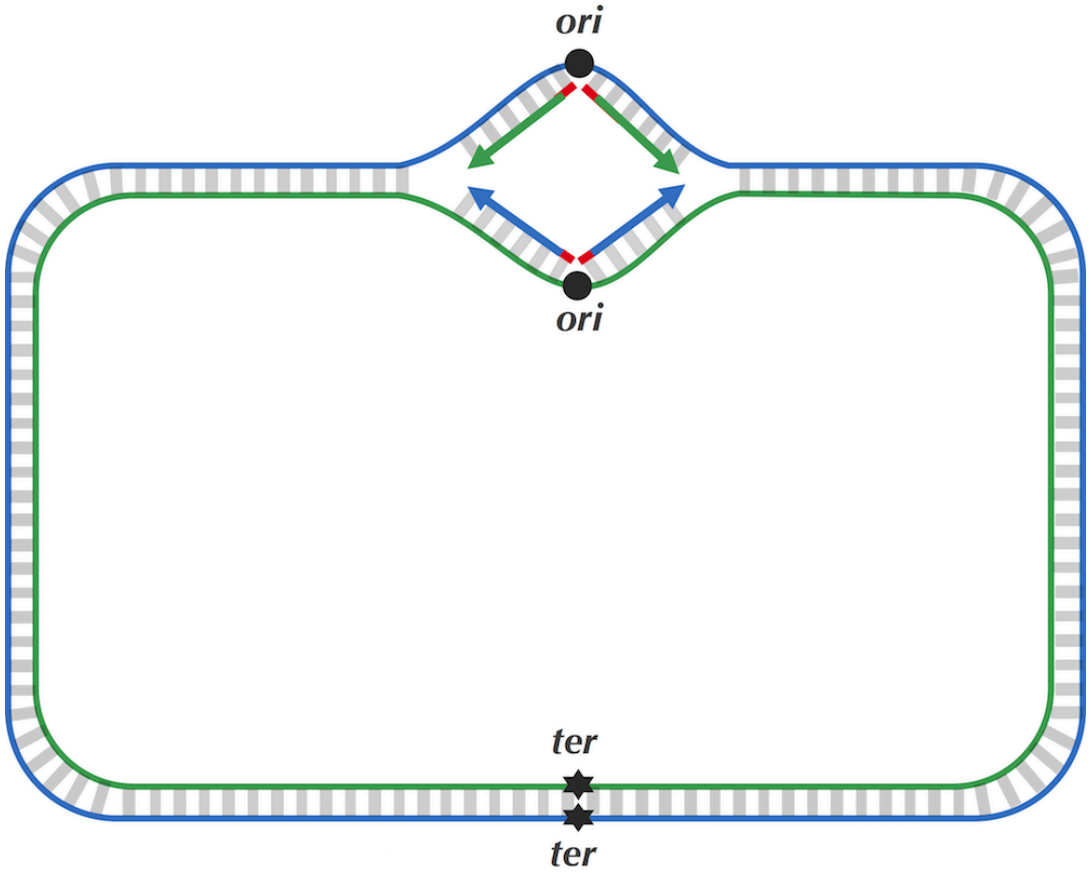
CMSC 423:
Finding Biological Signals

Part 4

Problem: Finding Hidden Messages in the Replication of Origin



- **Input**:

A string *Text* (representing the replication origin of a genome).

- **Output**:

A hidden message in *Text*.

DnaA: protein that binds to a short segment within the *ori* to begin replication

***DnaA* box**: where *DnaA* binds, the "hidden" message within the *ori*

ori of *Vibrio cholerae*

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

k	3	4	5	6	7	8	9
count	25	12	8	8	5	4	3
***k*-mers**	tga	atga	gatca	tgatca	atgatca	atgatcaa	atgatcaag
			tgatc				cttgatcat
							tcttgatca
							ctcttgatc

Are any of the most frequent 9-mers in the *ori* more surprising than the others?

```
atgatcaag
cttgatcat
tcttgata
ctcttgatc
```

Knowledge of biology is helpful

- DNA is double-stranded

Are any of the most frequent 9-mers in the *ori* more surprising than the others?

atgatcaag
cttgatcat
tcttgata
ctcttgatc

Knowledge of biology is helpful

- DNA is double-stranded
- *k*-mer can occur in either strand
- Algorithms stay the same, but need to run twice (for each strand + it's reverse complement)

