Physica 10D (1984) 157–167
North-Holland, Amsterdam

TOWARDS MODELING DNA SEQUENCES AS AUTOMATA

Christian BURKS
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, University of California, Los Alamos, NM 87545, USA

and

Doyne FARMER
Center for Nonlinear Studies, Los Alamos National Laboratory, University of California, Los Alamos, NM 87545, USA

We seek to describe a starting point for modeling the evolution and role of DNA sequences within the framework of cellular automata by discussing the current understanding of genetic information storage in DNA sequences. This includes alternately viewing the role of DNA in living organisms as a simple scheme and as a complex scheme; a brief review of strategies for identifying and classifying patterns in DNA sequences; and finally, notes towards establishing DNA-like automata models, including a discussion of the extent of experimentally determined DNA sequence data present in the database at Los Alamos.

1. Introduction

The discovery of the DNA molecule has revolutionized our way of thinking about biological evolution. Indeed, it is one of the best examples of the power of reductionism. However, for complex non-linear systems such as DNA, there are limits to the reductionist approach: Knowledge of the way in which individual parts fit together eventually fails to provide an adequate understanding of the system as a whole. These limits are currently evident in that, although over two million base pairs have been sequenced, the rate of interpretation of these data is lagging behind the rate of acquisition. Even were we able to completely sequence a cell's DNA and understand all the details of the attendant biochemistry, it is not clear that we would automatically emerge with an understanding of the global software design principles underlying the functioning of the DNA molecule.

The problem of understanding DNA might be likened to understanding the functioning of a computer, for someone who has never seen one before. We now understand the basic principles of the underlying hardware, and through our knowledge of the genetic code, the basics of how the program is run. We are perhaps close to understanding the "microcode" of this program, but the basic flow chart is still a mystery. Eventually we might hope to construct a digital model of the DNA molecule and its attendant hardware, and thus actually run a simulated "genetic program". The level of knowledge and the computation power that are prerequisites for this approach are likely to be some time in coming, however.

Another quite different approach to this problem, exemplified by the work of Holland [1] and Kauffman [2], is to study the generic properties of models whose properties and size are simpler and smaller than those of the full DNA molecule. If universal patterns emerge from these simplified models, we can hope that they will also be present in more complicated systems. Given the design constraints dictated by organic chemistry, there may be a limited number of different solutions to a given problem. Furthermore, the patterns of these solutions may not be specific to the details of organic chemistry, but rather only to the broad outlines. Exploring simple qualitative models based on DNA may give us insight and help illuminate the best methods for the higher level interpretation of DNA sequences. The converse is also true; as we learn more about DNA, we may hope that this knowledge will lend insight into good designs for adaptive automata, which in turn have many potential practical applications for artificial intelligence.

This situation is somewhat analogous to the problem of fluid turbulence. Although the equations for fluid flow have been known for more than 100 years, these equations by themselves tell us very little about the nature of turbulence. That is, although the equations are easily written down, exact solutions are not possible for turbulent flows. Furthermore, working directly from the equations, even the correct qualitative behavior has not been forthcoming. Simulating these equations in detail remains out of the reach of current computer technology. In recent years, however, by studying very simple nonlinear models, significant progress has been made toward understanding the qualitative nature of turbulence, especially near the transition. This progress was made by studying a variety of simple models until patterns in their behavior emerged, and then searching for these same patterns in experimental data generated by fluid flows. Conversely, studies of the behavior of fluid flows with nonlinear dynamics in mind have stimulated and guided a good deal of work in nonlinear dynamics.

The DNA molecule can be viewed as a one-dimensional lattice, with four states per lattice site, that is capable of producing copies of itself. This immediately suggests that automata might prove to be a valuable tool for the modeling of DNA. The view stated above, however, is at best a first approximation. The DNA molecule is more than just a linear string of base pairs, and has a variety of other mechanisms for information storage. Furthermore, for most purposes, the DNA molecule cannot be modeled in isolation; it is necessary to also take into account the functioning of the attendant biochemical machinery. If an approach such as that described above for fluid turbulence is to be successful in the study of DNA, we must determine the right class of models to examine, and the properties of these models must be firmly based on the properties of real DNA molecules. Those who wish to explore simple models of DNA should be properly apprised of recent developments concerning the higher level structures present in DNA before they begin their explorations.

This paper is modest in intent. With the purpose stated above in mind, we present a brief review of the dynamic properties of the DNA molecule, discussed in the context of automata. We have two goals in doing this: First, for non-biologists, we hope to provide an easily digestible introduction to the DNA molecule, written in non-technical terms; secondly, for both biologists and non-biologists, by including a simultaneous translation of properties of DNA into the language of automata theory, we hope to suggest strategies for modeling DNA in terms of automata. By sketching a few directions that the modeling of DNA might take, we hope to stimulate further research in this direction.

