1. *Initialization of the matrix.* Setting the values of the top row and left column. In the Smith-Waterman method the top row and left column are set to 0. As a result, either sequence can slide along the other before alignment starts, without incurring any gap penalty against the residues it leaves behind.
2. *Filling in the matrix.* In the example of global alignment, at each step, a choice is forced among match, insertion or deletion, even if none of these choices is attractive and even if a succession of unattractive choices degrades the score along a path containing a well-fitting local region. The Smith-Waterman method adds the fourth option: end the region being aligned.
3. *Scoring and traceback.* The score of a global alignment is the number in the matrix element at the lower right. In the Smith-Waterman method it is the optimal value encountered, wherever in the matrix it appears. For global alignment, traceback to determine the actual alignment starts at the lower-right cell. In the Smith-Waterman method it starts at the cell containing the optimal value and continues back only as far as the region of local similarity continues.

The Smith-Waterman method would report a unique global optimum for our example:

```
ggaatgg

  atg
```

Note that no gaps appear outside the region matched.
(Example adapted from: Tyler, E.C., Horton, M.R. and Krause, P.R. (1991) 'A review of algorithms for molecular sequence comparison,' *Comp. Biomed. Res*. 24, 72-96.)

[2]Optional section. Readers in doubt may consider the remarks in Lesk, A.M. (1988). 'TATA for now...', *Trends Biochem. Sci*. 13, 410.

Significance of alignments

Suppose alignment reveals an intriguing similarity between two sequences. Is the similarity significant or could it have arisen by chance? (We raised this question in Chapter 1.) For some simple phenomena - tossing a coin or rolling dice - it is possible to calculate exactly the expected distribution of results, and the likelihood of any particular result. For sequences it is not trivial to define the population from which the alignment is selected. For instance, to take random strings of nucleotides or amino acids as controls ignores the bias arising from non-random composition.
A practical approach to the problem is as follows: If the score of the alignment observed is no better than might be expected from a *random permutation* of the sequence, then it is likely to have arisen by chance. We may randomize one of the sequences, many times, realign each result to the second sequence (held fixed), and collect the distribution of resulting scores. Figure 4.4 shows a typical result. For database searches, use the population of results returned from entire database as the population with which to measure the statistics.

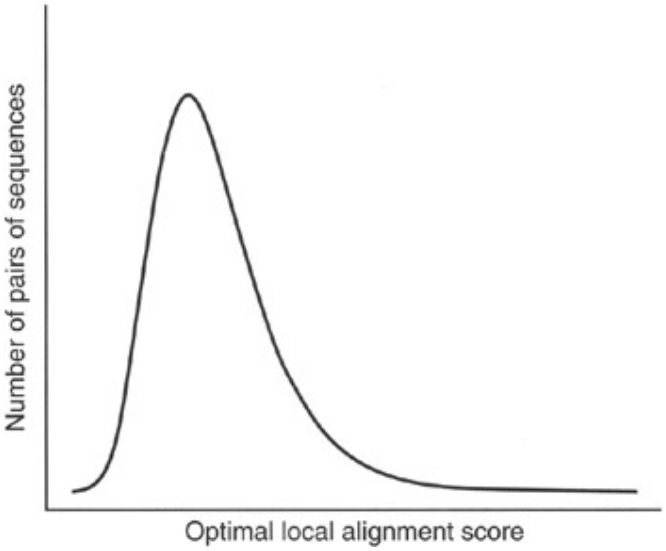

Figure 4.4: Optimal local alignment scores for pairs of random amino acid sequences of the same length follow an extreme-value distribution. For any score x, the probability of observing a score $\geq x$ is: $P(\text{score} \geq x) = 1 - \exp(-Ke^{-\lambda x})$, where K and λ are parameters related to the position of the maximum and the width of the distribution. Note the long tail at the right. This means that a score several standard deviations above the mean has a higher probability of arising by chance (that is, it is less significant) than if the scores followed a normal distribution.

Clearly, if the randomized sequences score as well as the original one, the alignment is unlikely to be significant. We can measure the mean and standard deviation of the scores of the alignments of randomized sequences, and ask whether the score of original sequence is unusually high. The Z-score reflects the extent to which the original result is an outlier from the population:

$$\text{Z-score} = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$$

A Z-score of 0 means that the observed similarity is no better than the average of random permutations of the sequence, and might well have arisen by chance. Other values used as measures of significance are P = the probability that the observed match could have happened by chance, and, for database searching, E = the number of matches as good as the observed one that would be expected to appear by chance in a database of the size probed (see Box, page 186).

Many 'rules of thumb' are expressed in terms of percent identical residues in the optimal alignment. If two proteins have over 45% identical residues in their optimal alignment, the proteins will have very similar structures, and are very likely to have a common or at least a similar function. If they have over 25% identical residues, they are likely to have a similar general folding pattern. On the other hand, observations of a lower degree of sequence similarity cannot rule out homology. R.F. Doolittle defined the region of 18–25% sequence identity as the 'twilight zone' in which the suggestion of homology is tantalizing but dangerous. Below the twilight zone is a region where pairwise sequence alignments tell very little. Lack of significant sequence similarity does not preclude similarity of structure.

Although the twilight zone is a treacherous region, we are not entirely helpless. In deciding whether there is a genuine relationship, the 'texture' of the alignment is important - are the similar residues isolated and scattered throughout the sequence; or are there 'icebergs' - local regions of high similarity (another term of Doolittle's), which may correspond to a shared active site? We may need to rely on other information about shared ligands or function. Of course, if the structures are known, we could examine them directly.

