The human sequence (red) differs from the chimpanzee sequence (blue) in only one amino acid in a protein chain of 153 residues for myoglobin.

http://www.ted.com/talks/jane_goodall_on_what_separates_us_from_the_apes.html
Like members of a human family, members of molecular families often have features in common.

Such family resemblance is most easily detected by comparing three-dimensional structure, the aspect of a molecule most closely linked to function.

Consider as an example ribonuclease from cows, which was introduced in our consideration of protein folding (Section 2.6).

Comparing structures reveals that the three-dimensional structure of this protein and that of a human ribonuclease are quite similar (Figure 6.1).
Although the degree of overlap between these two structures is not unexpected, given their nearly identical biological functions, similarities revealed by other such comparisons are sometimes surprising.

For example, angiogenin, a protein that stimulates the growth of new blood vessels, also turns out to be structurally similar to ribonuclease—so similar that it is clear that both angiogenin and ribonuclease are clearly members of the same protein family.

Angiogenin and ribonuclease must have had a common ancestor at some earlier stage of evolution.
Three-dimensional structures have been determined for only a small proportion of the total number of proteins.

In contrast, gene sequences and the corresponding amino acid sequences are available for a great number of proteins, largely owing to the tremendous power of DNA cloning and sequencing techniques including applications to complete genome sequencing.

Evolutionary relationships also are manifest in amino acid sequences.

For example, 35% of the amino acids in corresponding positions are identical in the sequences of bovine ribonuclease and angiogenin.

Is this level sufficiently high to ensure an evolutionary relationship?

In this chapter, we shall examine the methods that are used to compare amino acid sequences and to deduce such evolutionary relationships.

Sequence-comparison methods have become one of the most powerful tools in modern biochemistry.
By examining the footprints present in modern protein sequences, the biochemist can become a molecular archeologist able to learn about events in the evolutionary past.

Sequence comparisons can often reveal pathways of evolutionary descent and estimated dates of specific evolutionary landmarks.

Molecular evolution can also be studied experimentally.

In some cases, DNA from fossils can be amplified by PCR methods and sequenced, giving a direct view into the past.

In addition, investigators can observe molecular evolution taking place in the laboratory, through experiments based on nucleic acid replication. The results of such studies are revealing more about how evolution proceeds.
6.1 Homologs Are Descended from a Common Ancestor

The most fundamental relationship between two entities is **homology**; two molecules are said to be homologous if they have been derived from a common ancestor.

Homologous molecules, or homologs, can be divided into two classes.

**Paralogs** are homologs that are present within one species. Paralogs often differ in their detailed biochemical functions.

**Orthologs** are homologs that are present within different species and have very similar or identical functions.

*How can we tell whether two human proteins are paralogs or whether a yeast protein is the ortholog of a human protein?*
Both nucleic acid and protein sequences can be compared to detect homology.

Because nucleic acids are composed of fewer building blocks than proteins (4 bases versus 20 amino acids), the likelihood of random agreement between two DNA or RNA sequences is significantly greater than that for protein sequences.

For this reason, detection of homology between protein sequences is typically far more effective.
Let us consider a class of proteins called the **globins**. **Myoglobin** is a protein that binds oxygen in muscle, whereas **hemoglobin** is the oxygen-carrying protein in blood (Chapter 7).

**Both** proteins cradle a **heme** group, an iron-containing organic molecule that binds the oxygen.

Each human **hemoglobin** molecule is composed of four heme-containing polypeptide chains, two identical a chains and two identical β-chains.

Here, we consider only the **α-chain**.

To examine the similarity between the amino acid sequence of the human α-chain and that of human myoglobin (Figure 6.4),

we apply a method, referred to as a **sequence alignment**, in which the two sequences are systematically aligned with respect to each other to identify regions of significant overlap.
How can we tell where to align the two sequences?

In the course of evolution, the sequences of two proteins that have an ancestor in common will have diverged in a variety of ways.

Insertions and deletions may have occurred at the ends of the proteins or within the functional domains themselves.

Individual amino acids may have been mutated to other residues of varying degrees of similarity.

To understand how the methods of sequence alignment take these potential into account,
let us first consider the simplest approach, where we **slide one sequence past the other**, one amino acid at a time, and **count the number of matched residues**, or sequence identities.

For α-hemoglobin and myoglobin, the **best alignment** reveals **23 sequence identities**, spread throughout the central parts of the sequences.