DnaA box in *Vibrio cholerae*

atgatcaag/cttgatcat

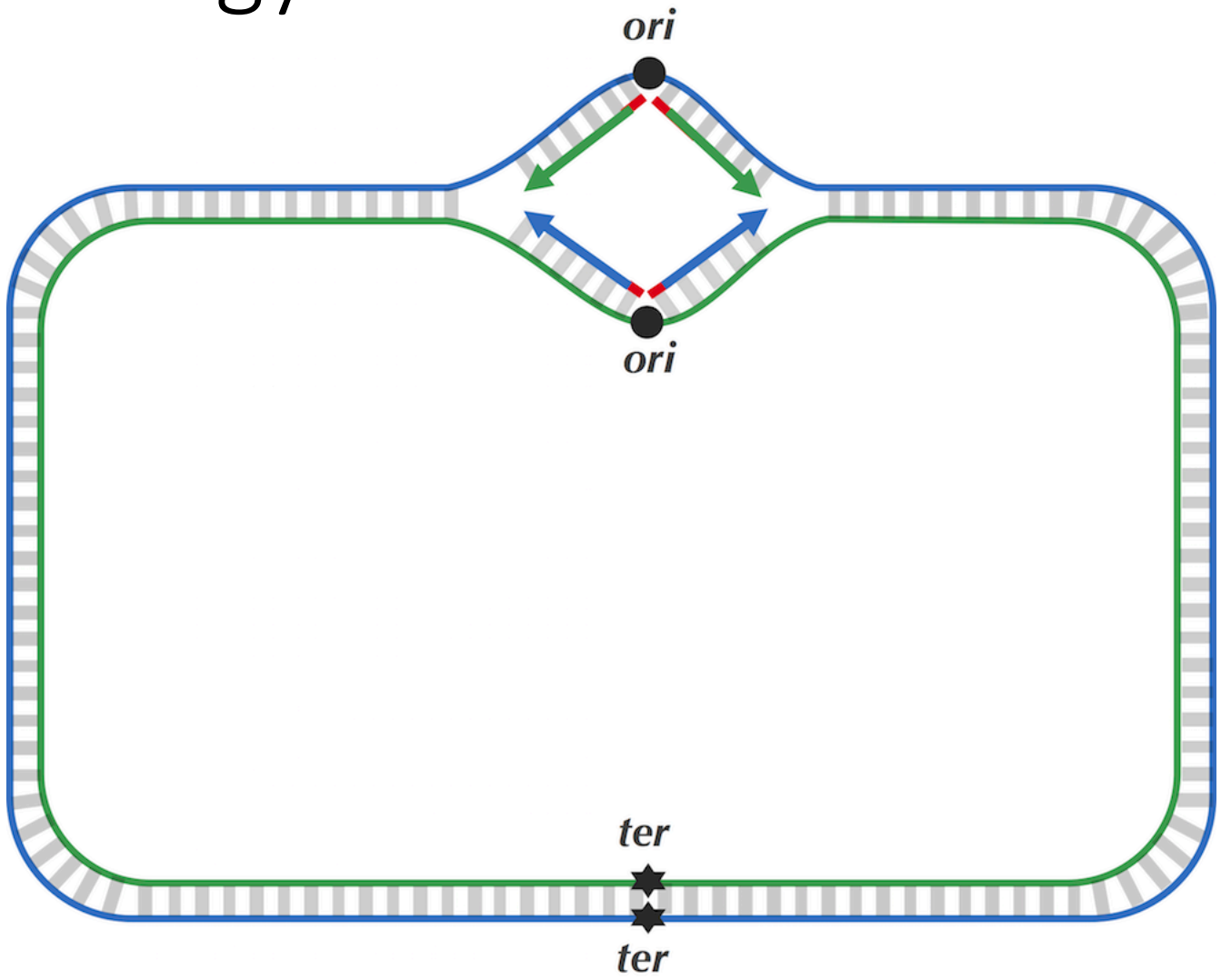
atgatcaag occurs at 17 spots in the genome:

116556, 149355, **151913**, **152013**, **152394**, 186189, 194276, 200076, 224527, 307692, 479770, 610980, 653338, 679985, 768828, 878903, 985368

