DNA Sequence Data Analysis

Steps Toward Computer Analysis of Nucleotide Sequences

Thomas R. Gingeras and Richard J. Roberts

Summary. Advances in recombinant DNA technology have allowed the isolation of large numbers of biologically interesting fragments of DNA. Concomitant improvements in methods for nucleic acid sequencing have led many investigators to characterize their clones by sequencing them. This has resulted in the accumulation of such large amounts of sequence data that computer-assisted methods, with programs directed toward the manipulation of nucleic acid sequences, have become indispensable during the collection and analysis of that data.

The search for rules to describe the relation between structures and their properties is a fundamental goal of scientific endeavor. For the molecular biologist, this means understanding the information encoded in the sequences of nucleic acid molecules, a quest requiring both the elucidation and analysis of these sequences. A problem that long thwarted these efforts was the extreme length of most DNA molecules. Fortunately, the discovery of restriction endonucleases enabled these molecules to be dissected in an orderly fashion (*1*) and, more recently, recombinant DNA techniques have facilitated the purification and characterization of individual restriction fragments from within extremely complex mixtures. So successful has this new technology been that we are now confronted with large numbers of biologically interesting pieces of DNA, and there is feverish activity to determine their sequences. Again, major technical advances have been made for the determination of these sequences. Whereas 8 years ago the longest known DNA sequence was a 20-base pyrimidine tract from the filamentous coliphage fd (*2*), now achievements are measured in kilobases or even tens of kilobases.

In the last few years it was recognized that manual methods are inadequate for the manipulation and analysis of this extraordinary amount of data and so began the alliance of computers and nucleic acid sequences. Without the extensive use of computer methods, it seems implausible that we would be able to progress much further in the collation of our sequence data, much less be in a position to analyze it adequately. It is the intent of this article to report on the developing role of computer technology in this field.

Assembling DNA Sequences

Sequencing strategy. Most sequencing projects have begun with the construction of a map of restriction enzyme sites in the DNA segment of interest. The detail deemed necessary in such a map has, to some extent, been influenced by the choice of sequencing method. By far the most used technique has been the chemical method of Maxam and Gilbert (*3*). Because the technical manipulations associated with this method are quite time-consuming, the construction of detailed restriction enzyme maps has been advantageous in allowing a sensible but limited choice of fragments to be labeled and sequenced. The alternative, the chain terminator method, developed by Sanger and his colleagues (*4*), is less demanding in terms of the technical manipulations required, but it could not easily be applied to double-stranded DNA molecules before the discovery that exonuclease III could be used to prepare templates (*5*). Given the ease of performing the sequencing reactions by this technique, the requirement for prior and extensive restriction enzyme mapping is diminished. The map can be deduced as sequencing proceeds and, if overlapping stretches of sequence result, this merely adds to the confidence with which the final sequence can be viewed.

Restriction enzyme mapping. The construction of restriction enzyme maps by conventional techniques has already been discussed (*1*) and, in general, is easy when just a few fragments need to be ordered; it becomes progressively more difficult as the number of fragments increases. Stefik (*6*) described an algorithm that uses a model-driven approach to construct restriction enzyme maps. The procedure requires only the sizes of all fragments produced by one or more restriction enzyme digestions. The program, called GA1, solves the mapping problem by inferring structures through use of an exhaustive model generator. Most of the models are eliminated by a pruning process that derives its rules from the data supplied by the user (that is, the number and size of fragments generated by each single, double, or triple enzyme digest). Many of the ideas inherent in this approach are similar to those used by the set of programs (DENDRAL) developed to predict the molecular structures of organic molecules (*7*).

Thomas R. Gingeras is a staff investigator and Richard J. Roberts is a senior staff investigator at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724.

0036-8075/80/0919-1322$01.75/0

In principle, by performing all possible combinations of digests with a few restriction enzymes, and measuring the lengths of the fragments produced, it should be possible to produce unambiguous maps by this approach. A major difficulty, which also plagues the conventional methods of mapping, lies in the accurate determination of fragment lengths. If these are known exactly, then mapping reduces to a simple problem of addition. Unfortunately, this is not the case. Existing gel systems, upon which these fragments are fractionated, allow only rough estimates of fragment lengths. Occasionally, variation in the base composition of individual fragments precludes even the relative ordering of fragments according to size.

One approach that attempts to overcome these difficulties is a computer program described by Schroeder and Blattner (*8*). It uses a least-squares method to improve the accuracy with which the size of each restriction fragment is known. From these improved estimates, restriction enzyme maps can be deduced. By an iterative procedure, the map can be refined so that the map position for each restriction site minimizes the sum of the squares of the fractional (rather than absolute) deviations of each measured fragment size from its predicted value.

It would be misleading to infer that a unique restriction map for any enzyme on any substrate can be the guaranteed product of either one or both of these programs at the present time. Rather, the programs should be viewed as providing a small number of possibilities for each map by using some well-defined mathematical principles. A few select experiments can then be performed to decide among the candidates.

