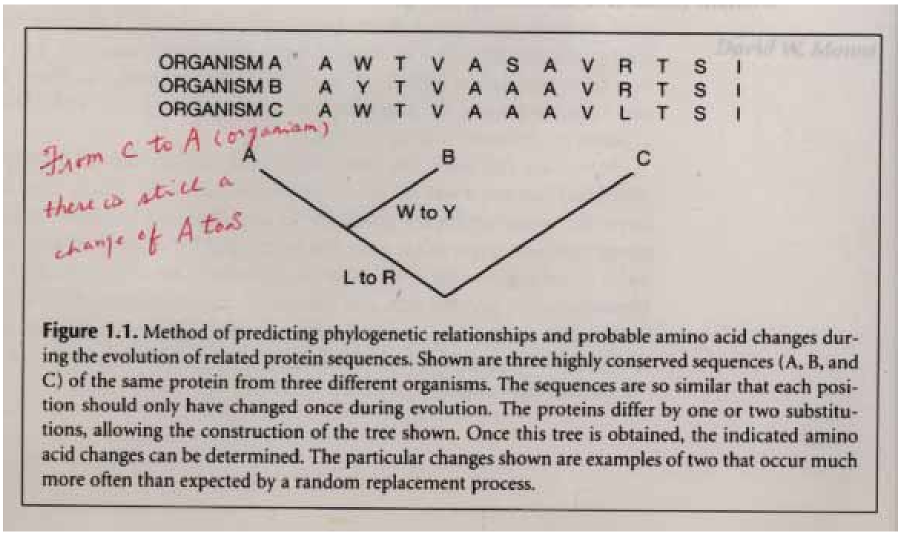
Sequence Analysis

I. Historical Introduction and Overview

1. THE FIRST SEQUENCES TO BE COLLECTED WERE THOSE OF PROTEINS
   a. The development of protein-sequencing methods (Sanger and Tuppy 1951) led to the sequencing of representatives of several of the more common protein families such as cytochromes from a variety of organisms.
   b. Margaret Dayhoff (1972, 1978) and her collaborators at the National Biomedical Research Foundation (NBRF), Washington, DC, were the first to assemble databases of these sequences into a protein sequence atlas in the 1960s, and their collection center eventually became known as the Protein Information Resource (PIR). The NBRF maintain the database from 1984, and in 1988, the PIR-International Protein Sequence Database was established as a collaboration of NBRF, the Munich Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID).
   c. Dayhoff and her coworkers organized the proteins into families and superfamilies based on the degree of sequence similarity.
   d. Tables that reflected the frequency of changes observed in the sequences of a group of closely related proteins were derived. Proteins that were less than 15% different were chosen to avoid the chance that the observed amino acid changes reflected two sequential amino acid changes instead of only one.
   e. From aligned sequences, a phylogenetic tree was derived showing graphically which sequences were most related and therefore shared a common branch on the tree. Once these trees were made. They were used to score the amino acid changes that occurred during evolution of the genes for these proteins in the various organisms from which they originated. See Figure 1.



ORGANISM A	A	W	T	V	A	S	A	V	R	T	S	I
ORGANISM B	A	Y	T	V	A	A	A	V	R	T	S	I
ORGANISM C	A	W	T	V	A	A	A	V	L	T	S	I

Figure 1.1. Method of predicting phylogenetic relationships and probable amino acid changes during the evolution of related protein sequences. Shown are three highly conserved sequences (A, B, and C) of the same protein from three different organisms. The sequences are so similar that each position should only have changed once during evolution. The proteins differ by one or two substitutions, allowing the construction of the tree shown. Once this tree is obtained, the indicated amino acid changes can be determined. The particular changes shown are examples of two that occur much more often than expected by a random replacement process.

Figure 1

f. Subsequently, a set of matrices (tables)—the percent amino acid mutations accepted by evolutionary selection or PAM tables—which showed the probability that one amino acid changed into any other in these trees was constructed, thus showing which amino acids are most conserved at the corresponding position in two sequences.

g. The rule used is that the more identical and conserved amino acids that there are in two sequences, the more likely they are to have been derived from a common ancestor gene during evolution.

h. If the sequences are very much alike, the proteins probably have the same biochemical function and three dimensional structure folds.

i. Thus Dayhoff and her colleagues contributed in several ways to modern biological sequence analysis by providing the first protein sequence database as well as PAM tables for performing protein sequence comparisons.

j. Amino acid substitution tables are routinely used in performing sequence alignments and data base similarity searches.