We begin by briefly reviewing the so-called Central Dogma of molecular biology. We then present a series of amendments to and extensions of the Central Dogma demonstrating that the mechanisms for storing information in the patterns of DNA sequences can be quite complex, and should be included as modifications of any simple model of the evolution of DNA sequences. We conclude with some remarks concerning approaches to modeling, including a summary of the type of information collected in the nucleic acid sequence database at Los Alamos, and a review of current approaches to searching for and comparing patterns in DNA sequences.

Wolfram [3–5] provides an introduction to cellular automata; for an additional review of biological applications in particular, see Ransom [6]. Presentations of molecular biology not otherwise referenced in this article are drawn from several texts that would provide useful starting points for the reader interested in background material; they are, in order of increasing detail, Crick [7], Watson [8], Lewin [9–12] and Nover et al. [13]. Additional citations are more often recent reviews than primary sources.

2. The Central Dogma

The building blocks of the DNA polymer are nucleotides, which in turn consist of a phosphate group, a sugar ring group and either a purine or a pyrimidine base group (see fig. 1). The two possible purines are guanine (G) and adenine (A); the two possible pyrimidines are thymine (T) and cytosine (C). The backbone of the DNA strand is formed by covalent bonds connecting alternating sugars and phosphates. The patterns present on a DNA strand are then the sequential arrangements of A, C, G and T along the strand. DNA isolated from cells is found to be an antiparallel double-stranded helix (see fig. 2), with the alignment of the two strands mediated through hydrogen bonding between a purine or a pyrimidine on one strand and a pyrimidine or a purine on the other. Furthermore, this base pairing is quite specific: A is always paired with T, and G is always paired with C. Thus, the base sequence one one strand completely determines the base sequence on the other, complementary strand, with the exception of transient mispairings.

The Central Dogma (see fig. 3) is a description of the way in which a DNA sequence specifies both its own regeneration (replication) and the synthesis of proteins (transcription and translation). Replication is accomplished with high fidelity by virtue of the base pairing complementarity of the DNA double strand: the cellular DNA-synthesizing machinery reads each strand to form its complementary strand. The reading of DNA to synthesize proteins is acomplished in two steps. In the first step, transcription, a messenger ribonucleic acid strand (mRNA)* is formed, again, as a complement to one of the DNA strands. In the second step, translation, the cellular protein-synthesizing machinery reads a section of the mRNA strand to form a protein strand. This mRNA base sequence is read as non-overlapping contiguous triplets, call codons. Proteins are composed of amino acids: there are 20 possible amino acids and 64 possible

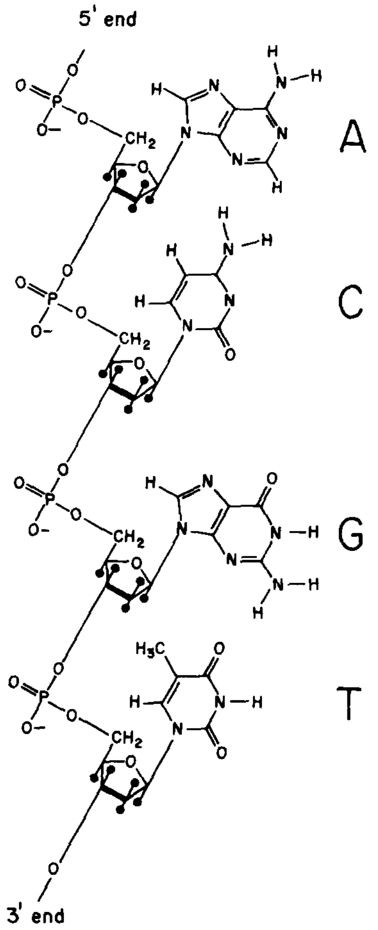

Fig. 1. *The DNA Polymer.* Each monomer consists of a phosphate group, a sugar ring group and one of the following bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The polarity of the polymer is determined by the arrangement of phosphate links with the 5' and 3' carbons of the sugar ring. The sequence of bases is read from the 5' end to the 3' end; this sequence would be read as ACGT. This figure is drawn from fig. 3.10 in Watson [8].

* RNA is similar to DNA, except that it is usually single stranded, the sugar has one less hydroxyl group, and uracil (U) is substituted for thymine (T). There are several different functional types of RNA, such as messenger, transfer (tRNA) and ribosomal (rRNA).

