matrix is thus appropriate for comparisons of closely related proteins. Next consider pairwise alignments of two relatively divergent proteins, human beta globin and alpha globin (NP_000549) (Fig. 3.16, red line). The PAM70 matrix yields the highest score. Lower PAM matrices (e.g., PAM10 to PAM60) produce lower bit scores because the sequences share only 42% amino acid identity, and mismatches are assigned large negative scores. We conclude that different scoring matrices vary in their sensitivity to protein sequences (or DNA sequences) of varying relatedness. When you compare two sequences you may need to repeat the search using several different scoring matrices. Alignment programs cannot be preset to choose the right matrix for each pair of sequences. Instead they begin with the most broadly useful scoring matrix such as BLOSUM62, which we describe next.

Important Alternative to PAM: BLOSUM Scoring Matrices

In addition to the PAM matrices, another very common set of scoring matrices is the blocks substitution matrix (BLOSUM) series. Henikoff and Henikoff (1992, 1996)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-2-	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

FIGURE 3.15. Log-odds matrix for PAM10. Low PAM values such as this are useful for aligning very closely related sequences. Compare this with the PAM250 matrix (Fig. 3.14) and note that there are larger positive scores for identical matches in this PAM10 matrix and larger penalties for mismatches.

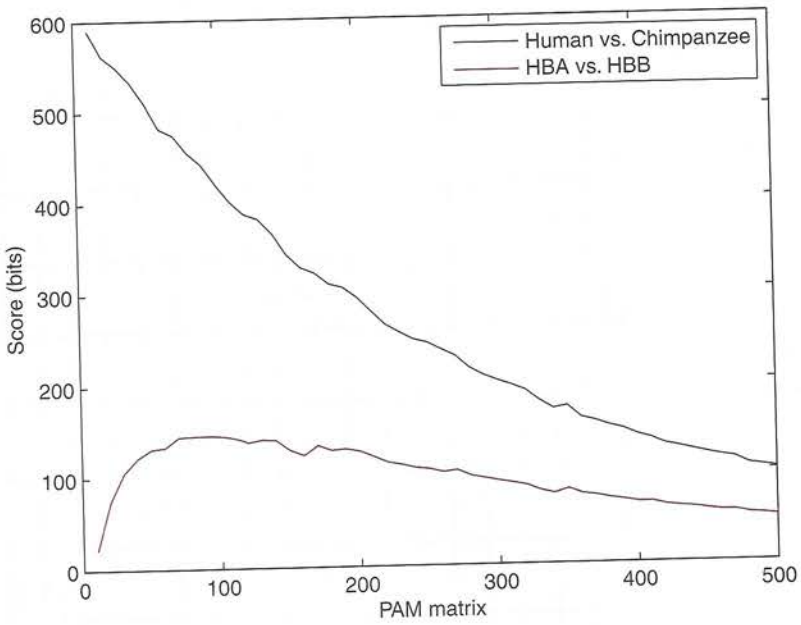


FIGURE 3.16. Global pairwise alignment scores using a series of PAM matrices. Two closely related globins (human and chimpanzee beta globin; black line) were aligned using a series of PAM matrices (x axis) and the bit scores were measured (y axis). For two distantly related globins (human alpha versus beta globin; red line) the bit scores are smaller for low PAM matrices (such as PAM1 to PAM20) because mismatches are severely penalized.

used the BLOCKS database, which consisted of over 500 groups of local multiple alignments (blocks) of distantly related protein sequences. Thus the Henikoffs focused on conserved regions (blocks) of proteins that are distantly related to each other. The BLOSUM scoring scheme employs a log-odds ratio using the base 2 logarithm:

$$s_{ij} = 2 \times \log_2\left(\frac{q_{ij}}{p_{ij}}\right) \tag{3.3}$$

Equation 3.3 resembles Equation 3.1 in its format. Karlin and Altschul (1990) and Altschul (1991) have shown that substitution matrices can be described in general in a log-odds form as follows:

$$s_{ij} = \left(\frac{1}{\lambda}\right) \ln\left(\frac{q_{ij}}{p_i p_j}\right) \tag{3.4}$$

The PAM matrix is given as 10 times the log base 10 of the odds ratio. The BLOSUM matrix is given as 2 times the log base 2 of the odds ratio. Thus, BLOSUM scores are not quite as large as they would be if given on the same scale as PAM scores. Practically, this difference in scales is not important because alignment scores are typically converted from raw scores to normalized bit scores (Chapter 4).

Here s_{ij} refers to the score of amino acid i aligning with j. q_{ij} are the positive target frequencies; these sum to 1. λ is a positive parameter that provides a scale for the matrix. We will again encounter λ when we describe the basic statistical measure of a BLAST result (Chapter 4, Equation 4.5).

