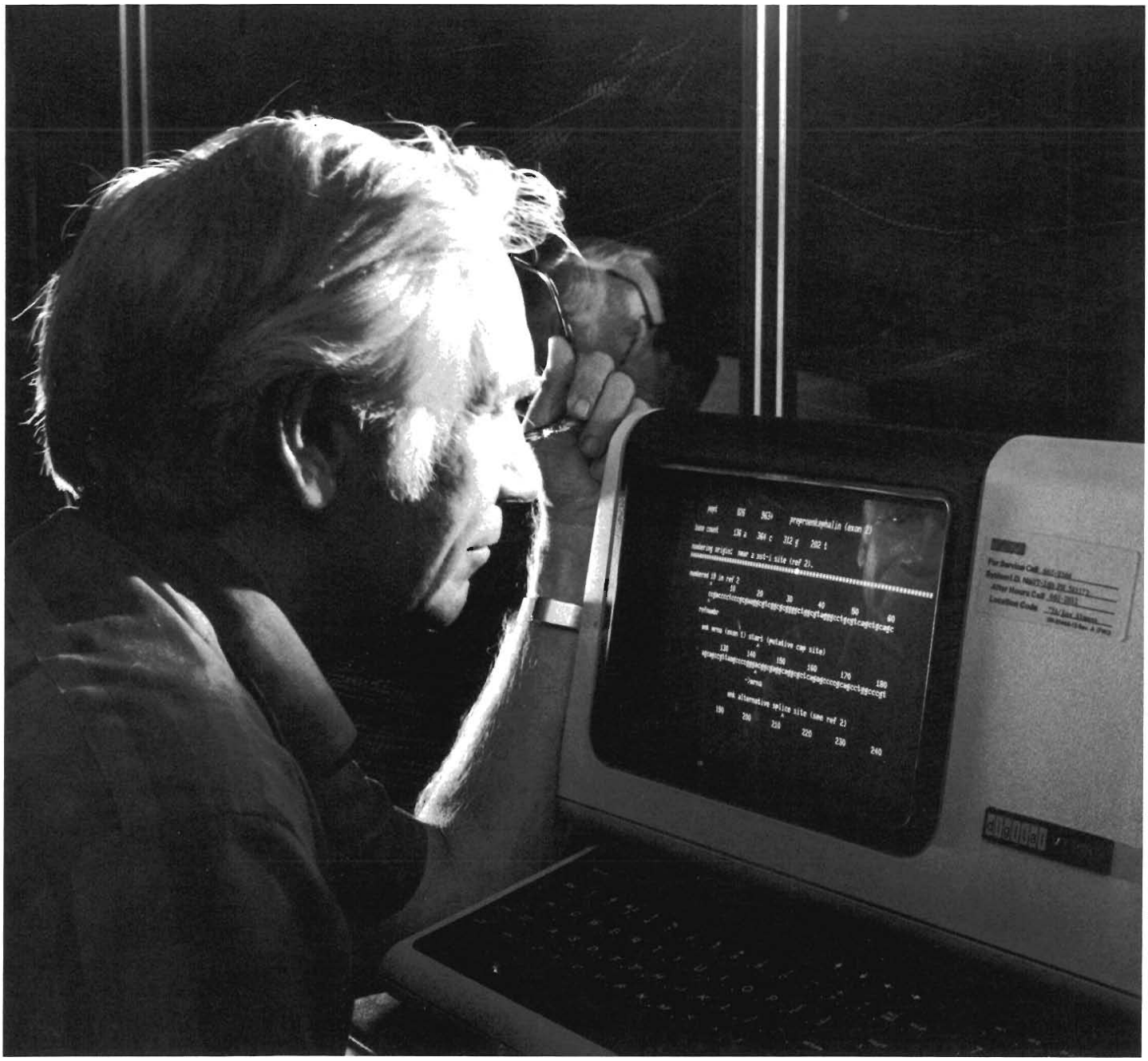
ACGGCGAGGCAGGCGCT..... So starts one of the sequences of bases in human DNA that encode enkephalin, an essential brain hormone and regulator of mood. Such sequences are being determined in great and growing numbers. A data bank that makes this information accessible to computer-aided analysis is becoming an important tool in the worldwide campaign to unravel the mechanisms of life and its evolution.



GenBank

by Walter B. Goad

To understand the significance of the information stored in GenBank, you need to know a little about molecular genetics. What that field deals with is self-replication—the process unique to life—and mutation and recombination—the processes responsible for evolution—at the fundamental level of the genes in DNA. This approach of working from the blueprint, so to speak, of a living system is very powerful, and studies of many other aspects of life—the process of learning, for example—are now utilizing molecular genetics.

Molecular genetics began in the early '40s and was at first controversial because many of the people involved had been trained in the physical sciences rather than the biological sciences, and yet they were answering questions that biologists had been asking for years. Max Delbrück, for instance, a very prominent early figure, was trained as a theoretical physicist. He and his group worked with bacteriophage as the simplest systems in which to study replication on a molecular level. Bacteriophage are just at the boundary of life and can replicate themselves only in host bacteria, but what Delbrück and his group learned about the mechanics of replication and the structure of the genes in these viruses turned out to be relevant to all living things.

