6

Multiple Sequence Alignment

Introduction

When we consider a protein (or gene), one of the most fundamental questions is what other proteins are related. Biological sequences often occur in families. These families may consist of related genes within an organism (paralogs), sequences within a population (e.g., polymorphic variants), or genes in other species (orthologs). Sequences diverge from each other for reasons such as duplication within a genome or speciation leading to the existence of orthologs. We have studied pairwise comparisons of two protein (or DNA) sequences (Chapter 3), and we have also seen multiple related sequences in the form of profiles or as the output of a BLAST or other database search (Chapters 4 and 5). We will also explore multiple sequence alignments in the context of molecular phylogeny (Chapter 7), protein domains (Chapter 10), and protein structure (Chapter 11).

In this chapter, we consider the general problem of multiple sequence alignment from three perspectives. First, we describe five approaches to making multiple sequence alignments from a group of homologous sequences of interest. Second, we discuss multiple alignment of genomic DNA. This is typically a comparative genomics problem of aligning large chromosomal regions from different species. Third, we explore databases of multiply aligned sequences, such as Pfam, the protein family database. While multiple sequence alignment is commonly performed for both protein and DNA sequences, most databases consist of protein families only. Nucleotides corresponding to coding regions are typically less well conserved than

Bioinformatics and Functional Genomics, Second Edition. By Jonathan Pevsner

proteins because of the degeneracy of the genetic code. Thus they can be harder to align with with high confidence.

Multiple sequence alignments are of great interest because homologous sequences often retain similar structures and functions. Pairwise alignments may suffice to create links between structure and function. Multiple sequence alignments are very powerful because two sequences that may not align well to each other can be aligned via their relationship to a third sequence, thereby integrating information in a way not possible using only pairwise alignments. We can thus define members of a gene or protein family, and identify conserved regions. If we know a feature of one of the proteins (e.g., RBP4 transports a hydrophobic ligand), then when we identify homologous proteins, we can predict that they may have similar function. The overwhelming majority of proteins have been identified through the sequencing of genomic DNA or complementary DNA (cDNA; Chapter 8). Thus, the function of most proteins is assigned on the basis of homology to other known proteins rather than on the basis of results from biochemical or cell biological (functional) assays.

Definition of Multiple Sequence Alignment

Domains or motifs that characterize a protein family are defined by the existence of a multiple sequence alignment of a group of homologous sequences. A multiple sequence alignment is a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned. Homologous residues are aligned in columns across the length of the sequences. These aligned residues are homologous in an evolutionary sense: they are presumably derived from a common ancestor. The residues in each column are also presumed to be homologous in a structural sense: aligned residues tend to occupy corresponding positions in the three-dimensional structure of each aligned protein.

Multiple sequence alignments are easy to generate, even by eye, for a group of very closely related protein (or DNA) sequences. We have seen an alignment of closely related sequences (Fig. 3.7, GAPDH). As soon as the sequences exhibit some divergence, the problem of multiple alignment becomes extraordinarily difficult to solve. In particular, the number and location of gaps is difficult to assess. We saw an example of this with kappa caseins (Fig. 3.8), and in this chapter we will examine a challenging region of five distantly related globins. Practically, you must (1) choose homologous sequences to align, (2) choose software that implements an appropriate objective scoring function (i.e., a metric such as maximizing the total score of a series of pairwise alignments), and (3) choose appropriate parameters such as gap opening and gap extension penalties.

There is not necessarily one "correct" alignment of a protein family. This is because while protein structures tend to evolve over time, protein sequences generally evolve even more rapidly than structures. Looking at the sequences of human beta globin and myoglobin, we saw that they share only 25% amino acid identity (Fig. 3.5), but the three-dimensional structures are nearly identical (Fig. 3.1). In creating a multiple sequence alignment, it may be impossible to identify the amino acid residues that should be aligned with each other as defined by the three-dimensional structures of the proteins in the family. We often do not have high-resolution structural data available, and we rely on sequence data to generate the alignment. Similarly, we often do not have functional data to identify domains (such as the specific amino acids that form the catalytic site of an enzyme), so again we rely on sequence data. It is possible to compare the results of multiple sequence alignments that are generated solely from sequence data and to then examine known structures for those proteins. For a given pair of divergent but significantly related protein sequences (e.g., for two proteins sharing 30% amino acid identity), Chothia and Lesk (1986) found that about 50% of the individual amino acid residues are superposable in the two structures.

Aligned columns of amino acid residues characterize a multiple sequence alignment. This alignment may be determined because of features of the amino acids such as the following:

- There are highly conserved residues such as cysteines that are involved in forming disulfide bridges.
- There are conserved motifs such as a transmembrane domain or an immunoglobulin domain. We will encounter examples of protein domains and motifs (such as the PROSITE dictionary) in Chapter 10.
- There are conserved features of the secondary structure of the proteins, such as residues that contribute to α helices, β sheets, or transitional domains.
- There are regions that show consistent patterns of insertions or deletions.

Typical Uses and Practical Strategies of Multiple Sequence Alignment

When and why are multiple sequence alignments used?

- If a protein (or gene) you are studying is related to a larger group of proteins, this group membership can often provide insight into the likely function, structure, and evolution of that protein.
- Most protein families have distantly related members. Multiple sequence alignment is a far more sensitive method than pairwise alignment to detect homologs (Park et al., 1998). Profiles (such as those described for PSI-BLAST and hidden Markov models in Chapter 5) depend on accurate multiple sequence alignments.
- When one examines the output of any database search (such as a BLAST search), a multiple sequence alignment format can be extremely useful to reveal conserved residues or motifs in the output.
- If one is studying cDNA clones, it is common practice to sequence them. Multiple sequence alignment can show whether there are any variants or discrepancies in the sequences. Alignments of genomic DNA containing single nucleotide polymorphisms (SNPs; Chapter 16) are of interest, for example, in the identification of nonsynonymous SNPs.
- Analysis of population data can provide insight into many biological questions involving evolution, structure, and function. The PopSet portion of Entrez (described below) contains nucleotide (and protein) population data sets that are viewed as multiple alignments.
- When the complete genome of any organism is sequenced, a major portion of the analysis consists of defining the protein families to which all the gene products belong. Database searches effectively perform multiple sequence alignments, comparing each novel protein (or gene) to the families of all other known genes.

- We will see in Chapter 7 how phylogeny algorithms begin with multiple sequence alignments as the raw data with which to generate trees. The most critical part of making a tree is to produce an optimal multiple sequence alignment.
- The regulatory regions of many genes contain consensus sequences for transcription factor-binding sites and other conserved elements. Many such regions are identified based on conserved noncoding sequences that are detected using multiple sequence alignment.

Benchmarking: Assessment of Multiple Sequence Alignment Algorithms

We will describe five different approaches to creating multiple sequence alignments. How can we assess the accuracy and performance properties of the various algorithms? The performance depends on factors including the number of sequences being aligned, their similarity, and the number and position of insertions or deletions (McClure et al., 1994).

A convincing way to assess whether a multiple sequence alignment program produces a "correct" alignment is to compare the result with the alignment of known three-dimensional structures as established by x-ray crystallography (Chapter 11). Several databases have been constructed to serve as benchmark data sets. These are reference sets in which alignments are created from proteins having known structures. Thus, one can study proteins that are by definition structurally homologous. This allows an assessment of how successfully assorted multiple sequence alignment algorithms are able to detect distant relationships among proteins. For proteins sharing about 40% amino acid identity or more, most multiple sequence alignment programs produce closely similar results. For more distantly related proteins, the programs can produce markedly different alignments, and benchmarks are useful to compare accuracy.

The performance of a multiple sequence alignment algorithm relative to a benchmark data set is measured by some objective scoring function. One commonly used metric is the sum-of-pairs score (Box 6.1). This involves counting the number of

Box 6.1
Evaluating Multiple Sequence Alignments

Thompson et al. (1999) described two main ways to assess multiple sequence alignments. The first is the sum-of-pairs scores (SPS). This score increases as a program succeeds in aligning sequences relative to the BAliBASE or other reference alignment. The SPS assumes statistical independence of the columns. For an alignment of N sequences in M columns, the ith column is designated A_{i1}, $A_{i2}, \ldots, A_{iN}$. For each pair of residues A_{ij} and A_{ik}, a score of 1 is assigned ($p_{ijk} = 1$) if they are also aligned in the reference, and a score of 0 is assigned if they are not aligned ($p_{ijk} = 0$). Then for the entire ith column, the score S_i is given by:

$$S_i = \sum_{j=1, j\#k}^{N} \sum_{k=1}^{N} p_{ijk}$$

For the entire multiple sequence alignment, the SPS is given by:

$$\mathrm{SPS} = \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{Mr} S_{ri}}$$

Here S_{ri} is the score S_i for the ith column in the reference alignment, and Mr corresponds to the number of columns in the reference alignment.

