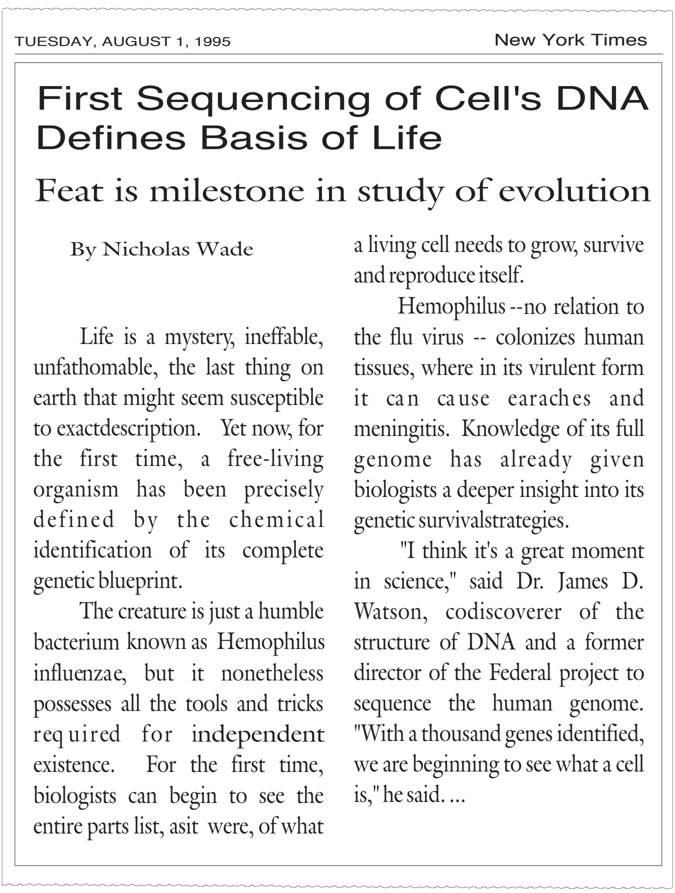
4

Patterns in DNA



[1]

TUESDAY, AUGUST 1, 1995 New York Times

First Sequencing of Cell's DNA Defines Basis of Life

Feat is milestone in study of evolution

By Nicholas Wade

Life is a mystery, ineffable, unfathomable, the last thing on earth that might seem susceptible to exactdescription. Yet now, for the first time, a free-living organism has been precisely defined by the chemical identification of its complete genetic blueprint.

The creature is just a humble bacterium known as Hemophilus influenzae, but it nonetheless possesses all the tools and tricks required for independent existence. For the first time, biologists can begin to see the entire parts list, asit were, of what a living cell needs to grow, survive and reproduce itself.

Hemophilus --no relation to the flu virus -- colonizes human tissues, where in its virulent form it can cause earaches and meningitis. Knowledge of its full genome has already given biologists a deeper insight into its genetic survivalstrategies.

"I think it's a great moment in science," said Dr. James D. Watson, codiscoverer of the structure of DNA and a former director of the Federal project to sequence the human genome. "With a thousand genes identified, we are beginning to see what a cell is," he said. ...

[1] Reprinted by permission.

Introduction

The human cytomegalovirus (CMV) is a potentially life-threatening disease for people with suppressed or deficient immune systems. To develop strategies for combating the virus, scientists study the way in which the virus replicates. In particular, they are in search of a special place on the virus' DNA that contains instructions for its reproduction; this area is called the origin of replication.

A virus' DNA contains all of the information necessary for it to grow, survive and replicate. DNA can be thought of as a long, coded message made from a four-letter alphabet: A, C, G, and T. Because there are so few letters in this DNA alphabet, DNA sequences contain many patterns. Some of these patterns may flag important sites on the DNA, such as the origin of replication. A complementary palindrome is one type of pattern. In DNA, the letter A is complementary to T, and G is complementary to C, and a complementary palindrome is a sequence of letters that reads in reverse as the complement of the forward sequence (e.g., GGGCATGCCC).

The origin of replication for two viruses from the same family as CMV, the herpes family, are marked by complementary palindromes. One of them, Herpes simplex, is marked by a long palindrome of 144 letters. The other, the Epstein–Barr virus, has several short palindromes and close repeats clustered at its origin of replication. For the CMV, the longest palindrome is 18 base pairs, and altogether it contains 296 palindromes between 10 and 18 base pairs long. Biologists conjectured that clusters of palindromes in CMV may serve the same role as the single long palindrome in Herpes simplex, or the cluster of palindromes and short repeats in the Epstein–Barr virus' DNA.

