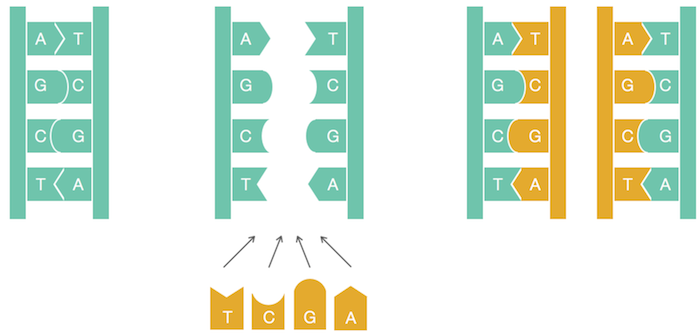
**Genome replication** is one of the most important tasks carried out in the cell. Before a cell can divide, it must first replicate its genome so that each of the two daughter cells inherits its own copy. In 1953, James Watson and Francis Crick completed their landmark paper on the DNA double helix with a now-famous phrase:

*"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."*

They conjectured that the two strands of the parent DNA molecule unwind during replication, and then each parent strand acts as a template for the synthesis of a new strand. As a result, the replication process begins with a pair of complementary strands of DNA and ends with two pairs of complementary strands, as shown in the figure below.



**Figure:** A naive view of DNA replication. Nucleotides adenine (A) and thymine (T) are complements of each other, as are cytosine (C) and guanine (G). Complementary nucleotides bind to each other in DNA.

Although this figure models DNA replication on a simple level, the details of replication turned out to be much more intricate than Watson and Crick imagined; as we will see, an astounding amount of molecular logistics is required to ensure DNA replication.

At first glance, a computer scientist might not imagine that these details have any computational relevance. To mimic replication algorithmically, we only need to take a string representing the genome and return a copy of it! Yet if we take the time to review the underlying biological process, we will be rewarded with new algorithmic insights into analyzing replication.

Replication begins in a genomic region called the **replication origin** (denoted *ori*) and is performed by molecular copy machines called **DNA polymerases**. Locating *ori* presents an important task not only for understanding how cells replicate but also for various biomedical problems. For example, some gene therapy methods use genetically engineered mini-genomes, which are called **viral vectors** because they are able to penetrate cell walls (just like real viruses). Viral vectors carrying artificial genes have been widely used in agriculture, such as to engineer frost-resistant tomatoes and pesticide-resistant corn. In 1990, gene therapy was first successfully performed on humans when it saved the life of a four year-old girl suffering from an immunodeficiency disorder; the girl had been so vulnerable to infections that she was forced to live in a sterile environment.

The idea of gene therapy is to intentionally infect a patient who lacks a crucial gene with a viral vector containing an artificial gene that encodes a therapeutic protein. Once inside the cell, the vector replicates and eventually produces many molecules of the therapeutic protein, which in turn treats the patient’s disease. To ensure that the vector actually replicates inside the cell, biologists must know where *ori* is in the vector’s genome and ensure that the genetic manipulations that they perform do not affect it.

In the following problem, we assume that a genome has a single *ori* and is represented as a **DNA string**, or a sequence of nucleotides from the four-letter alphabet {A, C, G, T}.

**Finding Origin of Replication Problem:**  
 **Input:**A DNA string Genome.  
 **Output:** The location of *ori* in Genome.

**STOP and Think:** Does this biological problem represent a clearly stated computational problem?

**Note:**﻿"STOP and Think" questions are opportunities for you to think about an aspect of the discussion before proceeding.  You do not need to answer them anywhere, but you may like to post your thoughts to the discussion forum at the bottom of the page and interact with other students.

Although the Finding Origin of Replication Problem asks a legitimate biological question, it does not present a well-defined computational problem. Indeed, biologists would immediately plan an experiment to locate *ori*: for example, deleting various short segments from the genome and eventually finding a segment whose deletion stops replication. Computer scientists, on the other hand, would shake their heads and demand more information before even thinking about the problem.

Why should biologists care what computer scientists think? Computational methods are now the only realistic way to answer many questions in modern biology. First, these methods are much faster than experimental approaches; second, the results of many experiments cannot be interpreted without computational analysis. In particular, existing experimental approaches to *ori* prediction are rather time consuming. As a result, *ori* has only been experimentally located in a handful of species. Thus, we would like to design a computational approach to find *ori* so that biologists are free to spend their time and money on other tasks.