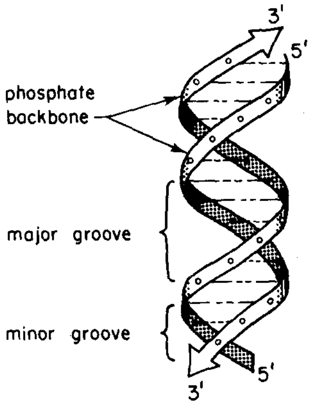

Fig. 2. *The Double Helix.* The solid horizontal lines represent bases, the horizontal dotted lines represent hydrogen bonds, and the helical bands represent the sugar-phosphate backbone of DNA. This is the antiparallel, right-handed, B-form double helix. The phosphate backbone is highly negatively charged, and the major and minor grooves provide access to the chemical groups on the individual bases. The conformation of the backbone and the degree of access provided by the grooves alters considerably in other conformation of DNA such as A and Z. This figure is drawn from fig. 4.15 in Watson [8].

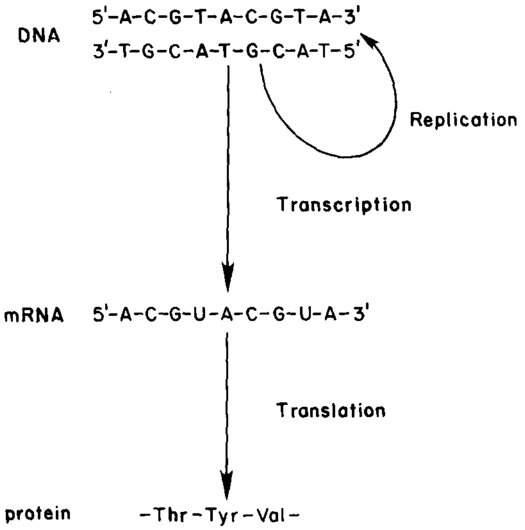

Fig. 3. *The Central Dogma.* Each DNA strand codes for its complement during replication. One DNA strand codes for a complementary mRNA strand synthesized during transcription (U in mRNA corresponds to T in DNA). Base triplets in mRNA code (5' to 3') for a protein's amino acids (polymerized during translation) as determined by the genetic code: in this example, ACG codes for threonine (Thr), UAC for tyrosine (Tyr), GUA for valine (Val). See text for further discussion.

codons, and therefore some redundancy in the translation, or mapping, of specific codons to specific amino acids. This mapping is called the genetic code. Thus, a sequence of bases along the DNA strand codes specifically for a sequence of amino acid groups along the protein strand.

3. Mechanisms modifying and extending the Central Dogma

The description of genetic information storage and regeneration in the Central Dogma suggests the appropriateness of modeling the evolution of DNA sequences within the framework of cellular automata: the sugar and phosphate groups define a one-dimensional lattice (DNA strand) of cells, and the state of any given cell is determined by an element from the four value nucleotide alphabet. Changes (mutations) in the configuration of the system (sequence) become evident after replication, which may be thought of as an iteration of the automaton.

However, in the years since the Central Dogma was realized, our increased knowledge of the mechanisms of biology at the molecular level has led to at least two elaborations that greatly qualify any simple model for the evolution of patterns in DNA sequences. First, not all DNA sequences code for protein; many sequences contain other types of information, for example regulatory signals controlling the synthesis of proteins. Secondly, the capacity for information storage appears to involve more than the one dimensional pattern of bases; for example, dynamic information is stored in the three-dimensional conformation of DNA. We will discuss these two elaborations below, touching on a representative set of examples.

To the extent that interactions between DNA and itself or other molecular components of the cell (membranes, fibers, proteins in solution, etc.) are accomplished through chemical contact, they involve base sequences mediating the ionic, hydrogen bonding and van der Waals interaction potentials of the DNA proximate to the contact site.

Thus, in addition to coding for protein sequences, the DNA must code for its involvement with the cellular machinery. This includes instructions concerning the regulation and execution of protein synthesis, as well as instructions for the packaging, storing, and manipulation of DNA within the cell.

We will examine a single stretch of DNA (see fig. 4) to get a sense of these several hierarchies of information storage. Fig. 4 presents a short section of DNA from the lac operon of E. coli [14]. An operon is a gene (protein coding sequence) plus the local regions governing the transcription of that gene. This gene codes for a protein that participates in the digestion of lactose, an alternative sugar source to glucose. In addition to the gene, the operon consists of a variety of recognition, binding, initiation, and termination sites that specify the correct beginning and ending positions for the operations needed to synthesize the protein. All of these are specified by patterns along the DNA strand. In addition, the operon contains regulatory sites that affect the level of transcription of this gene as a function of the availability of both glucose and lactose.