However, we see that another alignment, featuring **22 identities**, is nearly as good.

This alignment is shifted by six residues and yields identities that are concentrated toward the amino-terminal end of the sequences.
By introducing a *gap* into one of the sequences, the identities found in both alignments will be represented.

**Hemoglobin α**
VLSPADKTNVKA AWGKVGAHAGEY GAEALERMF S FPTTKTYFPHF D

**Myoglobin**
GLSEGEWQLVLNWKGKVEADIPGHGQEVILIRLFKGHPETL EKFKD KFKHLKSED

*LSHSGSAQVKGHKK KVA DA LTNA VAHVDDMPNAL SALS DLH AHKL RVDPVNKKLL EMKASE DLKKHGATVLTALGG I LKKKGHHEAEI KPLAQSHATKH KIPVKYLE F

**Figure 6.6**
*Biochemistry, Seventh Edition*
© 2012 W. H. Freeman and Company

38 identities : 38 x (+10) = 380
1 gap : 1 x (-25) = 25
Score: 355

**Insertion of gaps** allows the alignment method to compensate for the *insertions* or *deletions* of nucleotides that may have *taken place* in the gene for one molecule but not the other in *the course of evolution*.

These methods use **scoring systems** to compare different alignments, including **penalties for gaps** to prevent the insertion of an unreasonable number of them.

Next, *we must determine the significance of this score and level of identity*.
The Statistical Significance of Alignments Can Be Estimated by Shuffling

The similarities in sequence in Figure 6.6 appear striking, yet there remains the possibility that a grouping of sequence identities has occurred by chance alone.

Hence, we can assess the significance of our alignment by “shuffling,” or randomly rearranging, one of the sequences, repeat the sequence alignment, and determine a new alignment score.

**Figure 6.7**
*Biochemistry, Seventh Edition*
© 2012 W. H. Freeman and Company
When this procedure is applied to the sequences of myoglobin and a-hemoglobin the authentic alignment clearly stands out (Figure 6.8).

Its score is far above the mean for the alignment scores based on shuffled sequences.

The probability that such a deviation occurred by chance alone are approximately $1 \text{ in } 10^{20}$.

Thus, we can comfortably conclude that the two sequences are genuinely similar; the simplest explanation for this similarity is that these sequences are homologous—that is, that the two molecules have descended by divergence from a common ancestor.
Distant Evolutionary Relationships Can Be Detected Through the Use of Substitution Matrices

No credit is given for any pairing that is not an identity.

**Not all substitutions are equivalent.**

Amino acid changes classified as **structurally conservative** or **nonconservative**.

A **conservative substitution** replaces one amino acid with another that is similar in size and chemical properties. Conservative substitutions may have only minor effects on protein structure and often can be tolerated without compromising protein function.

In contrast, in a **nonconservative substitution**, an amino acid is replaced by one that is structurally dissimilar.

Conservative and single-nucleotide substitutions are likely to be more common than substitutions with more radical effects.

*How can we account for the type of substitution when comparing sequences?*
### Blosum-62 substitution matrix

Blosum: *Blocks of amino acid substitution matrix*

![Blosum-62 substitution matrix diagram](image_url)

**Figure 6.9**

*Biochemistry, Seventh Edition*  
© 2012 W. H. Freeman and Company
Notice that scores are not the same for each residue, owing to the less frequently occurring amino acids such as **cysteine (C)** and **tryptophan (W)** will align by chance less often than the more common residues align.

Furthermore, structurally conservative substitutions such as **lysine (K)** for **arginine (R)** and **isoleucine (I)** for **valine (V)** have relatively high scores, whereas nonconservative substitutions such as lysine for tryptophan result in negative scores.

In addition, the introduction of a single residue gap lowers the alignment score by **12 points** and the extension of an existing gap costs **2 points per residue**.
With the use of this scoring system, the alignment shown in Figure 6.6 receives a score of 115.

In many regions, most substitutions are conservative (defined as those substitutions with scores greater than 0) and relatively few are strongly disfavored types (Figure 6.11).

This scoring system detects homology between less obviously related sequences with greater sensitivity than would a comparison of identities only.
Consider, for example, the protein **leghemoglobin**, an oxygen-binding protein found in the roots of some **plants**.

Scoring based on identities by chance alone are *1 in 20*. In contrast, users of the **substitution matrix** the odds of the alignment occurring by chance approximately *1 in 300*. 