To find the origin of replication, DNA is cut into segments and each segment is tested to determine whether it can replicate. If it does not replicate, then the origin of replication must not be contained in the segment. This process can be very expensive and time consuming without leads on where to begin the search. A statistical investigation of the DNA to identify unusually dense clusters of palindromes can help narrow the search and potentially reduce the amount of testing needed to find the origin of replication. In practice, the CMV DNA was examined statistically for many different kinds of patterns. However, for this lab, the search will be restricted to looking for unusual clusters of complementary palindromes.

Data

Chee et al. ([CBB$^+$90]) published the DNA sequence of CMV in 1990. Leung et al. ([LBBK91]) implemented search algorithms in a computer program to screen the sequence for many types of patterns. Altogether, 296 palindromes were found that were at least 10 letters long. The longest ones found were 18 letters long. They occurred at locations 14719, 75812, 90763, and 173863 along the sequence.

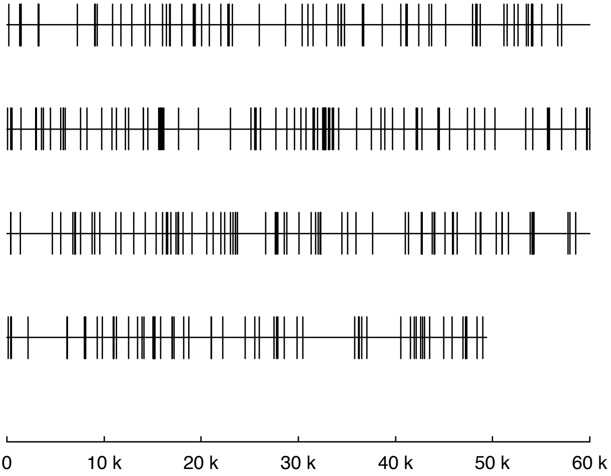


FIGURE 4.1. Diagram of the 296 palindrome locations for the CMV DNA (Chee et al. [CBB$^+$90]).

Palindromes shorter than 10 letters were ignored, as they can occur too frequently by chance. For example, the palindromes of length two — AT, TA, GC and CG — are quite common.

Altogether, the CMV DNA is 229,354 letters long. Table 4.1 contains the locations of the palindromes in the DNA that are at least 10 letters long. Notice that the very first palindrome starts at position 177, the second is at position 1321, the third at position 1433, and the last at position 228953. Each palindrome is also located on a map of the DNA sequence in Figure 4.1. In this figure, a palindrome is denoted by a vertical line; clusters of palindromes appear as thick lines on the map.

Background

DNA

In 1944, Avery, MacLeod, and McCarty showed that DNA was the carrier of hereditary information. In 1953, Franklin, Watson, and Crick found that DNA has

TABLE 4.1. CMV palindrome locations for the 296 palindromes each at least ten base pairs long (Chee et al. [CBB+90]).

177	1321	1433	1477	3248	3255
3286	7263	9023	9084	9333	10884
11754	12863	14263	14719	16013	16425
16752	16812	18009	19176	19325	19415
20030	20832	22027	22739	22910	23241
25949	28665	30378	30990	31503	32923
34103	34398	34403	34723	36596	36707
38626	40554	41100	41222	42376	43475
43696	45188	47905	48279	48370	48699
51170	51461	52243	52629	53439	53678
54012	54037	54142	55075	56695	57123
60068	60374	60552	61441	62946	63003
63023	63549	63769	64502	65555	65789
65802	66015	67605	68221	69733	70800
71257	72220	72553	74053	74059	74541
75622	75775	75812	75878	76043	76124
77642	79724	83033	85130	85513	85529
85640	86131	86137	87717	88803	89586
90251	90763	91490	91637	91953	92526
92570	92643	92701	92709	92747	92783
92859	93110	93250	93511	93601	94174
95975	97488	98493	98908	99709	100864
102139	102268	102711	104363	104502	105534
107414	108123	109185	110224	113378	114141
115627	115794	115818	117097	118555	119665
119757	119977	120411	120432	121370	124714
125546	126815	127024	127046	127587	128801
129057	129537	131200	131734	133040	134221
135361	136051	136405	136578	136870	137380
137593	137695	138111	139080	140579	141201
141994	142416	142991	143252	143549	143555
143738	146667	147612	147767	147878	148533
148821	150056	151314	151806	152045	152222
152331	154471	155073	155918	157617	161041
161316	162682	162703	162715	163745	163995
164072	165071	165883	165891	165931	166372
168261	168710	168815	170345	170988	170989
171607	173863	174049	174132	174185	174260
177727	177956	178574	180125	180374	180435
182195	186172	186203	186210	187981	188025
188137	189281	189810	190918	190985	190996
191298	192527	193447	193902	194111	195032
195112	195117	195151	195221	195262	195835
196992	197022	197191	198195	198709	201023
201056	202198	204548	205503	206000	207527
207788	207898	208572	209876	210469	215802
216190	216292	216539	217076	220549	221527
221949	222159	222573	222819	223001	223544
224994	225812	226936	227238	227249	227316
228424	228953				