The primary data. The nature of the primary data from which nucleic acid sequences are assembled is shown in Fig. 1, an autoradiograph of a sequencing gel that contains a ladder-like pattern of bands displayed in four basic channels. Each channel contains a band that corresponds to a particular base located at a particular position in the DNA sequence. In reality, the band measures the distance between some fixed point in the DNA sequence (most often the 5′ end of a restriction fragment) and a particular base in that sequence. Since the resolving power of the gel enables fragments that differ in length by one nucleotide to be clearly resolved from one another, it is possible to read the sequence directly from the gel merely by noting the position among the four channels of the next longest fragment. The development of very thin urea-containing polyacrylamide gels (*9*) has permitted the resolution of products, resulting from a single sequencing reaction, up to a chain length of 250 to 300 nucleotides per loading. The partial sequences, read from this and similar gels, are then recorded and compared in order to reconstruct the entire sequence. Two sources of trivial error are associated with this stage: one involves careless reading of the gel (for example, mistaking channels, skipping a band); the other involves errors introduced when manually recording the data.

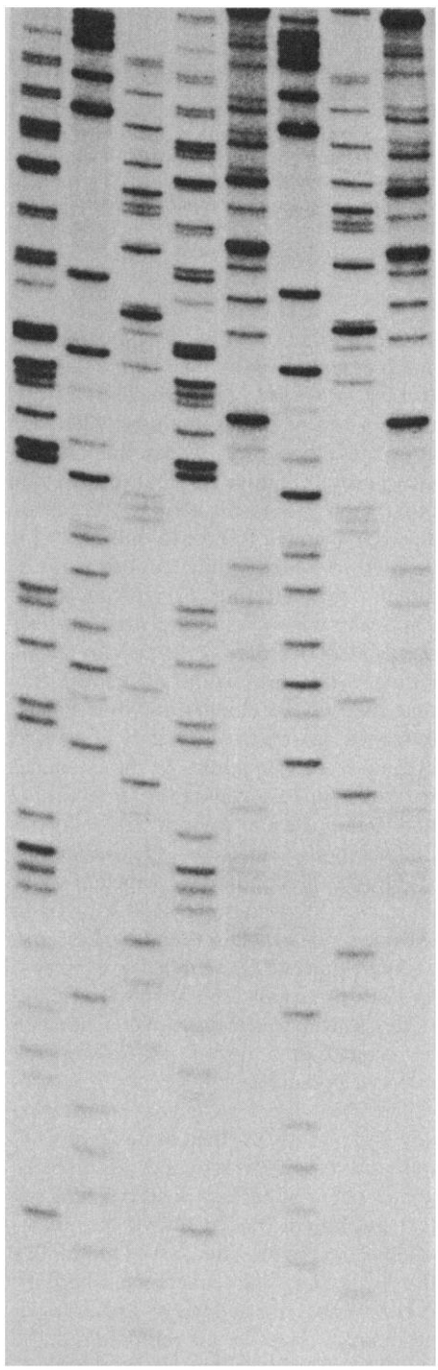

Fig. 1. An autoradiograph of a sequencing gel. The data were produced by using the chain termination procedure (*4*). A small restriction fragment (primer) was denatured and annealed to a DNA template prepared by resection with exonuclease III (*5*). The primer was then extended with the use of the Klenow fragment of DNA polymerase I in the presence of four ^{32}P-labeled deoxynucleoside triphosphates and one unlabeled dideoxynucleoside triphosphate. Incorporation of the dideoxynucleoside causes chain termination and results in a DNA fragment of unique length. One end is defined by the 5′ terminus of the priming restriction fragment, and the other is defined by the incorporation of the chain terminator. In the channels labeled A, the chain terminator was dideoxyadenosine triphosphate and each band corresponds to the location of an adenosine residue in the sequence. The sequence is read from the bottom to the top of the figure by noting the channel containing the next highest (longer) band. Duplicate channels for each base facilitate the correct ordering of the bands.

We have been attempting to overcome errors at this stage by automatically transferring data directly from the original autoradiograph into a computer. Our approach uses a digitizing tablet (Fig. 2) that allows the sequence to be read directly into the memory of the computer. The tablet operates by sending to the computer the location of any point on the surface of the pad once this point has been touched by a signal pen. The location is represented in the form of a digitized set of X and Y coordinates. The autoradiograph is placed on the pad and the channels are identified by touching each of their four corners with the pen. For each channel this defines a rectangle such that any location in the rectangle subsequently touched by the pen is automatically assigned the appropriate base. The gel is then read by touching each band with the pen and the location is recorded as the corresponding base. By repeating this process several times, the readings can be compared and discrepancies highlighted by the computer, thus allowing immediate checking of the appropriate region of the gel. Other areas of the digitizing tablet, which are not covered by the autoradiograph, can be used to send signals to the computer to invoke various useful functions. For instance, they may call an editor to correct the sequence just read, request programs to compare the newly entered sequence with other blocks already resident, or identify previous gels that contain sequences complementary to or homologous with the newly entered data.