It's amazing that so much of what today's biology major knows of genetics was discovered so recently. For example, it wasn't recognized until 1944 that DNA, which had been known since 1869, is the carrier of the genes. And not until 1953 was a structure suggested for DNA that explained its ability to transmit hereditary information. That year Watson and Crick proposed that DNA consists of two long nucleotide chains bound together as a double helix by the attractive forces between pairs of complementary bases. This binding of complementary bases is the key physical process in replication. It is physically a very weak association—in fact, two complementary bases bind to each other only a little more strongly than each binds to water when free in solution. Why this association should be such a very strong ordering factor in living systems is a puzzling question, but that it is so is an overwhelming fact.

Another important event in molecular genetics was the cracking of the genetic code, which you might say relates the stuff of memory—DNA—to the stuff of activity—proteins. The idea that somehow the bases in DNA determine the amino acids in proteins had been around for some time. In fact, George Gamow suggested in 1954, after learning about the structure proposed for DNA, that a triplet of bases corresponded to an amino acid. That suggestion was shown to be true, and by 1965 most of the genetic code had been deciphered. Also worked out in the '60s were many details of what Crick called the central dogma of molecular genetics—the now firmly established fact that DNA is not translated directly to proteins but is first transcribed to messenger RNA. This molecule, a nucleic acid like DNA, then serves as the template for protein synthesis.

Walter Goad, leader of the group responsible for the Laboratory's role in GenBank, at work on one of the entries in this national repository for nucleic acid sequence data.

These great advances prompted a very distinguished molecular geneticist to predict, in 1969, that biology was just about to end since everything essential was now known. It's ironic that within a year of that prediction, restriction enzymes were first isolated. These enzymes are the key to using recombinant DNA techniques to make billions or trillions of copies of a DNA segment. With that many copies and well-known chemical and physical techniques one can then sequence the segment, that is, determine the exact order of the bases it contains.

Before sequencing was possible, almost all the advances in molecular genetics were based on making a single change, with radiation or chemicals, in the DNA of an organism and seeing what happens. For example, you alter the DNA of a phage, allow the phage to multiply in its host bacterium, and examine the consequences of the change in the large population of identical descendants. Incidentally, phage are particularly convenient subjects because they multiply so rapidly. In these experiments you are essentially asking yes or no questions the way you do in the game of twenty questions. With twenty well-chosen questions you can narrow a million possibilities to just one very rapidly. But sequence data is a tool for analysis that is finer than asking yes or no questions.

People working on problems throughout biology—problems like hereditary diseases, cancer, evolutionary relationships—immediately saw that they could make very good use of this tool, and as soon as it became clear that sequencing could be done with facility, it also became clear that sequence data would accumulate at a very great rate. To discuss how these data could be managed and exploited, a meeting was organized at Rockefeller University in the summer of 1979. No one questioned but that computers and sequence data were made for each other. Transmitting a long, seemingly random

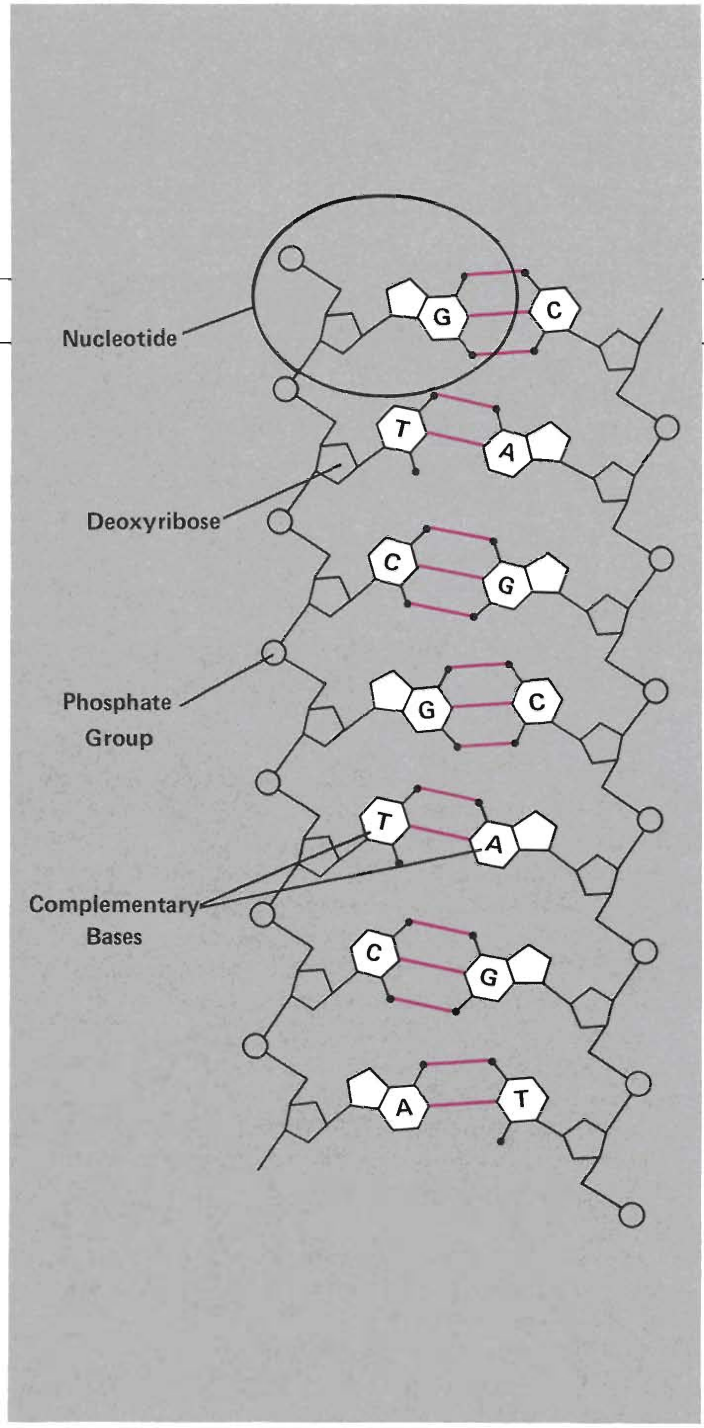
sequence of four letters from one person to another without errors is hardly possible except by putting the information on a computer-readable medium. The need for a data bank was in the air.