Transcription and translation of this gene proceed as follows: the mRNA-synthesizing machinery, called RNA *polymerase* (a protein conglomerate), binds loosely to the DNA strand at the recognition site and diffuses along the DNA until it binds tightly at the binding site. The RNA polymerase moves further along the DNA until it reaches the transcription initiation site, at which point it begins to read the DNA and synthesize mRNA; this continues until it reaches the transcription termination site (not shown in fig. 4). The protein-synthesizing machinery, called a *ribosome* (a protein and RNA conglomerate), binds to the mRNA at the ribosomal binding site; the ribosome then moves along the DNA until it reaches the translation initiation site (the start codon), at which point it begins to synthesize protein; this continues until the ribosome reaches the translation termination site (the stop codon – not shown in fig. 4). In the absence of lactose, a repressor protein binds to a short stretch of specifically recognized DNA, the repressor binding site, and prevents the movement of RNA polymerase along the DNA (therefore blocking transcription). In the absence of glucose, the CAP protein binds to the CAP binding site, facilitating the binding of RNA polymerase (therefore promoting transcription). The net effect of these mechanisms is that the gene is transcribed only when both lactose is present and glucose is absent. Again, the specificities of these sites are determined by subsequences of DNA that are not protein-coding sequences.

Thus we see how one of the building blocks of a chemical computer function. The lac operon can be summarized naturally in digital-mechanical terms. It consists of (1) a list of information stored in the gene, containing an abstract blueprint for the protein; (2) machinery to synthesize the protein; and (3) an input line capable of flipping a switch that determines whether or not the protein

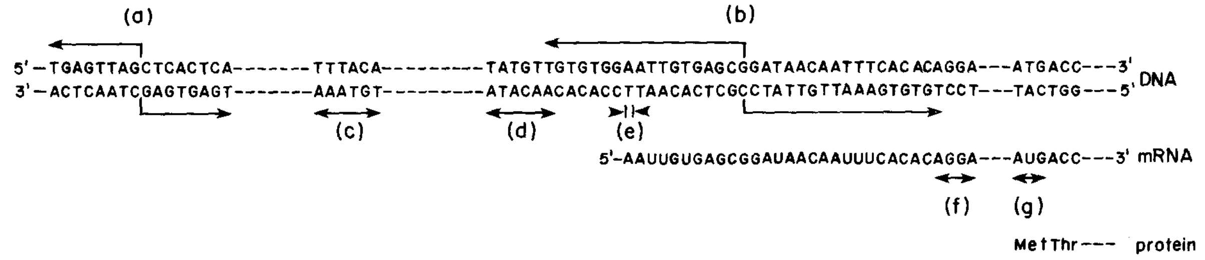
(a)
(b)
5'–TGAGTTAGCTCACTCA-------TTTACA---------TATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGA---ATGACC---3'
3'–ACTCAATCGAGTGAGT-------AAATGT---------ATACAACACACCTTAACACTCGCCTATTGTTAAAGTGTGTCCT---TACTGG---5'
DNA
(c)
(d)
(e)
5'-AAUUGUGAGCGGAUAACAAUUUCACACAGGA---AUGACC---3' mRNA
(f)
(g)
MetThr---
protein

Fig. 4. *The Lac Operon.* The hyphens indicate sections of sequences left out for simplicity – each hyphen represents two bases. The arrowed spans designate the extent of specific sites; the vertical segments of spans (a) and (b) designate axes of local palindromic symmetry in the DNA sequences. (a) CAP protein binding site; (b) repressor protein binding site; (c) RNA polymerase recognition site; (d) RNA polymerase binding site; (e) transcription initiation site; (f) ribosome binding site; (g) translation initiation site. See discussion in text. This figure is drawn from fig. 4.10 in Nover [14].

is synthesized. There are several other operons in E. coli known to function similarly to lac; in higher organisms, for instance vertebrates, there are analogous regulatory units, though these do not appear to be organized identically to those in bacteria.

To a first approximation the base sequence is fixed, and thus might be compared to the ROM (read-only-memory) of a digital computer. There are, however, methods for the dynamic storage of information in the DNA molecule that are effectively the RAM (read-write) memory of the DNA. This dynamic information storage operates through a variety of different mechanisms.

One such mechanism involves the chemical modification of bases along the DNA strand. In prokaryotes, for example, the restriction defense system [10], consists of pairs of enzymes, one of which chemically modifies specific short (4–8 bases pairs) sequences, and the other of which cuts the same sequence when it is unmodified; while the self-DNA is modified, and thus protected from cutting, the unmodified foreign DNA is chopped up by the cell's restriction enzymes. This system protects cells from the invasion of foreign DNA. In eukaryotes, the methylation of cytosine [15, 16] can be used to control the transcription level of various genes during the cell's development – highly methylated genes are not transcribed. In this case, the mechanism by which chemical modification regulates transcription has not been clearly established.