A second approach is to create a column score (CS). For the ith column, $C_i = 1$ if all the residues in the column are aligned in the reference, and $C_i = 0$ if not.

$$\mathrm{CS} = \sum_{i=1}^{M} \frac{C_i}{M}$$

Sum-of-pairs scores and column scores have been used to assess the performance of multiple sequence alignment algorithms. Gotoh (1995) and others further described weighted sum-of-pairs scores that correct for biased contributions of sequences caused by divergent members of a group being aligned. Lassmann and Sonnhammer (2005) note that a column score becomes zero if even a single sequence is misaligned; thus it may be too stringent.

pairs of aligned residues that occur in the target and reference alignment, divided by the total number of pairs of residues in the reference.

Benchmark data sets may contain separate categories of multiple sequence alignments, such as those having proteins of varying length, varying divergence, insertions or deletions (indels) of various lengths, and varying motifs (such as internal repeats). Investigators routinely employ benchmark data sets to assess the performance of alignment algorithms (e.g., Morgenstern et al., 1996; McClure et al., 1994; Thompson et al., 1999; Gotoh, 1996; Briffeuil et al., 1998). Blackshields et al. (2006) compared the properties of six benchmark datasets (Table 6.1).

Another approach to benchmarking is to use a program such as ROSE (Stoye et al., 1998) that simulates the evolution of sequences. We introduced ROSE in

You can examine typical benchmark entries for the globins and the lipocalins from the HOMSTRAD database (Mizuguchi et al., 1998) in Web documents 6.1 and 6.2 at ► http://www.bioinfobook.org/chapter6. HOMSTRAD (the homologous structure alignment database) contains aligned three-dimensional structures of homologous proteins from over 1000 families. Later in this chapter, studying the T-Coffee suite of programs, we will introduce a new approach to benchmarking that is based on structural data but does not employ a benchmark database.

TABLE 6-1 Benchmark Data Sets to Assess Multiple Sequence Alignment Accuracy

Database	Reference	URL
BAliBASE	Thompson et al. (2005)	http://www-bio3d-igbmc.u-strasbg.fr/balibase/
HOMSTRAD	Mizuguchi et al. (1998)	http://www-cryst.bioc.cam.ac.uk/~homstrad/
IRMBASE	Subramanian et al. (2005)	http://dialign-t.gobics.de/main
OxBench	Raghava et al. (2003)	http://www.compbio.dundee.ac.uk/Software/Oxbench/oxbench.htm
Prefab	Edgar (2004b)	http://www.drive5.com/muscle/prefab.htm
SABmark	Van Walle et al. (2005)	http://bioinformatics.vub.ac.be/databases/content.html

ROSE software is available at ► http://bibiserv.techfak.uni-bielefeld.de/rose/.

Chapter 5 as a benchmark for analyzing genomic alignment software. It has also been used to assess multiple sequence alignment software such as Kalign (Lassmann and Sonnhammer, 2005) and MUSCLE (Edgar, 2004a).

Five Main Approaches to Multiple Sequence Alignment

There are many approaches to multiple sequence alignment; in the past decade many dozens of programs have been introduced. We may consider five algorithmic approaches: (1) exact methods, (2) progressive alignment (e.g., ClustalW), (3) iterative approaches (e.g., PRALINE, IterAlign, MUSCLE), (4) consistency-based methods (e.g., MAFFT, ProbCons), and (5) structure-based methods that include information about one or more known three-dimensional protein structures to facilitate creation of a multiple sequence alignment (e.g., Expresso). The programs we will describe in categories (3) to (5) are often overlapping; for example, all rely on progressive alignment and some combine iterative and structure-based approaches. All the programs offer trade-offs in speed and accuracy. MUSCLE and MAFFT are fastest, and are thus most useful for aligning large numbers of sequences. ProbCons and T-Coffee, although slower, are more accurate in many applications.

We will explore how one set of globin sequences can be aligned differently using various programs, and we will try to assess which alignments are most accurate. A related question is the consequence of a misalignment. Potentially, the conservation of critical residues (such as active site amino acids of an enzyme, the heme-binding residues of a globin, or conserved residues that cause disease when mutated) may be missed. Phylogenetic inference (Chapter 7) may be compromised because all molecular phylogeny algorithms depend on a multiple sequence alignment as input. Protein structure prediction (Chapter 11) is severely compromised by faulty multiple sequence alignment, which is often a first step in homology-based modeling.

We will explore sets of distantly and closely related globin sequences in the FASTA format. These are available as web documents 6.3 and 6.4 at ► http://www.bioinfbook.org/chapter6. There are many ways that you can easily obtain a group of sequences in the FASTA format. Examples include HomoloGene at NCBI (for eukaryotic proteins), or you can select any subset of the results of a BLAST search and view the sequences in Entrez Protein (or Entrez Nucleotide) in the FASTA format.

The programs we will explore can be used by web interfaces, although local installation of the programs typically allows you access to a more complete package of options. All the web interfaces allow you to paste in a set of DNA, RNA, or protein sequences in the FASTA format, or to upload a text file containing these sequences.

Exact Approaches to Multiple Sequence Alignment

Dynamic programming as described by Needleman and Wunsch (1970) for pairwise alignment is guaranteed to identify the optimal global alignment(s). Exact methods for multiple sequence alignment employ dynamic programming, although the matrix is multidimensional rather than two-dimensional. The goal is to maximize the summed alignment score of each pair of sequences. Exact methods generate optimal alignments but are not feasible in time or space for more than a few sequences. For N sequences, the computational time that is required is $O(2^N L^N)$ where N is the number of sequences and L is the average sequence length. An exact multiple sequence alignment of more than four or five average sized proteins would consume prohibitively too much time. Nonexact methods, which we will discuss next, are computationally feasible. For example, ClustalW has time complexity $O(N^4 + L^2)$ and MUSCLE has time complexity $O(N^4 + NL^2)$. Although they are faster, these heuristic approaches are not guaranteed to produce optimal alignments.

Progressive Sequence Alignment

The most commonly used algorithms that produce multiple alignments are derived from the progressive alignment method. This was proposed by Fitch and Yasunobu (1975) and described by Hogeweg and Hesper (1984) who applied it to the alignment of 5S ribosomal RNA sequences. The method was popularized by Feng and Doolittle (1987, 1990). It is called "progressive" because the strategy entails calculating pairwise sequence alignment scores between all the proteins (or nucleic acid sequences) being aligned, then beginning the alignment with the two closest sequences and progressively adding more sequences to the alignment. A benefit of this approach is that it permits the rapid alignment of even hundreds of sequences. A major limitation is that the final alignment depends on the order in which sequences are joined. Thus, it is not guaranteed to provide the most accurate alignments.

Note that while most database searches such as BLAST rely on local alignment strategies, many multiple sequence alignments focus on global alignments, or a combination of global and local strategies.

Perhaps the most popular web-based program for performing progressive multiple sequence alignment is ClustalW (Thompson et al., 1994). There are many ways to access the program (Box 6.2). The ClustalW algorithm proceeds in three stages. We can illustrate the procedure by aligning five distantly related globins, selected from Entrez and pasted into a text document in the FASTA format (Fig. 6.1). The results are shown in Figs. 6.2 and 6.3. Later we will also align five closely related globins (Figs. 6.4 and 6.5). In this particular example we select proteins for which the corresponding three-dimensional structure has been solved by x-ray crystallography. This will help us to interpret the accuracy of the alignment from a structural perspective as well as an evolutionary perspective.

1. In stage 1, the global alignment approach of Needleman and Wunsch (1970; Chapter 3) is used to create pairwise alignments of every protein that is to be included in a multiple sequence alignment (Fig. 6.2, stage 1). As shown in the figure, for an alignment of five sequences, 10 pairwise alignment scores are generated.

Algorithms that perform pairwise alignments generate raw similarity scores. Note that for the default setting of ClustalW the scores are simply the percent identities. Many progressive sequence alignment algorithms including ClustalW use a distance matrix rather than a similarity matrix to describe the relatedness of the proteins. The conversion of similarity scores for each pair of sequences to distance scores is outlined in Box 6.3. The purpose of generating distance measures is to generate a guide tree (stage 2, below) to construct the alignment.

For N sequences that are multiply aligned, the number of pairwise alignments that must be calculated for the initial matrix equals $\frac{1}{2}(N - 1)(N)$. For five proteins, 10 pairwise alignments are made. For a multiple sequence alignment of 500 proteins, (499)(500)/2 = 12,250 pairwise alignments are made; this is why the speed of an algorithm can be a concern. ClustalW is slow relative to other approaches such as MUSCLE, described below, but for most typical applications its speed is quite reasonable.