How We Got Involved

Mike Waterman and Temple Smith from Los Alamos went to that meeting because for a number of years they had been working with Bill Beyer and Stan Ulam on recognition of patterns in the sequences of amino acids in proteins. Nucleic acid sequencing was just breaking on the scene, but some hundreds of protein sequences were already known, at least partially. Mike and Temple came back from the meeting and talked to people here who were interested in sequence data and their analysis with computers.

I was one of those people, although my training had not been in biology. My early work at the Laboratory was in physics—neutron transport and hydrodynamics—and in molecular physics in the sense of computing equations of state for various materials. I first ventured into biology in 1960 or thereabout. Jim Tuck had met Leonard Lerman, one of the small pioneering group studying bacteriophage, at a cocktail party in Boulder and invited him to Los Alamos. Lerman had discovered that the simplest electronic excitation of DNA—by ultraviolet light—affected its genetic behavior in some more complicated way than would be expected if the ultraviolet light simply altered a single base. To explain this he suggested that perhaps the excitation energy migrated as excitons, which were somewhat the rage in solid-state physics at the time. I talked to Lerman and became quite fascinated by the opportunities for studying a living system on the molecular level. It was so unlike anything known in physics that a single molecular change in DNA could be duplicated faithfully and its consequences examined. You might say that biology offers to physics a molecular amplifier. I hadn't any idea that such a thing existed.

So Lerman and I did some work on the migration of excitons in DNA and its genetic interpretation. As a consequence, I became involved with Ted Puck's group at the University of Colorado Medical Center in Denver. They were looking at genetics at a higher level—at chromosomal abnormalities in newborns. There seemed to be epidemics of them, and I tried to determine how likely it was that we were seeing purely chance behavior instead of epidemics. Then in 1970 I spent a year with Crick at the Laboratory of Molecular Biology in Cambridge, where Max Perutz was working out the structural and functional details of hemoglobin. As a result I did some work on conformational changes in proteins. These changes in shape allow proteins to act as adapters to bring about interactions between molecules that have no specific chemical relation to each other. A classic example is the interaction that hemoglobin induces among four oxygen molecules. I also worked with several molecular



A small segment of a DNA molecule. For clarity the two strands of nucleotides are shown uncoiled from the usual double helical configuration. Each nucleotide is made up of a phosphate group, deoxyribose, and one of four nitrogenous bases. The two nucleotide strands are bound to each other by hydrogen bonds (red) between two possible pairs of structurally complementary bases: guanidine (G) and cytosine (C), or thymine (T) and adenine (A). This pairing of bases is the mechanism directing both replication and transcription of DNA. The DNA of one organism differs from that of another by the sequence of bases along the strands. DNAs of viruses contain tens of thousands of nucleotide pairs, those of bacteria a few million, and those of higher plants and animals billions.