***DnaA* boxes**

In the rest of this chapter, we will focus on the relatively easy case of finding *ori* in bacterial genomes, most of which consist of a single circular chromosome. Research has shown that the region of the bacterial genome encoding *ori* is typically a few hundred nucleotides long. Our plan is to begin with a bacterium in which *ori* is known, and then determine what makes this genomic region special in order to design a computational approach for finding *ori* in other bacteria. Our example is *Vibrio cholerae*, the pathogenic bacterium that causes cholera; here is the nucleotide sequence appearing in the *ori* of *Vibrio cholerae*:

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt tccttaaccctctattttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

How does the bacterial cell know to begin replication exactly in this short region within the much larger *Vibrio cholerae* chromosome, which consists of over a million nucleotides? There must be some “hidden message” in the *ori* region ordering the cell to begin replication here. Indeed, we know that the initiation of replication is mediated by ***DnaA***, a protein that binds to a short segment within the *ori* known as a ***DnaA* box**. You can think of the *DnaA* box as a message within the DNA sequence telling *DnaA*: “bind here!” The question is how to find this hidden message without knowing what it looks like in advance — can you find it? In other words, can you find something that stands out in *ori*? This discussion motivates the following problem.

**Hidden Message Problem:** *Find a “hidden message” in the replication origin.*  
 **Input:** A string Text.  
 **Output:** A hidden message in Text.

**STOP and Think:** Does the Hidden Message Problem represent a clearly stated computational problem?

## Hidden messages in “The Gold-Bug”

Although the Hidden Message Problem poses a legitimate intuitive question, it again makes absolutely no sense to a computer scientist because the notion of a “hidden message” is not precisely defined. The ori region of Vibrio cholerae is currently just as puzzling as the parchment discovered by William Legrand in Edgar Allan Poe's story "The Gold-Bug". Written on the parchment was the following:

53‡‡†305))6·;4826)4‡.)4‡);806·;48†8^60))85;161;:‡·8

†83(88)5·†;46(;88·96·?;8)·‡(;485);5·†2:·‡(;4956·2(5

·—4)8^8·;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4

)485†528806·81(‡9;48;(88;4(‡?34;48)4‡;1‡(;:188;‡?;

Upon seeing the parchment, the narrator remarks, "Were all the jewels of Golconda awaiting me upon my solution of this enigma, I am quite sure that I should be unable to earn them." Legrand retorts, "It may well be doubted whether human ingenuity can construct an enigma of the kind which human ingenuity may not, by proper application, resolve." He reasons that the three consecutive symbols **;48** appear with surprising frequency on the parchment.

53‡‡†305))6·**;48**26)4‡.)4‡);806·**;48**†8^60))85;161;:‡·8

†83(88)5·†;46(;88·96·?;8)·‡(**;48**5);5·†2:·‡(;4956·2(5

·—4)8^8·;4069285);)6†8)4‡‡;1(‡9**;48**081;8:8‡1**;48**†85;4

)485†528806·81(‡9**;48**;(88;4(‡?34**;48**)4‡;1‡(;:188;‡?;

Legrand had already deduced that the pirates spoke English; he therefore assumed that the high frequency of **;48** implied that it encodes the most frequent English word, **THE**. Substituting **;** for T, 4 for H, and 8 for E, Legrand had a slightly easier text to decipher (shown below), which would eventually lead him to the buried treasure. Can you decode this message too?

53‡‡†305))6·THE26)H‡.)H‡)TE06·THE†E^60))E5T161T:‡·E  
†E3(EE)5·†TH6(TEE·96·?TE)·‡(THE5)T5·†2:·‡(TH956·2(5  
·—H)E^E·TH0692E5)T)6†E)H‡‡T1(‡9THE0E1TE:E‡1THE†E5TH

)HE5†52EE06·E1(‡9THET(EETH(‡?3HTHE)H‡T1‡(T:1EET‡?T

**Note:** Don't spend too much time on this task; if you would like to see the answer, it is available at [Wikipedia](https://en.wikipedia.org/wiki/The_Gold-Bug#The_cryptogram).

## Counting words

Operating under the assumption that DNA is a language of its own, let’s borrow Legrand’s method and see if we can find any surprisingly frequent “words” within the *ori* of *Vibrio cholerae*. We have added reason to look for frequent words in the *ori* because for various biological processes, certain nucleotide strings appear surprisingly often in small regions of the genome. This is often because certain proteins can only bind to DNA if a specific string of nucleotides is present, and if there are more occurrences of the string, then it is more likely that binding will successfully occur. (It is also less likely that a mutation will disrupt the binding process.)