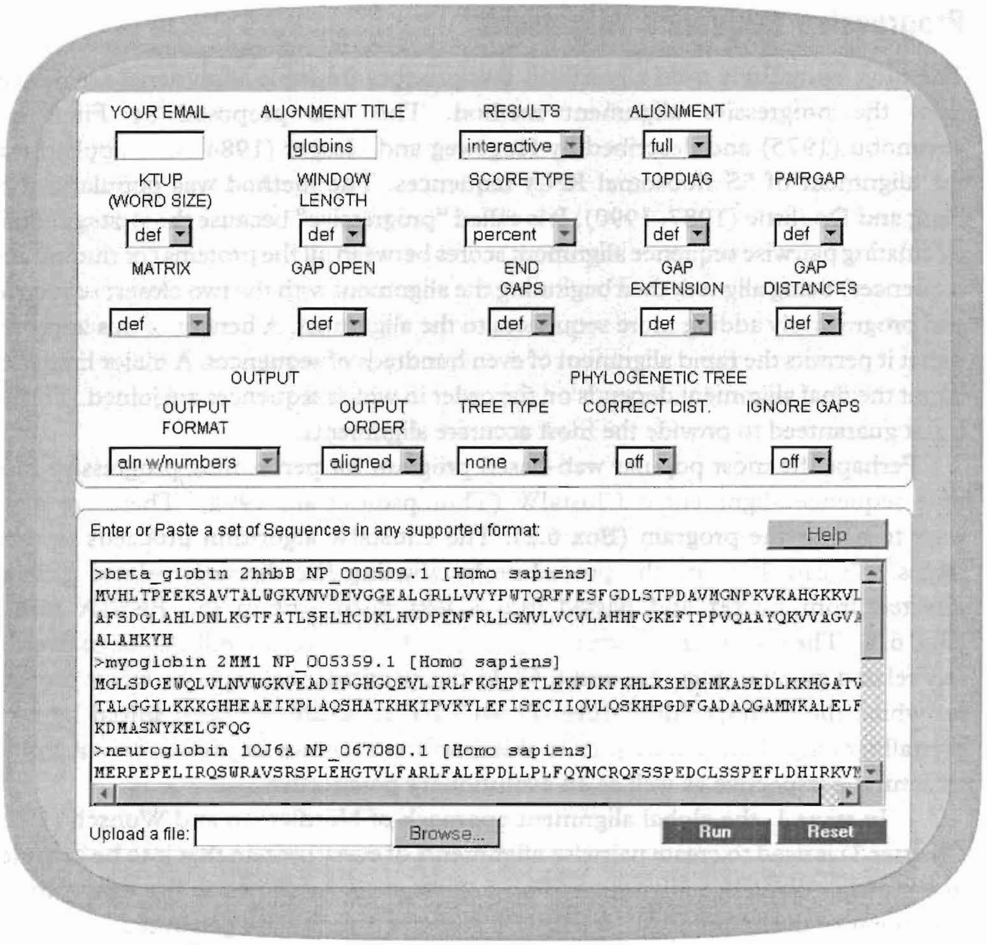
To confirm that the ClustalW scores are percent identities, perform pairwise alignments between any two of the sequences in Fig. 6.2 or 6.4 using BLAST at NCBI (Chapter 3).

Box 6.2
Using ClustalW

ClustalW is accessed online at many servers, including ► http://www.ebi.ac.uk/clustalw/, where it is hosted by the European Bioinformatics Institute.

Another way to access ClustalW is through the EMBOSS program emma. A variety of EMBOSS servers hosting emma are available, including ► http://phytophthora.vbi.vt.edu/EMBOSS/, ► http://bioportal.cgb.indiana.edu/cgi-bin/emboss/emma and ► http://embossgui.sourceforge.net/demo/emma.html.

ClustalX is a downloadable stand-alone program related to ClustalW (Thompson et al., 1997). ClustalX offers a graphical user interface for editing multiple sequence alignments. You can obtain ClustalX at ► http://bips.u-strasbg.fr/fr/Documentation/ClustalX/. An introductory tutorial for using ClustalX in conjunction with phylogeny software has been written by Hall (2001).



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT
	globins	interactive	full

KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVL
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVA
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATV
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELF
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVM
```

Upload a file:

FIGURE 6.1. Multiple sequence alignment of five distantly related globins using the ClustalW server at EBI (▶ http://www.ebi.ac.uk/clustalw/). Five distantly related globin proteins were pasted in using the FASTA format from Entrez (NCBI).

Stage 1: generate a series of pairwise alignments

```
SeqA Name          Len(aa)  SeqB Name          Len(aa)  Score
=============================================================
1    beta_globin   147      2    myoglobin     154      25
1    beta_globin   147      3    neuroglobin   151      15
1    beta_globin   147      4    soybean       144      13
1    beta_globin   147      5    rice          166      21
2    myoglobin     154      3    neuroglobin   151      16
2    myoglobin     154      4    soybean       144      8
2    myoglobin     154      5    rice          166      12
3    neuroglobin   151      4    soybean       144      17
3    neuroglobin   151      5    rice          166      18
4    soybean       144      5    rice          166      43 ◀—— 1
=============================================================
```

Stage 2: create a guide tree, calculated from a distance matrix

```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```

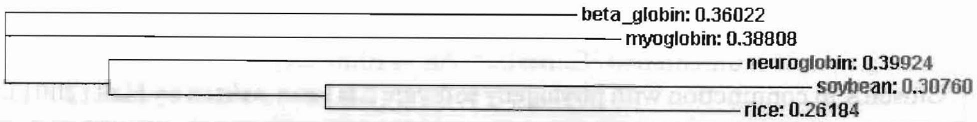
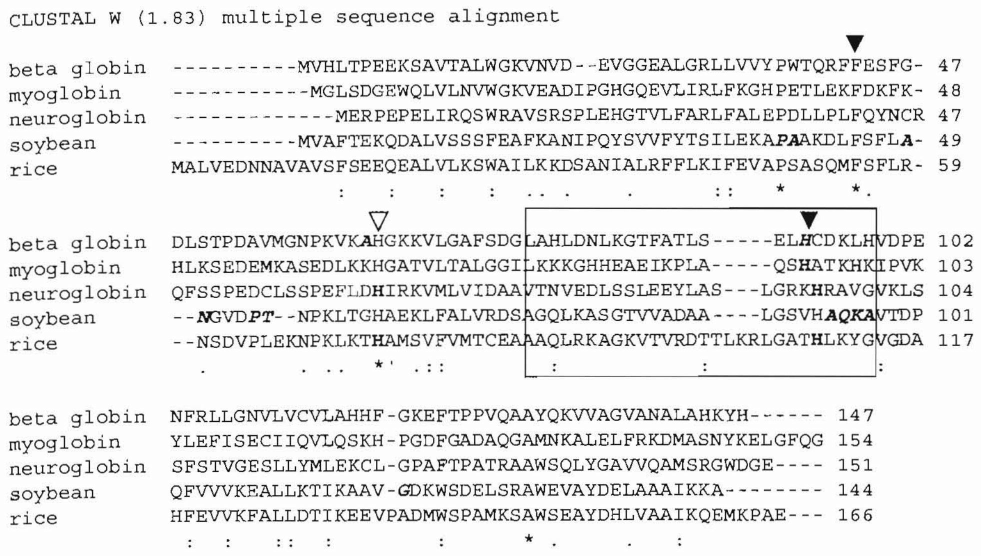


FIGURE 6.2. Progressive alignment method of Feng and Doolittle (1987) used by many multiple alignment programs such as ClustalW. In stage 1, a series of pairwise alignments is generated for five distantly related globins (see Fig. 6.1). Note that the best score is for an alignment of two plant globins (score = 43; arrow 1). In stage 2, a guide tree is calculated describing the relationships of the five sequences based on their pairwise alignment scores. A graphical representation of the guide tree is shown using the JalView tool at the ClustalW web server. Branch lengths (rounded off) reflect distances between sequences and are indicated on the tree; compare to Fig. 6.4.