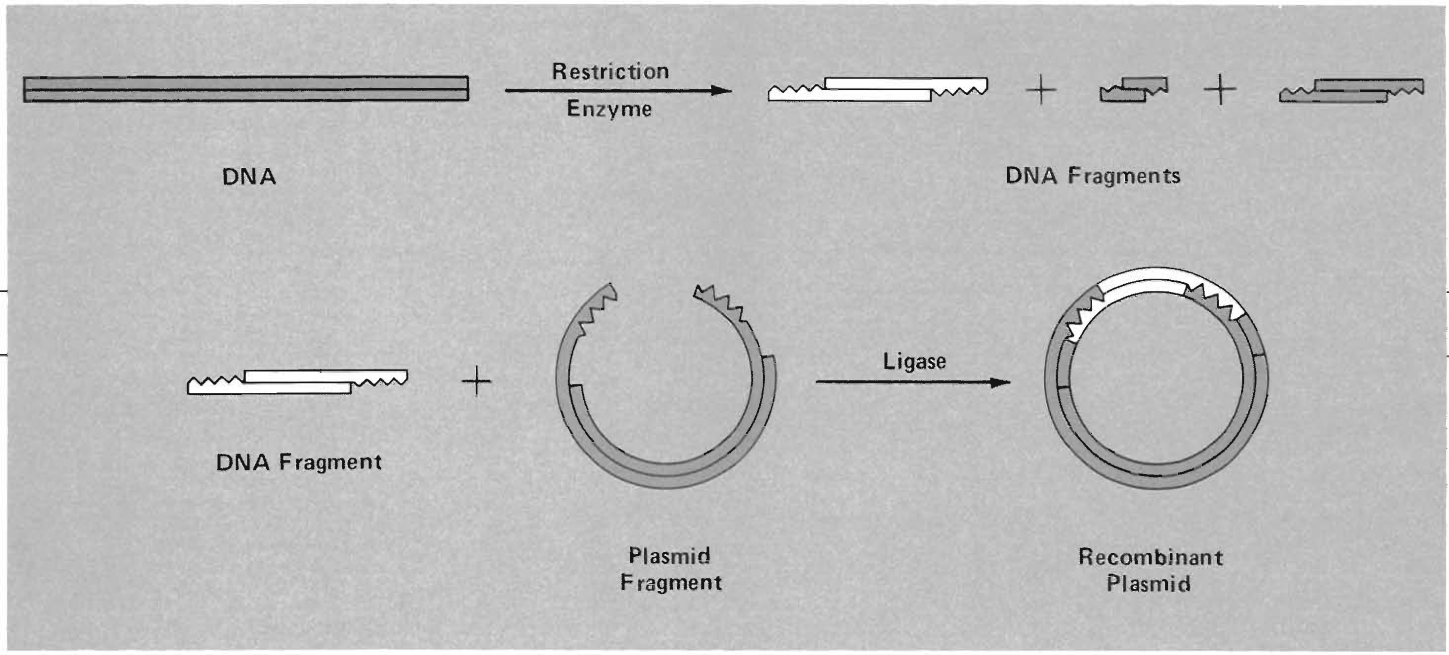
TTT	phenylalanine	TCT	serine	TAT	tyrosine	TGT	cysteine
TTC	phenylalanine	TCC	serine	TAC	tyrosine	TGC	cysteine
TTA	leucine	TCA	serine	TAA	end	TGA	end
TTG	leucine	TCG	serine	TAG	end	TGG	tryptophan
CTT	leucine	CCT	proline	CAT	histidine	CGT	arginine
CTC	leucine	CCC	proline	CAC	histidine	CGC	arginine
CTA	leucine	CCA	proline	CAA	glutamine	CGA	arginine
CTG	leucine	CCG	proline	CAG	glutamine	CGG	arginine
ATT	isoleucine	ACT	threonine	AAT	asparagine	AGT	serine
ATC	isoleucine	ACC	threonine	AAC	asparagine	AGC	serine
ATA	isoleucine	ACA	threonine	AAA	lysine	AGA	arginine
ATG	methionine	ACG	threonine	AAG	lysine	AGG	arginine
GTT	valine	GCT	alanine	GAT	aspartic acid	GGT	glycine
GTC	valine	GCC	alanine	GAC	aspartic acid	GGC	glycine
GTA	valine	GCA	alanine	GAA	glutamic acid	GGA	glycine
GTG	valine	GCG	alanine	GAG	glutamic acid	GGG	glycine

Dictionary of the genetic code, listing the codons, or triplets of bases, in DNA that code for amino acids in proteins. The code is degenerate in the sense that, with the exception of tryptophan and methionine, each amino acid is specified by more than one codon. Synthesis of a protein involves transcription of a DNA segment to a single-stranded chain of nucleotides known as messenger RNA (mRNA) and translation of the mRNA to the protein. The bases in mRNA are complementary to those in one strand of the DNA and are assembled in the order given by that strand. The transcribed codons are then translated to the sequence of amino acids in the protein. The codons TAA, TAG, and TGA act as signals for terminating protein synthesis, and the codon for methionine, ATG, acts as a signal for initiating the synthesis of nearly all proteins.

geneticists on the details of how bacterial DNA is replicated and how its expression is turned on and off. These experiences led rather naturally to my being interested in DNA sequence data.

That 1979 meeting set several of us to thinking about a data bank. It was clear that just accumulating the data in a computer was not particularly interesting. What was interesting were questions about organizing, managing, and analyzing the data. We started collecting sequence data and writing software for analysis. We worked up a proposal and presented it to NIH and NSF and anyone else who would read it. That proposal included, in addition to a data bank, an analysis center that would provide access to software running on our computers. Thousands of experts around the country have questions they would like to ask about sequence data. Many of them will find collaborators who are expert in applying computers to answer their questions. But the ideal situation is for the person with a question to sit down at a terminal and ask it himself. There are many barriers to doing that, but one that can be removed is the necessity of writing your own software or of rewriting some existing software to suit your own computer. It turns out that very little software is really portable despite all the talk about the issue. But given modern telecommunications, there is no reason not to use existing software on the machine it was developed for. We were excited about an analysis center because it would put us in touch with many imaginative, talented people.

We contacted other groups—in the States, in Europe, and in Israel—that were also interested in sequence data. We talked, for instance, with Margaret Dayhoff's group at the National Biomedical Research Foundation in Washington. Margaret for many years led their data bank for protein sequences; unfortunately she died not too long ago. Her group started collecting data on nucleic acid sequences about the same time we did. We exchanged ideas with Margaret's group about organizing and managing the data, which were