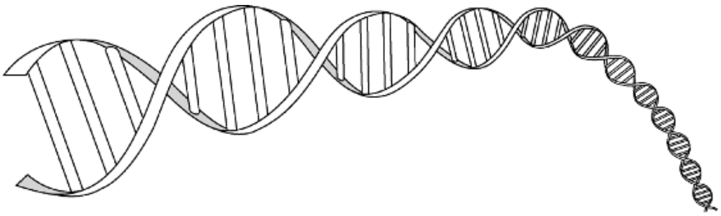


FIGURE 4.2. Paired ribbons of DNA forming the double helix structure.

a double helical structure (Figure 4.2) composed of two long chains of nucleotides. A single nucleotide has three parts: a sugar, a phosphate, and a base. All the sugars in DNA are deoxyribose — thus the name deoxyribose nucleic acid, or DNA. The bases come in four types: adenine, cytosine, guanine, and thymine, or A, C, G, T for short. As the bases vary from one nucleotide to another, they give the appearance of a long, coded message.

The two strands of nucleotides are connected at the bases, forming complementary pairs. That is, the bases on one strand are paired to the other strand: A to T, C to G, G to C, and T to A. Therefore, one strand "reads" as the complement of the other. This pairing forms a double helix out of the two strands of complementary base sequences.

The CMV DNA molecule contains 229,354 complementary pairs of letters or base pairs. In comparison, the DNA of the *Hemophilus influenzae* bacterium has approximately 1.8 million base pairs, and human DNA has more than 3 billion base pairs.

Viruses

Viruses are very simple structures with two main parts: a DNA molecule wrapped within a protein shell called a capsid. The DNA stores all the necessary information for controlling life processes, including its own replication. The DNA for viruses typically ranges up to several hundred thousand base pairs in length. According to *The Cartoon Guide to Genetics* ([GW91]), the replication of the bacteria *E. coli* happens as follows:

> In *E. coli* replication begins when a "snipping" enzyme cuts the DNA strand apart at a small region called the *origin*. In the neighborhood are plenty of free nucleotides, the building blocks for the new strands. When a free nucleotide meets its complementary base on the DNA, it sticks, while the "wrong" nucleotides bounce away. As the snipping enzyme opens the DNA further, more nucleotides are added, and a clipping enzyme puts them together.

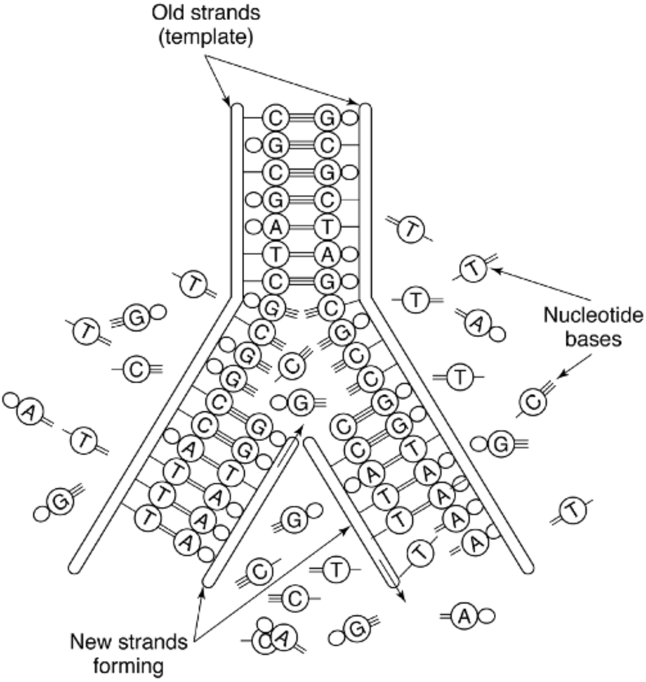


FIGURE 4.3. A sketch of DNA replication.