There are several significant advantages in this approach, not the least of which is the removal of trivial errors that accompany the manual reading of auto-

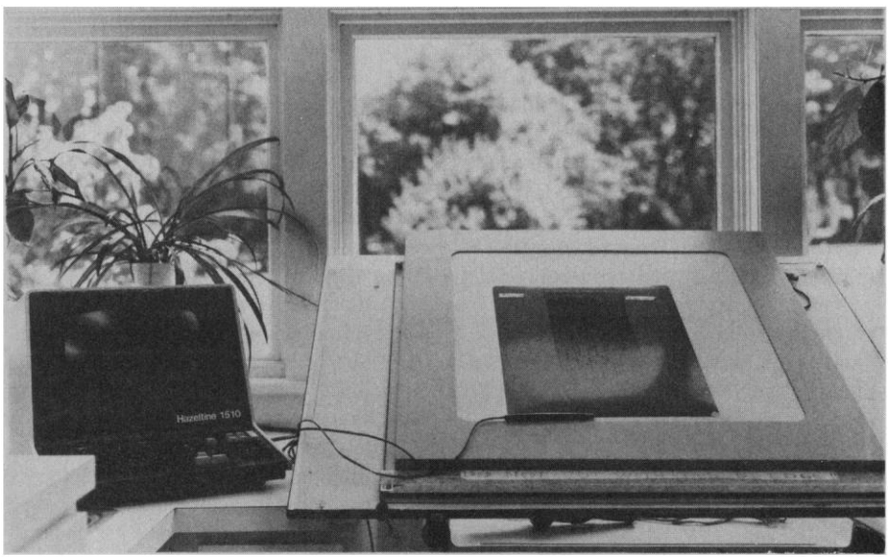

Fig. 2. A simple, semiautomated gel reading station. This station has a translucent digitizing tablet and a cathode ray tube terminal used for display. An autoradiograph is positioned on the surface of the tablet and the sequence is read from the autoradiograph (by the signal pen) directly into memory. The nucleotide sequence is displayed on the terminal screen. The limit of resolution for this tablet is 0.1 mm. A menu of other functions (editing, homology searches) is encoded on the surface of the tablet in an area not covered by the autoradiograph. These functions can be activated by touching the designated areas with the signal pen.

radiographs. One important feature, which we consider highly desirable, is that the experimenter remains in close proximity to his data. The actual reading still depends on his experience and wisdom. Those areas of the gel in which compression of bands occurs, usually attributed to secondary structure, can be noted and the unexpected results, which so often lead to new insights, may still be found. This would not be true if a completely automated gel reading system were devised. We have explored an automated gel scanning device, one designed to read two-dimensional protein gels (*10*), but for the present we view this alternative with caution because it effectively separates the investigator from the examination of his primary data.

Reconstruction of large sequences. The primary data from a set of sequencing reactions consist of a number of short sequence stretches that may be up to 250 nucleotides in length. The next stage is to order these short segments to find overlapping stretches of common nucleotides and, finally, to join them together into one long block of sequence that represents the primary structure of the original molecule.

Two sets of computer programs (*11*, *12*) have been written to aid in this process. These programs provide three essential functions. The first is to direct each new piece of sequence data into a master archive, which can be used for future reference. The second function allows the identification of those blocks of sequence that contain homologous or complementary stretches—the overlaps. In both programs, the stringency of the overlap can be set by the user. The program developed by Staden (*11*) has a very useful feature in that characters other than the usual adenine (A), cytosine (C), guanine (G), or thymine (T) can be used to represent nucleotides that are difficult to determine. For instance, occasionally it is difficult to distinguish real bands from artifacts. In the chain termination procedure, some structural feature may cause elongation to stop prematurely, giving rise to bands in several channels. Alternatively, in the chemical method, the discrimination between C and T may be too poor to allow for an unambiguous decision. These ambiguities can be denoted by a special set of characters (for example, R is A or G) in the sequence and can then be taken into account during the search for overlaps. In both programs, the strings of nucleotides sharing overlapping sequence can be printed in a format that aligns the areas of homology (Fig. 3).

The third function is that of melding any two strings of nucleotides that contain overlapping sequence. If discrepancies occur at certain positions, diacritical marks are placed above the nucleotides concerned. This serves to identify those positions in a sequence where either new data are needed or some reevaluation of the old data is required. As the sequencing project continues, the original stretches of short sequence gradually become melded into longer and longer blocks until the reconstruction of the original molecule is accomplished.

Checking the sequence. A most important step in nucleotide sequence determination comes when a complete sequence has been assembled and its accuracy must be assessed. Since DNA is double stranded, sequence can be obtained independently from both strands and thus, even during the assembly of the sequence, its accuracy can be continually checked. If a discrepancy appears in the sequence of either strand, it is, of course, unclear which strand is incorrect. Such discrepancies can result from a number of causes, some technical, in which case the experiment may have to be repeated, and some systematic. For example, certain stretches of sequence are able to form very strong secondary structures that result in anomalous mobilities (compressions) during gel electrophoresis. The exact positions and extent of these compressions often vary from one strand to another such that perfectly good sequence may be read from one strand while the other is indecipherable. Alternatively, the relative location of available restriction sites may preclude sequence data from being gathered from both strands; in this case the sequence assignment must be based, for short regions, on only one strand of data.

In any event, there is an alternative way to assess the accuracy of the sequence by using restriction enzymes. First, restriction enzyme maps deduced before sequencing should be confirmed by the final sequence. Second, other restriction enzyme sites used for fragment preparation during the sequencing operation should also be confirmed. Finally, and perhaps most importantly, with a complete sequence in hand, it is possible to predict the fragmentation patterns that should be observed upon digestion of the original molecule with the known restriction enzymes. These patterns can be generated experimentally very rapidly and can serve to randomly check the integrity of the sequence. Computer programs that scan the sequence and identify the location of restriction enzymes sites abound and represent one of the more straightforward and practical uses of the computer in nucleic acid sequence analysis.