Many copies of a DNA fragment are needed to determine its sequence of bases by standard chemical or biochemical methods. These copies are provided by recombinant DNA techniques made possible by the action of restriction enzymes, which recognize certain short sequences of bases in a DNA molecule and cut the molecule at those sequences. Such an enzyme cuts each DNA molecule in a sample containing many molecules of the DNA into the same set of fragments. The desired fragments can then be separated from the others by length. Many restriction enzymes cut DNA so as to leave the fragments with "sticky" ends; that is, a short stretch of bases left on one end of a fragment is complementary to a short stretch of bases left on the other end. If the restriction enzyme used does not act in this way, sticky ends can be added to the fragments by other enzymes. The DNA fragments are then mixed with plasmids (small circular pieces of DNA found in bacteria) whose circles have been opened by an appropriate restriction enzyme and equipped, one way or the other, with complementary sticky ends. Some of the opened plasmid circles associate with and are closed by the added DNA fragments, whereas others simply close again. Treatment with still another enzyme, a ligase, re-establishes the covalently linked circles of DNA, which can now infect their bacterial hosts and be replicated rapidly. Finally the added DNA fragments are removed from the multiplied plasmids, again by action of an appropriate restriction enzyme and separation by length. A great many ingenious refinements have been devised to enhance the efficiency and ease of these procedures. For example, some of the bacteria will lack an infecting plasmid, and some of the plasmids will lack an added DNA fragment. Those bacteria infected by a "recombinant" plasmid can be made to replicate to a greater degree by arranging that the recombinant plasmids carry some property advantageous to the bacteria.

particularly important issues in light of the anticipated volume of the data. And we were in touch with people at Stanford who had feet in both the computer science department and the biochemistry and molecular genetics departments, some of whom were especially interested in the question of interfaces between people with questions and the computers and software that could answer them.

Many people we talked with agreed that a data bank would be most useful if combined with an analysis center. But partly because of the ever-present shortage of funds, partly because of the difficulty of specifying what an analysis center should be like, and partly because of politics, NIH decided to put off an analysis center and to invite proposals only for a data bank to be cosponsored by a number of organizations, including NSF, DOE, and DOD.

Oddly enough, we submitted two proposals, each a joint proposal with one of two firms we were already involved with. One of the firms was IntelliGenetics, which was formed by the Stanford group to offer analytical services to the biotechnology industry; the other was Bolt Beranek and Newman Inc. BBN had distributed our collection of nucleic acid sequence data through their PROPHET system, which provides dial-up software to bench biochemists and biologists.

Both BBN and IntelliGenetics approached us independently and pointed out that a joint proposal would make sense. We would collect and manage and organize the data, and the company would handle the distribution. That sounded good to us. For one thing, a joint proposal with a commercial firm solved the practical problem of collecting user fees, which are probably necessary if for no other reason than to discourage abuse of the system.

However, as part of a national laboratory we could not offer our collaboration to one firm and not the other. So we ended up with two joint proposals to write. They were quite different, though, because BBN is a spin-off from MIT and IntelliGenetics is a spin-off from Stanford. Anyone familiar with the artificial intelligence community knows that MIT and Stanford differ a great deal in style.

After having the proposals reviewed—there were three major ones, our two and a third by Margaret Dayhoff's group—and asking a great many questions, NIH selected our joint proposal with BBN. We heard the news in September of last year and felt nicely rewarded for all the effort on the proposals. But then the work began in earnest.

How the Data Bank Operates

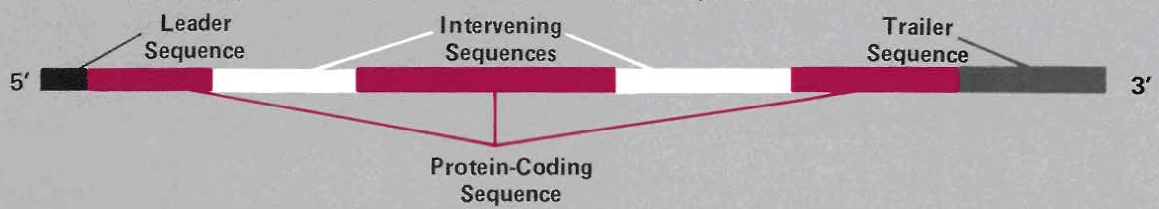
The first job we face is collecting the sequence data. This is straightforward although physically very demanding. We've been working very hard just to get caught up with the data that was in existence when we started. Our estimate of that was based on scanning some journals and looking through lists of titles. Unfortunately, the estimate was low because the titles of many articles with sequences in them don't reveal that. Recently, more and more sequences are being sent directly to us by the authors, and we are trying to get support from the journals to encourage or enforce that practice.

Our contract calls for us to collect all nucleic acid sequences containing more than fifty bases. The data bank now contains about two million bases, and that number is increasing at the rate of about a million bases per year, roughly what we expected. Most of the sequences are on the order of a thousand bases long, but the longest—the entire genome of the lambda bacteriophage—contains about 50,000 bases. About 200 species are represented, although the usual laboratory species, drawn from viruses, bacteria, fruit flies, and small rodents, predominate. We do have a few unusual species—some plants, apes, and an East Indian deer that happens to have a very small number of chromosomes. However, for only a very, very few of these species is the entire genome known. In fact, even for that most studied of all bacteria, *E. coli*, only about 3.5 percent of its genome has been sequenced, and for *Homo sapiens* the fraction is less by a factor of about a thousand.