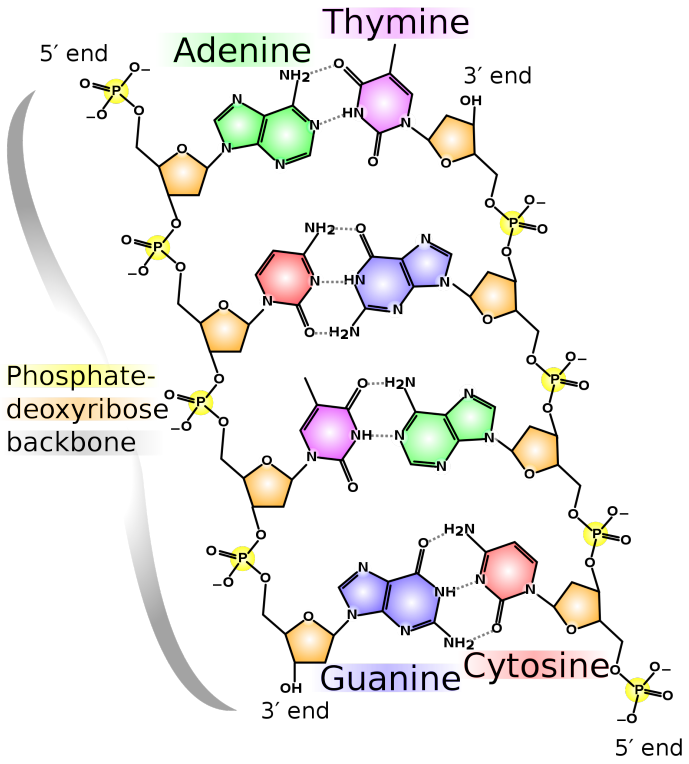
Clump Finding

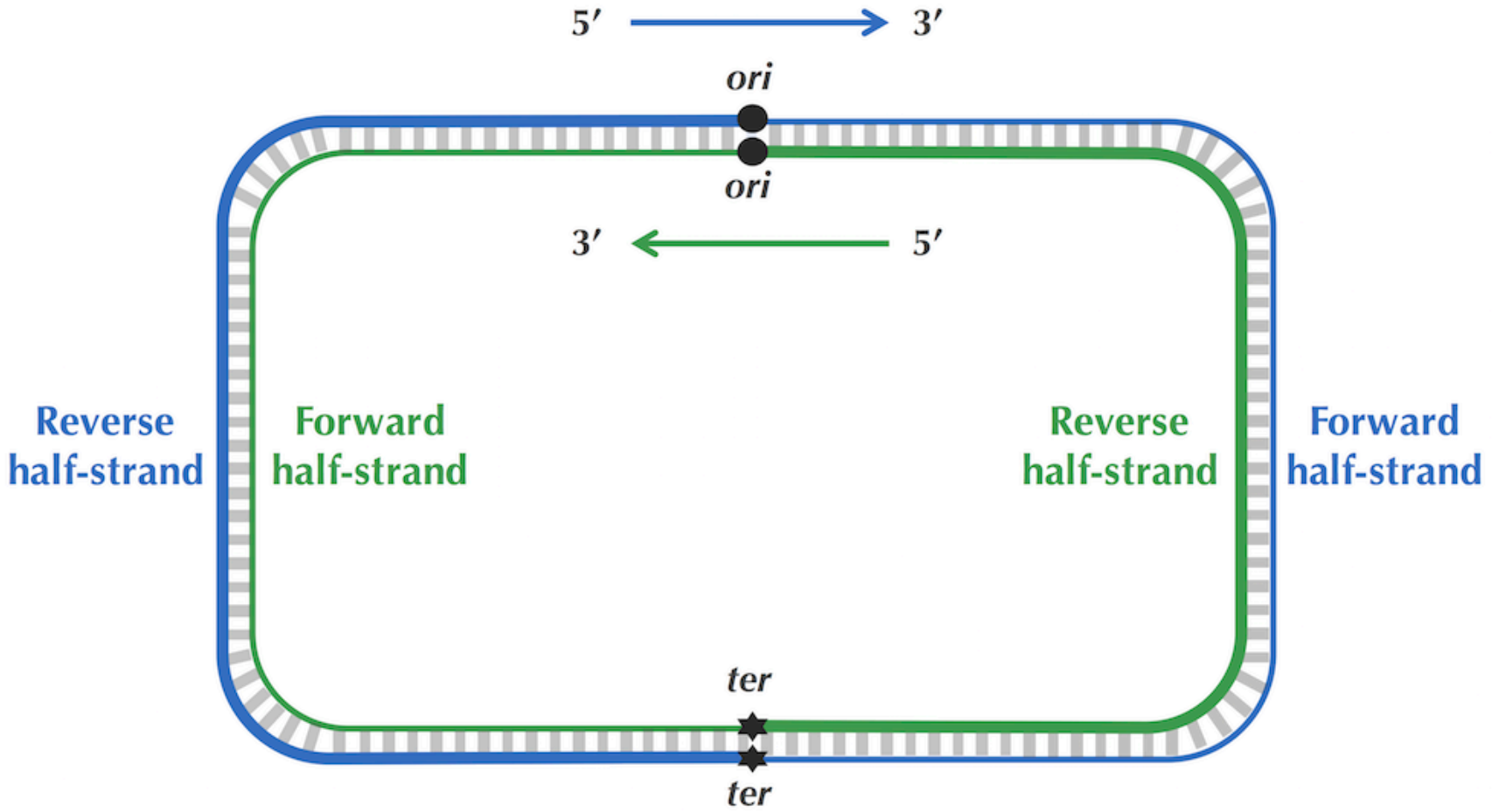
- Find every *k*-mer that forms a “clump” in the genome
- *L* = window of fixed length being searched
- (*L*, *t*)-clump = a *k*-mer that appears at least *t* times in a window of size *L*
  - `atgatcaag` forms a (500, 3)-clump in Vibrio cholerae
- Doesn’t always work– we find hundreds of 9-mers in E coli forming (500, 3)-clumps