Although the traditional view of DNA structure as being both locally and globally monotonous in the form of the right-handed B-form double helix is still valid as a crude approximation, the geometrical form, or *conformation* of the molecule is increasingly recognized as being both base pattern dependent and dynamic [17, 18]. Conformation depends on local base sequence, protein interactions, solvent conditions and twisting and bending strains that may originate at sites distant along the DNA helix axis. Thus, changes in the conformation provide a means of inputting information to the DNA, as well as a means of dynamically storing it. Conformational changes can generate non-local interactions that may be used to bring normally distant parts of the DNA together, or activate or deactivate particular sequences. These changes modulate and regulate the functioning of the DNA.

Crystal and fiber diffraction studies have demonstrated the existence of at least three major helical conformations of DNA: right-handed A and B, and left-handed Z. The accessability and chemical characteristics of the major groove, minor groove and phosphate backbone vary markedly from helix form to helix form. These changes can have strong effects on the specificity of interactions between DNA and other macromolecules such as proteins. Even within one of the helical forms, there are sequence-dependent local variations in helix parameters such as twist angle, etc. Local regions of palindromic symmetry in the base pattern may lead to cruciform structures where local intrastrand double helices are formed. These structures create helix junctions that interrupt motion of DNA-binding proteins along the helix; in addition, they alter the super-helical stress on the DNA molecule, thus affecting DNA conformation at distant sites along the helix. The double-stranded DNA fluctuates in time between various conformational states as a function of both super-helical stress on the DNA and whether or not proteins and other regulatory molecules are bound to the DNA.

DNA does not exist as a random coil in cells – it is packed tightly. For instance, the DNA double helix in eukaryotes is super-helically wrapped around a series of protein cores; this strand of "beads on a chain" is then further condensed in an even higher-order conformation [14, 19]. The extent to which a local region of DNA is thus packed can effect its accessibility to cellular proteins such as RNA polymerase. The packing of DNA is also a dynamic process, as is clear during mitosis and meiosis. Thus, regulation of the packing patterns is another mechanism through which DNA can modulate its own functions.

Segments of DNA can be shuffled; that is, two pieces of DNA that are far apart, on different strands, or in a particular mutual orientation can, at a different point in time, be close together, on the same strand or in a different mutual orientation. For instance, when, prior to germ-line cell division (meiosis), the two parental copies of a cell's chromosomes are aligned, DNA strands commonly exchange corresponding but non-identical segments of DNA [8, 20]; this is called recombination, and leads to the mixing of genes from both parents in a single chromosome. In another example, the cells that are responsible for synthesizing antibody proteins create a final coding sequence by randomly combining individual partial sequences found in dispersed collections containing multiple variants of the partial sequences [21–23]. This is one of the mechanisms that provide an almost unlimited repertoire of potential antibody responses to foreign antigens, as compared with the alternative of coding for variant antibodies with single, complete sequences.

Because of the contiguous triplet nature of the genetic code, any given DNA sequence has six possible reading frames. For instance, consider the DNA sequence in fig. 3. This sequence could be read alternately on the top strand as ACG–TAC–GTA, XAC–GTA–CGT–AXX or XXA–CGT–ACG–TAX, and on the complementary strand as TAC–GTA–CGT, XTA–CGT–ACG–TXX or XXT–ACG–TAC–GTX. Of course, since those sequences actually coding for viable proteins are limited to certain patterns of amino acids, trying to arrange overlapping coding sequences puts severe constraints on the base sequence of the overlapping region. Most often, only one reading frame is actually read and transcribed, but there are cases, especially with organisms constrained to small genomes, such as viruses, where two or even three of the reading frames are transcribed [19]. Thus, a single stretch of DNA can code for two or more different proteins. In addition, a common occurence in eukaryotes is the interruption of reading frames along the DNA sequence by intervening non-coding sequences, called introns [24, 25]; these sequences are transcribed as they appear on the DNA, but then excised from the mRNA prior to translation. Introns are thought to be remnants of previous shuffling events involving sequences coding for portions of proteins.

Finally, DNA can move from organism to organism. Several eukaryotic viruses temporarily splice portions of their genome into the host genome (see, for example, Varmus [26]); recent evidence suggests that the spliced-in viral DNA occasionally becomes a permanent part of the host genome that is passed intact to the host progeny.

4. Modeling in terms of automata

The form that an automaton model of DNA might take will vary greatly according to the purpose for which it is intended. At the outset it is necessary to distinguish between modeling replication and modeling translation and transcription. This also might be thought of as the difference between modeling the evolution of germ-line DNA (phylogeny) and modeling the developmental changes in somatic cell DNA (ontogeny); note that somatic DNA also undergoes replication. The purpose behind developing these two types of models is quite different.