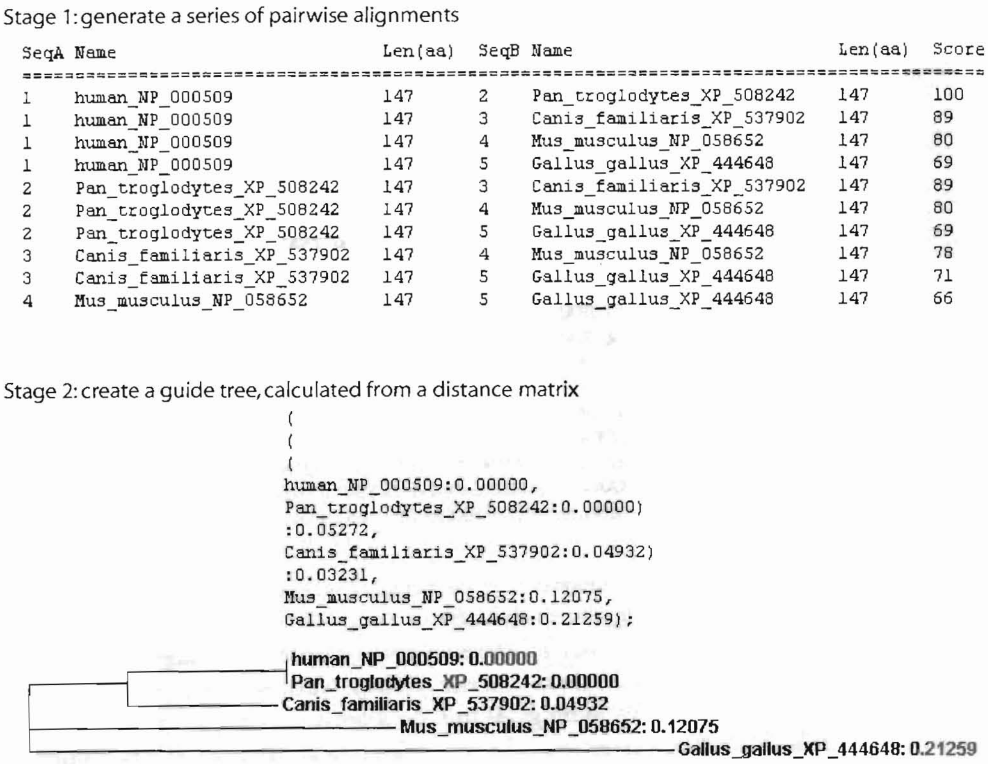
```
CLUSTAL W (1.83) multiple sequence alignment

                                                      ▼
beta globin     ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin       -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin     -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean         ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice            MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                             :   :   :   :   .. .   .      ::   *     *.

                                ▽                                 ▼
beta globin     DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin       HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean         --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice            --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                  .       . ..  *' .::        :           :             :

beta globin     NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin       YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean         QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice            HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
                 :  :   ::  :        :      * .     .   :
```

FIGURE 6.3. Multiple sequence alignment of five distantly related globins. The output is from ClustalW using the progressive alignment algorithm of Feng and Doolittle (1987). In stage 3, a multiple sequence alignment is created by performing progressive sequence alignments. First, the two closest sequences are aligned (soybean and rice globins). Next, further sequences are added in an order based on their position in the guide tree. An asterisk indicates positions in which the amino acid residue is 100% conserved in a column; a colon indicates conservative substitutions; a dot indicates less conservative substitutions. The proteins are human beta globin (accession NP_000509; Protein Data Bank identifier 2hhb), human myoglobin (NP_005359; 2MM1), human neuroglobin (NP_067080; 1OJ6A), leghemoglobin (from the soybean Glycine max*; 1FSL), and nonsymbiotic plant hemoglobin (from rice; 1D8U). Regions of alpha helices (defined in Chapter 11) based on x-ray crystallography are indicated in red letters. Three highly conserved residues are indicated by arrowheads: phe44 of myoglobin (red arrowhead), his65 (open arrowhead); and his93 (black arrowhead). These two histidines are important in coordinating protein binding to the heme group. A box surrounds the second histidine including five amino acids downstream (to the carboxy-terminal) and 17 amino acids upstream (to the end of an alpha helical region). We will discuss the alignment within this box for ClustalW in comparison to other alignment programs (Fig. 6.6).*

In our example, note that the best pairwise global alignment score is for rice versus soybean hemoglobin (Fig. 6.2, arrow 1). For a group of closely related beta globins, all have high scores (Fig. 6.4), even for sequences from avian and mammalian species that diverged over 300 million years ago.

2. In the second stage, a guide tree is calculated from the distance (or similarity) matrix. There are two principal ways to construct a guide tree: the unweighted pair group method of arithmetic averages (UPGMA) and the neighbor-joining method. We will define these algorithms in Chapter 7. The two main features of a tree are its topology (branching order) and branch lengths (which can be drawn so that they are proportional to evolutionary distance). Thus, the tree reflects the relatedness of all the proteins to be multiply aligned.

In ClustalW, the tree is described with a written syntax called the Newick format, as well as with a graphical output (Figs. 6.2 and 6.4, stage 2). The chicken sequence has the lowest score relative to the human, chimpanzee, dog, and mouse beta globins, and this is reflected in its position in the guide tree (Fig. 6.4, stages 1 and 2). A tree can also be displayed graphically at the ClustalW site by using the JalView option.



Stage 1:generate a series of pairwise alignments

```
SeqA Name                         Len(aa)  SeqB Name                         Len(aa)  Score
=========================================================================================
1    human_NP_000509              147      2    Pan_troglodytes_XP_508242    147      100
1    human_NP_000509              147      3    Canis_familiaris_XP_537902   147      89
1    human_NP_000509              147      4    Mus_musculus_NP_058652       147      80
1    human_NP_000509              147      5    Gallus_gallus_XP_444648      147      69
2    Pan_troglodytes_XP_508242    147      3    Canis_familiaris_XP_537902   147      89
2    Pan_troglodytes_XP_508242    147      4    Mus_musculus_NP_058652       147      80
2    Pan_troglodytes_XP_508242    147      5    Gallus_gallus_XP_444648      147      69
3    Canis_familiaris_XP_537902   147      4    Mus_musculus_NP_058652       147      78
3    Canis_familiaris_XP_537902   147      5    Gallus_gallus_XP_444648      147      71
4    Mus_musculus_NP_058652       147      5    Gallus_gallus_XP_444648      147      66
```

Stage 2: create a guide tree, calculated from a distance matrix

```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```

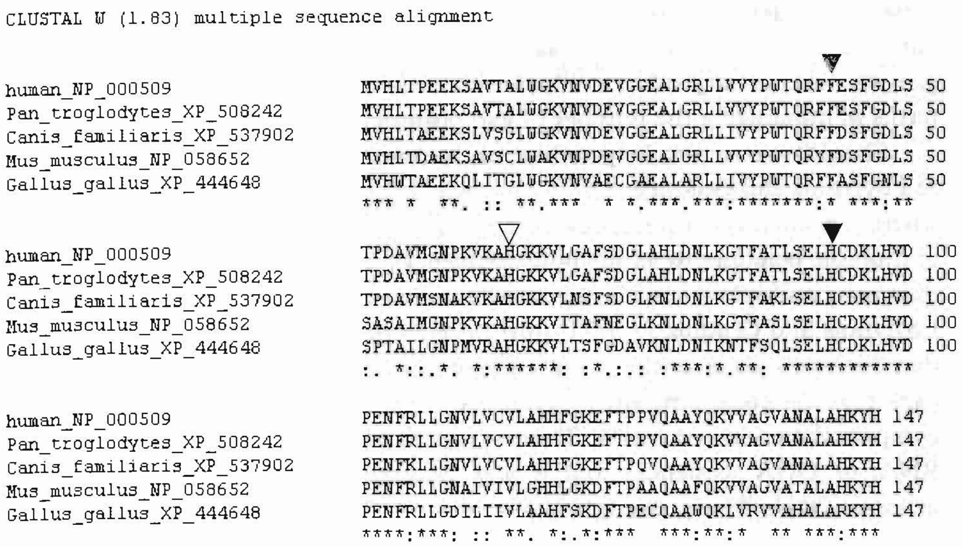
human_NP_000509: 0.00000
Pan_troglodytes_XP_508242: 0.00000
Canis_familiaris_XP_537902: 0.04932
Mus_musculus_NP_058652: 0.12075
Gallus_gallus_XP_444648: 0.21259

FIGURE 6.4. *Example of a multiple sequence alignment of closely related globin proteins using the progressive sequence aligment method of Feng and Doolittle (1987) as implemented by ClustalW. Compare these scores to those for distantly related proteins (Fig. 6.2), and note that the pairwise alignment scores are consistently higher and the distances (reflected in branch lengths on the guide tree) are much shorter.*

Guide trees are usually not considered true phylogenetic trees, but instead are templates used in the third stage of ClustalW to define the order in which sequences are added to a multiple alignment. A guide tree is estimated from a distance matrix based on the percent identities between sequences you are aligning. In constrast, a phylogenetic tree almost always includes a model to account for multiple substitutions that commonly occur at the position of aligned amino acids (or nucleotides), as discussed in Chapter 7.

3. In stage 3, the multiple sequence alignment is created in a series of steps based on the order presented in the guide tree. The algorithm first selects the two most closely related sequences from the guide tree and creates a pairwise alignment. These two sequences appear at the terminal nodes of the tree, that is, the locations of extant sequences. For example, rice globin and soybean globin are aligned. The next sequence is either added to the pairwise alignment (to generate an aligned group of three sequences, sometimes called a profile) or used in another pairwise alignment. At some point, profiles are aligned with profiles. The alignment continues progressively until the root of the tree is reached, and all sequences have been aligned. At this point a full multiple sequence alignment is obtained (Figs. 6.3 and 6.5, stage 3).

In the alignment of five distantly related globins, we can note that a highly conserved phenylalanine is aligned (Fig. 6.3, red arrowhead) as is a histidine that coordinates heme binding in most globins (open arrowhead). However, an even more highly conserved histidine (black arrowhead) is aligned in beta globin and myoglobin, but is placed in a separate column for neuroglobin and two plant globins. This represents a misalignment, and we will explore how other programs treat this region. For a group of closely related globins, the level of conservation is so high that there are no gaps and thus no ambiguities about how to perform the alignment (Fig. 6.5).



