SEQUENCE ANALYSIS

Contributions by Ulam to Molecular Genetics

by Walter B. Goad

Lord Rayleigh once introduced a key idea with "It is tolerably obvious once remarked. ." Yes, I think now, that is just how it was—Stan Ulam providing us with a steady stream of ideas and observations "tolerably obvious" only in retrospect, and then striking in the way they became integral to one's tangible world of evolved and evolving forms and actions. Here I would like to sketch ideas developed during the sixties and seventies as an avalanche of detail, still growing, gathered about the way sequences of nucleotide bases in DNA encode instructions for development and propagation of living organisms. Stan showed us a very general way of thinking precisely about relationships among sequences, in particular, how to devise quantitative measures of relationship that. together with the computer, are of immense help in ferreting out meaning in the very great quantities of data now pouring forth.

I met Stan soon after arriving in Los Alamos at the end of 1950. I came ostensibly to finish a thesis begun at Duke under Lothar Nordheim, who had arrived several months earlier while I stayed in Durham awaiting security clearance. At last a telegram came from Carson Mark that read, "Your clearance not available." An anxious telephone call established that the "not" had been garbled in transit from "now." I was immediately swept up in the thermonuclear program, kept busy with the rest dissecting schemes and designs, and sometimes new phenomena, usually standing around a blackboard. Introducing the right factors, right at least in order of magnitude, was both vital and enjoyably competitive, laced with humor-esoteric, malicious, or plain—and an occasional flash of ego. The key, of course, was to discern the dominant phenomenon and to estimate its role in the matter at hand. One always had a feeling, almost visceral, as to how deeply an argument was rooted in the web of our knowledge of physics and mathematics. Stan habitually turned things to view from a variety of directions, much as he would see an algebraic structure topologically, and vice versa, and often supplied the connection that dispelled a gathering fog.

Around 1960 Jim Tuck invited Leonard Lerman, who was in the thick of the gathering revolution in biology and then at the University of Colorado, to visit Los Alamos. The "phage group" gathered loosely around Max Delbruck had established a mode of analysis that is still driving the biological revolution: Changes in a single DNA molecule are amplified by biological reproduction, usually in a microorganism, to the macroscopic level; there the consequences of those changes, however ramified, can be studied with the resources of physics and chemistry. The amplification is made possible by an immensely powerful, and growing, armory of molecular tools based on enzymes that carry out specific operations on specific DNA'S. As we grasped those ideas from Leonard and began to see the clarity and concreteness with which the mechanisms of life would emerge from such analysis, many of us were galvanized. We soon responded in a way typical of the culture, organizing a seminar, hungrily seeking out the many aspects of the subject. As I recall, the seminar continued through the sixties and early seventies with a varying membership but with Stan, Jim Tuck, George Bell, and me as regulars. We were frequently visited, and enormously encouraged, by Ted Puck, who has built a distinguished school of molecular and cell biology at the University of Colorado and who was, and is, exceedingly optimistic about the contribution systematic theory can make to biology.

A quick tour of systematic theory inevitably would start with Darwin's grand synthesis. For physicists a key way point would be the publication in 1944 of Erwin Schrodinger's short book *What Is Life?,* which equates that grand question with one congenial to physicists: What generates "negentropy," the high degree of order that living systems are continually creating from the environment? Ever since, theorists of all kinds have looked to the formulation of some powerful physical theory of life. Short of that, what we do know is that living systems escape from the determinism of ordinary chemistry by interposing molecular adaptors to control molecular interactions. An example is provision by the complex protein structure of hemoglobin of an effective interaction between 0_2 molecules that is completely unrelated to their interactions as free molecules: Within a hemoglobin molecule up to four 0_2's bind at distinct sites and thus effectively stick together. Furthermore, three or four stick more tightly than one or two. So, where there is much oxygen, four are tightly bound; where there is little, departure of one causes the others to more easily depart. Invoking the adaptor principle, Francis Crick predicted the existence of what are now called transfer RNA'S—small RNA molecules, a particular species of which adapts each three-base codon to molecules of a particular amino acid. A Zen-like consciousness of physical necessity—for the way in which electrons and nuclei, and thus atoms and molecules, do what they must—leads first to puzzlement at living systems and then to resolution: Molecular adaptors free the logic of higher levels of organization to adopt and express a logic of their own, exploiting, not circumventing, physical necessity.