2. DNA SEQUENCE DATABASE

a. DNA sequence databases were first assembled at Los Alamos National Laboratory (LANL), New Mexico, by Walter Goad and colleagues in the GenBank database and at the European

Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.

b. GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov).

c. The EMBL Data Library was founded in 1980 (http://www.ebi.ac.uk). In 1984 the DNA DataBank of Japan (DDBJ), Mishima, Japan, came into existence (http://www.ddbj.nig.ac.jp).

d. GenBank, EMBL, and DDBJ have now formed the International Nucleotide Sequence Database Collaboration (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis.

e. A sequence entry includes a computer file name, DNA or protein sequence files, functions, mutations, encoded proteins, regulatory sites, and references. This information was organized into a database format that could be readily searched for many types of information.

3. SEQUENCE RETRIEVAL FROM PUBLIC DATABASE

a. An important step in providing sequence database access was the development of Web pages that allow queries to be made of the major sequence databases (GenBank, EMBL, etc.).

b. An early example of this technology was a menu-driven program called GENINFO developed by D. Benson, D. Lipman, and colleagues. Subsequently, a derivative program called ENTREZ (http://www.ncbi.nlm.nih.gov/Entrez) with a simple window-based interface, and eventually a Web-based interface, was developed at NCBI. The idea behind these programs was to provide an easy-to-use interface with a flexible search procedure to the sequence databases.

c. ENTREZ searches for similar or related terms, or complex searches composed of several choices, with great ease and lists the found items in the order of likelihood that they matched the original query.

d. ENTREZ allows straightforward access to a number of databases such as DNA sequence database, protein sequence database, Medline (the full bibliographic database of the National Library of Medicine), a phylogenetic database of organisms, and a protein structure database, etc. This access is provided without cost to any user—private, government, industry, or research—a decision by the staff of NCBI that has provided a stimulus to biomedical research that cannot be underestimated.

4. SEQUENCE ANALYSIS PROGRAMS

As more DNA sequences became available in the 1970s, interests also increased in developing computer programs to analyze these sequences in various ways. Programs were then developed for large mainframe computers down to the then-new microcomputers. These computer programs were shared on a no-cost or low-cost basis. Some of them became commercialized and are still widely used. Websites are available to perform many types of sequence analyses; they are free to academic institutions or are available at moderate cost to commercial users.

Following is a brief review of the development of methods for sequence analysis.

5. THE DOT MATRIX OR DIAGRAM METHOD FOR COMPARING SEQUENCES

In 1970, A.J. Gibbs and G.A. McIntyre described a new method for comparing two amino acid and nucleotide sequences in which a graph was drawn with one sequence written across the page and the other down the left-hand side. Whenever the same letter appeared in both sequences, a dot was placed at the intersection of the corresponding sequence positions on the graph (Figure 2). The resulting graph was then scanned for a series of dots that formed a diagonal, which revealed similarity, or a string of the characters, between the sequences. Long sequences can also be compared in this manner on a single page by using smaller dots.

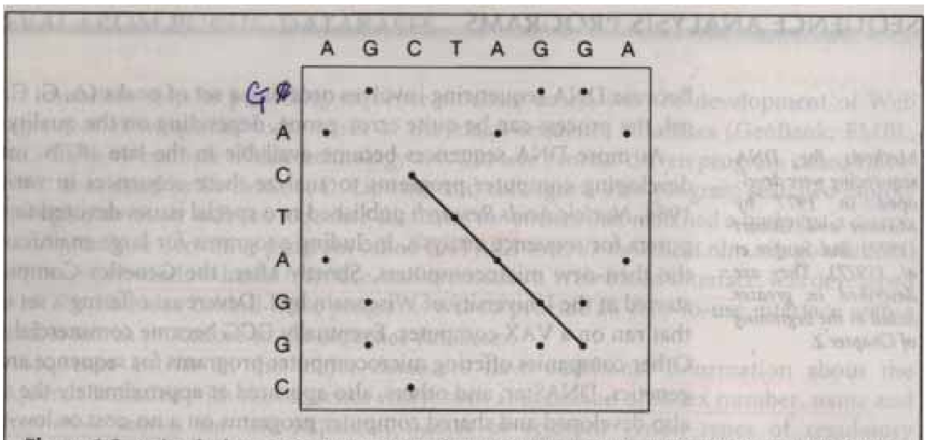


Figure 1.2. A simple dot matrix comparison of two DNA sequences, AGCTAGGA and GACTAGGC. The diagonal of dots reveals a run of similar sequence CTAGG in the two sequences.