The data bank entry for a sequence includes, in addition to the sequence itself, a name up to ten characters long that tries to speedily identify the sequence by at least suggesting species and function, the

The weekly meeting of the group is a time for discussing problems about the GenBank entries.

```
locus        humhba2      1138 bp                    updated   06/30/82
definition   human alpha2-globin gene and flanks.  1138bp
source       human.
reference    1  (bases 1 to 1138)
  authors    liebhaber,s.a., goossens,m.j. and kan,y.w.
  journal    proc nat acad sci usa 77, 7054-7058 (1980)
reference    2  (bases 1 to 1054)
  authors    proudfoot,n.j. and maniatis,t.
  journal    cell 21, 537-544 (1980)
reference    3  (bases 98 to 929)
  authors    orkin,s.h., goff,s.c. and hechtman,r.l.
  journal    proc nat acad sci usa 78, 5041-5045 (1981)
reference    4  (bases 802 to 837)
  authors    proudfoot,n.j. and longley,j.i.
  journal    cell 9, 733-746 (1976)
comment      formerly humanhba2. compared with humhbaps in ref 2. ref 3 reports
             sequence of alpha-thalassemic individual with complete absence of
             alpha2 mrna. ref 3 suggests that cttgg rather than cctgg at base
             660 is inconsistent with the presence of a bstni restriction site.
             ref 4 compares with rabbit alpha- and beta-globins.
features        from        to        description
    pept         135       230        alpha2-globin
                 348       551
                 692       820
base count       183 a     411 c     351 g     193 t
origin       approximately 100bp 5-prime to bstni site
        1 aggccgcgcc ccgggctccg cgccagccaa tgagcgccgc ccggccgggc gtgcccccgc
       61 gccccaagca taaaccctgg cgcgctcgcg gcccggcact cttctggtcc ccacagactc
      121 agagagaacc caccatggtg ctgtctcctg ccgacaagac caacgtcaag gccgcctggg
      181 gtaaggtcgg cgcgcacgct ggcgagtatg gtgcggaggc cctggagagg tgaggctccc
      241 tcccctgctc cgacccgggc tcctcgcccg cccggaccca caggccaccc tcaaccgtcc
      301 tggccccgga cccaaacccc acccctcact ctgcttctcc ccgcaggatg ttcctgtcct
      361 tccccaccac caagacctac ttcccgcact tcgacctgag ccacggctct gcccaagtta
      421 agggccacgg caagaaggtg gccgacgcgc tgaccaacgc cgtggcgcac gtggacgaca
      481 tgcccaacgc gctgtccgcc ctgagcgacc tgcacgcgca caagcttcgg gtggacccgg
      541 tcaacttcaa ggtgagcggc gggccgggag cgatctgggt cgaggggcga gatggcgcct
      601 tcctctcagg gcagaggatc acgcgggttg cgggaggtgt agcgcaggcg gcggcgcggc
      661 ttgggccgca ctgaccctct tctctgcaca gctcctaagc cactgcctgc tggtgaccct
      721 ggccgcccac ctccccgccg agttcacccc tgcggtgcac gcttccctgg acaagttcct
      781 ggcttctgtg agcaccgtgc tgacctccaa ataccgttaa gctggagcct cggtagccgt
      841 tcctcctgcc cgctgggcct cccaacgggc cctcctcccc tccttgcacc ggccccttcct
      901 ggtctttgaa taaagtctga gtgggcggca gcctgtgtgt ggctgggttc tctctgtccc
      961 ggaatgtgcc aacaatggag gtgtttacct gtctcagacc aaggacctct ctgcagctgc
     1021 atggggctgg ggagggagaa ctgcagggag tatgggaggg gaagctgagg tgggcctgct
     1081 caagagaagg tgctgaacca tcccctgtcc tgagaggtgc cagcctgcag gcagtggc
sites        key          span  description
       98 ->mrna             1  mrna cap site
       98 refnumbr           1  numbered 1 in ref 1; zero not used
      135 ->pept             1  hba2 cds start
      138 pept/pept          0  hba2 mature protein start
      139 refnumbr           1  numbered 1 in ref 2
      231 pept/ivs           0  ivs1 start
      231 variation          5  tgagg deleted in ref 3
      348 ivs/pept           0  ivs1 end
      416 conflict           1  a in refs 1 & 2; g in ref 3
      552 pept/ivs           0  ivs2 start
      655 conflict           1  c in refs 1 & 2; ct in ref 3
      659 conflict           4  gctt in refs 1 & 2; ggcct in ref 3
      692 ivs/pept           0  ivs2 end
      763 conflict           1  t in refs 1 & 2; c in ref 3
      820 pept<-             1  hba2 cds end
      930 mrna<-             1  poly a addition site
      930 refnumbr           1  numbered 1 in refs 4;3'to 5'
```



The GenBank entry for a gene specifying the alpha-2 polypeptide chain of human hemoglobin. This aberrant gene is responsible for a particular form of thalassemia, that is, for one of a group of hereditary anemias. We have added to the entry a schematic representation of the gene and its environs, which is color-keyed to the sites of interest and their descriptions given in the entry. The sequence specifying the leader region of the mRNA begins at base 98 with a site to which is added a "cap" (containing a methylated guanine bound to the mRNA by phosphate groups) that may promote translation of the mRNA. The polypeptide coding, which begins at base 135 with the initiation codon ATG, is interrupted by two intervening sequences beginning at bases 231 and 552. These sequences are transcribed but are spliced out before translation. The sequence specifying the trailer region of the mRNA ends with a site to which is added a "poly-A tail" (of many adenylate residues) that may protect the mRNA from degradation.

source of the sequence data, and a succinct description of the biological function and setting of the sequence. The sequences are catalogued according to these three items, which are roughly equivalent to title, author, and subject of a book. We also note features that have been determined to be of biological interest—but not those that are only speculated to be so. We update the entries, of course, as new information comes our way. For example, many of the sequences determined early in the game have since been redone. We make no judgment about the accuracy or reliability of the data; that is a matter for investigators, referees, and users to thrash out. We are, however, very concerned about handling the entries consistently because the data are to be used primarily as input for computer programs, which are not at all tolerant of inconsistency. At our weekly meetings problems of consistency take a great fraction of our time.

