# Biology Meets Programming: Bioinformatics for Beginners

Week 1

1. Where in the Genome Does Replication Begin? (Part 1)

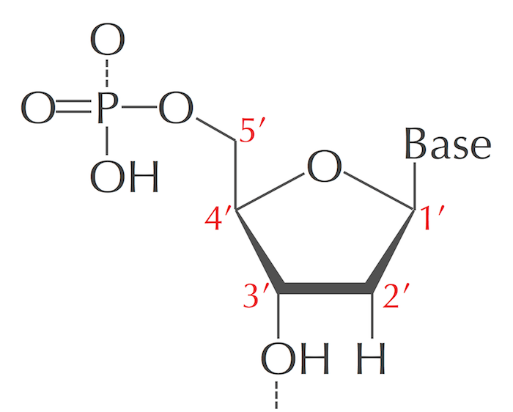
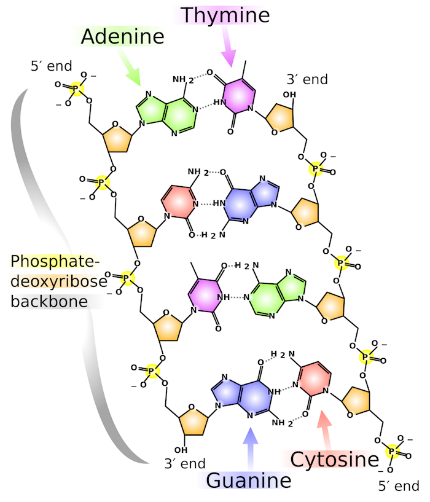
Replication begins in a genomic region called the replication origin (denoted ***ori***) and is performed by molecular copy machines called DNA polymerases. Locating ori presents an important task not only for understanding how cells replicate but also for various biomedical problems. For example, some 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.

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. 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.

Research has shown that the region of the bacterial genome encoding ori is typically a few hundred nucleotides long. 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.

Experiments have revealed that bacterial DnaA boxes are usually 9 nucleotides long. 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.

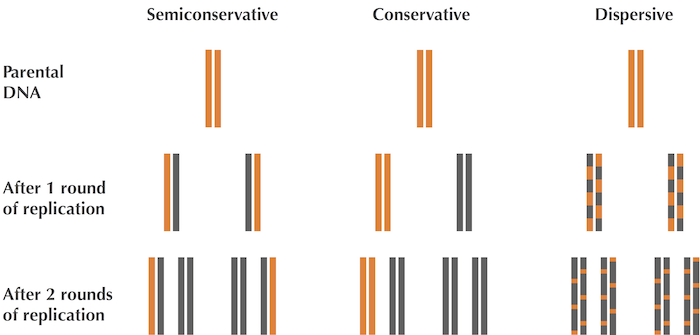
The beginning and end of a DNA strand are denoted 5’ (pronounced “five prime”) and 3’ (pronounced “three prime”), respectively. Each DNA strand is read in the 5' 🡪 3' direction, and the complementary strand runs in the opposite direction to the template strand.

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

"ATGATCAAG"와 그 역상보서열은 *Vibrio cholerae*의 전체 genome에서 17번 나타나지만 ori 영역에서만 서로 가까운 위치에 나타난다. 하지만 이는 다른 박테리아에서도 동일한 것은 아니다.