```
CLUSTAL W (1.83) multiple sequence alignment

                                                                              ▼
human_NP_000509                     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Pan_troglodytes_XP_508242           MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Canis_familiaris_XP_537902          MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
Mus_musculus_NP_058652              MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
Gallus_gallus_XP_444648             MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLS 50
                                    *** *  **. :: **.***  * *.***.***:*******:* ***:**

                                                 ▽                            ▼
human_NP_000509                     TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Pan_troglodytes_XP_508242           TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Canis_familiaris_XP_537902          TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
Mus_musculus_NP_058652              SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100
Gallus_gallus_XP_444648             SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100
                                    :. *::.*. *:******: :*.:.: :***:*.**: ************

human_NP_000509                     PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Pan_troglodytes_XP_508242           PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Canis_familiaris_XP_537902          PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
Mus_musculus_NP_058652              PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
Gallus_gallus_XP_444648             PENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147
                                    ****:***: :: **. *:.*:***  ***:**:*  ** ***:***
```

FIGURE 6.5. *Multiple sequence of five closely related beta globin orthologs (see Fig. 6.4). The output is a screen capture from ClustalW using the progressive alignment algorithm of Feng and Doolittle. The arrowheads (red, open, and black) correspond to the human beta globin phe44, his72, and his104 residues, respectively. These are highly conserved among the globin superfamily.*

Box 6.3
Similarity versus Distance Measures

Trees that represent protein or nucleic acid sequences usually display the differences between various sequences. One way to measure distances is to count the number of mismatches in a pairwise alignment. Another method, employed by the Feng and Doolittle progressive alignment algorithm, is to convert similarity scores to distance scores. Similarity scores are calculated from a series of pairwise alignments among all the proteins being multiply aligned. The similarity scores S between two sequences (i, j) are converted to distance scores D using the equation

$$D = -\ln S_{eff}$$

where

$$S_{eff} = \frac{S_{\text{real}(ij)} - S_{\text{rand}(ij)}}{S_{\text{iden}(ij)} - S_{\text{rand}(ij)}} \times 100$$

Here, $S_{\text{real}(ij)}$ describes the observed similarity score for two aligned sequences i and j, $S_{\text{iden}(ij)}$ is the average of the two scores for the two sequences compared to themselves (if score i compared to i receives a score of 20 and score j compared to j receives a score of 10, then $S_{\text{iden}(ij)} = 15$); $S_{\text{rand}(ij)}$ is the mean alignment score derived from many (e.g., 1000) random shufflings of the sequences; and S_{eff} is a normalized score. If sequences i, j have no similarity, then $S_{eff} = 0$ and the distance is infinite. If sequences i, j are identical, then $S_{eff} = 1$ and the distance is 0.

The Feng–Doolittle approach includes the rule "once a gap, always a gap." The most closely related pair of sequences is aligned first. As further sequences are added to the alignment, there are many ways that gaps could be included. The rationale for the "once a gap, always a gap" rule is that the two most closely related sequences that are initially aligned should be weighted most heavily in assigning gaps. ClustalW dynamically assigns position-specific gap penalties that increase the likelihood of having a new gap occur in the same position as a preexisting gap. That serves to give the overall alignment a block-like structure that often appears efficient in terms of minimizing the number of gap positions.

Should an insertion be penalized the same amount as a deletion? No, according to Loytynoja and Goldman (2005): a single deletion event is typically penalized once where it occurs, but a single insertion event that occurs once inappropriately results in multiple penalties to all the other sequences. The result of these high penalties is that many multiple sequence alignments are unrealistically aligned with too few gaps. Loytynoja and Goldman (2005) introduced a pair hidden Markov model approach that distinguishes insertions from deletions. They showed that their method creates gaps that are consistent with phylogeny, even though the alignments appear less compact than with ClustalW. Their approach applies to the alignment of protein, RNA, or DNA sequences, but it may be especially useful for the alignment of genomic DNA. There, overfitting may occur with traditional progressive alignment, for example when one sequence has long insertions. The approach of Loytynoja and Goldman (2005), reviewed in Higgins et al. (2005), provides multiple sequence alignments that have more gaps but are likely to be more accurate, based on criteria such as correct alignment of exons.

ClustalW implements a series of additional features to optimize the alignment (Thompson et al., 1994). The distance of each protein (or DNA) sequence from the root of the guide tree is calculated, and those sequences that are most closely related are downweighted by a multiplicative factor. This adjustment assures that if an alignment includes a group of very closely related sequences as well as another group of divergent sequences, the closely related ones will not overly dominate the final multiple sequence alignment. Other adjustments include the use of a series of scoring matrices that are applied to pairwise alignments of proteins depending on their similarity, and compensation for differences in sequence length.

The website ► http://msa.cgb.ki.se includes Kalign for alignment, Kalignvu as a viewer, and Mumsa to assess the quality of a multiple sequence alignment (Lassmann and Sonnhammer, 2006). Kalign is also offered through the European Bioinformatics Institute (► http://www.ebi.ac.uk/kalign/).

Many other algorithms use variants of progressive alignment. For example, Kalign employs a string-matching algorithm to achieve speeds ten times faster than ClustalW (Lassmann and Sonnhammer, 2005). Kalign aligns 100 protein sequences of length 500 residues in less than a second.

Iterative Approaches

Iterative methods compute a suboptimal solution using a progressive alignment strategy, and then modify the alignment using dynamic programming or other methods until a solution converges. Thus, they create an initial alignment and then modify it to try to improve it. Progressive alignment methods have the inherent limitation that once an error occurs in the alignment process it cannot be corrected, and iterative approaches can overcome this limitation. In standard dynamic programming the branching order of the guide tree may be suboptimal, or the scoring parameters may cause gaps to be misplaced. Iterative refinement can search for more optimal solutions stochastically (seeking higher maximal scores according to some metric such as the sum-of-pairs scores; Box 6.1) or by systematically extracting and realigning sequences from an initial profile that is generated. Examples of programs employing iterative approaches are MAFFT (Multiple Alignment using Fast Fourier Transform) (Katoh et al., 2005), Iteralign (Karlin and Brocchieri, 1998), Praline (Profile ALIgNmEnt) (Heringa, 1999; Simossis and Heringa, 2005), and MUSCLE (MUltiple Sequence Comparison by Log-Expectation) (Edgar, 2004a, 2004b).

MAFFToffers a suite of tools with choices of more speed or accuracy. The fastest version involves progressive alignment using matching 6-tuples (strings of six residues) to calculate pairwise distances. This approach is called *k*-mer counting. A *k*-mer (also called a *k*-tuple or word) is a contiguous subsequence of length *k*. *k*-mer counting is extremely fast because it requires no alignment. The initial distance matrix can optionally be recalculated once all pairwise alignments are calculated, yielding a more reliable progressive alignment. In the iterative refinement step, a weighted sum-of-pairs score is calculated and optimized. MAFFT allows options including global or local pairwise alignment.

MAFFT and PRALINE can both incorporate information from homologous sequences that are analyzed in addition to those you submit for multiple sequence alignment. These sequences are used to improve the multiple sequence alignment; in the case of MAFFT, the extra sequences are then removed. PRALINE performs a PSI-BLAST search (Chapter 5) on the query protein sequences and then performs progressive alignment using the PSI-BLAST profiles. PRALINE also permits the incorporation of predicted secondary structure information.

MAFFT is available at the EBI website, ► http://www.ebi.ac.uk/mafft/, or with more options from its project home page, ► http://align.bmr.kyushu-u.ac.jp/mafft/software/. PRALINE can be accessed from ► http://zeus.cs.vu.nl/programs/pralinewww/.