More biology



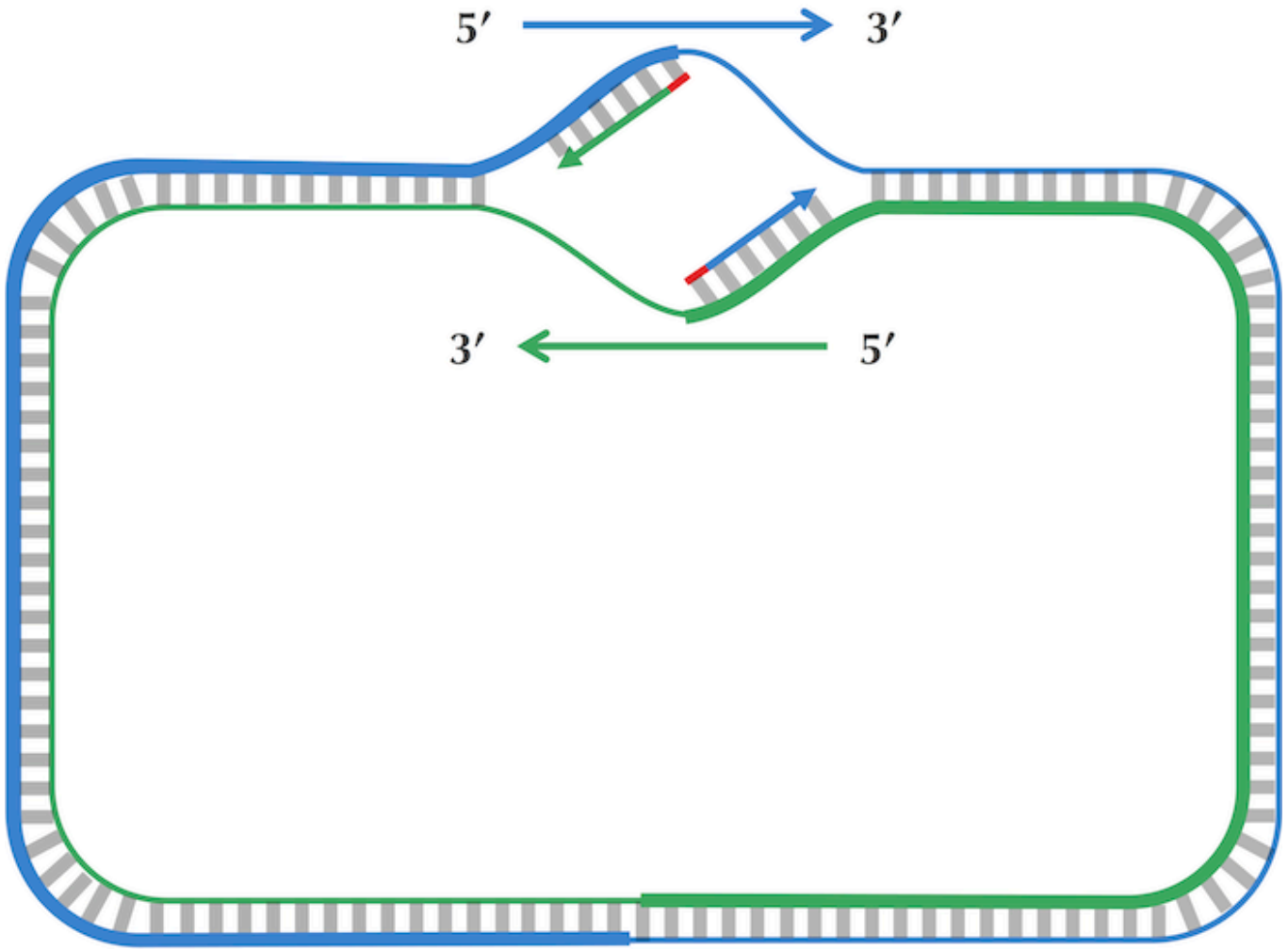
More biology: DNA polymerases are unidirectional