The most likely purpose for modeling replication is the study of natural selection or prebiotic evolution. In principle the basic idea is simple: A one-dimensional automaton can be used to represent the germ-line DNA or pre-DNA macromolecule. Iterations of the automaton represent the replication process. The rules of the automaton can be picked in order to mimic base substitution, base deletion or insertion, recombination and other changes that occur in DNA over a period of time. By iterating the automaton one can watch the nature of the genetic population evolve.

One example of this type is due to Holland [1]. He invents a simplified one-dimensional chemistry, with only two "bases", two amino acids, two types of bonds, a simple notion of mutation, and a primative notion of catalysis. He proves that, starting with a random ensemble of initial elements, catalysis eventually dominates, and the arrangement of the amino acids always organizes into groupings that are functionally equivalent to enzymes.

The natural approach in models of this type is to introduce a stochastic element into the iteration of the automaton. Holland, for example, accomplishes this through a random shuffling operator. In a more realistic model this random element should depend on chemical constraints as well. The likelihood of a mutation varies markedly for different sequences of base pairs, and this should be reflected in realistic models.

One problem that is difficult to treat in general is a proper notion of selection. In the prebiotic stages, selection can be simply phrased in terms of competition for raw materials among the various enzymes or chemical structures. This is the approach taken by both Holland [1] and by Eigen et al. [27]. To represent later stages of evolution, this seems unlikely to be an adequate criterion, since it is clear that more highly developed organisms have a phenotype that is quite different in form than the genotype, and many evolutionary criteria act on the phenotype rather than the geneotype. Incorporating an appropriate notion of selection is one of the important unsolved problems in this field.

It is also possible to use automata to make a model of the process of translation and transcription. The most obvious purpose of a model of this type is to attempt to understand the global mechanisms underlying the functioning of the somatic DNA. This basic approach might also be useful in understanding problems underlying development. In such a model, incorporating dynamical information storage is essential. The basic iteration of the automaton corresponds to times at which conformational changes and other changes described in the previous section take place, as well as possible replications. This is an interesting problem from the point of view of automata theory, since the necessity of dealing with conformation changes injects an unusual aspect into the construction of the automaton; while the automaton is basically a one dimensional object, it must also have a notion of its own geometry in three dimensions. Changes in this geometry are one of the primary changes taking place as the automaton is iterated. Thus, the resulting automaton is intermediate between a one and a three dimensional object.

One very successful example of a model of the global aspects of the translation and transcription process is due to Kauffman [2]. Without considering any of the dynamic information storage mechanisms explicitly, Kauffman lumps them all together into a genetic switching network, and assumes that the action of each gene can be digitally switched on and off in more or less the same manner as the lac operon. He searches for generic properties of random switching networks, and shows that the correct qualitative properties are obtained only if the random connections are constrained to satisfy a certain "canalizing property". He is thus able to make a significant prediction about the possible structure of the global rules governing interactions between the various operons. As experiments reach the level where we can examine the functioning of individual operons, but also collect information about the way operons effect each other, it will be interesting to discover whether Kauffman's prediction of canalizing structures is borne out. If so, this is one of the key design elements in the global programming structure underlying the functioning of DNA, and a good example of the usefulness of the qualitative modeling approach advocated in the introduction.

A positive side effect of the type of qualitative modeling exemplified by the work of Holland and Kauffman is a contribution to our understanding of adaptive automata in general. Although initially motivated by a desire to mimic aspects of DNA, these works also yield an understanding of universal principles that adaptive automata must satisfy in any context. This promises to have valuable spinoff into the field of artificial intelligence. This should not be surprising, since the DNA molecule is certainly the most sophisticated computer of which we are aware (at least when structures such as the brain that are coded for by DNA are included).

5. Efforts at collecting and interpreting DNA sequences

Any future efforts at modeling DNA as automata are likely to be aided by current programs for collecting what is known about DNA in an organized fashion. In particular, the nucleic acid sequence database at Los Alamos provides a valuable tool that those interested in automata models of DNA should be aware of. With this purpose in mind we present a brief review of this database.

In response to the rapidly increasing number of experimentally-determined sequences available, and the obvious advantages of centralizing these data, several groups worldwide began, several years ago, to accumulate the known DNA sequences into computer databases. The effort at Los Alamos National Laboratory was designated, in collaboration with Bolt, Beranek and Newman, Inc., as the National Nucleic Acid Sequence Database (GenBank) by the National Institutes of Health in 1982*. A broad view of the kind and quantity of DNA sequences in the database suggests that the available data provides a useful vantage point for recognizing and interpreting DNA sequence patterns.