Since its introduction in 2004, the MUSCLE program of Robert Edgar (2004a, 2004b) has become popular because of its accuracy and its exceptional speed, especially for multiple sequence alignments involving large numbers of sequences. For example, 1000 protein sequences of average length 282 residues were aligned in 21 seconds on a desktop computer (Edgar, 2004a). MUSCLE operates in a series of three stages. First, a draft progressive alignment is generated. To achieve this, the algorithm calculates the similarity between each pair of sequences using either the fractional identity (calculated from a global alignment of each pair of sequences), or *k*-mer counting. Based on the similarities, MUSCLE calculates a triangular distance matrix, then constructs a rooted tree using UPGMA or neighbor-joining (see Chapter 7). Sequences are added progressively to the multiple sequence alignment following the branching order of the tree. In the second stage, MUSCLE improves the tree and builds a new progressive alignment (or a new set of alignments). The similarity of each pair of sequences is assessed using the fractional identity, and a tree is constructed using a Kimura distance matrix (discussed in Chapter 7). In a comparison of two sequences there is some likelihood that multiple amino acid (or nucleotide) substiutions occurred at any given position, and the Kimura distance matrix provides a model for such changes. As each tree is constructed it is compared to the tree from stage 1, and the process results in an improved progressive alignment. In stage 3 the guide tree is iteratively refined by systematically partitioning the tree to obtain subsets; an edge (branch) of the tree is deleted to create a bipartition. Next, MUSCLE extracts a pair of profiles (multiple sequence alignments), and realigns them (performing profile-profile alignment; see Box 6.4). The algorithm accepts or rejects the newly generated alignment based on whether the sum-of-pairs score increases. All edges of the tree are systematically visited and deleted to create bipartitions. This iterative refinement step is rapid and had been shown earlier to increase the accuracy of the multiple sequence alignment (Hirosawa et al., 1995).

The idea of a triangular distance matrix in stage 1 is that the distance measure between sequences (A,B) equals the distance of (A,C) plus (B,C). This is a good approximation for closely related sequences, but the accuracy is further increased using the Kimura distance correction in stage 2.

MUSCLE can be downloaded or accessed via web servers at ► http://www.drive5.com/muscle/ or at the European Bioinformatics website, ► http://www.ebi.ac.uk/muscle/.

The alignments of five distantly related globins using PRALINE (Fig. 6.6a) and MUSCLE (Fig. 6.6b) show a somewhat different result than we saw with ClustalW (Fig. 6.3). In the boxed region there are only 10 total gaps with PRALINE and 4 with MUSCLE, compared with 17 using ClustalW. This reflects a more compact overall alignment. Both these programs still fail to align the highly conserved histidine (Fig. 6.6a and b, black arrowhead).

Box 6.4
Profile-Profile Alignment with the MUSCLE Algorithm

The name MUSCLE (multiple sequence comparison by log expectation) includes the phrase "log expectation." Like ClustalW, MUSCLE measures the distance between sequences (Edgar, 2004a, 2004b). In its third stage, MUSCLE iteratively refines a multiple sequence alignment by deleting the edge of the guide tree to form a bipartition, then extracting a pair of profiles and realigning them. It does this using several scoring functions to optimally align pairs of columns. For amino acid types i and j, p_i is the background probability of i, p_{ij} is the joint probability of i and j being aligned, S_{ij} is the score from a substitution matrix, f_i^x is the observed frequency of i in column x of the first profile, f_G^x is the observed frequency of gaps in column x, and α_i^x is the estimated probability of observing residue i in position x in the family based on the observed frequencies f. (Note that $S_{ij} = \log(p_{ij}/p_i p_j)$ as discussed in Chapter 3.) MUSCLE, ClustalW, and MAFFT use a profile sum-of-pairs (PSP) scoring function:

$$PSP^{xy} = \sum_i \sum_j f_i^x f_j^y S_{ij}$$

PSP is a sequence-weighted sum of substitution matrix scores for each pair of letters (one from each column that is being aligned in a pairwise fashion). The PSP function maximizes the sum-of-pairs objective score. MUSCLE applies two PAM matrices for its PSP function. MUSCLE also employs a novel log-expectation (LE) score that is defined as follows:

$$LE^{xy} = (1 - f_G^x)(1 - f_G^y) \log \sum_i \sum_j f_i^x f_j^y \frac{p_{ij}}{p_i p_j}$$

The factor $(1 - f_G)$ is the occupancy of a column. This promotes the alignment of columns that are highly occupied (i.e., that have fewer gaps) while downweighting column pairs with many gaps. Edgar (2004a) reported that this significantly improved the accuracy of the alignment.

Consistency-Based Approaches

In progressive alignments using the Feng–Doolittle approach, pairwise alignment scores are generated and used to build a tree. Consistency-based methods adopt a different approach by using information about the multiple sequence alignment as it is being generated to guide the pairwise alignments. We will discuss two consistency-based multiple sequence alignment programs: ProbCons (Do et al., 2005) and T-Coffee (Notredame et al., 2000). The MAFFT program also includes an iterative refinement approach with consistency-based scores (Katoh et al., 2005).

The idea of consistency is that for sequences x, y, and z, if residue x_i aligns with z_k and z_k aligns with y_j, then x_i should align with y_j. Consistency-based techniques score pairwise alignments in the context of information about multiple sequences, for example, adjusting the score of x_i to y_i based on the knowledge that z_k aligns to both x_i and to y_i. This approach is distinctive because it incorporates evidence from multiple sequences to guide the creation of a pairwise alignment (Do et al., 2005). Using the notation given in a review by Wallace and colleagues (2005), the likelihood that residue i from sequence x and residue j from sequence y are aligned, given the sequences of x and y, is given by:

$$P(x_i \sim y_j | \, x, y) \tag{6.1}$$

This is the posterior probability, and it is calculated for each pair of amino acids. The consistency transformation further incorporates data from additional residues to improve the estimate of two residues aligning (that is, given information about how x and y each align with z):

$$P(x_i \sim y_j | \, x, y, z) \approx \sum_k P(x_i \sim z_k \,|\, x, z) P(y_i \sim z_k \,|\, y, z) \tag{6.2}$$

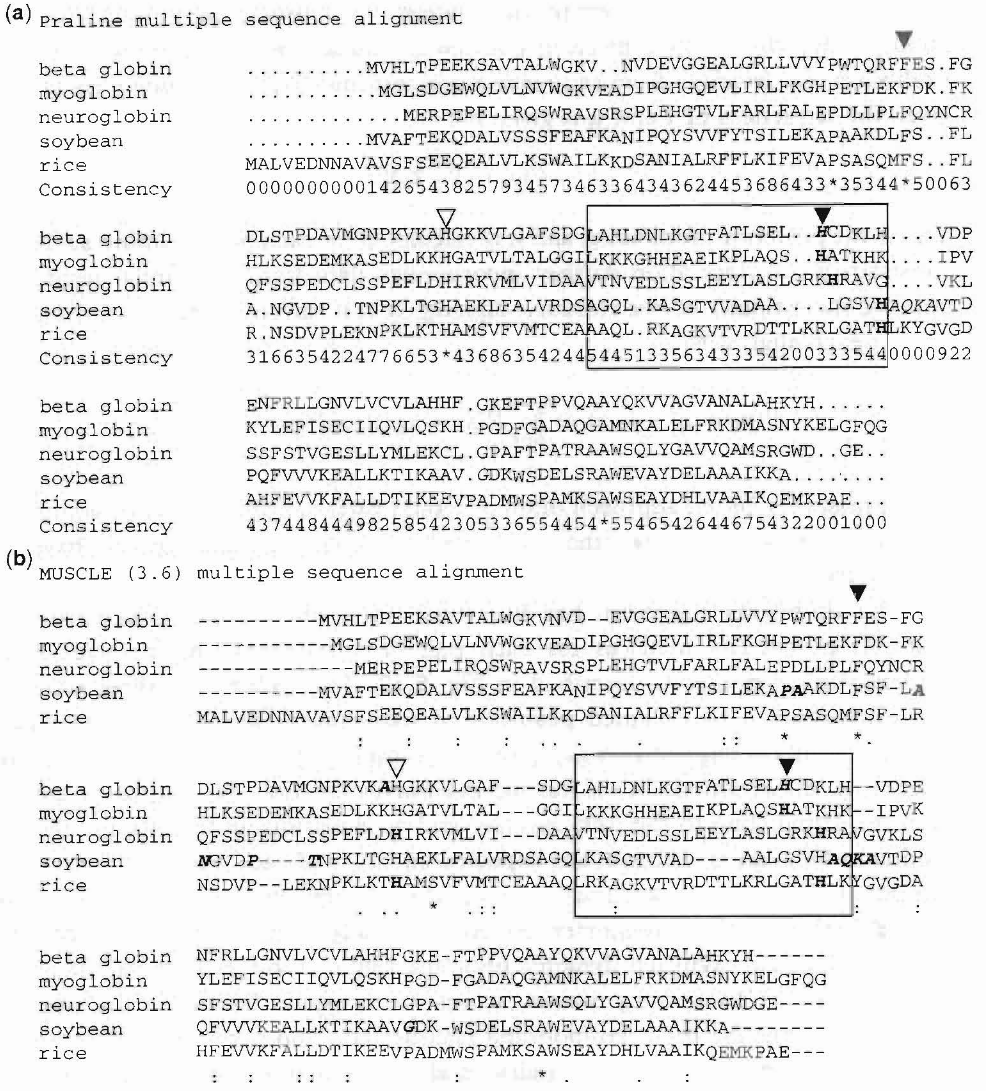
The consistency-based approach often generates final multiple sequence alignments that are more accurate than those achieved by progressive alignments, based on benchmarking studies.

The ProbCons algorithm has five steps. First, the algorithm calculates the posterior probability matrices for each pair of sequences. This involves a pair hidden Markov model as described in Fig. 5.12. This HMM has three states: M (corresponding to two aligned positions of sequences x and y), I_x (a residue in sequence x that is aligned to a gap), and I_y (a residue in y that is aligned to a gap). There is an initial probability of starting in a particular state, a transition probability from the initial state to the next residue, and an emission probability for the next residue to be aligned. Second, the expected accuracy of each pairwise alignment is computed. The expected accuracy is the number of correctly aligned pairs of residues divided by the length of the shorter sequence. The alignment is performed according to the Needleman–Wunsch dynamic programming method, but instead of using a PAM or BLOSUM scoring matrix, scores are assigned based on the posterior probability terms for the corresponding residues and gap penalties are set to zero. Third, the quality scores for each pairwise alignment are reestimated by applying a "probabilistic consistency transformation." This step applies information about conserved residues that were identified through all the pairwise alignments, resulting in the use of more accurate substitution scores. Fourth, an expected accuracy guide tree is constructed using hierarchical clustering (similar to the approach adopted by ClustalW). The guide tree is based on similarities (rather than distances). Fifth, the sequences are progressively aligned (as in ClustalW) by following the order specified by the guide tree. Further iterative refinements may be applied. Do et al. (2005) reported that ProbCons outperformed six other multiple sequence alignment programs, including ClustalW, DIALIGN, T-Coffee, MAFFT, MUSCLE, and Align-m, based on testing on the BAliBASE, PREFAB, and SABmark benchmark databases.

ProbCons is available at ► http://probcons.stanford.edu/.

T-Coffee is an acronym for tree based consistency objective function for alignment evaluation. T-Coffee first computes a library consisting of pairwise alignments. By default these include all possible pairwise global alignments of the input sequences (using the Needleman–Wunsch algorithm), and the ten highest-scoring local alignments. Every pair of aligned residues is assigned a weight. These weights are recalculated to generate an "extended library" that serves as a position-specific substitution matrix. The program then computes a multiple sequence alignment by progressive alignment, creating a distance matrix, calculating a neighbor-joining

T-Coffee was developed by Cédric Notredame, Desmond Higgins, Jaap Heringa, and colleagues. It is available at ► http://www.tcoffee.org. It is also mirrored at the European Bioinformatics Institute (► http://www.ebi.ac.uk/t-coffee/), the Swiss Institute of Bioinformatics, and the Centre National de la Recherche Scientifique (Paris).



(a)
```
Praline multiple sequence alignment