Figure 4.3 illustrates the replication process. The *origin* described in Gonick and Wheelis ([GW91]), where the snipping enzyme starts to cut apart the DNA strands, is the object of the search in this lab.

Human Cytomegalovirus

CMV is a member of the Herpes virus family. The family includes Herpes simplex I, chicken pox, and the Epstein–Barr virus. Some Herpes viruses infect 80% of the human population; others are rare but debilitating. As for CMV, its incidence varies geographically from 30% to 80%. Typically, 10 – 15% of children are infected with CMV before the age of 5. Then the rate of infection levels off until young adulthood, when it again increases ([Rya94, pp. 512–513]). While most CMV infections in childhood and adulthood have no symptoms, in young adults CMV may cause a mononucleosis-like syndrome.

Once infected, CMV typically lays dormant. It only becomes harmful when the virus enters a productive cycle in which it quickly replicates tens of thousands of copies. In this production cycle, it poses a major risk for people in immune-depressed states such as transplant patients who are undergoing drug therapy to suppress the immune system or people with Acquired Immune Deficiency Syndrome (AIDS). For these people, if the virus is reactivated, it can cause serious infections in internal organs. For example, CMV pneumonia is the leading cause of death among patients receiving bone marrow transplants. In AIDS patients, CMV infection often leads to neurological disorders, gastrointestinal disease and pneumonia. In addition, CMV is the most common infectious cause of mental retardation and congenital deafness in the United States.

Locating the origin of replication for CMV may help virologists find an effective vaccine against the virus. Research on the DNA for other Herpes viruses has uncovered the origin of replication for Herpes simplex I and Epstein–Barr. As stated earlier, the former is marked by one long palindrome of 144 base pairs, and the latter contains several short patterns including palindromes and close repeats. In earlier research, Weston ([Wes88]) found that a cluster of palindromes in the CMV DNA in the region 195,000 to 196,000 base pairs (see Figure 4.1) marked the site of another important function, called the enhancer.

Genomics

Recent advances in recombinant DNA and in machines that automate the identification of the bases have led to a burgeoning new science called genomics (Waterman [Wat89]). Genomics is the study of living things in terms of their full DNA sequences. Discoveries in genomics have been aided by advances in the fields of computer science, statistics, and other areas of mathematics, such as knot theory. For example, computer algorithms are being designed to search long sequences of DNA for patterns, information theory is facing the challenge of how to compress and manage these large databases, statistics and probability theory are being developed for matching sequences and identifying nonrandom structure in sequences, and knot theory has provided insights into the three-dimensional structure and molecular dynamics of DNA.

Investigations

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

- *Random scatter.*
  To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA. Of course, a random uniform scatter does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes. To look for structure, examine the locations of the palindromes, the spacings between palindromes, and the counts of palindromes in nonoverlapping regions of the DNA. One starting place might be to see first how random scatter looks by using a computer to simulate it. A computer can simulate 296 palindrome

sites chosen at random along a DNA sequence of 229,354 bases using a pseudorandom number generator. When this is done several times, by making several sets of simulated palindrome locations, then the real data can be compared to the simulated data.

- *Locations and spacings*. Use graphical methods to examine the spacings between consecutive palindromes and sums of consecutive pairs, triplets, etc., spacings. Compare what you find for the CMV DNA to what you would expect to see in a random scatter. Also, consider graphical techniques for examining the locations of the palindromes.
- *Counts*. Use graphical displays and more formal statistical tests to investigate the counts of palindromes in various regions of the DNA. Split the DNA into nonoverlapping regions of equal length to compare the number of palindromes in an interval to the number that you would expect from uniform random scatter. The counts for shorter regions will be more variable than those for longer regions. Also consider classifying the regions according to their number of counts.
- *The biggest cluster.* Does the interval with the greatest number of palindromes indicate a potential origin of replication? Be careful in making your intervals, for any small, but significant, deviation from random scatter, such as a tight cluster of a few palindromes, could easily go undetected if the regions examined are too large. Also, if the regions are too small, a cluster of palindromes may be split between adjacent intervals and not appear as a high-count interval. These issues are discussed in more detail in the Extensions section of this lab.

How would you advise a biologist who is about to start experimentally searching for the origin of replication? Write your recommendations in the form of a memo to the head biologist of a research team of which you are a member.