DNA replication – semiconservative model

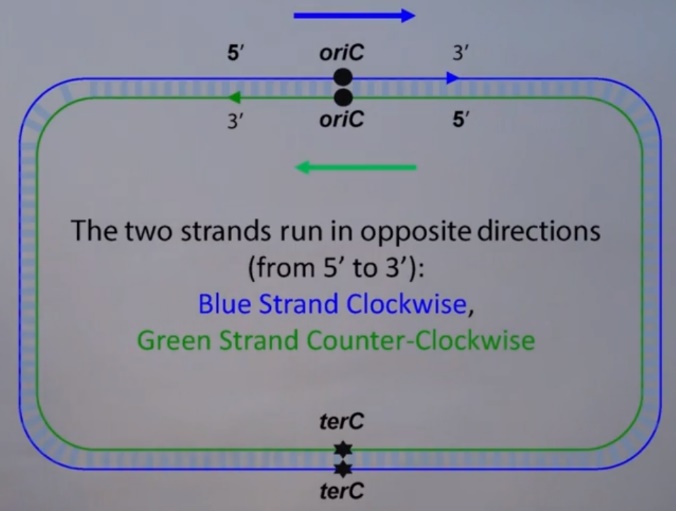


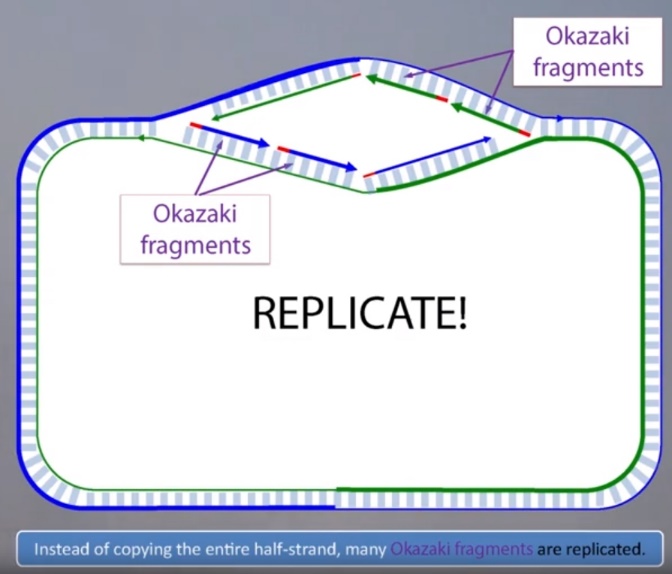
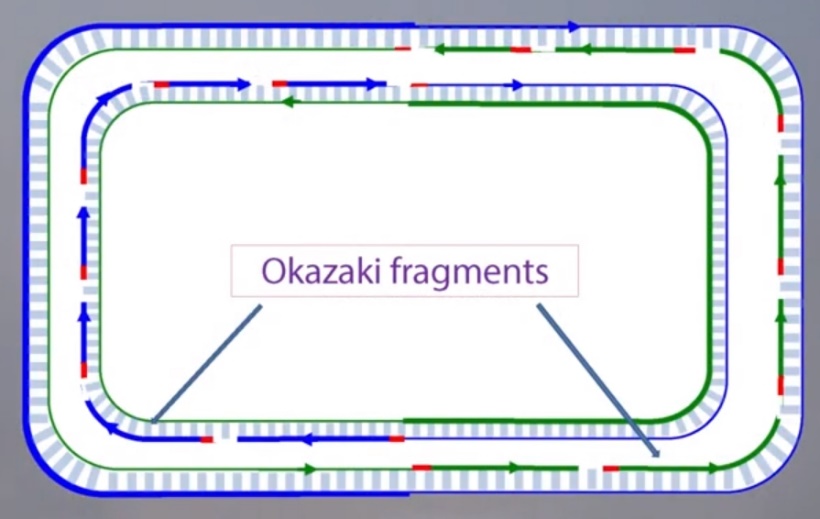
-------------------------------------------------------------------------------------------

Week 2

1. Where in the Genome Does Replication Begin? (Part 2)

(1) Asymmetry of Replication



The two complementary DNA strands running in opposite directions around a circular chromosome unravel, starting at *ori*. As the strands unwind, they create two replication forks, which expand in both directions around the chromosome until the strands completely separate at the replication terminus (denoted *ter*). The replication terminus is located roughly opposite to ori in the chromosome.

An important thing to know about replication is that a DNA polymerase does not wait for the two parent strands to completely separate before initiating replication; instead, it starts copying while the strands are unraveling.

Since a DNA polymerase can only move in the reverse (3' 🡪 5') direction, it can copy nucleotides non-stop from ori to ter along reverse half-strands. However, replication on forward half-strands is very different because a DNA polymerase cannot move in the forward (5' 🡪 3') direction; on these half-strands, a DNA polymerase must replicate backwards toward ori.

Finally, consecutive Okazaki fragments must be sewn together by an enzyme called DNA ligase, resulting in two intact daughter chromosomes, each consisting of one parent strand and one newly synthesized daughter strand.

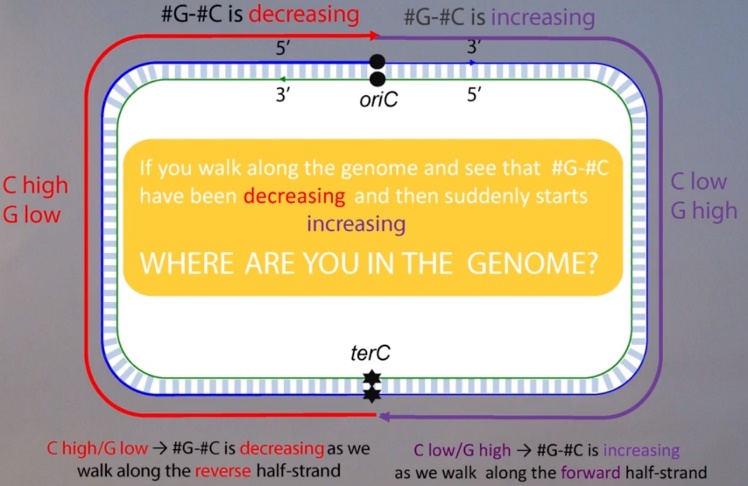
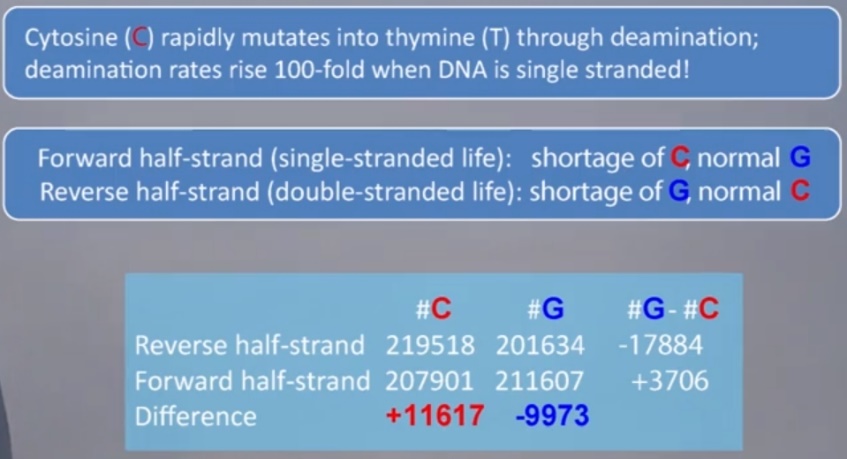
(2) Peculiar Statistics of the Forward and Reverse Half-Strands : Skew Diagram