Figure 2

a The dot matrix method quite readily reveals the presence of insertions or deletions between sequences because they shift the diagonal horizontally or vertically by the amount of change.

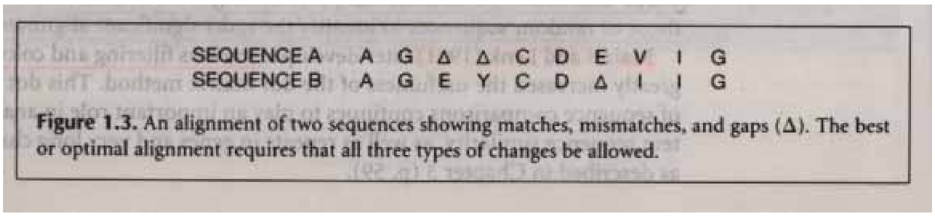
b Comparing a single sequence to itself can reveal the presence of a repeat of the same sequence in the same (direct repeat) or reverse (inverted repeat or palindrome) orientation. This self-comparison can reveal several features, such as similarity between chromosomes, tandem genes, repeated domains in a protein sequence, regions of low sequence complexity where the same characters are often repeated, or self-complementary sequences in RNA that can potentially base-pair to give a double-stranded structure.

c This dot matrix representation of sequence comparisons continues to play an important role in analysis of DNA and protein sequence similarity, as well as repeats in genes and very long chromosomal sequences.

6. ALIGNMENT OF SEQUENCES BY DYNAMIC PROGRAMMING

a The dot matrix method can be used to detect sequence similarity but fails to resolve similarity that is interrupted by regions that do not match very well or that are present in only one of the sequences (e.g., insertions or deletions). Therefore we need to devise a method that is able to provide the very best possible alignment, called an optimal alignment, between the two sequences.

b Such an alignment can be represented by writing the sequences on successive lines across the page, with matching characters placed in the same column and unmatched characters placed in the same column as a mismatch or next to a gap as an insertion (or deletion in the other sequence), as shown in Figure 3.



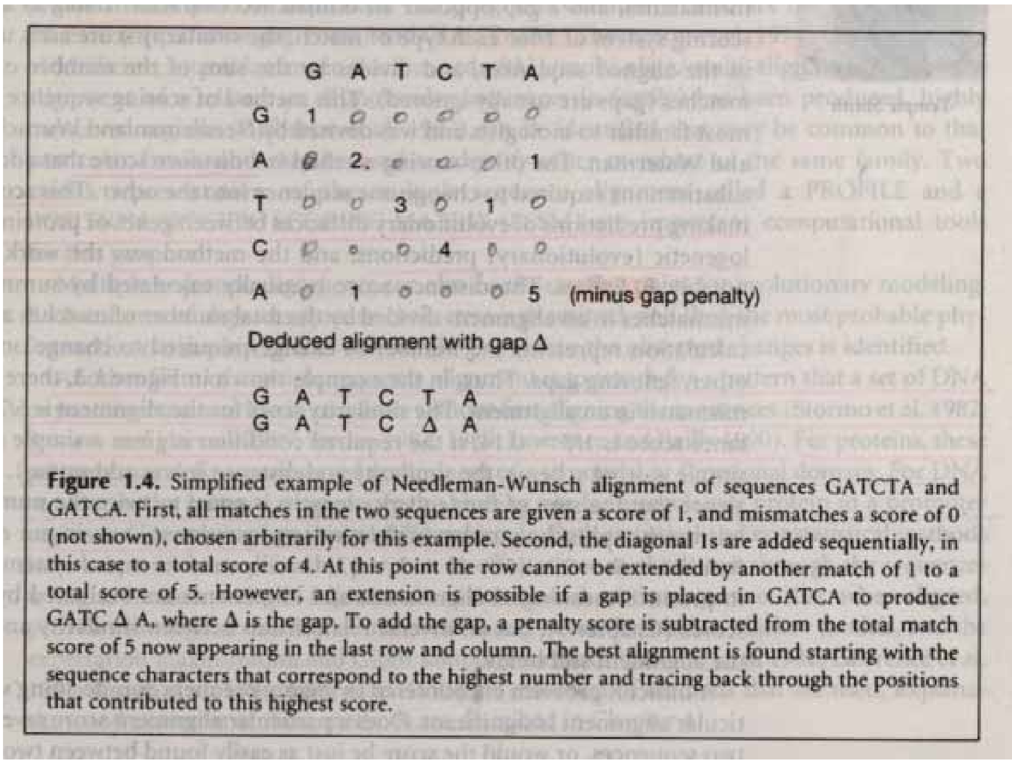
SEQUENCE A	A	G	Δ	Δ	C	D	E	V	I	G
SEQUENCE B	A	G	E	Y	C	D	Δ	I	I	G