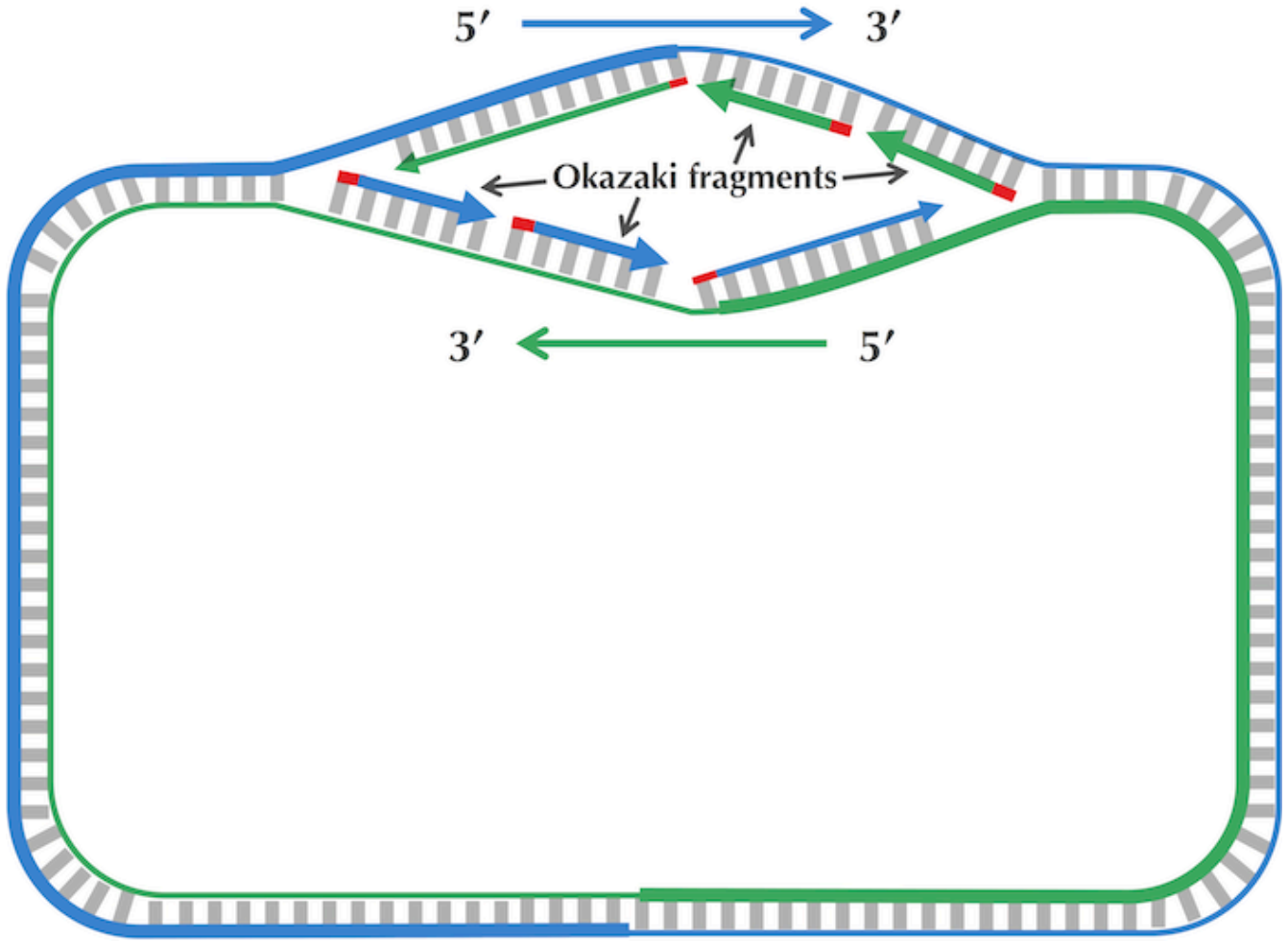


https://en.wikipedia.org/wiki/Directionality_(molecular_biology)