The BLOSUM62 matrix is the default scoring matrix for the BLAST protein search programs at NCBI. It merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. If a block of aligned globin orthologs includes several that have 62%, 80%, and 95% amino acid identity, these would all be weighted (grouped) as one sequence. Substitution frequencies for the BLOSUM62 matrix are weighted more heavily by blocks of protein sequences having less than 62% identity. (Thus, this matrix is useful for scoring proteins that share less than 62% identity.) The BLOSUM62 matrix is shown in Fig. 3.17.

Henikoff and Henikoff (1992) tested the ability of a series of BLOSUM and PAM matrices to detect proteins in BLAST searches of databases. They found that

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

FIGURE 3.17. The BLOSUM62 scoring matrix of Henikoff and Henikoff (1992). This matrix merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. BLOSUM62 performs better than alternative BLOSUM matrices or a variety of PAM matrices at detecting distant relationships between proteins. It is thus the default scoring matrix for most database search programs such as BLAST (Chapter 4). Used with permission.

BLOSUM62 performed slightly better than BLOSUM60 or BLOSUM70 and dramatically better than PAM matrices at identifying various proteins. Their matrices were especially useful for identifying weakly scoring alignments. BLOSUM50 and BLOSUM90 are other commonly used scoring matrices in BLAST searches. (For an alignment of two proteins sharing about 50% identity, try using the BLOSUM50 matrix. The fasta family of sequence comparison programs use BLOSUM50 as a default.)

The relationships of the PAM and BLOSUM matrices are outlined in Fig. 3.18. To summarize, BLOSUM and PAM matrices both use log-odds values in their scoring systems. In each case, when you perform a pairwise sequence alignment (or when you search a query sequence against a database), you specify the exact matrix to use based on the suspected degree of identity between the query and its matches. PAM matrices are based on data from the alignment of closely related protein families, and they involve the assumption that substitution probabilities for highly related proteins (e.g., PAM40) can be extrapolated to probabilities for distantly related proteins (e.g., PAM250). In contrast, the BLOSUM matrices are based on empirical observations of more distantly related protein alignments. Note that a PAM30 matrix, which is available as an option on standard blastp searches at NCBI (Chapter 4), may be

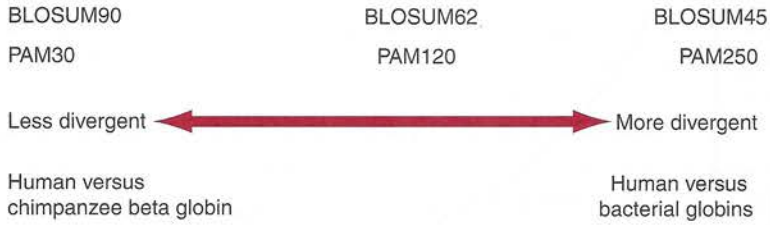


FIGURE 3.18. Summary of PAM and BLOSUM matrices. High-value BLOSUM matrices and low-value PAM matrices are best suited to study well-conserved proteins such as mouse and rat globins. BLOSUM matrices with low numbers (e.g., BLOSUM45) or high PAM numbers are best suited to detect distantly related proteins. Remember that in a BLOSUM45 matrix all members of a protein family with greater than 45% amino acid identity are grouped together, allowing the matrix to focus on proteins with less than 45% identity.

useful for identifying significant conservation between two closely related protei However, a BLOSUM matrix with a high value (such as the BLOSUM80 ma that is available at the NCBI blastp site) is not necessarily suitable for scoring clos related sequences. This is because the BLOSUM80 matrix is adapted to regions sequences that share up to 80% identity, but beyond that limited region two prote may share dramatically less amino acid identity (Pearson and Wood, 2001).

Pairwise Alignment and Limits of Detection: The "Twilight Zone"

A hit is a change in an amino acid residue that occurs by mutation. We discuss mutations (including multiple hits at a nucleotide position) in Chapter 7 (see Fig. 7.11). We discuss mutations associated with human disease in Chapter 20.

When we compare two protein sequences, how many mutations can occur betwe them before their differences make them unrecognizable? When we compar glyceraldehyde-3-phosphate dehydrogenase proteins, it was easy to see their relati ship (Fig. 3.7). However, when we compared human beta globin and myoglobin, t relationship was much less obvious (Fig. 3.5). Intuitively, at some point two homo gous proteins are too divergent for their alignment to be recognized as significa

The best way to determine the detection limits of pairwise alignments is throu statistical tests that assess the likelihood of finding a match by chance. These described below, and in Chapter 4. In particular we will focus on the expect value. It can also be helpful to compare the percent identity (and percent divergen of two sequences versus their evolutionary distance. Consider two protein sequenc each 100 amino acids in length, in which various numbers of mutations are int duced. A plot of the two diverging sequences has the form of a negative exponen (Fig. 3.19) (Doolittle, 1987; Dayhoff, 1978). If the two sequences have 100% am acid identity, they have zero changes per 100 residues. If they share 50% amino ac identity, they have sustained an average of 80 changes per 100 residues. One mig have expected 50 changes per 100 residues in the case of two proteins that sh 50% amino acid identity. However, any position can be subject to multiple hi Thus, percent identity is not an exact indicator of the number of mutations th have occurred across a protein sequence. When a protein sustains about 250 h

The plot in Fig. 3.19 reaches an asymptote below about 15% amino acid identity. This asymptote would reach about 5% (or the average background frequency of the amino acids) if no gaps were allowed in the comparison between the proteins.