Analysis of Nucleic Acid Sequences

Once a large segment of accurate sequence is available, it is usually desirable and necessary to identify those features of the sequence that determine its biolog-

```
OVERLAPPING SEGMENTS C-1(09/18/79/)                AND   C-44B(10/08/79)
   7
   1    42    44
  48    88    21
  70   109     6
  77   115     5
  84   124    11
  96   135    16
 112   152     8
*                                                                      *
* <                                                    >GATTAAG       *
* <ATTAATGAC CAGACACGTC CTGAGTGTGT TACTTTTCAG CA>GATTAAG              *
*                    *                                               *
* GATAATTGCG CTAATGANCT TGATCTGCTG GCGCAGA<AG T>ATTCCATA              *
* GATAATTGCG CTAATGAGCT TGATCTGCTG GCGCAGA<TG  >ATTCCATA              *
*                                                                     *
* GAGCAGCTGA CCA<C>TTAC TG<G>CTGCA <GC  >GGGG ATGATTT<T>              *
* GAGCAGCTGA CCA< >TTAC TG< >CTGCA <AGCA>GGGG ATGATTT< >              *
*                                                                     *
* GAGGAGGCTA TTAGGG< >T ATATGCA<AA GGTGCATTAG GCAGATTGCA              *
* GAGGAGGCTA TTAGGG<A>T ATATGCA<                                      *
*
* GTACACAT>
*         >
```

Fig. 3. A sample printout of the results generated by the nucleotide sequence assembly program called ASSEMBLER (*12*). A table detailing the positions of the overlapping sequences is presented above the printed sequences. In this example, this table indicates that there are seven regions of homology between the two sequences (C-1 and C-44B) that are being compared. These overlapping sequences begin with nucleotide 1 of C-1 and nucleotide 42 of C-44B. This match lasts for a length of 44 bases when it is interrupted by a mismatch of three nucleotides at position 45 of C-1. Such areas of mismatch are denoted in the aligned sequences printed below this table by a set of brackets. The rest of the overlapping segments between C-1 and C-44B are summarized by the remaining lines of the table. An asterisk is placed above the sequence whenever an N (an unidentified nucleotide) occurs in either or both sequence elements.

ical properties. Such analysis falls into two discrete stages: The first involves the straightforward search for sequences with known properties, such as the identification of start points and stop points for transcription or translation and perhaps, also, RNA processing sites. Usually, additional experimental data are needed to accurately determine these positions. The second stage, which is both more interesting and more challenging, involves attempts to detect subtle and less straightforward sequence patterns. These patterns may reveal controlling elements such as promoters for RNA transcription, or previously unsuspected properties, for instance, additional RNA species or protein products. These predictions can then be tested experimentally. The final result will be a catalog of sequence patterns that are related to function and that can then be compared to other such catalogs. Perhaps common features will be apparent and new hypotheses will emerge and become the basis for yet more experimentation. As the total amount of sequence data from many sources grows, it can serve as an experimental resource, both for deriving new hypotheses and also for testing them. Computer programs have aided in both stages of this analytical process.

Straightforward analysis. Virtually all aspects of sequence analysis begin with simple search routines to locate specific patterns of interest. Algorithms for this purpose are easily devised and many laboratories have developed their own. The end product is a list noting the locations at which the pattern occurs, and output may vary from a simple table to a detailed graphic display. Two fairly comprehensive collections of programs, which include such cataloging features, have been described. One of these, developed by Korn and Queen (*13*), includes such features as printing the sequence in a format that indicates the occurrences of various patterns, either from a standard list or as directed by the user, and translating all or part of the sequence into its polypeptide equivalent. One useful feature of these search programs is that, although perfect matches with the input sequence are always found, imperfect matches may also be located at the user's discretion. Furthermore, the probability of finding any given sequence by chance is assessed by the computer, thus giving a preliminary indication of the possible significance of nonrandom distribution of particular patterns. This last portion of the program has been significantly modified by others (*14*) so as to reflect the base composition of the input sequence.

During the sequencing of ϕX174, the first set of computer programs designed to assist a DNA sequencing project was written (in ANS COBOL) by McCallum and Smith (*15*). These were followed by an extensive collection (in FORTRAN) (*16*) containing many of the basic features alluded to above. These programs provide useful facilities for storing, editing, and manipulating long strings of sequence, and they use common subroutines, each of which represents a basic component in the analytical process. This feature is valuable because many of the routines can be used as the basis for more sophisticated programs. Since these programs are written in FORTRAN, they are easily transported from one machine to another and can be readily understood and modified by anyone with a little experience in FORTRAN. We have found this latter aspect most helpful and have often taken advantage of various subroutines in our own programs. We would not wish to give the impression that FORTRAN is a language ideally suited for the analysis of nucleic acid sequences. There are other languages, such as SNOBAL (used for character string manipulation) or Pascal (noted for its high degree of structure), that may be more suited for this type of analysis. However, FORTRAN has been most widely used thus far because it is a language with which most people are familiar. As more people become interested in sequence analysis and use their own machines for this purpose, it is likely that new algorithms will be developed, and many will wish to incorporate them into their programming reservoir. Such analytical programs should be viewed as another tool for molecular biology and, as with all tools, their communication and improvement are an important consideration.