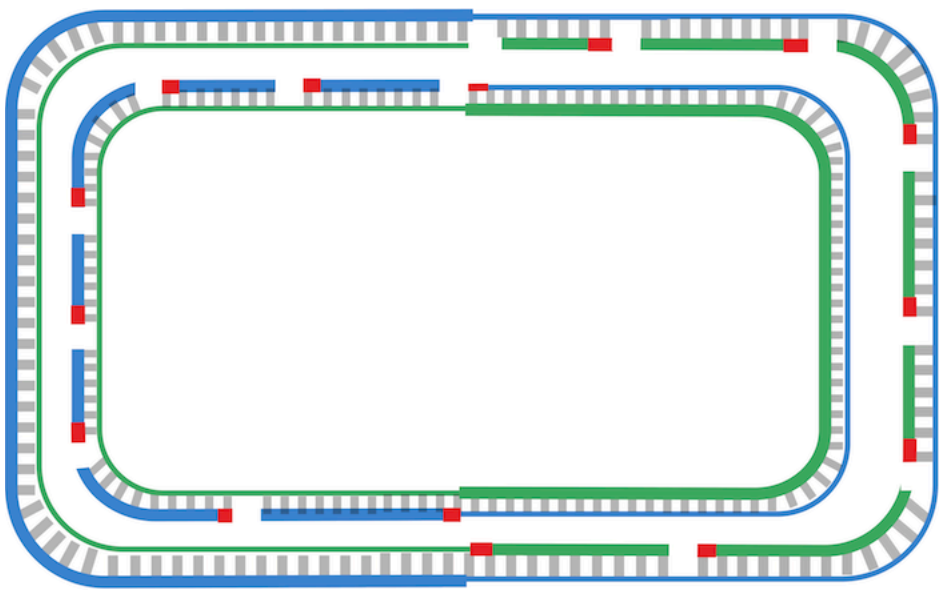
More biology: Replication is asymmetric

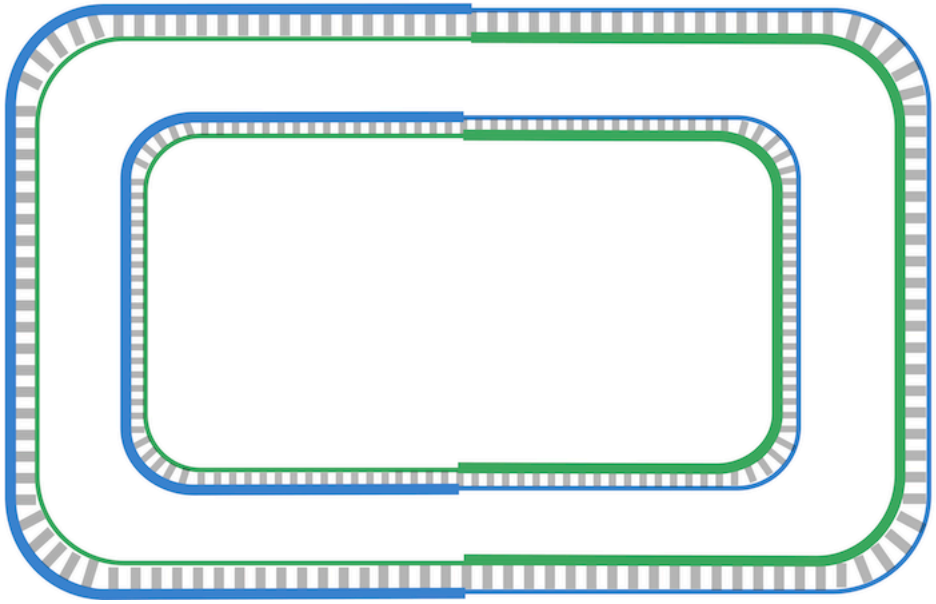


More biology: Replication is asymmetric



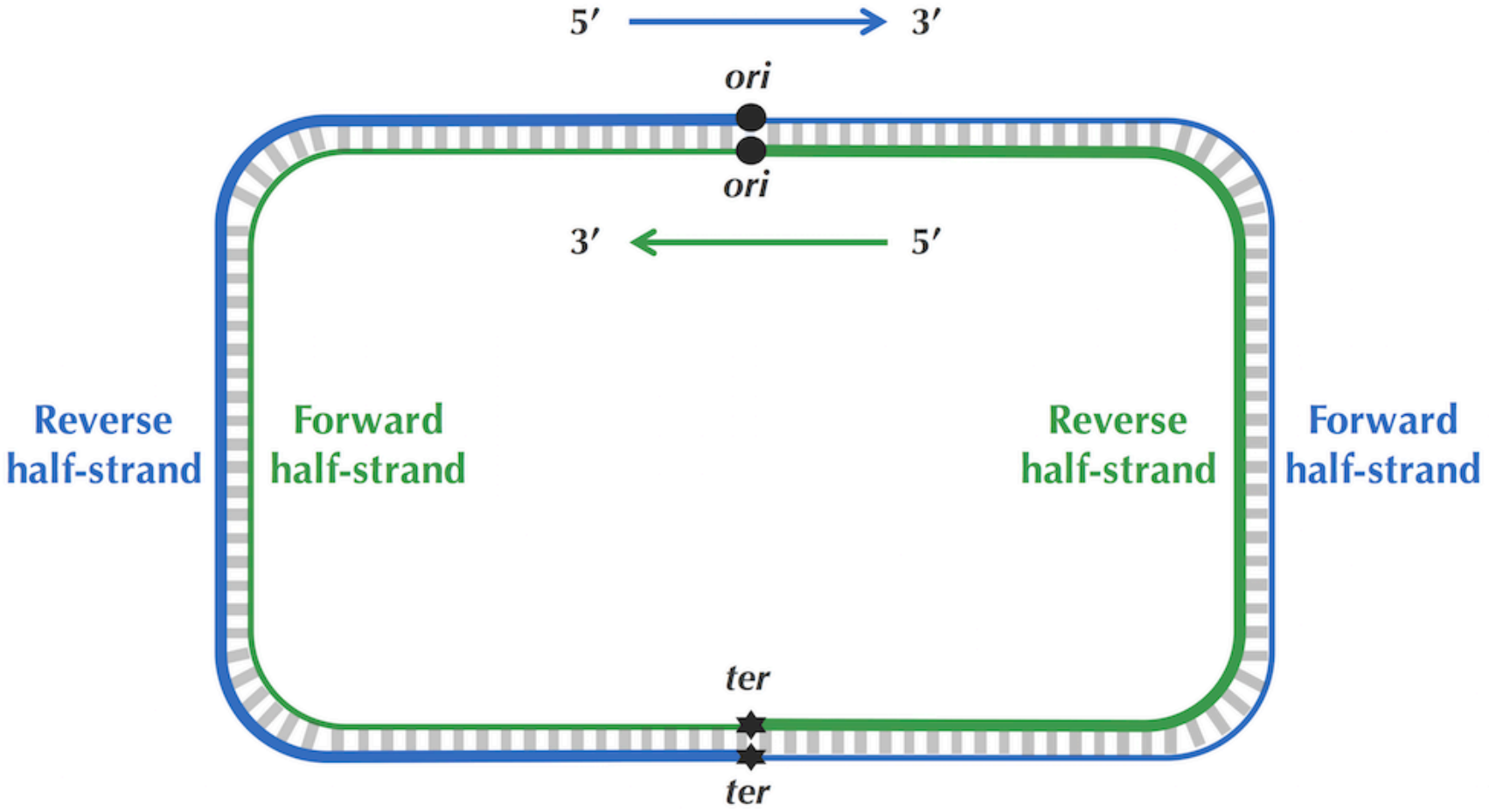
