Quantitative Comparison

SCIENCE IDEAS

by William A. Beyer, Christian Burks, and Walter B. Goad

Although DNA sequences are replicated and passed on to future generations with great fidelity, changes do, of course, occur. They provide mutations, the raw material for evolution, as well as causation for disease and death. Three kinds of localized change can occur: replacement of one base by another, deletion of a base, or insertion of a base. In addition. a number of adjacent bases may be simultaneously deleted or inserted. The probabilities of these various changes are not known in general, and their determination is an outstanding problem.

The idea of comparing sequences quantitatively—in this case the sequences of amino acids in proteins—goes back to 1963. Then Linus Pauling and Emile Zuckerkandl suggested the possibility of reconstructing the course of evolution by examining the relations among the sequences of hemoglobin proteins in extant vertebrate organisms. And in 1967 W. M. Fitch and E. Margoliash constructed an evolutionary tree by measuring "distances" among the cytochrome c proteins of various organisms. Unfortunately, some of the distances in the tree were negative! Then in 1968 Stan Ulam, in conversation with Temple Smith, both at the University of Colorado, suggested that the relatedness of two sequences be measured by use of a distance that fulfills the criteria of a metric: a binary relation that is real-valued, positive-definite, symmetric. and satisfies the triangle inequality. In terms of the changes that occur in the evolution of protein or nucleic acid sequences. these properties of a metric make biological sense, excepting perhaps the symmetry property. This distance between two sequences was defined as the minimum total of localized changes—replacements, insertions. and deletions—that would transform one sequence into the other.

Another measure of relatedness of sequences is called similarity. The properties of similarity have never been made precise. Presumably similarity should be a binary, positive-valued, symmetric relation and should in some unspecified sense be complementary to a metric distance. That is. a small distance should correspond to a high similarity and a large distance to a low similarity.

Now if you imagine comparing two sequences by, say, writing them on paper tapes and sliding one along relative to the other, you will quickly see that to find by trial and error the minimum number of changes—an optimal alignment of the two sequences—generally requires considerable effort. You have to be prepared to snip out a base from one tape or the other, see whether the resulting alignment is improved, and repeat this operation many times. In 1970 two biologists. Needleman and Wunsch, then at Northwestern University, devised a procedure for finding the optimal alignment (calculating the similarity) on a computer. Their method proceeds by induction, that is, by assuming that the optimal alignment of the first n bases of one sequence with the first m bases of another is constructible from the optimal alignments of shorter segments of the two sequences. The resulting algorithm requires on the order of nm operations.

Also in 1970 Bill Beyer of Los Alamos, Smith, and Ulam commenced work on refinements of the idea of distance between sequences and on applications of those distances to studies of evolution. They developed a mathematical theory in which biological sequences were regarded as words of finite length over a finite alphabet. (The alphabets for DNA and protein sequences consist of four bases and twenty amino acids, respectively.) Smith made use of a suggestion by Fitch that local closeness of two sequences could be detected by comparing all possible subsequences of one sequence with all possible subsequences of the other sequence and then comparing the sums of certain differences with those expected for two random sequences. Beyer developed a method for applying linear programming to the construction of evolutionary trees based on distances between contemporary protein sequences. This method, together with a metric of Smith's, was used to produce evolutionary trees based on cytochrome c sequences. Most of the computer calculations were done by Myron Stein on the MANIAC computer.

In 1974 Peter Sellers, a mathematician at Rockefeller University, after hearing a talk there by Ulam. developed a theory of metrics among sequences and an algorithm. related to a 1972 algorithm by David Sankoff of University de Montreal, to calculate one of Ulam's metrics. (It was not until 1981 that Smith and Mike Waterman showed that, under a certain relation between similarity and distance, the Needleman-Wunsch and the Sellers algorithms are equivalent.)

The Needleman-Wunsch algorithm, and its refinements, finds the optimal overall alignment of two fixed sequences. However, one of the key discoveries of recent work in molecular genetics is the frequency and great biological importance of events in which substantial pieces of DNA are moved from one place to another in the genome of an organism or from one organism to another. To locate such DNA segments, algorithms are needed that find locally close subsequences embedded within otherwise unrelated sequences. Sellers devised one solution to this problem in 1979, and later in the same year Goad and Minoru Kanehisa and, independently, Smith and Waterman devised another that provides a more controlled "sieve." The latter finds all pairs of subse-

of DNA Sequences

quences whose distances fall below a prescribed threshold.