Proteins and RNA'S provide an array of complex and highly specific adaptors, and their structures are encoded in sequences of nucleotide bases in DNA. To a large extent the double-helical structure of DNA wraps the information-conveying part of the DNA into a protected interior and so in the main removes chemical constraints on the propagation and selection of sequences.

Working on DNA as a substrate, evolution has produced the marvelously complex web of living systems we see today. The working hypothesis, to which no exception is yet known, is that all of the information for propagation and development of individual organisms is encoded somehow in the sequence of four bases adenine (A), thymine (T), guanine (G), and cytosine (C) along the DNA molecules (or, in some cases, RNA molecules) that compose its genome. The "somehow" includes the great triumphs of the past two decades, the present frontiers of molecular biology, and, undoubtedly, a great deal that we do not now even glimpse. Less than a decade after Watson and Crick determined the structure of DNA, researchers at the laboratories of Nirenberg, Khorana, and Ochoa fully worked out the "genetic code" by which the base sequences of particular segments of DNA—genes—are translated into sequences of amino acids that fold up as particular proteins. For a few years many people felt that, in principle, DNA function was now completely understood. But in the mid seventies methods were worked out for determining sequences of bases in DNA, and it amost immediately emerged that not even the sequences that are translated into proteins are simple, continuous coding sequences. The last few years have seen the discovery of a great many distinct "signals" that control the replication of DNA and the expression of genes. However, it is not yet known how the action of those signals is coordinated, as it must be, to yield the patterns seen during reproduction and development. On the other hand, an outline is emerging of the organization within DNA of repetitive sequences, which make up a substantial fraction of the genome in higher organisms. That organization may or may not have signaling capabilities, but it is almost surely important in evolution. Perhaps most striking of all is the growing knowledge of phenomena—such as the mobility and duplication of pieces of DNA and its rearrangement-that introduce into the genome a degree of dynamism far beyond what classical genetics had led us to suspect.

Most of this was yet to come in the late sixties, when the amino-acid sequences of a few proteins were the only biological sequences known. However, it was already clear that the information on which a cell acts is encoded in sequences of bases, and the question of how to characterize relationships among sequences hundreds or thousands of bases long was at hand. With his almost visceral feeling **for representation of natural phenomena** by general mathematical structures, Stan immediately framed the question in terms of defining a distance between sequences or, more generally, of defining a usable metric space of sequences (Ulam 1972). This he did by considering certain elementary base changes by which one sequence might be transformed into a second: Replacement of one base by another and insertion or deletion of a base. (Combinations of these changes can result from errors in DNA replication, chromosomal crossover during meiosis, insertion of viral or other DNA, or the action of mutagens.) Obviously, one sequence can be transformed into another by more than one set of elementary changes, as shown in the accompanying figure. What Stan proposed was to compute a measure, a “size,” for each such set and to define as the distance between the sequences the minimum value of the measure.

In simplest form the measure is a sum of weights, one for each of the elementary changes that compose a transformation set. The set corresponding to the minimum measure—the distance between the two sequences-can be interpreted as the minimal mutational path by which one sequence could have evolved from the other. In 1974 Stan, with Bill Beyer, Temple Smith, and Myron Stein applied the idea of distance to discerning evolutionary relationships among various species from variations in the amino-acid sequences of a protein they all share. Also in 1974 Peter Sellers, after hearing Stan talk at Rockefeller University, proved that such a distance can indeed satisfy the conditions of a metric, the most demanding of which is satisfaction of a triangle inequality. Without that, one’s sense of what it means for some among several sequences to be close and others distant would be quite unreliable.