Less straightforward analysis. A second aspect to the analysis of nucleic acid sequences arises when simple search and catalog functions become intertwined with other routines. Often, all possible occurrences of a specific type of pattern will need to be examined. Some rules will be applied, and the particular subset of occurrences that obeys the rules will be selected. Further pruning may then occur until one, or a few, possibilities remain. These will then be used either to test the validity of the rules or to provide some new hypothesis, or perhaps both. A good example of this is found in the computer programs written to generate secondary structure models for RNA transcripts (*17–20*). By using the well-known base pairing rules for polynucleotides and measurements of the free energy contributions of these paired structures (*21*), programs have been written that will construct models to describe the possible configurations of a variety of single-stranded RNA molecules. Such models can then form a basis for hypotheses relating secondary structure to biological properties and so lead to new lines of experimentation.

With the exception of a few specific in-

Table 1. Computer programs used during nucleic acid sequencing.

Function	Program name*	Program language	Reference
	Presequencing preparations		
Restriction enzyme mapping	"GA1"†	SAIL	(*6*)
	Least-squares method for restriction mapping‡	FORTRAN	(*8*)
Reverse translation	"REVTRANS"§	FORTRAN	
	Nucleic acid sequence analysis‖	PL/1	(*13*)
	Collection and assembly of sequences		
Semi-automated autoradiograph reading	"READ"¶	FORTRAN	
Automated autoradiograph reading	Scanning program**		(*30*)
Sequence assembly	"ASSEMBLER"††	FORTRAN	(*12*)
	Overlap-meld‡‡	FORTRAN	(*11*)
	Analysis of nucleotide sequences		
Printing, editing, storage, and manipulation	Nucleic acid sequence analysis‖	PL/1 and SAIL†††	(*13*)
Search routines (restriction enzyme sites, direct repeats, true and dyad symmetries)	DNA-handling program‡‡	FORTRAN	(*16*)
Translation			
Restriction enzyme recognition site predictions	"MONITOR"††, ***	FORTRAN	(*25*)
Transfer RNA gene prediction	"tRNA"‡‡	FORTRAN	(*28*)
Secondary structure prediction	Secondary structure program§§	FORTRAN	(*18*)
	Secondary structure program ‖ ‖	APL, FORTRAN	(*17, 19*)
Tertiary structure modeling	3D Molecular modeling¶¶	FORTRAN and BLISS	(*40*)

*Listing in quotation marks is a program title; the other names are brief descriptions. †Sumex System, Stanford University, P. Friedland, D. Brutlag, L. Kedes. ‡University of Wisconsin, J. L. Schroeder, F. Blattner. §Cold Spring Harbor Laboratory, R. Blumenthal, R. J. Roberts (unpublished). ‖National Institutes of Health, C. Queen, L. Korn. ¶Cold Spring Harbor Laboratory, T. R. Gingeras, P. Rice, R. J. Roberts (unpublished). **European Molecular Biology Laboratory, Heidelberg, Germany, S. Provencher, R. Vogel, V. Dovi, H. Lehrach. ††Cold Spring Harbor Laboratory, T. R. Gingeras, J. Milazzo, R. J. Roberts. ‡‡MRC Laboratory of Molecular Biology, Cambridge, England, R. Staden. §§Syracuse University, G. Pavlakis, J. Vournakis. ‖ ‖University of California at Los Angeles, G. M. Studnick G. M. Rahn, I. W. Cummings, W. A. Salser. ¶¶National Institutes of Health, R. J. Feldmann. ***University of Wisconsin, C. Fuchs, E. C. Rosenvold, A. Honigman, W. Szybalski. †††This version is available from the Sumex System at Stanford University.

stances—the correlation between the computer-generated secondary structure model for transfer RNA (tRNA), the x-ray crystallographic results (*20*), and the use of S1 and T1 nucleases as probes for the presence of single-stranded regions predicted from computer-generated secondary structure models (*22*)—rather little has been done to demonstrate the actual existence of many of the secondary structure models proposed in the literature. Moreover, the rules for assigning free energy contributions for RNA secondary structures are, at best, approximations, and it is clear that the predictions generated by the existing computer programs must be viewed with great caution. The tRNA cloverleaf, first proposed by Holley *et al.* (*23*), does seem to have withstood the test of time. However, the fascinating three-dimensional structure revealed by x-ray crystallography (*24*) most surely points to the importance of structural considerations other than simple base pairing, and it would be quite beyond our present capabilities to accurately predict such a structure when only the primary sequence is known. More frustrating is the dearth of experimental evidence allowing an accurate description of secondary structure in solution, let alone evidence implicating such structure in a biological function. If such experimental evidence could prove that a certain secondary structure feature exists in a long RNA sequence, it should be possible to derive the rules that allow this structure to form rather than others. The greatest hope for the application of computer programs in this area may lie in designing algorithms aimed at improving the rules that predict secondary structures.