Figure 1.3. An alignment of two sequences showing matches, mismatches, and gaps (Δ). The best or optimal alignment requires that all three types of changes be allowed.

Figure 3

To find an optimal alignment in this way is computationally so difficult that for proteins of length 300, 10^{88} comparisons will have to be made (Waterman 1989).

c To simplify the task, Needleman and Wunsch (1970) broke the problem down into a progressive building of an alignment by comparing two amino acid pair at a time. During the process of comparison they allowed various combinations of matched pairs, mismatched pairs, or extra amino acids in one sequence (insertion or deletion). This is the so called "dynamic programming". This approach generates (1) every possible alignment, each one including every possible combination of match, mismatch, and single insertion or deletion, and (2) a scoring system to score the alignment. Every match in a trial alignment was given a score 1, every mismatch a score 0, and individual gaps a penalty score. These numbers were then added across the alignment to obtain a total score for the alignment. The alignment with the highest possible score was defined as the optimal alignment. A simplified example of the Needleman-Wunsch alignment of sequences GATATA and GATCA is shown in Figure 4.



	G	A	T	C	T	A	
G	1	0	0	0	0	0	
A	0	2	0	0	0	1	
T	0	0	3	0	1	0	
C	0	0	0	4	0	0	
A	0	1	0	0	0	5	(minus gap penalty)

Deduced alignment with gap Δ

```
G A T C T A
G A T C Δ A
```

Figure 1.4. Simplified example of Needleman-Wunsch alignment of sequences GATCTA and GATCA. First, all matches in the two sequences are given a score of 1, and mismatches a score of 0 (not shown), chosen arbitrarily for this example. Second, the diagonal 1s are added sequentially, in this case to a total score of 4. At this point the row cannot be extended by another match of 1 to a total score of 5. However, an extension is possible if a gap is placed in GATCA to produce GATC Δ A, where Δ is the gap. To add the gap, a penalty score is subtracted from the total match score of 5 now appearing in the last row and column. The best alignment is found starting with the sequence characters that correspond to the highest number and tracing back through the positions that contributed to this highest score.

Figure 4

d The alignment described above is the global alignment.

7. FINDING LOCAL ALIGNMENTS BETWEEN SEQUENCES

a Smith and Waterman (1981) recognized that the most biologically significant regions in DNA and protein sequences were subregions that align well and that the remaining regions made up of less-related sequences were less significant. Therefore, they developed an important modification of the Needleman-Wunsch algorithm, called the local alignment or Smith-Waterman algorithm, to locate such regions. They also recognized that insertions or deletions of any size are likely to be found as evolutionary changes in sequences, and therefore adjusted their method to accommodate such changes. Finally they provided mathematical proof that the dynamic programming method is guaranteed to provide an optimal alignment between sequences.

b Two complementary measurements had been devised for scoring an alignment of two sequences, a similarity score and a distance score. The method of scoring sequence similarity is the one most familiar to biologists and was devised by Needleman and Wunsch and used by Smith and Waterman. The distance score is most useful for making predictions of evolutionary distances between genes or proteins to be used for phylogenetic (evolutionary) predictions, and the method was the work of mathematicians, notably P. Sellers.

8. MULTOPLE SEQUENCE ALIGNMENT

a In addition to aligning a pair of sequences, methods have been developed for aligning three or more sequences at the same time. These methods are computer-intensive and usually are based on a sequential aligning of the most-alike pairs of sequences.

b Once the alignment of a related set of molecular sequences (a family) has been produced, highly conserved regions can be identified that may be common to that particular family and may be used to identify other members of the same family. These alignments can be the starting point for evolutionary modeling.