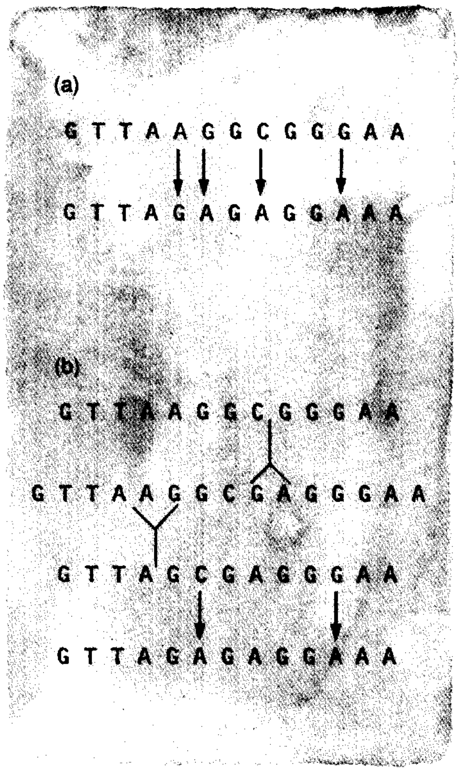
Finding the distance between two sequences of length N by brute force, that is, by computing the measures for all the possible sets of elementary changes, requires on the order of N ! computer operations. An algorithm for determining the distance in N* operations was discovered by the biologists Saul Needleman and Christian Wunsch in 1970 and independently by Sellers in 1974. Essentially, the algorithm proceeds by induction: The minimal set of changes needed to transform the first *n* bases of one sequence into the first ***m*** bases of the other is found by extending already computed minimal transformations of shorter subsequences, then ***n* and *m* are** increased, and so on until the ends of the sequences have been reached.

By the end of the 1970s, it was apparent that DNA sequencing would take off, and that investigators from all areas of biology, biomedicine, and bioagriculture would increasingly apply it to their particular research problems. It was also obvious that computer manipulation and analysis of sequences, much of it flowing from Stan’s idea for a metric, would play an increasingly large role in exploiting the information. Mike Waterman had joined Beyer and Smith in working on sequence analysis, and Minoru Kanehisa, a postdoc from Japan, and I made genetic sequences and their analysis our principal preoccupation from then on. In 1982 a consortium of federal agencies funded GenBank, the national genetic-sequence data bank. Los Alamos collects and organizes the sequence data, and Bolt Beranek and Newman Inc. distribute them to users. By the end of 1986, DNA sequences totaling about 15 million bases, from several hundred species, had been deposited in GenBank.

In the 1980s a series of problems in sequence comparison have been faced with varying degrees of success. One problem now solved concerns global versus local closeness (closeness, that is, in the sense of a distance between sequences). Often of interest are sequences that are close to each other although em-

DISTANCE BETWEEN DNA SEQUENCES

Consider the two short DNA sequences GTTAAGGCGGGAA and GTTAGAGAGGAAA. As shown in (a), one of these can be transformed into the other by four base substitutions. If the “weight” assigned to a base substitution is x, then the “measure” of the set of changes in (a) is 4x. Alternatively, as shown in (b), one sequence can be transformed into the other by two base insertions, two base deletions, and two base substitutions. Since base insertions (deletions) occur less frequently than do base substitutions, the weight y assigned to an insertion (deletion) is different from that assigned to a substitution; in particular y is assigned a value greater than that of x. The measure of the set of changes illustrated in (b) is 2x+4y, which is greater than 4x. The distance between the two given sequences is defined as the minimum of the measures calculated for all possible sets of elementary changes that transform one sequence into the other.