More biology: Replication is asymmetric





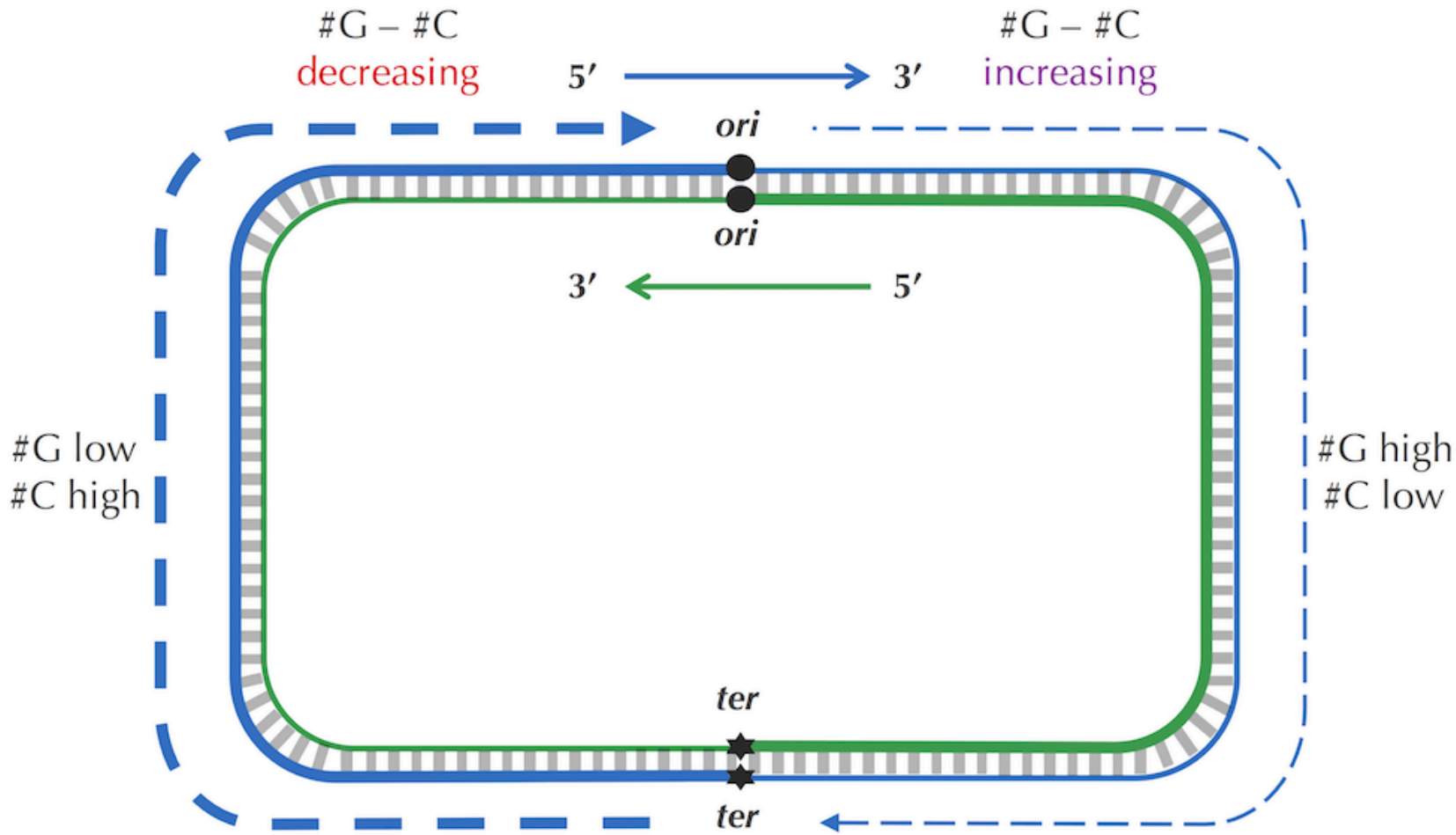
More biology: Replication is asymmetric

Deamination: C -> T mutation occurs more frequently in single stranded DNA
Also occurs with time (ancient DNA)



