

homologs and their domain structures will be conserved, even though the proteins may have diverged in sequence. Structure can be used as a criterion for identifying homologs in a test set.

A “good” alignment program meets at least two criteria: it maximizes the number of homologs found (true positives), and it minimizes the number of nonhomologous proteins found (false positives). Another way to describe these criteria is in terms of *sensitivity* and *specificity*, which are discussed in more detail in Chapter 9. In this context, sensitivity is a measure of the fraction of actual homologs that are identified by the alignment program, and the specificity is a measure of the fraction of HSPs that are not actually homologs. Brenner et al. (1998) tested a number of different alignment approaches, including Smith-Waterman, FASTA, and an early version of BLAST. They discovered that, at best, only about 35% of homologs were detectable at a false positive error frequency of 0.1% per query sequence.

An intuitive measure of homology employed in the past was the percentage of sequence identity. The rule of thumb was that sequence identities of 25%–30% in an alignment signified true homology. Brenner et al. employed a database of known proteins annotated with respect to homology/non-homology relationships to test the relationship between sequence identity and homology. Their results are shown in Fig. 7.6. Figure 7.6B shows percentage identity plotted against alignment length for proteins that are *not* homologs. For comparison, a threshold percentage identity taken to imply similar structure is plotted as a line (see Brenner et al., 1998 for details). The point is that for alignments 100 residues in length, about half of the *nonhomologous* proteins show *more* than 25% sequence identity. At  $50 \pm 10$  residues of alignment length, there are a few nonhomologous proteins having over 40% sequence identity. A particular example of this is shown in Fig. 7.6A. This serves as a reminder of why methods providing detailed statistical analysis of HSPs are required (Section 7.4.2).

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.

---

**Fig. 7.6 (opposite page).** The limitations of sequence identity as an indicator of homology. Panel A: Unrelated proteins that have 40% sequence identity over a segment of approximately 60 residues. Panel B: Scores of unrelated, nonhomologous proteins as a function of alignment length. The line indicates the sequence identity cutoff, sometimes taken as an indicator of homology. Reprinted, with permission, from Brenner SE et al. (1998) *Proceedings of the National Academy of Sciences USA* 95:6073-6078. Copyright 1998 National Academy of Sciences, USA.