beta globin        ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin          ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin        .............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean            ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice               MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency        000000000014265438257934573463364343624453686433*35344*50063

beta globin        DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin          HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin        QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean            A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice               R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency        3166354224776653*4368635424454451335634333542003335440000922

beta globin        ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin          KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin        SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean            PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice               AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency        43744844498258542305336554454*55465426446754322001000
```

(b)
```
MUSCLE (3.6) multiple sequence alignment

beta globin        ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin          -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin        -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean            ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice               MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                                :   :   :   :  .. .    .      ::   *     *.

beta globin        DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin          HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin        QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean            NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice               NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                                . ..   *  .::         :                   :    :

beta globin        NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin          YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin        SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean            QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
rice               HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
                    :  :   ::  :        :      * .     .   :
```

FIGURE 6.6. Multiple sequence alignment of five distantly related globins using four different programs. The alignments were performed with (a) PRALINE, (b) MUSCLE, (c) ProbCons, and (d) T-Coffee. The proteins used to make the alignments and the symbols used to illustrate the figure are the same as those described in Fig. 6.3. Note that the programs differ in their abilities to align corresponding regions of alpha helical secondary structure (red lettering); in their alignment of a highly conserved histidine residue (black arrowhead); and in the number and placement of gaps (see boxed regions).

guide tree, and using dynamic programming and the substitution matrix derived from the extended library.

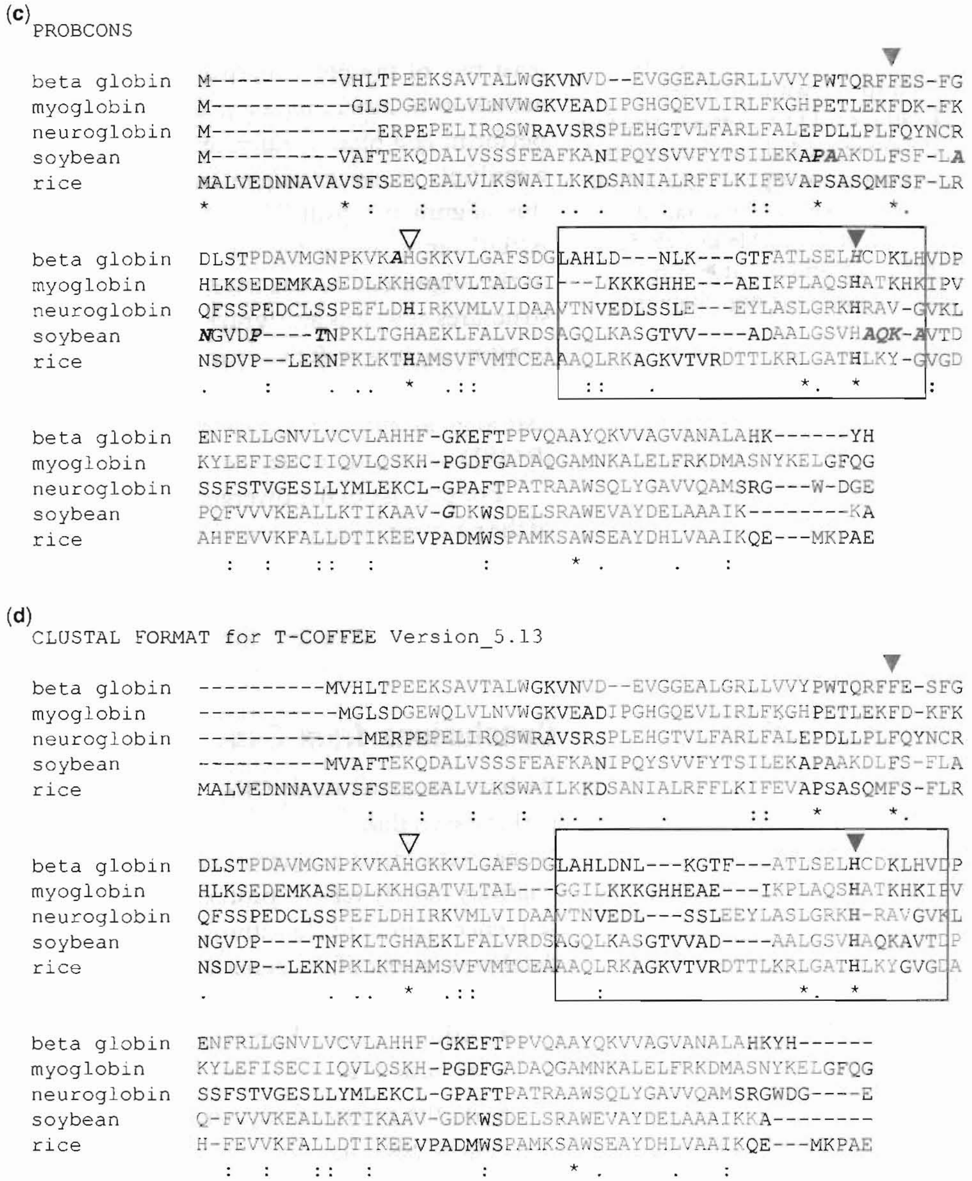
You can see an output of the five distantly related globins using M-Coffee in web document 6.5.

T-Coffee includes a suite of related alignment and evaluation tools. M-Coffee (Meta-Coffee) combines the output of as many as 15 different multiple sequence alignment methods (Wallace et al., 2006; Moretti et al., 2007). These include T-Coffee, ClustalW, MAFFT, MUSCLE, and ProbCons. M-Coffee employs a consistency-based approach to estimate a consensus alignment that is more accurate than any of the individual methods. By adding structural information (discussed next), even further accuracy is achieved.

Structure-Based Methods

PipeAlign is available at ► http://bips.u-strasbg.fr/PipeAlign/.

Tertiary structures evolve more slowly than primary sequences. Thus, for example, human beta globin and myoglobin share limited sequence identity (in the "twilight zone") yet share structures that are clearly related. It is possible to improve the accuracy of multiple sequence alignments by including information about the three-dimensional structure of one or more members of the group of proteins being



(c)