Restriction enzyme recognition sites. We and others (*25*) have attempted to use the known sequences of viral genomes to predict the recognition sequences for new restriction endonucleases. This is achieved by cleaving DNA's of known sequence with the new restriction endonuclease and measuring the length of the fragments produced. The computer is then asked to produce a set of theoretical fragmentation patterns for that DNA sequence by using all possible tetranucleotide, pentanucleotide, and hexanucleotide sequence combinations. Many of these combinations of nucleotides already define the recognition sequences of existing restriction enzymes (*26*). By comparing the observed pattern with these theoretical patterns and eliminating those that differ significantly, one is usually left with a small number of potential recognition sequences. Ordinarily, the larger the number of fragments produced by the enzyme, the greater the likelihood that this program will arrive at a unique candidate sequence. The predicted recognition sequence can then be tested experimentally either by cleaving another substrate of known sequence or perhaps by mapping one or more of the cleavage sites in the known sequence. The value of this approach is that new and potentially useful recognition sites can be identified quickly, and often simple experiments can be devised to prove (or disprove) the predicted sequence.

The tRNA genes. The sequences of many tRNA molecules are known (*27*) and show certain characteristic features. In particular, they share a similar secondary structure (the cloverleaf) and a constant number of bases in particular portions of the molecules. Staden (*28*) has devised a program that can identify putative tRNA genes in a long DNA sequence by searching for stretches of sequence displaying these common features. This has been applied to stretches of DNA sequence from the human mitochondrial genome as determined in Sanger's laboratory. It is known that this genome encodes many tRNA genes, and two of these have been located by the use of this program (*29*). It is rather interesting to note that these mitochondrial tRNA genes differ in several significant respects, although they also bear some structural resemblances to other tRNA genes. These differences might have precluded their detection had the set of rules used by the program been overly stringent.

It will be clear from the foregoing discussion that although computers have now become an integral part of nucleic acid sequence analysis, only limited use has been made, thus far, of their analyti-

cal abilities. A summary of programs currently available and known to us either through publication or personal communication is presented in Table 1. Several obvious gaps exist; however, some of the missing programs may already exist in unpublished form. Many groups who have written programs for their own use believe that these programs are too trivial to warrant publication. This is unfortunate since it probably means that there will be much duplication of effort, and many useful algorithms will enjoy only limited circulation. More disturbing is the realization that many of the algorithms, which have been freshly developed for the analysis of DNA sequences, are ones used commonly in other disciplines (for example, pattern recognition).

Future directions. In recent years there have been dramatic improvements in optical scanning devices, and it would seem obvious that this technology could be applied to the automated reading of sequencing gels. Such development, both in scanning devices and computer software, is currently proceeding (*30*). Although this would certainly ease the burden on the investigator, it does carry inherent disadvantages by separating the experimenter from his data. Nevertheless, as larger and larger sequencing projects are attempted, it seems likely that automated methods will become a necessity, if only to relieve the boredom associated with the collection of sequence data.

Within this same context, there is a great need for more flexible programs to analyze, sort, and store the primary data during the assembly of DNA sequences. Quite aside from the sequence data itself, there is a wealth of additional information that can help in its ordering and can be used to direct further experiments. This is particularly true for the chain termination method where the size of the primer fragment, the nature of the restriction enzyme that produced it, and the strandedness of the sequence are all known. At present this information may be used manually, but more often it is discarded or considered only after the complete sequence has been deduced. It seems likely that the computer can be used effectively in the initial stages to analyze and correlate a great deal of random sequence information and then indicate those experiments that can most profitably be performed to fill remaining gaps. The increasing need for this type of program is well illustrated by a recent development in sequencing technology that uses the M13 cloning system to obtain primary sequence data by a "shotgun" approach (*31*).

Undoubtedly, the greatest scope for computer-assisted method lies in the actual analysis of complete sequences. At present we know few of the rules that dictate the biological activity of nucleotide sequence. Indeed, given an unknown piece of DNA sequence, we would be hard pressed to predict whether it came from a eukaryotic or prokaryotic source, much less the RNA molecules transcribed from it or the polypeptides it might encode. Until recently, this type of problem was rarely encountered because sequence determination was sufficiently difficult that it was usually undertaken only after the particular RNA or protein encoded by that sequence was known, and details of the DNA sequence were required to provide a structural basis for its expression. That situation has now changed. Although most sequence projects begin with some prior knowledge of properties, there has been an increasing awareness that surrounding (and intervening) sequences are also important. Frequently a sequence of several kilobases will be determined, and only a small fraction can be immediately associated with function. Thus there is great need for further programs to assist in the analysis of these sequences. Such programs could make limited predictions and hence suggest further experiments to test the validity of those predictions.

One area in which we are actively engaged concerns the phenomenon of RNA splicing in eukaryotes. From the existing data, the only common features that occur at all splicing sites are the presence of a GU (U, uracil) dinucleotide at the 5′ end of the intervening sequence and an AG dinucleotide at the 3′ end of that sequence (*32*). It is clear that other information, be it primary sequence or some structural feature dependent upon that primary sequence, is also necessary since not all GU, AG combinations are joined. We are approaching this problem by using the computer to generate potential messenger RNA's (mRNA's) from a DNA sequence by making all possible pairwise combinations of GT and AG. The task then becomes one of finding a rule or rules to apply to distinguish correct splicing events from incorrect ones. The program is highly interactive and has two distinct aspects. On the one hand, if the actual splice points are unknown at the nucleotide level, but known approximately by electron microscopic measurements, then only a subset of all possible mRNA's are generated on the basis of the approximate positions known to be involved. Each mRNA is then translated and the molecular weights of the predicted polypeptides may be used to further limit the search if, for instance, the actual size of the protein product is known. Yet further restrictions may be placed by requiring that certain sequences be present or absent from the final mRNA or that certain tryptic peptides be present. The second part of the program is still under development and is used when the actual splice points in the sequence are known. At this point, the task is to find sequence elements that provide a unique environment for the two ends of the splice junction and would thus allow discrimination between the splice points and the rest of the sequence. The essence of the idea is to remove from consideration any features that occur both at the splice point and elsewhere in the sequence.