![Figure 6.12](Image)
Thus, an analysis performed by using the substitution matrix reaches a much firmer conclusion about the evolutionary relationship between these proteins.

Experience with sequence analysis has led to sequence identities greater than 25% are probably homologous. less than 15% identical is unlikely to indicate statistically significant similarity. between 15 and 25% identical, further analysis is necessary to determine the statistical significance of the alignment.

It must be emphasized that the lack of a statistically significant degree of sequence similarity does not rule out homology.
Databases Can Be Searched to Identify Homologous Sequences

When the sequence of a protein is first determined, *comparing it with all previously characterized sequences* can be a source of tremendous insight into its evolutionary relatives and, hence, its structure and function. The sequence-alignment methods just described are used to compare an individual sequence with all members of a database of known sequences.

Database searches are most often accomplished by using resources available on the Internet at the National Center for Biotechnology Information (www.ncbi.nih.gov).

The procedure used is referred to as a **BLAST (Basic Local Alignment Search Tool)** search.

An amino acid sequence is typed or pasted into the Web browser, and a search is performed, most often against a non-redundant database of all known sequences. At the end of **2009**, this database included more than **10** million sequences.

A BLAST search yields a list of sequence alignments, each accompanied by an estimate giving the likelihood that the alignment occurred by chance (Figure 6.14).
BLASTP 2.2.10 [Oct-19-2004]

Query= gi:12517444|gb|AAG58041.1|AE005521_9 ribosephosphate isomerase, constitutive [Escherichia coli O157:H7] (219 letters)

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples 2,205,431 sequences; 747,548,157 total letters

Distribution of 268 Blast Hits on the Query Sequence

Sequence producing significant alignments:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>score</th>
<th>E value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi:42943</td>
<td>emb</td>
<td>CAA47309.1</td>
</tr>
<tr>
<td>gi:12517444</td>
<td>gb</td>
<td>AAG58041.1</td>
</tr>
<tr>
<td>gi:33126317</td>
<td>gb</td>
<td>AAK95569.1</td>
</tr>
<tr>
<td>gi:29897136</td>
<td>gb</td>
<td>AAP10413.1</td>
</tr>
</tbody>
</table>

Amino acid Sequence being queried:

Score = 117 bits (294), Expect = 2e-25
Identities = 82/224 (36%), Positives = 118/224 (52%), Gaps = 15/224 (6%)

Query: 4

Sequence of homologous protein from Homo sapiens:

Length = 237

Query: 60

Sequence of homologous protein from Homo sapiens:

Query: 120

Query: 170

Figure 6.14

Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

In 1995, investigators reported the first complete sequence of the genome of a free-living organism, the bacterium *Haemophilus influenzae*.

With the sequences available, they performed a **BLAST search** with each deduced protein sequence.

Of 1,743 identified protein-coding regions, also called **open reading frames**, 1,007 (58%) could be linked to some protein of known function that had been previously characterized in another organism.

An additional 347 open reading frames could be linked to sequences in the database for which no function had yet been assigned ("hypothetical proteins").

The remaining 389 sequences did not match any sequence present in the database at that time.

Thus, investigators were able to identify likely functions for more than half the proteins within this organism solely by sequence comparisons.

*Sequence-comparison methods have become one of the most powerful tools in modem biochemistry.*
6.3 Examination of Three-Dimensional Structure Enhances Our Understanding of Evolutionary Relationships

Sequence comparison is a powerful tool for extending our knowledge of protein function and kinship.

However, biomolecules generally function as intricate three-dimensional structures rather than as linear polymers.

**Mutations** occur at the level of sequence, but the effects of the mutations are at the level of function, and function is directly related to tertiary structure.

Consequently, to gain a deeper understanding of evolutionary relationships between proteins, we must examine three-dimensional structures, especially in conjunction with sequence information.

The techniques of structural determination are presented in Chapter 3.
Tertiary structures of the globins are extremely similar even though the similarity between human myoglobin and lupine leghemoglobin is just barely detectable at the sequence level and that between human α-hemoglobin and lupine leghemoglobin is not statistically significant (15.6% identity).

This structural similarity firmly establishes that the framework that binds the heme group and facilitates the reversible binding of oxygen has been conserved over a long evolutionary period.
In a growing number of other cases, however, a comparison of three-dimensional structures has revealed **striking similarities between proteins** that were *not expected to be related*.

