→ **MILESTONE 15**

# BLAST-off for genomes

The generation of genome sequences from a wide range of organisms has opened the field of comparative genomic analysis, assisting the annotation of individual genomes, and bringing new insights into genome evolution. Central to the field of comparative genomics have been programs used to search and align protein or DNA sequences based on a measure of similarity.

Sequence alignment can involve either global alignment, in which the two sequences are aligned over their entire length, or local alignment, in which subregions of the two sequences are aligned. The latter has been more widely applied, as DNA sequences generally show isolated regions of similarity.

An exact solution to the global alignment problem was developed by Saul Needleman and Christian Wunsch in 1970, by applying dynamic programming to find the optimal alignment between two sequences. In 1981, Temple Smith and Michael Waterman extended this dynamic programming approach to solve the local alignment problem. These exact solutions placed sequence comparisons on a firm mathematical grounding, and formed the basis for the early alignment search algorithms.

However, the exact solutions proved slow in practice, especially for searching large databases, spurring the development of faster heuristic approaches. One early successful heuristic algorithm to enable efficient searching of large databases, FASTA, was presented by David Lipman and William Pearson in 1988. This simplified the problem by searching for short regions of exact match and then extending them. In 1990, Stephen Altschul and colleagues presented the basic local alignment search tool (BLAST), which instead searched for all short matches above a given scoring threshold, and showed that this improved speed. In addition, Altschul developed a statistical framework for sequence alignment that provided a conceptual basis for understanding similarity measures, and a method for assessing the statistical significance of a given alignment. BLAST indexes the query sequence and scans against a database, whereas Jim Kent in 2002 showed how the reverse could increase speed (at tolerably reduced sensitivity), with the BLAST-like alignment tool (BLAT).

The sequencing of full-length genomes increased interest in developing tools for genome-wide multiple alignments. Paving the way, in the 1980s, Clustal drew on information in phylogenetic trees to go from pairwise to multiple sequence alignments, although mainly for protein sequences. MUMmer, which in 1999 was one of the first to move to whole-genome multiple alignments, proved useful for aligning bacterial genomes. This was soon followed by BLASTZ and MultiZ from Webb Miller and colleagues; these were useful for cross-species comparisons, allowing alignment and searching of mammalian chromosome-length sequences.

*Orli Bahcall, Associate Editor,*
Nature Genetics

> "…BLAST indexes the query sequence and scans against a database."



**ORIGINAL RESEARCH PAPERS** Smith, T. F., & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981) | Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**, 237–244 (1988) | Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988) | Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990) | Altschul, S. F. & Gish, W. Local alignment statistics. *Methods Enzymol.* **266**, 460–480 (1996) | Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999) | Kent, W. J. BLAT: the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002) | Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003)