In some cases we have to decide whether two sequences thought to be from different genetic regions are actually from the same region and vice versa. These situations can arise from sequencing errors, from working with different strains of the same organism or with different alleles of the same gene, or from the fact that DNA is not always copied exactly and some of the inexact copies are still being replicated. We try to resolve conflicts of this nature with the authors, and if it is determined that two sequences are indeed the same, we combine them in a single entry and note the differences.

The work involved in collecting and annotating the sequences will soon be somewhat reduced because of an arrangement we have worked out with the data bank for nucleic acid sequences at the European Molecular Biology Laboratory in Heidelberg. That data bank and ours have been communicating since the beginning but until now have been randomly duplicating rather than sharing each other's efforts. But the sharing must be done in a way that preserves the integrity of each data bank, since each has a different constituency and different support. The Japanese are also moving toward creating a data bank, and we believe they will be a third collaborating entity. We hope this collaboration will permit us to spend less time just hurrying to keep up and more on improving our operation and getting on with our research.

One improvement we have been able to implement recently is the transfer of data entry and management activities to a microcomputer. Previously we had been using the Laboratory's mainframe computers, which certainly have their virtues but offer only line-editing capability. With screen editing we should all be more productive. Using a microcomputer has another advantage—it puts us more in touch with the people who are doing the sequencing, who seldom use big computers but are turning increasingly to microcomputers. We can easily make the software developed for our microcomputer compatible with other commonly used ones. The authors will, we think, find this software useful and will be more likely to send their sequence data to us already in our format.

But we love big computers, too. For detailed comparisons of sequences, a Cray is the machine to use because it's twenty times faster and, more important for us, ten times cheaper. We feel that each new entry should be routinely compared with existing entries and think we can develop a much cheaper way to do that. But unfortunately, our preoccupation with the primary job delays work on many such improvements.

The sequence data are available to users primarily through BBN. Once a month we send the data base to them, and they issue it on tape. A few people subscribe to the data base on a regular basis, but most receive it only now and again. The data base is also available over the telephone through a computer at BBN. National Biomedical and IntelliGenetics distribute the data base to their customers, and several computer centers at universities maintain it for a large number of users. We also offer our software to users on a dial-up basis. About seventy people around the country take advantage of this service now and then. We regard this as a way of encouraging people to submit data to us directly, to find errors in the entries, and to criticize and make suggestions about our operation. We would like to have more funds to support this aspect of GenBank, but it smacks of an analysis center, which is still hanging fire.

I'd like to mention our splendid crew, which includes people from mathematics and from biology and certainly represents one of the few instances where these two fields really coalesce. The staff members are Jim Fickett, Christian Burks, and Minoru Kanehisa. Minoru helped create the data bank; he is now physically at NIH and works only half-time for us. Temple Smith and Ruth Nussinov have been involved as visiting staff members. And Gerry Myers,

a molecular biologist and tutor at St. John's College, is a consultant and major contributor. We rely for help on Graduate Research Assistants—young people between college and graduate school—with majors in biology or computer science. Since they write the first drafts of the sequence annotations, they learn quite a bit about molecular genetics, not to mention about computers. They generally stay for about a year, and all have gone on to either graduate or medical school. Right now we have with us Bryan Bingham, Ute Elbe, Leslie Kay, Randy Linder, and Debra Nelson. Carol England is both secretary and coordinator of data entry, and Mia McLeod has just joined us as a data analyst.

What to Look for in the Data

Analysis of the sequences is, of course, the ultimate objective of their being determined and our collecting them. But those of you who may have seen sequences in journals will agree that picking out features of interest or making comparisons requires a specialist—or a computer fed very clever software. Since sequence data have neither the internal order of numerical data—the order of the numbers themselves—nor that of textual data—the grammar and syntax of the language—their analysis calls for a different approach. We have developed some interesting programs for analysis. For example, Jim Fickett discovered that the portions of a sequence that code for proteins have a regularity, a periodicity in the statistical sense, that is absent in noncoding portions. This fact can be used to find the protein-coding segments, and then it is relatively trivial to translate the bases in the segments to the amino acids in the proteins. Incidentally, more and more protein sequences are being predicted from nucleic acid sequences because it's much easier determining them that way than from the proteins themselves. People are particularly interested in the predicted amino acid sequences of membrane proteins because these proteins are of special importance in cellular activities.

As I mentioned before, comparison of sequences is a significant aspect of sequence analysis. [See the sidebar "Quantitative Comparison of DNA Sequences" for more detail about this subject.] What one wants to know is in what way and by how much two sequences differ, as a whole or over certain portions. Generally sequences differ a great deal, but the lowest level of difference between sequences from closely related DNAs is the replacement of one base by another here and there along the sequences. Such base replacements may in some instances be inconsequential because of the degeneracy of the genetic code. A higher level of difference arises from additions or deletions of bases, and we and other people have developed very efficient algorithms for spotting such differences.