A case in point is the protein **actin**, a major component of the cytoskeleton, and **heat shock protein 70** (Hsp-70), which assists protein folding inside cells.

These two proteins were found to be **noticeably similar in structure** despite only **15.6% sequence identity**.

On the basis of their three-dimensional structures, actin and Hsp-70 are **paralogs**.
Knowledge of Three-Dimensional Structures Can Aid in the Evaluation of Sequence Alignments

The sequence-comparison methods described thus far treat all positions within a sequence equally.

However, we know from examining families of homologous proteins for which at least one three-dimensional structure is known that regions and residues critical to protein function are more strongly conserved than are other residues.

For example, each type of globin contains a bound heme group with an iron atom at its center.

A histidine residue that interacts directly with this iron (residue 64 in human myoglobin) is conserved in all globins.

After we have identified key residues or highly conserved sequences within a family of proteins, we can sometimes identify other family members even when the overall level of sequence similarity is below statistical significance.

Thus it may be useful to generate a sequence template—a map of conserved residues that are structurally and functionally important and are characteristic of particular families of proteins.
Repeated Motifs Can Be Detected by Aligning Sequences with Themselves

More than 10% of all proteins contain sets of two or more domains that are similar to one another.

Sequence search methods can **often detect internally repeated sequences** that have been characterized in other proteins.

Their presence can be detected by attempting to **align a given sequence with itself**.

For the **TATA-box-binding protein**, a key protein in controlling gene transcription such an alignment is highly significant: **30% of the amino acids are identical**.
The determination of the **three-dimensional structure** of the TATA-box-binding protein **confirmed the presence of repeated structures**; the protein is formed of two nearly identical domains.

*The evidence is convincing that the gene encoding this protein evolved by duplication of a gene encoding a single domain.*
Thus far, we have been exploring proteins derived from common ancestors – that is, through *divergent evolution*.

Other cases have been found of proteins that are structurally similar in important ways but are not descended from a common ancestor.

*How might two unrelated proteins come to resemble each other structurally?*

Two proteins evolving independently may have converged on a similar structure to perform a similar biochemical activity.

Perhaps *that structure was an especially effective solution to a biochemical problem* that organisms face.

The process by which very different evolutionary pathways lead to the same solution is called *convergent evolution*.
One example of convergent evolution is found among the **serine proteases**. These enzymes, to be considered in more detail in Chapter 9, **cleave peptide bonds by hydrolysis**.

Figure 6.18 shows the **structure of the active sites**—that is, the sites on the proteins at which the hydrolysis reaction takes place—for two such enzymes, chymotrypsin and subtilisin.

**These active-site structures are remarkably similar.** In each case, a **serine** residue, a **histidine** residue, and an **aspartic acid** residue are positioned in space in nearly identical arrangements.

![Figure 6.18](image)

However, the key serine, histidine, and aspartic acid residues **do not** occupy similar positions or even **appear in the same order** within the two sequences.
As we will see, this is the case because chymotrypsin and subtilisin use the same mechanistic solution to the problem of peptide hydrolysis. At first glance, this similarity might suggest that these proteins are homologous. However, striking differences in the overall structures of these proteins make an evolutionary relationship extremely unlikely (Figure 6.19).

Whereas chymotrypsin consists almost entirely of β-sheets, subtilisin contains extensive α-helical structure.

It is extremely unlikely that two proteins evolving from a common ancestor could have retained similar active-site structures while other aspects of the structure changed so dramatically.
Comparison of RNA Sequences Can Be a Source of Insight into RNA Secondary Structures

Homologous RNA sequences can be compared in a manner similar to that already described. Such comparisons can be a source of important insights into evolutionary relationships; in addition, they provide clues to the three-dimensional structure of the RNA itself.

In a family of sequences that form similar base-paired structures, base sequences may vary, but base-pairing ability is conserved.