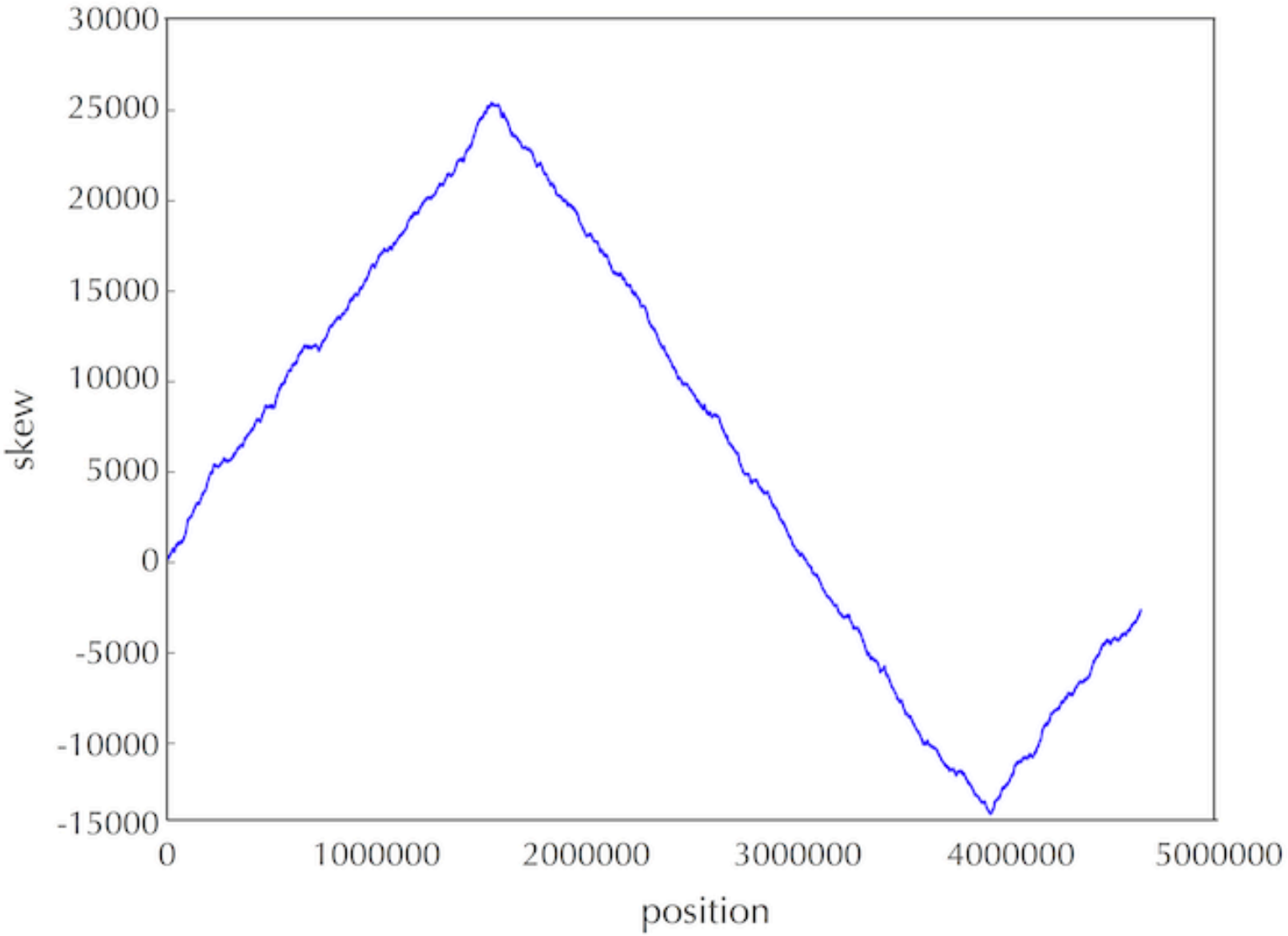
Deamination occurs more on the forward half-strand

Use “skew” between G and C to find the *ori*



Use “skew” between G and C to find the *ori*

ori = where skew attains a minimum



ori of *Vibrio cholerae*

atcaATGATCAACgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagATGATCAAGagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagCATGATCATggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctCTTGATCATcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctCTTGATCATgtt
tccttaaccctctatttttacggaagaATGATCAAGctgctgctCTTGATCATcgtttc

Knowledge of biology is helpful

- DNA is double-stranded
- *k*-mer can occur in either strand
- Algorithms stay the same, but need to run twice (for each strand + it's reverse complement)
- Algorithms need to account for mismatches

Later in the class

Finding a pattern efficiently

Finding patterns with mismatches/errors

How fast?