A more challenging prospect concerns the fact that both DNA and RNA are three-dimensional molecules. By depicting a DNA sequence as a featureless one-dimensional string, we often forget that this is a naïve perspective when seeking to explain its properties. The finding that $(dG\text{-}dC)_3$ (dG, deoxyguanylate; dC, deoxycytidylate) crystallizes with a left-handed helical conformation (*33*) provides an extreme example of the fact that the DNA double helix is not perfectly smooth and regular. Much experimental evidence exists to show that local distortions will result from the influence of primary sequence (*34, 35*), and these may be the very elements by which other macromolecules interact with DNA. It is already apparent that structures recognized by RNA polymerases can be defined by several quite different sequences (*36*). Undoubtedly, these can eventually be cataloged, but we must be aware that the "boxes" postulated by Pribnow or Hogness (*37*) may reflect our desire for simple rules. The experimental data concerning recognition by RNA polymerase (*36*) demonstrate quite clearly the importance of a three-dimensional approach and confirm ideas developed previously (*34, 38*). In a similar vein, the computer programs developed by Trifonov (*39*) address the question of DNA folding around the nucleosome and raise the possibility that patterns existing in the primary sequence have been designed to facilitate such folding.

The kinds of three-dimensional structures that can be formed by DNA molecules are severely constrained by the loss of flexibility associated with its double-stranded nature. This is not true

for single-stranded RNA molecules which, a priori, have many more structural possibilities because of their inherent flexibility. We are slowly becoming accustomed to the two-dimensional representations of RNA, although the actual structures taken up by these planar stems and loops are difficult to visualize. The need for better algorithms to predict secondary structures was mentioned above. It would also seem sensible to investigate the rules for additional interactions in light of the x-ray crystallographic results for tRNA (*24*).

Computer programs that examine the three-dimensional properties of macromolecules have been devised by Feldmann (*40*). However, at present, they rely upon x-ray crystallographic data to provide the basic parameters and use computer graphics to depict and transform the structures. Clearly, we are some way from being able to define the rules that relate primary sequence to three-dimensional structure; much further experimentation is necessary. However, if we are ever to understand the nature of the interactions that control the behavior of macromolecules it will be essential to expand both our experiments and our theories into this third dimension. Computer graphics is likely to prove an important tool in these endeavors.

Conclusions

Only a casual look at the literature is required to appreciate the growth of interest in DNA, RNA, and protein sequences. There is no indication that this interest is about to subside. Rather, it seems to be gaining momentum (*41*). This poses severe problems for those who wish to analyze the data since few, if any, of us have the capacity to absorb the subtle features inherent in each individual sequence. It is inevitable, therefore, that we should call upon computers to assist in this task. The programs developed so far have proved useful and valuable in analyzing individual sequences, but they represent only a first, hesitant step toward the more complex goal of correlating sequence with tertiary structure and eventually with function.