As of this writing (July, 1983), the database contains two million base pairs in 2000 sequences. Over 200 vertebrate, invertebrate eukaryotic, bacterial, bacteriophage and eukaryotic viral species are represented. We estimate that sequences are currently being reported in the literature at the rate of one million base pairs per year.

How does the available data compare with the number of base pairs required to determine an entire organism? We have the complete genome for several host-dependent species, including viroids (250–350 base pairs), influenza A virus (1.4×10^4 bp), mammalian mitochondria (1.6×10^4 bp), T7 bacteriophage (4.0×10^4 bp) and lambda bacteriophage (5.0×10^4 bp). For E. coli, on the other hand, we have only approximately 3.5% of the entire genome, and for human, we have about 0.004%. Clearly, with these latter examples, we are far short of the complete data set for genetically determining an organism. In the case of E. coli, given the publication rate of newly determined sequences, it is possible to envision the entire genome being known within another 10 years or so; however, it will clearly be much longer before close to the complete human genome is known.

6. Finding patterns in DNA sequences

In parallel with the increasing availability of known DNA sequences, increasing attention has been directed to identifying patterns of bases in DNA sequences; the goal of the effort can be either identifying unusual patterns along a single strand of DNA, or comparing two sequences so as to identify regions of similarity. The identification and classification of patterns in DNA sequences would be very important for identifying emerging regulatory and protein-coding sequences during iterations of an automaton model of DNA replication. One approach is based, not on the particular complete arrangement of A, C, G and T in the DNA, but on statistical analyses of the base distribution in the sequence. Such an approach uses one sequence set to build up statistical standards which are then used to analyze test sequences. Gold et al. [28] use a test of base compositional skew to characterize ribosome binding sites in prokaryotes; Fickett [29] uses base composition and positional correlation of bases along the sequence to distinguish protein-coding regions from regions that do not code for protein.

Another approach to recognizing patterns is the identification of conserved particular sequences of A, C, G and T among compared sequences. For example, the sequence ACAATTG would be observed to be similar (though not identical) to ACAAATG when the first base of the first sequence is aligned with the first base of the second sequence, etc. This kind of search has been useful

* The database can be acquired either on magnetic tape or on-line by contacting: Gen-Bank, c/o Computer Systems Division, Bolt, Beranek and Newman, Inc., 10 Moulton St., Cambridge, MA, 02238 (telephone: 617-491-1850). A hard-copy compendium will be published early in 1984.

in identifying, for example, ribosome binding sites [28], intron boundaries [25] and RNA polymerase binding sites [14, 19].

An outgrowth of this latter approach has been the need to formulate an analytically exact method for gauging the similarity of two sequences. In other words, given an alignment of two sequences, how can one translate the resulting matches, mismatches, deletions and insertions of bases into a quantitative similarity value? There has been considerable progress in this area [30, 31], with the result that one can routinely compare sequences using a well-defined metric [32, 33]. There is no complete solution to the problem of assessing the statistical significance of these found similarity values, but there are partial probabilistic [32] and empirical [34] guidelines that enable, in most cases, reliable estimates of significance.

The discussion thus far has focused on patterns of A, C, G and T in DNA sequences. However, these letters are only a short-hand representation of the actual DNA structure; our ultimate goal in identifying and classifying patterns will be to discuss the local structure of the double helix itself. With the recent solution of several crystal structures of short DNA double helices (as reviewed by Zimmerman [17] and Dickerson et al. [18]), attempts are now being made to correlate base sequences with local double helix conformation [35–39]. The eventual outcome of this line of research should be the ability, given the base sequence, to predict patterns of chemical characteristic along a DNA strand. This will provide a more powerful basis for pattern analysis, if for no other reason than compensating for cases where two sequences with very little base difference are conformationally quite distinct, or, on the other hand, where two sequences with extremely different base patterns are conformationally quite similar.

7. Future prospects

The database of experimentally determined DNA sequences will, of course, continually expand, as will the body of experimental work characterizing both the chemical basis of protein-DNA interactions and the structural correlates of particular base sequences. The convergence of these two bodies of data should both inspire and provide a rich testing ground for modeling of the evolution of patterns in DNA sequences. Continued elucidation of the chemical mechanisms for natural mutation of DNA sequences (which we have not touched on in this article) will, in particular, provide an additional framework for suggesting the "rules" governing the evolution in time of these patterns.

This work will facilitate the development of automata models of the self-organizing properties of DNA. As automata models are developed, they should in turn aid in the interpretation of DNA sequences, and possibly indicate new directions in the field of artificial intelligence.