```
PROBCONS

beta globin   M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin     M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin   M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
              *          * :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
              .    :    . ..  *  .::        ::   .           *.  *     :

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean       PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice          AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

(d)

```
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                           :   :   :   :   ...    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin     HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean       NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice          NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
              .         . ..  *  .::         :               *.  *

beta globin   ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin     KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin   SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean       Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice          H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                :  :   ::  :        :      * .     .   :
```

FIGURE 6.6. *(Continued)*

aligned. Programs that enable you to incorporate structural information include PRALINE (Simossis and Heringa, 2005), the T-Coffee module Expresso (Armougom et al., 2006b), and PipeAlign (Plewniak et al., 2003).

When you use the Expresso program at the T-Coffee website, you submit a series of sequences (typically in the fasta format). Each sequence is automatically searched by BLAST against the Protein Data Bank (PDB) database, and matches (sharing >60% amino acid identity) are used to provide a template to guide the creation of the multiple sequence alignment.

We described BLAST in Chapter 4, and we will describe PDB in Chapter 11.

Structural information can also be used to assess the accuracy of a multiple sequence alignment after it has been made. This is done in benchmarking studies (described above) for protein families having known structures. In another approach you can incorporate structural information and assess the quality of a protein multiple sequence alignment that you make at the iRMSD-APDB ("Analyze alignments with Protein Data Bank") server of the T-Coffee package (O'sullivan et al., 2003;

The iRMSD-APDB server is part of the T-Coffee suite of tools (► http://www.tcoffee.org). Examples of five divergent and five closely related globin sequences, formatted for input to the APDB server, as well as the detailed output, are available in web documents 6.6 and 6.7 at ► http://www.bioinfbook.org/chapter6.

Armougom et al., 2006c). It is necessary to obtain the accession numbers corresponding to the Protein Data Bank (PDB) file having the known structures of at least two of the proteins you are aligning. As an example, we can obtain the PDB accession numbers for each of the five distantly related globins described above by performing a blastp search at NCBI, restricting the output to PDB. Next, perform a multiple sequence alignment using T-Coffee or any other program. Finally, input this alignment (using the PDB accession number in place of the name) to the APDB server at the T-Coffee website. The output provides an analysis of the quality of the alignment on the basis of all pairwise comparisons of those sequences having structures as well as an average quality assessment for each protein. The main approach to assessing how well two structures align is to measure the root mean square deviation (RMSD) (see Chapter 11). The RMSD is a measure of how closely the alpha carbons of two aligned amino residues are positioned. Notredame and colleagues introduced iRMSD as an intra molecular RMSD measure (Armougom et al., 2006a).

For the case of five divergent globins analyzed with the iRMSD-APDB server, 79% of the pairwise columns could be evaluated, 51% of the columns were aligned correctly (according to APDB), and the average iRMSD over all the evaluated columns was 1.07 Ångstroms. This analysis did not depend on a reference alignment, but instead involved a calculation of the superposition of the structures in the alignment.

Conclusions from Benchmarking Studies

We have discussed some of the programs for making multiple sequence alignments, and we have seen that they can produce differing results for a set of distantly related globins. Nonetheless most programs produce reasonably consistent alignments, especially for relatively closely related protein or DNA sequences. Comparative studies of multiple sequence alignment algorithms have been performed based on tests against benchmark databases. Some of the general conclusions include the following.

- Adding more homologs to a multiple sequence alignment improves its accuracy (Katoh et al., 2005).
- As the group of sequences being multiply aligned begins to share less amino acid identity, the accuracy of the alignments decreases (Briffeuil et al., 1998; Blackshields et al., 2006). For groups of sequences that share less than 25% identity, the problem becomes especially severe. Thompson et al. (1999) found that the best programs available at the time (PRRP, ClustalX, and SAGA) aligned about 60% to 70% of the amino acid residues for groups of proteins with <25% identity. For multiple sequence alignments of proteins sharing more identity (20% up to 40%), they found that on average 80% of the residues were aligned properly (Thompson et al., 1999).
- For highly divergent DNA sequences, programs that use local alignment (such as DiAlign and LAGAN) perform better than those using global alignment (such as ClustalW) (Kumar and Filipski, 2007).
- Orphan sequences are proteins that are highly divergent members of a family. If we examined a multiple sequence alignment of retinol-binding protein (RBP) from 10 species, then added the distantly related odorant-binding protein (OBP) to that multiple sequence alignment, OBP would be considered an orphan. Orphans might be expected to disrupt the organization of a multiple sequence alignment, and yet they do not. Global alignment

algorithms outperform local alignment methods for the introduction of orphans to an alignment (Thompson et al., 1999).

- Separate multiple sequence alignments can be combined, such as a group of closely related myoglobins and a group of closely related neuroglobins. Iterative algorithms performed this task better than progressive alignment methods (Thompson et al., 1999). However, many programs have difficulty in accurately producing a single alignment from a subset of alignments.
- Often, some proteins in a family contain large extensions at the amino- and/ or carboxy-terminals. Overall, local alignment programs dramatically outperformed global alignment programs at this task. For most multiple sequence alignment applications, global alignments are superior.

Databases of Multiple Sequence Alignments

We have discussed different methods for creating multiple sequence alignments. We will next examine databases of precomputed multiple sequence alignments, many of which are available. These may be searched using text (i.e., a keyword search) or using any query sequence. The query may be an already known sequence (such as myoglobin or RBP) or any novel protein (such as the raw sequence of a new lipocalin or globin you have identified). In some databases, the query sequence you provide is incorporated into the multiple sequence alignment of a particular precomputed protein family.

Pfam is maintained by a consortium of researchers, including Alex Bateman, Ewan Birney, Lorenzo Cerrutti, Richard Durbin, Sean Eddy, and Erik Sonnhammer, and others. Five sites host Pfam: ▶ http://www.sanger.ac.uk/Software/Pfam/ (U.K.), ▶ http://pfam.janelia.org/ (U.S.), ▶ http://pfam.cgb.ki.se/ (Sweden), ▶ http://pfam.jouy.inra.fr/ (France), and ▶ http://pfam.ccbb.re.kr/index.shtml (South Korea). Version 23.0 (July 2008) has 10,340 protein families. Pfam is based on sequences in Swiss-Prot and SP-TrEMBL (Chapter 2). Currently (May 2007), 74% of the proteins in those databases have at least one domain that matches to a Pfam family.

Pfam: Protein Family Database of Profile HMMs

Pfam is one of the most comprehensive databases of protein families (Bateman et al., 2004; Finn et al., 2006). It is a compilation of both multiple sequence alignments and profile HMMs of protein families. The database can be searched using text (keywords or protein names) or by entering sequence data. Its combination of HMM-based approach and expert curation makes Pfam one of the most trusted and widely used resources for protein families.

Pfam consists of two databases. Pfam-A is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. HMMER software (Chapter 5) is used to perform searches. For each family, Pfam provides four features: annotation, a seed alignment, a profile HMM, and a full alignment. The full alignment can be quite large; currently the top 20 Pfam families each contain over 20,000 sequences in their full alignment. The seed alignments contain a smaller number of representative family members. Sequences in Pfam-A are grouped in families, assigned stable accession numbers (such as PF00042 for globins) and expertly curated. Additional protein sequences are automatically aligned and deposited in Pfam-B where they are not annotated or assigned permanent accession numbers. Pfam-B serves as a useful supplement that makes the database more comprehensive. For all Pfam families, the underlying HMM is accessible from the main output page.

We can see the main features of Pfam in a search for globins using the Wellcome Trust Sanger Institute site. There are three main ways to access the database: by browsing for families, by entering a protein sequence search (with a protein accession number or sequence), and by entering a text search. From the front page, select a text-based search and enter "globin." The results summary includes links to the Pfam entry and to related databases (InterPro, described below; the Protein Data