Some illustrative examples are:

- Sperm whale myoglobin and lupin leghaemoglobin have 15% identical residues in optimal alignment. This is even below Doolittle's definition of the twilight zone. But we also know that both molecules have similar three-dimensional structures, both contain a haem group, and both bind oxygen. They are indeed distantly-related homologues.
- The sequences of the N- and C-terminal halves of rhodanese have 11% identical residues in optimal alignment. If these appeared in independent proteins, one could not conclude from the sequences alone that they were related. However, their appearance in the same protein suggests that they arose via gene duplication and divergence. The striking similarity of their structures confirms their relationship.

- **How to play with matches but not get burnt**

- Pairwise alignments and database searches often show tenuous but tantalizing sequence similarities. How can we decide whether we are seeing a true relationship? Statistics cannot answer biological questions directly, but can tell us the likelihood that a similarity as good as the one observed would appear, just by chance, among unrelated sequences. To do this we want to compare our result with alignments of the same sequences to a large population. This 'control' population should be similar in general features to our aligned sequences, but should contain few sequences related to them. Only If the observed match stands out from the population can we regard it as significant.

- To what population of sequences should we compare our alignment? For pairwise alignments, we can pick one of the two sequences, make many scrambled copies of it using a random-number generator, and align each permuted copy to the second sequence. For probing a database, the entire database provides a comparison population.

- Alignments of our sequence to each member of the control population generates a large set of scores. How does the score of our original alignment rate? Several statistical parameters have been used to evaluate the significance of alignments.
    - The *Z*-score is a measure of how unusual our original match is, in terms of the mean and standard deviation of the population scores. If the original alignment has score S,

$$\text{Z-score of S} = \frac{\text{S} - \text{mean}}{\text{standard deviation}}$$

    - A *Z*-score of 0 means that the observed similarity is no better than the average of the control population, and might well have arisen by chance. The higher the *Z*-score, the greater the probability that the observed alignment has not arisen simply by chance. Experience suggests that *Z*-scores ≥ 5 are significant.
    - Many programs report *P* = the probability that the alignment is no better than random. The relationship between *Z* and *P* depends on the distribution of the scores from the control population, which do *not* follow the normal distribution.

A rough guide to interpreting *P* values:

$P \leq 10^{-100}$	exact match
P in range 10^{-100}-10^{-50}	sequences very nearly identical, for example, alleles or SNPs
P in range 10^{-50}-10^{-10}	closely related sequences, homology certain
P in range 10^{-5}-10^{-1}	usually distant relatives
$P > 10^{-1}$	match probably insignificant

    - For database searches, some programs (including PSI-BLAST) report *E*-values. The *E*-value of an alignment is the expected number of sequences that give the same *Z*-score or better if the database is probed with a random sequence. *E* is found by multiplying the value of *P* by the size of the database probed. Note that *E* but not *P* depends on the size of the database. Values of *P* are between 0 and 1. Values of *E* are between 0 and the number of sequences in the database searched.

A rough guide to interpreting *E* values:

$E \leq 0.02$	sequences probably homologous
E between 0.02 and 1	homology cannot be ruled out
$E > 1$	you would have to expect this good a match just by chance

- Statistics are a useful guide, but not a substitute for thinking carefully about the results, and further analysis of ones that look promising!
  - As a cautionary note, consider the proteinases chymotrypsin and subtilisin. They have 12% identical residues in optimal alignment. These enzymes have a common function, and a common catalytic triad. However, they have dissimilar folding patterns, and are not related. Their common function and mechanism is an example of convergent evolution. This case serves as a warning against special pleading for relationships between proteins with dissimilar sequences on the basis of similarities of function and mechanism!

Multiple sequence alignment

'One amino acid sequence plays coy; a pair of homologous sequences whisper; many aligned sequences shout out loud.' In nature, even a single sequence contains all the information necessary to dictate the fold of the protein. How does a multiple sequence alignment make the information more intelligible to us? Alignment tables expose patterns of amino acid conservation, from which distant relationships may be more reliably detected. Structure prediction tools also give more reliable results when based on multiple sequence alignments than on single sequences.

Visual examination of multiple sequence alignment tables is one of the most profitable activities that a molecular biologist can undertake away from the lab bench. Don't even THINK about not displaying them with different colours for amino acids of different physicochemical type. A reasonable colour scheme (not the only one) is:

Colour	Residue type	Amino acids
Yellow	Small nonpolar	Gly, Ala, Ser, Thr
Green	Hydrophobic	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Magenta	Polar	Asn, Gln, His
Red	Negatively charged	Asp, Glu
Blue	Postively charged	Lys, Arg

To be informative a multiple alignment should contain a distribution of closely- and distantly-related sequences. If all the sequences are very closely related, the information they contain is largely redundant, and few inferences can be drawn. If all the sequences are very distantly related, it will be difficult to construct an accurate alignment (unless all the structures are available), and in such cases the quality of the results, and the inferences they might suggest, are questionable. Ideally, one has a complete range of similarities, including distantly-related examples linked through chains of close relationships.

Structural inferences from multiple sequence alignments

Thioredoxins are enzymes found in all cells. They participate in a broad range of biological processes, including cell proliferation, blood clotting, seed germination, insulin degradation, repair of oxidative damage,