References and Notes

1. M. Zabeau and R. J. Roberts, in *Molecular Genetics*, J. H. Taylor, Ed. (Academic Press, New York, 1979), vol. 3, p. 1.
2. V. Ling, *J. Mol. Biol.* **64**, 87 (1972); *Proc. Natl. Acad. Sci. U.S.A.* **69**, 742 (1972).
3. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977); *Methods Enzymol.* **65**, 499 (1980).
4. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
5. A. J. H. Smith, *Nucleic Acids Res.* **6**, 831 (1979).
6. M. Stefik, *Artif. Intell.* **11**, 85 (1978).
7. B. G. Buchanan and E. A. Feigenbaum, *ibid.*, p. 5.
8. J. L. Schroeder and F. Blattner, *Gene* **4**, 167 (1978).
9. F. Sanger and A. R. Coulson, *FEBS Lett.* **87**, 107 (1978).
10. J. I. Garrells, *J. Biol. Chem.* **254**, 7961 (1979); J. Taylor, N. L. Anderson, B. P. Coulter, A. E. Scandira, N. G. Anderson, in *Electrophoresis "79,"* B. J. Radola, Ed. (Gruyter, Berlin, 1980).
11. R. Staden, *Nucleic Acids Res.* **6**, 2601 (1979).
12. T. R. Gingeras, J. P. Milazzo, D. Sciaky, R. J. Roberts, *ibid.* **7**, 529 (1979).
13. J. L. Korn, C. L. Queen, M. N. Wegman, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4401 (1977); C. L. Queen and L. J. Korn, *Methods Enzymol.* **65**, 595 (1980).
14. D. Brutlag, personal communication.
15. D. McCallum and M. Smith, *J. Mol. Biol.* **116**, 29 (1977).
16. R. Staden, *Nucleic Acids Res.* **4**, 4037 (1977); *ibid.* **5**, 1013 (1978).
17. W. M. Fitch, *J. Mol. Evol.* **1**, 185 (1972); J. M. Pipas and J. E. McMahon, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 2017 (1975); G. M. Studnicka, G. M. Rahn, I. W. Cummings, W. A. Salser, *Nucleic Acids Res.* **5**, 3365 (1978).
18. G. Pavlakis and J. Vournakis, personal communication.
19. R. Nussinov and G. Pieczenik, personal communication; L. Garber, G. Garber, R. Nussinov, G. Pieczenik, personal communication; R. Nussinov, G. Pieczenik, J. Griggs, D. Kleitman, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.* **35**, 68 (1978).
20. M. Philipp, D. Ballinger, H. Seliger, *Naturwissenschaften* **65**, 388 (1978).
21. J. Gralla and D. M. Crothers, *J. Mol. Biol.* **73**, 497 (1973); I. Tinoco, O. C. Uhlenbeck, M. Levine, *Nature (London)* **230**, 362 (1971); I. Tinoco, P. W. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, J. Gralla, *Nature (London) New Biol.* **246**, 40 (1973).
22. G. Pavlakis, R. E. Lockhard, N. Vamvakopoulos, L. Rieser, U. L. Rajbhandary, J. N. Vournakis, *Cell* **19**, 91 (1980).
23. R. W. Holley *et al.*, *Science* **147**, 1462 (1965).
24. J. D. Robertus, J. E. Ladner, J. T. Finch, D. Rhodes, R. S. Brown, B. F. C. Clark, A. Klug, *Nature (London)* **250**, 546 (1974); S. H. Kim *et al.*, *Science* **185**, 435 (1974).
25. T. R. Gingeras, J. P. Milazzo, R. J. Roberts, *Nucleic Acids Res.* **5**, 4105 (1978); C. Fuchs, E. C. Rosenvold, A. Honigman, W. Szybalski, *Gene* **4**, 1 (1978).
26. R. J. Roberts, *Nucleic Acids Res.* **8**, r63 (1980).
27. M. Sprinzel, F. Grueter, A. Spelzhaus, D. H. Gauss, *ibid.*, p. r1.
28. R. Staden, *ibid.*, p. 817.
29. B. G. Barrell, A. T. Bankier, J. Drouin, *Nature (London)* **282**, 189 (1979).
30. S. Provencher, R. Vogel, V. Dovi, H. Lehrach, personal communication.
31. P. H. Schreier and R. Cortese, *J. Mol. Biol.* **129**, 169 (1979); S. Anderson, M. J. Gait, L. Mayol, I. G. Young, *Nucleic Acids Res.* **8**, 1731 (1980); F. Sanger, A. R. Coulson, B. G. Barrell, A. J. H. Smith, B. A. Roe, in preparation.
32. R. Breathnach, J. D. Mandel, P. Chambon, *Nature (London)* **270**, 314 (1977).
33. A. H. J. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, A. Rich, *ibid.* **282**, 680 (1979).
34. R. D. Wells *et al.*, *CRC Crit. Rev. Biochem.* **4**, 305 (1977).
35. R. T. Simpson and P. Kunzler, *Nucleic Acids Res.* **6**, 1387 (1979); H. Shindo and S. B. Zimmerman, *Nature (London)* **283**, 690 (1980); M. A. Viswamitra, O. Kennard, P. G. Jones, G. N. Sheldrick, S. Salisbury, L. Falvello, Z. Shakked, *Nature (London)* **273**, 687 (1978).
36. U. Siebenlist, R. B. Simpson, W. Gilbert, *Cell* **20**, 269 (1980).
37. D. Pribnow, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 784 (1975); M. Goldberg, thesis, Stanford University (1979).
38. M. C. O'Neill, *Nucleic Acids Res.* **4**, 4439 (1977).
39. E. Trifonov, personal communication.
40. R. J. Feldmann, D. H. Bing, B. C. Furie, B. Furie, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 5409 (1978); M. Bina, R. J. Feldmann, R. G. Deeley, *ibid.* **77**, 1278 (1980).
41. At a meeting organized by NIH and held in Washington, D.C., on 14 and 15 July 1980, the question of organizing a central data bank for nucleic acid sequences was discussed. It was decided that until such time as a formal bank can be established, an interim bank of published (or in press) sequences could be started. Initially, the sequences would be gathered from individual collections already on computer tape and would be distributed through the MOLGEN group over the SUMEX network. Anyone who has such a collection of sequences is urged to send a copy to Dr. Elke Jordan, Genetics Program, National Institute of General Medical Science, Bethesda, Maryland 20205. At a similar meeting at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, 23 to 25 April 1980, it was also decided to establish a central data bank at the EMBL.

42. The authors thank J. Milazzo, who introduced them to the potentialities provided by the computer and wrote the code for several programs mentioned in this article, R. Blumenthal, J. Brooks, and G. Albrecht-Buehler for suggestions and critical readings of this manuscript, and M. Moschitta for help in preparing this manuscript. Supported by National Cancer Institute grant CA13106 and NIH grant 1RO1-CA27275-01.

20 June 1980