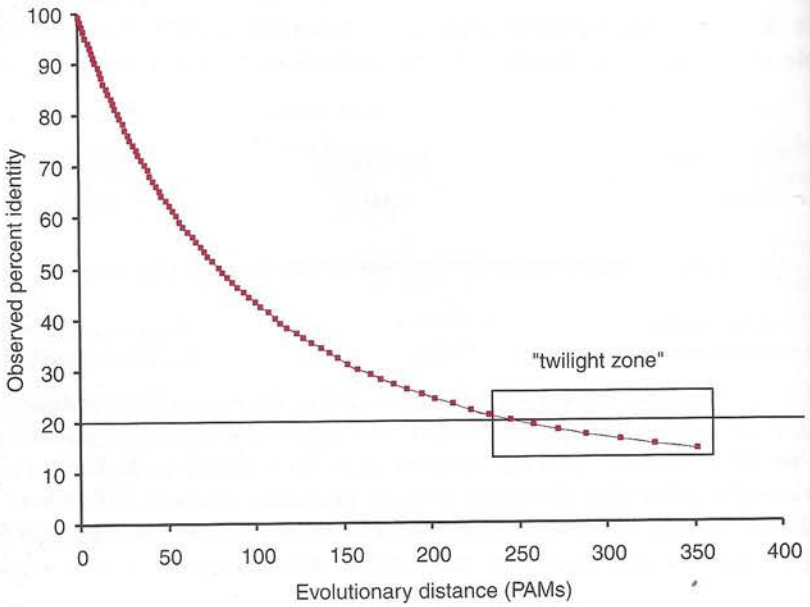


FIGURE 3.19. Two randomly diverging protein sequences change in a negatively exponential fashion. This plot shows the observed number of amino acid identities per 100 residues of two sequences (y axis) versus the number of changes that must have occurred (the evolutionary distance in PAM units). The twilight zone (Doolittle, 1987) refers to the evolutionary distance corresponding to about 20% identity between two proteins. Proteins with this degree of amino acid sequence identity may be homologous, but such homology is difficult to detect. This figure was constructed using MATLAB® software with data from Dayhoff (1978) (see Table 3.3).

TABLE 3-3 Relationship between Observed Number of Amino Acid Differences per 100 Residues of Two Aligned Protein Sequences and Evolutionary Difference[a]

Observed Differences in 100 Residues	Evolutionary Distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

[a]The number of changes that must have occurred, in PAM units.
Source: Adapted from Dayhoff (1978, p. 375). Used with permission.

per 100 amino acids, it may have about 20% identity with the original protein, and it can still be recognizable as significantly related. If a protein sustains 360 changes per 100 residues, it evolves to a point at which the two proteins share about 15% amino acid identity and are no longer recognizable as significantly related in a direct, pairwise comparison.

The PAM250 matrix assumes the occurrence of 250 point mutations per 100 amino acids. As shown in Fig. 3.19, this corresponds to the "twilight zone." At this level of divergence, it is usually difficult to assess whether the two proteins are homologous. Other techniques, including multiple sequence alignment (Chapter 6) and structural predictions (Chapter 11), are often very useful to assess homology in these cases. PAM matrices are available from PAM1 to PAM250 or higher, and a specific number of observed amino acid differences per 100 residues is associated with each PAM matrix (Table 3.3 and Fig. 3.19). Consider the case of the human alpha globin compared to myoglobin. These proteins are approximately 150 amino acid residues in length, and they may have undergone over 300 amino acid substitutions since their divergence (Dayhoff et al., 1972, p. 19). If there were 345 changes (corresponding to 230 changes per 100 amino acids), then an additional 100 changes would result in only 10 more observable changes (Dayhoff et al., 1972; Table 3.3).

There are about $2^{2n}/\sqrt{\pi n}$ possible global alignments between two sequences of length n (Durbin et al., 2000; Ewens and Grant, 2001). For two sequences of length 1000, there are about 10^{600} possible alignments. For two proteins of length 200 amino acid residues, the number of possible alignments is over 6×10^{58}.

Alignment Algorithms: Global and Local

Our discussion so far has focused on matrices that allow us to score an alignment between two proteins. This involves the generation of scores for identical matches, mismatches, and gaps. We also need an appropriate algorithm to perform the alignment. When two proteins are aligned, there is an enormous number of possible alignments.