Another type of analysis involves searching for strings of complementary bases along an RNA molecule that would permit it to fold and bind to itself—to form hairpin-like structures. A good deal of experimental evidence indicates that in many circumstances it is such a structure, rather than the sequence itself, that is recognized by an enzyme or a protein as a signal for some activity. Similarly, one can look for sequences of bases in a DNA molecule that would cause irregularities in its structure. Even today the structures of various DNAs are not precisely known because of the difficulty of obtaining large crystals. One idealizes the structure as perfectly regular, but considering that the sugar-phosphate backbone encloses four different bases, two of which are twice as large as the other two, certain sequences of bases are bound to cause irregularities in shape or mechanical properties. These irregularities may serve as signals for initiating certain processes or may play a role in the packaging of DNA. In higher organisms DNA does not exist simply as a random coil. Instead it is superwound around some proteins, and this superwound structure is itself superwound into a very complex structure.

In terms of analysis, one thing is clear: as more and more sequence data become available, more and more ideas about what to look for will be proposed. And there's no question but that this aspect of GenBank is the most challenging and the most rewarding.

What the Sequence Data Offer

There's also no question but that sequence data will answer—and raise—more and more questions about the mechanisms of life and its evolution. I don't mean, however, to imply that sequence data displaces everything else. Biochemistry, cell biology, organismic biology—these fields are as important as ever, but they are increasingly being propelled and unified by insights and techniques from molecular genetics. And much has been and can be learned about DNA without actually determining sequences. It is relatively simple experimentally to determine, for example, how many times a certain segment of DNA is repeated or the degree to which different organisms share a given gene. But sequence data provide the finest detail.

Classical evolutionary studies, for instance, rely on comparing characteristics such as anatomy or geographical distribution to find out how organisms are related. At the level of molecular genetics, you compare the sequences of bases in the genes of the organisms and the proteins dictated by those sequences. One of the things you find is that there is a tremendous range in how exactly the sequences of bases are conserved even between organisms that had a common ancestor not so long ago. What apparently is true is that if a small change in a gene causes a big disturbance in the organism, then the sequences differ by very little. But if a small change disturbs the organism very little or not at all—perhaps makes only a trivial change of an amino acid in a protein or no change whatever—then

the sequences will drift a lot. One of the exciting lines of inquiry is to try to understand, in terms of these big or small changes in the sequences, what evolutionary pressures were involved.

Comparison of sequences can give the history of evolution over various time scales. Over the whole of evolution, one looks at widely different organisms, say yeast and man, for genes that are common and genes that are entirely different. Over shorter evolutionary periods one compares, say, mice and rats. And over even shorter times one can examine the differences among the sequences of human ethnic and racial groups. In terms of the evolutionary tree, one can compare branches that are close together or one can compare the root with the top of the tree. Finding out what the similarities and differences are is very interesting for the light it sheds not only on how organisms evolve but also on how they had to evolve, that is, on what functions are essential.

Another intriguing aspect of DNA that sequence data has revealed with great clarity is how little of it codes for proteins—only about 20 percent in most organisms. In human DNA, for example, one short sequence of about 300 noncoding bases is repeated about 300,000 times, and at least four or five others are repeated a comparable number of times. Some noncoding sequences are repeated tens of thousands of times, some thousands of times, and some hundreds of times. The sequences are repeated not exactly but with a very high degree of fidelity. What is all that DNA doing? Some of it may be reproducing itself because it can, parasitically so to speak. And clearly some of it must be involved in controlling gene expression and, thereby, morphogenesis—the transformation of a single fertilized egg cell to a complex three-dimensional organism containing an enormous number of specialized cells.

I should point out that bacteria seem to be much more economical—only about 10 percent of their DNA does not code for proteins. Perhaps their high rate of reproduction precludes wasting energy by replicating nonessentials. But most of the DNA in all higher organisms does not code for protein. An extreme example is a certain crab, 95 percent of whose DNA consists solely of repetitions of the sequence AT.

Questions remain even about the DNA that does code for proteins: human DNA probably contains codes for about 100,000 different proteins and yet less than 1000 proteins have actually been identified. Many of the coded proteins are probably produced in very small numbers, a few molecules per cell. It is likely that different kinds of cells produce quite different arrays of these minor proteins, as is true for many proteins that are abundant enough that one can tell. We would like to know under what circumstances the minor proteins are synthesized and what their functions are. Undoubtedly, in a general sense they serve as control devices.

A most unexpected fact learned from sequence data is that DNA undergoes much more change in the course of development of an organism and in short-term evolution than anyone had anticipated. Classical genetics had established that genetic traits are very stable, changing only occasionally from one generation to another. One would expect this stability to be reflected on the molecular level. But that seems not always to be the case. It is clear that pieces of DNA move about from one part of the genome to another. Viruses may be responsible for some of this dynamism. It is certainly true that bacteria pass DNA around from one to another and indeed from one species to another. Another example of the fluidity of DNA is the difference between the DNA in those white blood cells that produce antibodies and the DNA in other cells of the same organism. In the course of becoming antibody-producing cells, they snip and splice portions of their DNA to form the code for the particular antibodies they produce.

I've mentioned only a few specific examples of what the sequence data offer to biology and, more broadly, to human thought. Hardly anything affects the way people think about their world more than detailed understanding of how living systems work according to the ordinary laws of physics and chemistry. My outlook is that mysticism about life is being crowded out by the greater joy of knowledge—thanks to molecular genetics and molecular biology in general. There is, after all, an immense difference between speculating about the way things *might* work and knowing how they *do* work. ■

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 Université de Montréal, 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 GenBank 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 GenBank 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 recoginition 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. ■