9. PREDICTION OF RNA SECONDARY STRUCTURE

a RNA secondary structure predictions on computers were also developed at an early time.

b If the complement of a sequence on an RNA molecule is repeated down the sequence in the opposite chemical direction, the regions may base-pair and form a hairpin structure, as illustrated in Figure 5.

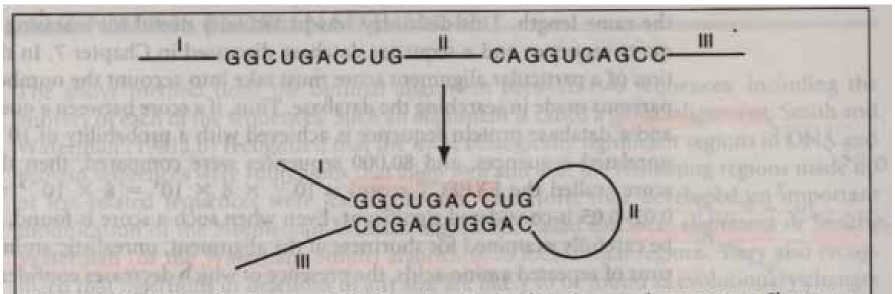


Figure 1.5. Folding of single-stranded RNA molecule into a hairpin secondary structure. Shown are portions of the sequence that are complementary: They can base-pair to form a double-stranded region. G/C base pairs are the most energetic due to 3 H bonds; A/U and G/U are next most energetic with two and one H bonds, respectively.

Figure 5

c Alignment, structural modeling, and phylogenetic analysis based on RNA sequences have made possible the discovery of evolutionary relationships among organisms.

10. IMPORTANCE OF DATABASE SEARCHES FOR SIMILAR SEQUENCES

a Genes with an important biological function could be sequenced with the hope of learning something about the biochemical nature of the gene product.

b An example was the retrovirus-encoded v-*sis* and v-*src* oncogenes, genes that cause cancer in animals. By comparing the predicted sequences of the viral products with all of the known protein sequences at a time, R. Doolittle and colleagues (1983) and W. Barger and M. Dayhoff (1982) both made the startling discovery that these genes appeared to be derived from cellular genes. The Sis protein had a sequence very similar to that of the platelet-derived growth factor (PDGF) from mammalian cells, and the Src protein to the catalytic chain of mammalian cAMP-dependent kinases. Thus, it appeared likely that the retrovirus had acquired the gene from the host cell as some kind of genetic exchange event and then had produced a mutant form of the protein that could compromise the function of the normal protein when the virus infected another animal.

c As molecular biologists analyzed more and more gene sequences, they discovered that many organisms share similar genes that can be identified by their sequence similarity.

d These searches have been greatly facilitated by having genetic and biochemical information from model organisms, such as the bacterium *Escherichia coli* and the budding yeast *Saccharomyces cerevisiae*. In these organisms extensive genetic analysis has revealed the function of genes, and the sequences of these genes have also been determined.

e Finding a gene in a new organism (e.g., a crop plant) with a sequence similar to a model organism gene (e.g., yeast) provides a prediction that the new gene has the same function as in the model organism.

f Such searches are becoming quite commonplace and are greatly facilitated by programs such as FASTA (Pearson and Lipman 1988) and BLAST (Altschul et al. 1990).

11. THE FASTA AND BLAST METHODS FOR DATABASE SEARCHES

a W. Pearson and D. Lipman (1988) developed a program called FASTA, which performed a database scan for similarity in a short enough time to make such scans routinely possible. FASTA provides a rapid way to find short stretches of similar sequence between a new sequence and any sequence in a database. The FASTA program has been continually improved by Pearson in 1990 and 1996.

b The BLAST program is even faster than the FASTA. This program was developed by S. Altschul et al. in 1990. This method is widely used from the Web site of the National Center for Biotechnology Information (NCBI) at the National Library of Medicine in Washington, DC. The BLAST server is probably the most widely used sequence analysis facility in the world and provides similarity searching to all currently available sequences.

12. PREDICTING THE SEQUENCE OF A PROTEIN BY TRANSLATION OF DNA SEQUENCES

a Protein sequences are predicted by translating DNA sequences that are cDNA copies of mRNA sequences from a predicted start and end of an open reading frame.

b For organisms that have few or no introns in their genomic DNA (such as bacteria genomes), the genomic DNA may be translated. For most eukaryotic organisms with introns in their genes, the protein-encoding exons must be predicted before translation.