When insertion and deletion of bases is allowed. any two sequences can be aligned in some way. To distinguish biologically important relationships, it becomes important to study the frequency with which subsequences of a given closeness occur in unrelated sequences—that is. by chance alone. Such a study was begun by Goad and Kanehisa in 1982 and is being continued by them. Earlier this year, Smith. Waterman, and Christian Burks completed an investigation of the statistics of close subsequences in the entire Gen Bank database. The results of this investigation provide an empirical basis for assessing the statistical significance of calculated similarities. However, establishing a biologically proper measure for statistical significance remains a critical problem.

The combination of the Gen Bank database and methods for determining similarities between sequences will provide a very useful tool to molecular biologists. For example, screening the database for similarities to a newly sequenced segment of DNA can reveal in the case of an extremely high similarity, that the new segment has been sequenced previously in either the same or a different genetic context. High similarity here means that the two sequences being compared are almost identical over a span of greater than fifty to one hundred nucleotides. A lower, though still statistically significant similarity may indicate that the two sequences share a common functional role in living cells, despite originating in different genetic locations. The distance algorithm can also be fruitful in comparing the sequence for one strand of a DNA segment with that for its own complementary strand. High similarities in this type of comparison can be used to trace regions of potential "hairpin" structures on the RNA transcribed from the DNA, Such structures, where the RNA folds and binds to itself, are in some cases known to be the basis for recognition by an enzyme. Kanehisa and Goad have developed an elaboration of the distance algorithm for this purpose. Self-comparison of sequences has also proved useful in catching the evidence left behind by a particular kind of experimental error. called loop-back, that often occurs during the process of biochemically determining nucleic acid sequences.

To enable and encourage searches of the entire database for similarities to a "query" sequence. Smith and Burks have worked on developing an implementation of the distance algorithm that will make such comparisons, which have not been practicable by hand or even on most computers, possible now and as the database continues to grow. The current program employs the following strategy. For every comparison of the query sequence with another sequence, the similarity score for the best local alignment of the two sequences is saved; after a run through the database, the statistically significant scores are printed out, together with the names of the corresponding sequences. This list can then guide a more focused examination of the similarity of the query sequence to others in the database. The program was written to take advantage of the vector architecture of Cray computers, and a recent run involving about 44,000 comparisons between pairs of vertebrate sequences, each several hundred nucleotides long, took 170 minutes on a Cray-1 at Los Alamos.

Scientists will continue to increase the speed of comparisons based on the concept of distance between sequences by developing more efficient algorithms and computer programs. For instance, Jim Fickett has developed an algorithm that. in most cases. increases the speed of the distance calculation by a factor of ten. Efforts in this direction will. of course. become more and more essential as the sequence data expand. But a more exciting direction now being explored is that of making the transition from basing the characterization of distance on the symbolic, or alphabetic. representations of sequences to basing this characterization on the physical structures of the DNA segments. An analogy with human language illustrates the need to extend the distance concept in this way.

Consider the words "leek" and "leak"; if we were comparing only the letters in this pair of homonyms, we would judge them to be almost identical. Or consider the words "'sanguine" and "cheerful"; on the same basis of comparison, these synonyms would be judged quite dissimilar. Of course, in terms of the role of words in allowing communication between people, the meaning of a word is a much more appropriate criterion for comparison than the symbols for that meaning. Now consider the following nucleotide sequences:

(1) ACACAC,
(2) ACAAAC,
(3) GTGTGT,

The distance algorithm discussed above would classify (1) and (2) as quite close (only a single mismatch among the six bases) and (1) and (3) as quite distant (six mismatches). However, extrapolation from recent x-ray crystallographic studies of DNA by Dickerson and coworkers at Caltech and by Rich and coworkers at MIT indicate that although (2) is found in the right-handed B-form double-helical structure suggested by Watson and Crick. (1) and (3) are both found in radically different left-handed Z-form double-helical structures. From the point of view of the proteins in living cells that have to communicate with DNA by making chemical contact with its nucleotide strings, (1) and (3) would be almost identical sequences, both quite different from (2). Thus. current attempts to extend the distance algorithm are anticipating and incorporating a variety of spectroscopic, crystallographic, and biochemical data that identify. on the basis of structure and function, homonyms and synonyms in nucleic acid sequences.

This work is an example of the evolution of biology itself from the qualitative studies of the pre-DNA days to the mathematical. highly quantitative studies of today. ■