Acknowledgements

We are grateful to J. Fickett for help in compiling the database statistics. This work was carried out under the auspices of the U.S. Department of Energy. GenBank is funded in part by the National Institutes of Health contract NO1-GM-2-2127, with joint support from The National Institute of General Medical Sciences, The National Institute of Cancer, The National Institute of Allergy and Infectious Diseases, The Division of Research Resources of The NIH, The National Science Foundation, The Department of Energy and The Department of Defense.

References

[1] J.H. Holland, in: Automata, Languages, Development, A. Lindenmayer and G. Rozenberg, eds. (North-Holland, Amsterdam, 1976), pp. 385–404.

[2] S. Kauffman, this issue.

[3] S. Wolfram, Caltech preprint CALT-68-938(3982).

[4] S. Wolfram, Rev. Mod. Physics 55 (1983) 601.

[5] S. Wolfram, Preface to these proceedings.

[6] R. Ransom, Computers and Embryos: Models in Devel-

opmental Biology (Wiley, New York, 1981), pp. 106–156, 184–189.

[7] F.H.C. Crick, in: The Great Ideas Today 1980, Encyclopedia Britannica, pp. 644–683.

[8] J.D. Watson, Molecular Biology of the Gene, 3rd ed. (Benjamin, Menlo Park, CA., 1976).

[9] B. Lewin, Genes (Wiley, New York, 1983).

[10] B. Lewin, Gene Expressions, vol. I, Bacterial Genomes (Wiley-Interscience, New York, 1974).

[11] B. Lewin, Gene Expressions, vol. II, Eukaryotic Genomes (Wiley-Interscience, New York, 1980).

[12] B. Lewin, Gene Expression, vol. III, Plasmids and Phages (Wiley-Interscience, New York, 1977).

[13] L. Nover, M. Luckner and B. Parthier, ed., Cell Differentiation: Molecular Basis and Problems (Springer, New York, 1982).

[14] L. Nover, in: Cell Differentiation: Molecular Basis and Problems, L. Nover, M. Luckner and B. Parthier, eds. (Springer, New York, 1982), pp. 99–256.

[15] A. Razin and A.D. Riggs, Science 210 (1980) 604.

[16] G. Felsenfeld and J. McGhee, Nature 296 (1982) 602.

[17] S.B. Zimmerman, Ann. Rev. Biochem. 51 (1982) 395.

[18] R.E. Dickerson, H.R. Drew, B.N. Conner, R.M. Wing, A.V. Fratini and M.L. Kopka, Science 216 (1982) 475.

[19] L. Nover and H. Reinbothe, in: Cell Differentiation: Molecular Basis and Problems, L. Nover, M. Luckner and B. Parthier, eds. (Springer, New York, 1982), pp. 23–74.

[20] D. Dressler and H. Potter, Ann. Rev. Biochem. 51 (1982) 727.

[21] S. Tonegawa, Nature 302 (1983) 575.

[22] P. Leder, Sci. Amer. 246 (May) (1982) 102.

[23] R.F. Richter and L. Nover, in: Cell Differentiation: Molecular Basis and Problems, L. Nover, M. Luckner and B. Parthier, eds. (Springer, New York, 1982), pp. 449–475.

[24] F.H.C. Crick, Science 204 (1979) 264.

[25] R. Breathnach and P. Chambon, Ann. Rev. Biochem. 50 (1981) 349.

[26] H.E. Varmus, Science 216 (1982) 812.

[27] M. Eigen, W. Gardner, P. Schuster and R. Winkler-Oswatitsch, Sci. Amer. 244 (April) (1981) 88.

[28] L. Gold, D. Pribnow, T. Schneider, S. Shmedling, B.S. Singer and G. Stormo, Ann. Rev. Microbiol. 35 (1981) 365.

[29] J. Fickett, Nucl. Acids Res. 10 (1982) 5303.

[30] T.F. Smith and C. Burks, Nature 301 (1983) 194.

[31] J.B. Kruskal, SIAM Rev. 25 (1983) 2011.

[32] W.B. Goad and M.I. Kanehisa, Nucl. Acids Res. 10 (1982) 247.

[33] T.F. Smith and M.S. Waterman, J. Mol. Biol. 141 (1981) 195.

[34] T.F. Smith, M.S. Waterman and C. Burks, LANL Preprint LA-UR-83-1661 (1983) Los Alamos National Laboratory.

[35] R.E. Dickerson and H.R. Drew, J. Mol. Biol. 149 (1981) 761.

[36] R.E. Dickerson and H.R. Drew, Proc. Natl. Acad. Sci. USA 78 (1981) 7318.

[37] C.R. Calladine, J. Mol. Biol. 161 (1982) 343.

[38] W. Kabsch, C. Sander and E.N. Trifinov, Nucl. Acids Res. 10 (1982) 1097.

[39] R.E. Dickerson, J. Mol. Biol. 166 (1983) 419.