13. PREDICTING PROTEIN SECONDARY STRUCTURE

a There are large number of proteins whose sequences are known, but very few whose structures have been solved. Solving protein structures involves the time-consuming and highly specialized procedures of X-ray crystallography and nuclear magnetic resonance (NMR).

b Consequently, there is much interest in trying to predict the structure of a protein, given its sequence.

c Proteins are synthesized as linear chains of amino acids; they then form secondary structures along the chain, such as ɑ helices, as a result of interactions between side chains of nearby amino acids. The region of the molecule with these secondary structures then folds back and forth on itself to form tertiary structures that include ɑ helices, β sheets comprising interacting β strands, and loops (see Figure 6).

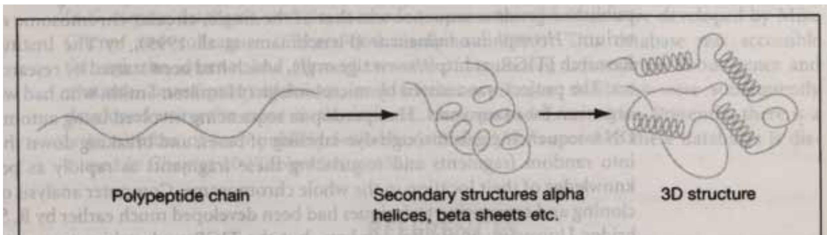


Figure 1.7. Folding of a protein from a linear chain of amino acids to a three-dimensional structure. The folding pathway involves amino acid interactions. Many different amino acid patterns are found in the same types of folds, thus making structure prediction from amino acid sequence a difficult undertaking.

Figure 6

d This folding often leaves amino acids with hydrophobic side chains facing into the interior of the folded molecule and polar amino acids that can interact with water and the molecular environment facing outside in loops.

e The amino acids sequence of the protein directs the folding pathway, sometimes assisted by proteins called chaperonins.

f Computational methods were used to find proteins that had a similar structural fold (the same arrangement of secondary structures connected by similar loops). These methods led to the discovery that as new protein structures were being solved, they often had a structural fold that was already known in a group of sequences.

g Proteins are found to have a limited number of ~500 folds (Chothia 1992), perhaps due to chemical restrains on protein folding or to the existence of a single evolutionary pathway for protein structure (Gibrat et al. 1996).

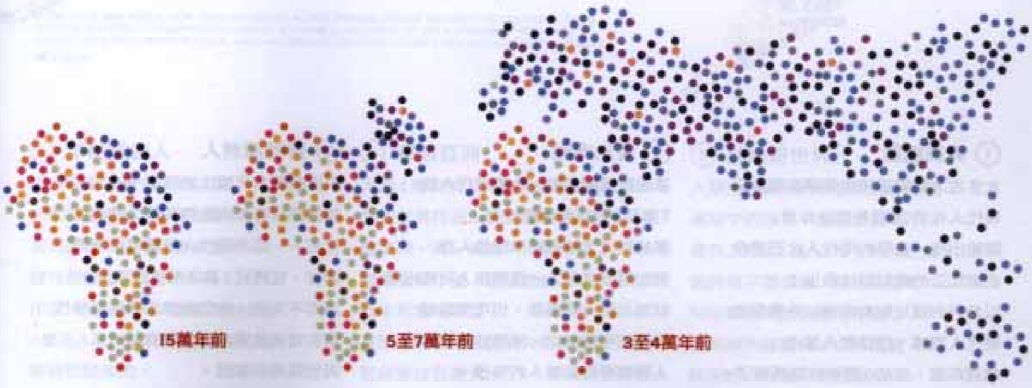
h Proteins without any sequence similarity could adopt the same fold. This fact complicates the prediction of structure from sequence. Methods for finding whether or not a given protein sequence can occupy the same three-dimensional conformation as another based on the properties of the amino acids have been devised (Bowie et al. 1991).

i Amos Bairoch (Bairoch et al. 1997) developed another method for predicting the biochemical activity of an unknown protein, given its sequence. He collected sequences of proteins that a common biochemical activity, for example an ATP-binding site, and deduced the pattern of amino acids that was responsible for that activity, allowing for some variability. These patterns were collected into the PROSITE database.

j Sophisticated statistical and machine-training techniques have been used in more recent protein structure prediction programs, and the success rate has increased. A recent advance in this active field of research is to organize proteins into groups or families on the bases of sequence similarity, and to find consensus patterns of amino acid domains characteristic of these families using some statistical methods.