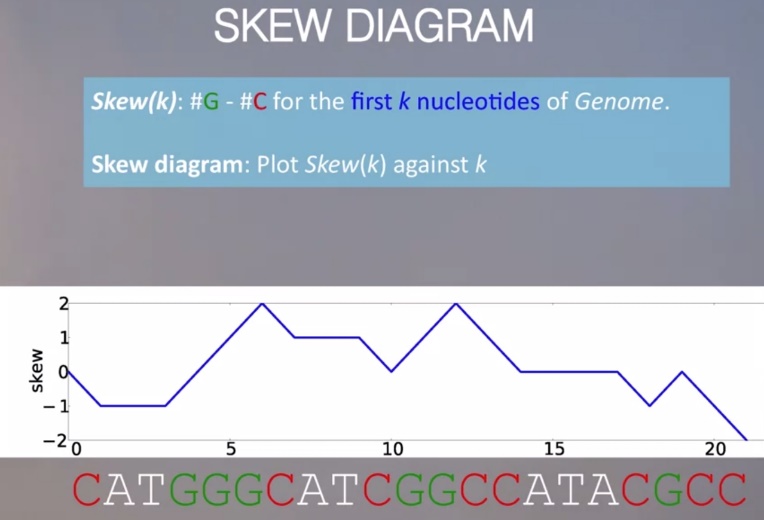
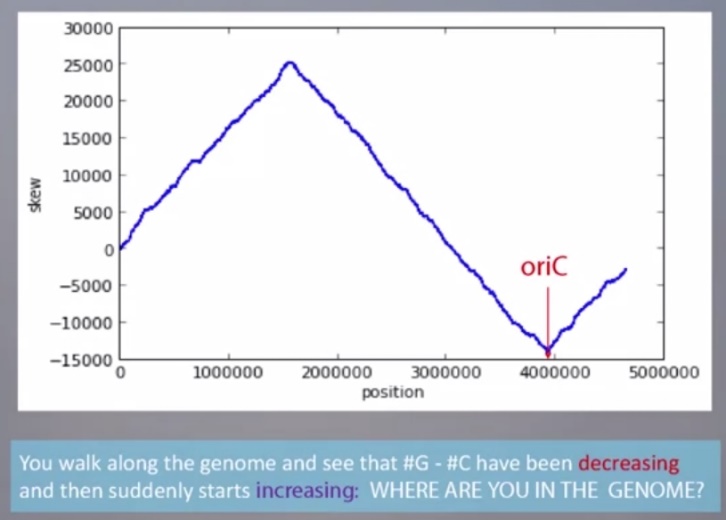
DNA polymerase synthesizes DNA quickly on the reverse half-strand but suffers delays on the forward half-strand.

정방향과 역방향 가닥의 복제 속도가 서로 다름 🡪 ori 영역을 찾는 G-C mutation rate 다름 🡪 ori 위치 탐색.

Cytosine (C) has a tendency to mutate into thymine (T) through a process called deamination. Deamination rates rise 100-fold when DNA is single-stranded, which leads to a decrease in cytosine (C) on the forward half-strand.

\* Thermotoga petrophila genome



skew 최소값 🡪 *ori* 영역의 위치

그림은 *E. coli*에 대한 skew diagram. 최대값은 위치 1600000 주변에서 발생하고 최소값은 위치 4000000 주변에서 발생. 따라서 역방향 절반 가닥은 위치 1600000 주변에서 시작하고 순방향 절반 가닥은 위치 4000000 주변에서 시작한다고 추론 할 수 있다. 즉 E. coli 게놈의 *ori*가 4000000 위치 근처에 있다는 것을 발견.

DnaA can bind not only to “perfect” DnaA boxes but to their slight modifications as well.

- ATGATCAAG and CTTGATCAT

- ATGATCAAC and CATGATCAT (additional occurrences)

두 문자열 p와 q 사이의 총 불일치 수 = Hamming distance.

Hamming distance를 이용해 약간 다른 조합 k-mer로 DnaA box를 발견할 수 있다.

\* Compications

- Some bacteria have fewer *DnaA* boxex.

- Terminus of replication is often not located directly opposite to *oriC*.

- The skew diagram is often more complex than in the case of *E. coli*.