(a)
G T T A A G G C G G G A A
G T T A G A G A G G A A A
(b)
G T T A A G G C G G G A A
G T T A A G G C G A G G G A A
G T T A G C G A G G G A A
G T T A G A G A G G A A A

bedded in otherwise unrelated longer sequences. Peter Sellers first introduced the important distinction between local and overall closeness in 1980. A measure suited to the local problem (essentially the number of weighted changes per base, formulated so that the algorithm of Needleman, Wunsch, and Sellers can still be used) was introduced in slightly different forms by Kanehisa and me in 1982 and by Smith and Waterman in 1981. Another class of problems stems from the sheer quantity of data-examining 15 million bases, even with an N^2 algorithm, requires hundreds of hours on a Cray. That problem has been reasonably successfully dealt with by presecreening sequences for likely candidates for significant relationships. A table of pointers to the locations of short subsequences (a simple hash table) is created and searched for short matching sequences. At this writing the method is being implemented with new hardware features of the Cray *XMP*. For a general review of sequence-comparison algorithms, see Goad 1986; for a review that emphasizes mathematical aspects, see Waterman 1984.

Devising a metric appropriate to the investigation at hand is probably not a problem that can be precisely posed, much less solved. A simple metric in which each elementary change is given the same weight may well suffice when the object of study is a virus under great pressure to preserve a small genome. But such a metric may show misleading relationships when applied to segments of DNA from a more complicated organism, as Fitch and Smith found in 1983 for mammalian hemoglobins. Some relationships may depend on similarities in three-dimensional structure of DNA that are preserved through a set of sequences, as may be the case for the elements that control initiation of expression of particular genes. To discover such relationships, one needs a measure of structural similarity, expressed of course in terms of sequences. That problem is just beginning to be faced. A good sense of the problem, and of the limitations of sequence comparison, is given by analogy to another idea of Stan's. He proposed that perception, and thought itself, be considered in terms of a metric space. This frames the question: How is the distance between the visual fields corresponding to, say, two tables—which will vary greatly with circumstances-computed in our brains so that it is small compared with the distance between the visual fields corresponding to a table and a chair? Clearly the metric appropriate to a particular class of problems depends on the mechanisms one hopes to discover or illuminate.

Mathematical analysis has spread into nearly every corner of molecular genetics; its spread and development is still accelerating. In early 1986 the Department of Energy took the initiative in seriously exploring sequencing of the complete human genome, some 3 billion bases. In that project computerized management and analysis of information will play a key role.

Speaking of sequence analysis, GenBank, and all that, Stan once said, "I started all this." Yes. ■

Further Reading

S. M. Ulam. 1972. Some ideas and prospects in biomathematics. *Annual Review of Biophysics and Bioengineering* 1: 277–291.

Willaim A. Beyer, Myron L. Stein, Temple F. Smith, and Stanislaw M. Ulam. 1974. A molecular sequence metric and evolutionary trees. *Mathematical Biosciences 19: 9–25.*

Peter H. Sellers. 1974. On the theory and computation of evolutionary distances. *SIAM Journal on Applied Mathematics 26: 787.*

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48: 443.*

Peter H. Sellers. 1980. The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms I: 359-373.*

Walter B. Goad and Minoru I. Kanehisa. 1982. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Research 10: 247.*

T. F. Smith and M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology 147: 195–197.*

Walter B. Goad. 1986. Computational analysis of genetic sequences. *Annual Review of Biophysics and Biophysical Chemistry 15: 79–95.*

Michael S. Waterman. 1984. General methods of sequence comparison. *Bulletin of Mathematical Biology 46: 473–500.*

Walter M. Fitch and Temple F. Smith. 1983. Optimal sequence alignments. *Proceedings of the National Academy Of Sciences of the United States of America 80:* 1382–1386.

Walter B. Goad received a B.S. in physics from Union College in 1945 and a Ph. D., also in physics, from Duke University in 1954. He has been a member of the staff of Los Alamos since 1950. In 1982 he received a Distinguished Performance Award from the Laboratory in recognition of his efforts at establishing GenBank, and in 1987 he was named a Fellow of the Laboratory. Until recently he directed the activities of GenBank, in which he continues to participate. His research focuses primarily on analysis of biological sequences. He is a Fellow of the American Physical Society and of the American Association for the Advancement of Science.