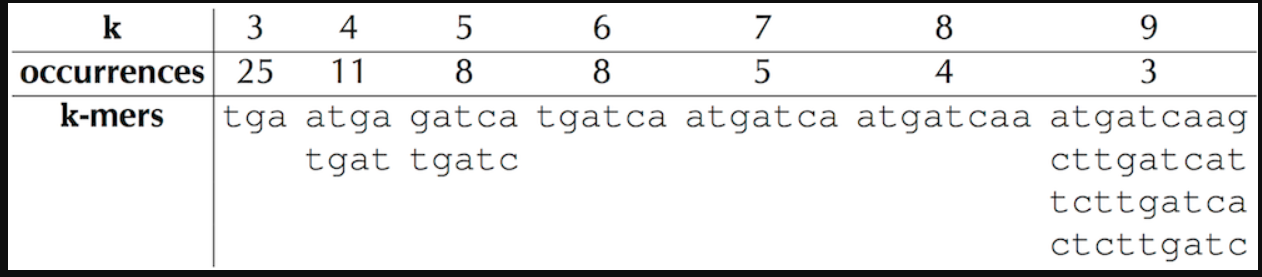
For example, "**ACTAT**" is a surprisingly frequent **substring** of

"ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT".

We use the term **k-mer** for a string of length k and define PatternCount(Pattern, Text) as the number of times that a k-mer Pattern appears as a substring of Text. Following the above example,

PatternCount("ACTAT", "ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT") = 3.

Note that PatternCount("ATA", "CG**ATATA**TCC**ATA**G") is equal to 3 (not 2) since we should account for overlapping occurrences of Pattern in Text.



The figure below reveals the most frequent k-mers in the *ori* region from *Vibrio cholerae*.

**STOP and Think:** Do any of the counts in the figure seem surprisingly large?

**Figure:** The most frequent k-mers in the *ori* region of *Vibrio cholerae* for *k* ranging from 3 to 9, along with the number of times that each k-mer occurs.

For example, the 9-mer "**ATGATCAAG**" appears three times in the *ori* region of *Vibrio cholerae* — is it surprising?

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac  
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca  
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt  
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt  
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga  
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat  
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag  
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt  
tccttaaccctctattttttacggaaga**ATGATCAAG**ctgctgctcttgatcatcgtttc

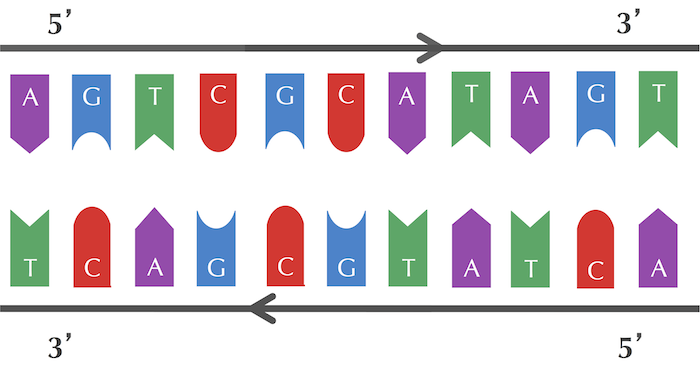
We highlight a most frequent 9-mer instead of using some other value of *k* because experiments have revealed that bacterial *DnaA* boxes are usually 9 nucleotides long. Furthermore, it is very unlikely that a 9-mer would appear three or more times in a randomly generated DNA string of length 500 due to random chance. In fact, there are four different 9-mers repeated three or more times in this region: **"ATGATCAAG"**, "CTTGATCAT", "TCTTGATCA", and "CTCTTGATC".

The low likelihood of witnessing even one repeated 9-mer in the *ori* region of *Vibrio cholerae* leads us to the working hypothesis that one of these four 9-mers may represent a potential *DnaA*box that, when appearing multiple times in a short region, jump-starts replication. But which one?

**STOP and Think:** Is any one of the four most frequent 9-mers in the *ori* of *Vibrio cholerae* “more surprising” than the others?

Recall that nucleotides **A** and **T** are complements of each other, as are **C** and **G**. Having one strand of DNA and a supply of “free floating” nucleotides as shown in the figure below, we can imagine the synthesis of a complementary strand on a template strand. This model of replication was confirmed by Meselson and Stahl in 1958 (see [DETOUR: The Most Beautiful Experiment in Biology](https://stepik.org/lesson/Detour-The-Most-Beautiful-Experiment-in-Biology-13)). The figure below shows a template strand "**AGTCGCATAGT**" and its complementary strand "**ACTATGCGACT**".

At this point, you may think that we have made a mistake, since the complementary strand in the figure below reads out "**TCAGCGTATCA**" from left to right rather than "**ACTATGCGACT**". **A** and **T** are complements of each other, as are **C** and **G**. The beginning and end of a DNA strand are denoted 5’ (pronounced “five prime”) and 3’ (pronounced “three prime”), respectively. Each DNA strand is read in the 5' → 3' direction, and the complementary strand runs in the opposite direction to the template strand. See [DETOUR: Directionality of DNA Strands](https://stepik.org/lesson/Detour-Directionality-of-DNA-Strands-14) to learn why biologists use 5’ and 3’ to refer to the beginning and end of a strand of DNA.