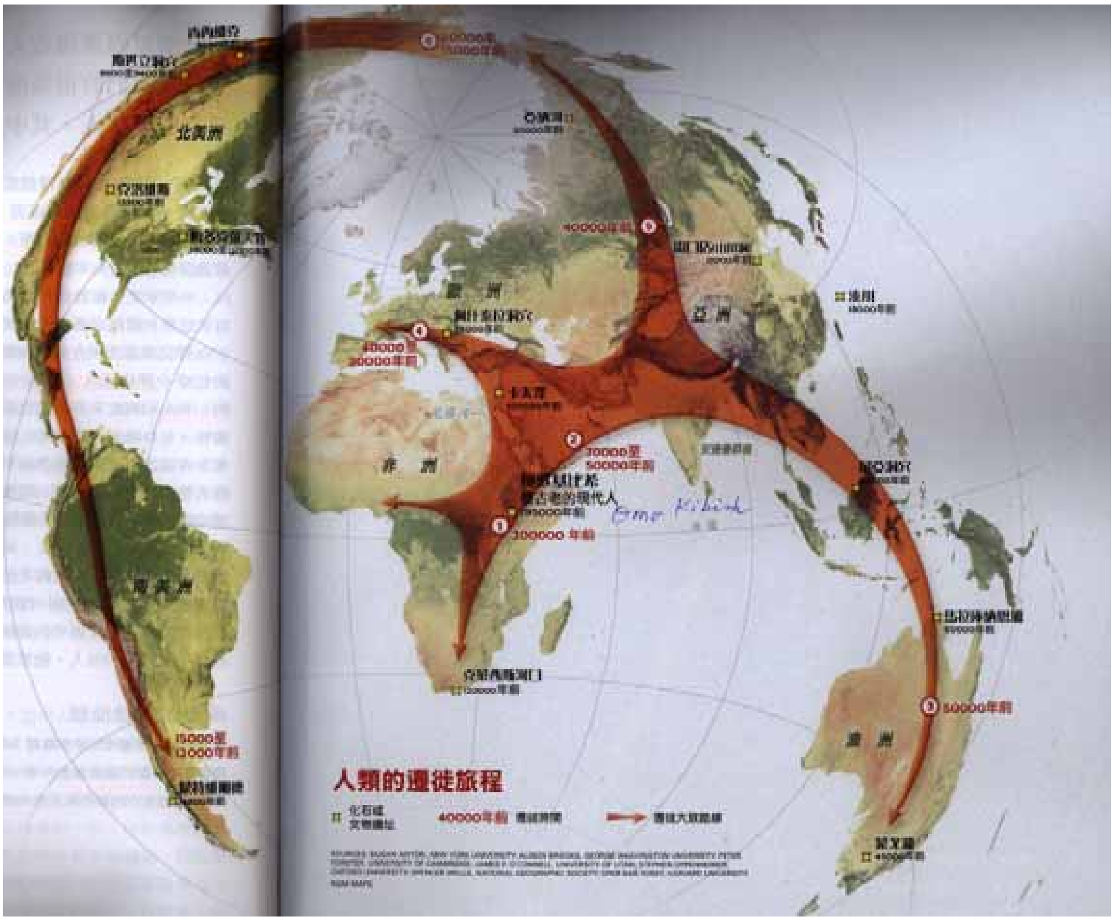
基因標誌的多樣性在非洲最高（地圖的彩色圓點），顯示這裡是現代人最早的家園。只有少數的人、帶著其中少數的標誌走出非洲（中），在接下來的幾萬年間散播到各地去（右）。

「世界其他地方的基因組合是非洲基因的子集。」耶魯大學的遺傳學家肯尼斯．基德說。



© KENNETH K. KIDD

人類的旅程 33


北美洲
南美洲
15000至
13000年前
歐洲
40000至
30000年前
卡夫澤
亞洲
非洲
70000至
50000年前
古老的現代人
200000 年前
Omo Kibish
40000年前
50000年前
澳洲
克萊西斯河口
人類的遷徙旅程
化石或
文物遺址
40000年前 遷徙時間
遷徙大致路線