Consider, for example, a region from a large RNA molecule present in the ribosomes of all organisms

\[\begin{align*}
\text{ARCHAEA} & \quad \text{Halobacterium halobium} & C & C & G & G & U & G & U & G & C & G & G & G & G & U & A & A & G & C & C & U & G & U & C & A & C & C & G & U \\
\text{EUKARYA} & \quad \text{Homo sapiens} & G & G & G & C & C & A & U & U & U & U & G & G & G & G & G & G & C & A & A & A & C & U & G & G & C & C & C \\
\end{align*}\]

reveals that the bases in positions 9 and 22 retain the ability to form a Watson-Crick base pair even though the identities of the bases in these positions vary.
The observation that homology is often manifested as sequence similarity suggests an evolutionary tree.
Such comparisons reveal only the relative divergence times for example, that myoglobin diverged from hemoglobin twice as long ago as the $\alpha$-chain diverged from the $\beta$-chain.

**How can we estimate the approximate dates** of gene duplications and other evolutionary events?

Evolutionary trees can be **calibrated by** comparing the deduced branch points with divergence times determined from the **fossil record**.

For example, the **duplication** leading to the two chains of hemoglobin appears to have **occurred 350 million years ago**.

This estimate is supported by the observation that **jawless fish** such as the lamprey, which diverged from bony fish approximately **400 million years ago**, contain **hemoglobin built from a single type of subunit** (Figure 6.22).
Two techniques of biochemistry have made it possible to examine the course of evolution more directly and not simply by inference.

The polymerase chain reaction allows the direct examination of ancient DNA sequences, releasing us, at least in some cases, from the constraints of being able to examine existing genomes from living organisms only.

Molecular evolution may be investigated through the use of combinatorial chemistry, the process of producing large populations of molecules en masse and selecting for a biochemical property.

This exciting process provides a glimpse into the types of molecules that may have existed very early in evolution.
Ancient DNA Can Sometimes Be Amplified and Sequenced

The tremendous chemical stability of DNA makes the molecule well suited to its role as the storage site of genetic information. So stable is the molecule that samples of DNA have survived for many thousands of years under appropriate conditions.

With the development of PCR methods, such ancient DNA can sometimes be amplified and sequenced. This approach has been applied to mitochondrial DNA from a Neanderthal fossil estimated at 38,000 years of age excavated from Vindija Cave, Croatia, in 1980.

Remarkably, investigators have completely sequenced the mitochondrial genome from this specimen.

Comparison of the Neanderthal mitochondrial sequence with those from Homo sapiens individuals revealed between 201 and 234 substitutions, considerably fewer than the approximately 1,500 differences between human beings and chimpanzees over the same regions.
Further analysis suggested that the **common ancestor of modern human beings and Neanderthals** lived approximately **660,000 years** ago.

An evolutionary tree constructed by using these data and others revealed that the Neanderthal was not an intermediate between chimpanzees and human beings but,

instead, was an **evolutionary "dead end"** that became extinct (Figure 6.23).

A few earlier studies claimed to determine the sequences of **far more ancient DNA such as that found in insects trapped in amber**, but these studies appear to have been **flawed**. The source of these sequences turned out to be contaminating **modern DNA**.

Successful sequencing of ancient DNA requires sufficient DNA for reliable amplification and the **rigorous exclusion of all sources of contamination**.
A Draft Sequence of the Neandertal Genome


Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.
**Molecular Evolution Can Be Examined Experimentally**

**Evolution** requires three processes:

1. the generation of a **diverse population**
2. the **selection** of members based on some criterion of **fitness**
3. **reproduction** to enrich the population in these more-fit members.

Nucleic acid molecules are capable of undergoing all three processes *in vitro* under appropriate conditions.

The results of such studies enable us to glimpse

*how evolutionary processes might have generated*

catalytic activities and specific binding abilities important biochemical functions in all living systems.
A diverse population of nucleic acid molecules can be synthesized in the laboratory by the process of combinatorial chemistry.

When an initial population has been generated, it is subjected to a selection process that isolates specific molecules with desired binding or reactivity properties.

Finally, molecules that have survived the selection process are replicated through the use of PCR;

Errors that occur naturally in the course of the replication process introduce additional variation into the population in each “generation.”
The new population was subjected to additional rounds of selection for ATP-binding activity.

After eight generations, members of the selected population were characterized by sequencing.

Seventeen different sequences were obtained, 16 of which could form the structure shown in Figure 6.25.

Each of these molecules bound ATP with high affinity, as indicated by dissociation constants less than 50 µM.
As expected, this 40-nucleotide molecule is composed of two Watson-Crick base-paired helical regions separated by an 11-nucleotide loop. This loop folds back on itself in an intricate way to form a deep pocket into which the adenine ring can fit. Thus, a structure had evolved that was capable of a specific interaction.