How do you spell DNA?

SIR — In a provocative column, you discussed the theory and practice of artificial intelligence[1], using spelling correction as an example of a task in which automated computer procedures should be of great value. In the past ten years, a number of tools have been developed to solve similar problems in molecular biology.

The usual connection between molecular biology and language is made by drawing parallels between the bases of DNA and the letters of the alphabet, between codons and words, and between proteins and sentences. Expanding on this analogy, mutations in DNA sequences can be thought of as spelling errors, creating new sequences from those already in existence. The processes of evolution, whether of DNA or language, select which of these new sequences or words are to survive.

What then can computer programs tell us about these molecular spelling errors? One approach, pioneered by Needleman and Wunsch and elaborated by Sankoff, Sellers and others, maximizes matches, subtracting a weighted sum of mismatches and insertion/deletions, between two DNA or protein sequences. For example, AATCAG and ATTCG might be related by

A A T C A G
A T T C * G

where there is one point mutation and one insertion/deletion. Two sequences of length N have of the order of 2^{2N} possible relationships. For $N = 1{,}000$, this number is approximately 10^{600} so that an exhaustive search is impossible. Nonetheless, rigorous algorithms and programs exist to locate rapidly the optimal relationships for sequences of 1,000 or more bases. Many modifications and improvements exist. Long insertions/deletions can be weighted according to length. Other mutational events, such as inversions, are being incorporated into these rapid comparison algorithms.

These algorithms can be extended in a novel way to predict RNA secondary structure; the base-paired regions correspond to matches in the sequence comparison algorithms. Indeed, these basic methods have been adapted and applied to speech recognition, geological strata, handwriting recognition, bird song and gas chromatography. Many such applications and associated theory appear in a recent book edited by D. Sankoff and J.B. Kruskal[2]. In addition to spelling correction, the analogue to a dictionary search for a word is the comparison of a new sequence to an existing DNA or protein data base. These searches locate library sequences with regions of similarity to the new sequence, combining the concepts of dictionary and imperfect spelling. New tests are being devised to estimate the statistical significance of the similarities. Methods to perform these large searches have been developed, and were successfully applied in the recent discovery of sequence similarity between the transforming protein of a primate sarcoma virus and a platelet-derived growth factor[3,4]. Another recent computer finding indicates that an oncogene product appears to have arisen as a result of recombination of two unrelated cellular genes[5].

One of the most intriguing areas of DNA sequence analysis parallels the concept of language interpretation. Patterns such as repetitive DNA are frequently noticed before their meaning is understood. An important case is the search for promoter sequences in *Escherichia coli*[6]. DNA upstream from *E. coli* coding regions is presumed to contain base sequences that spell "begin transcription". These are searches for patterns of unknown composition and length which we know must occur, however imperfectly, within specific regions of DNA. Text editors, even those equipped to search for regular expression patterns[7], are not adequate to this task. If useful and rigorous algorithms are developed for these tasks of locating imprecise words of unknown spelling, new and nontrivial insights could result.

New techniques of pattern recognition in DNA and protein sequences are resulting from creating and applying concepts of mathematics, statistics and computer science appropriate to specific questions of molecular biology. As often happens in science, these methods may turn out to have broad applicability.

MICHAEL S. WATERMAN*
Departments of Mathematics and of Biological Sciences,
University of Southern California,
Los Angeles, California 90089, USA

*This work was supported by a grant from the System Development Foundation.

1. *Nature* **306**, 637 (1983).
2. Sankoff, D. & Kruskal, J.B. (eds) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, London, 1983).
3. Weiss, R. *Nature* **304**, 12 (1983).
4. Doolittle, R.F. *et al. Science* **221**, 275 (1983).
5. Naharro, G., Robbins, K.C. & Reddy, E.P. *Science* **223**, 67 (1984).
6. Hawley, D.K. & McClure, W.R. *Nucleic Acids Res.* **11**, 2237 (1983).
7. Aho, A.V., Hopcroft, J.E. & Ullman, J.D. *The Design and Analysis of Computer Algorithms* (Addison-Wesley, London, 1974).