The reverse complement of a DNA string Pattern is the string formed by taking the complementary nucleotide of each nucleotide in Pattern, then reversing the resulting string. For example, the reverse complement of "**AGTCGCATAGT**" is "**ACTATGCGACT**". This leads us to the following computational problem.

**Reverse Complement Problem:**  *Find the reverse complement of a DNA string.*  
**Input:** A DNA string Pattern.  
**Output:** The reverse complement of Pattern.

Interestingly, among the four most frequent 9-mers in the *ori* region of *Vibrio cholerae*, "**ATGATCAAG**" and "**CTTGATCAT**" are reverse complements of each other, resulting in the following six occurrences of these strings.

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt tccttaaccctctattttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

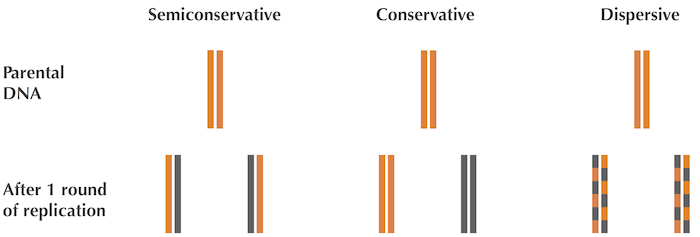
Finding a 9-mer that appears six or more times (either as itself or as its reverse complement) in a DNA string of length 500 is far more surprising than finding a 9-mer that appears three or more times alone. This statistical evidence leads us to the working hypothesis that "**ATGATCAAG**" and its reverse complement "**CTTGATCAT**" indeed represent *DnaA* boxes in*Vibrio cholerae*. Our computational conclusion makes sense biologically because the *DnaA* protein that binds to *Dna*A boxes and initiates replication does not care which of the two strands it binds to. For our purposes, both "**ATGATCAAG**" and "**CTTGATCAT**" represent *DnaA* boxes.

## Looking for hidden messages in multiple genomes

We should not jump to the conclusion that "**ATGATCAAG**"/"**CTTGATCAT**" is a hidden message for all bacterial genomes without first checking whether it even appears in known *ori* regions from other bacteria. After all, maybe the clumping effect of "**ATGATCAAG**"/"**CTTGATCAT**" in the *ori* region of *Vibrio cholerae* is simply a statistical fluke that has nothing to do with replication. Or maybe different bacteria have different *DnaA* boxes . . .

The Meselson-Stahl experiment, conducted in 1958 by Matthew Meselson and Franklin Stahl, is sometimes called "the most beautiful experiment in biology". In the late 1950s, biologists debated three conflicting models of DNA replication, illustrated in the figure below. The **semiconservative hypothesis** suggested that each parent strand acts as a template for the synthesis of a daughter strand. As a result, each of the two daughter molecules contains one parent strand and one newly synthesized strand. The **conservative hypothesis** proposed that the entire double-stranded parent DNA molecule serves as a template for the synthesis of a new daughter molecule, resulting in one molecule with two parent strands and another with two newly synthesized strands. The **dispersive hypothesis** proposed that some mechanism breaks the DNA backbone into pieces and splices intervals of synthesized DNA, so that each of the daughter molecules is a patchwork of old and new double-stranded DNA.

The three models are illustrated below, with parent DNA colored yellow and newly synthesized DNA colored black.



Meselson and Stahl's insight relied on the fact that one isotope of nitrogen, **Nitrogen-14** (14N), is lighter and more abundant than **Nitrogen-15** (15N). Knowing that the DNA molecular structure naturally contains 14N , Meselson and Stahl grew E. coli for many rounds of replication in a 15N medium, which caused the bacteria to gain weight as they absorbed the heavier isotope into their DNA.

When they were confident that the bacterial DNA was saturated with 15N, they transferred these heavy *E. coli* cells to a less dense 14N medium.

**STOP and Think:** What do you think happened when the “heavy” *E. coli* replicated in the “light” 14N medium?

Meselson and Stahl had rejected the conservative and dispersive hypotheses of replication, and yet they wanted to make sure that the semiconservative hypothesis was confirmed by further E. coli replication. This model predicted that after three rounds of replication, one-quarter of the DNA molecules should still have a 15N strand, causing 25% of the DNA to have an intermediate density, whereas the remaining 75% should be lighter, having only 14N. This is indeed what Meselson and Stahl witnessed in the lab, and the semiconservative hypothesis has stood strong to this day.