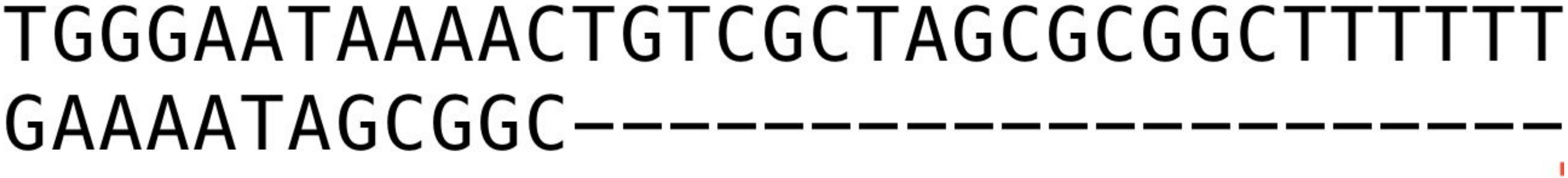
DNA Detectives: Data Cleaning!

Week 2

What’s wrong with this picture?

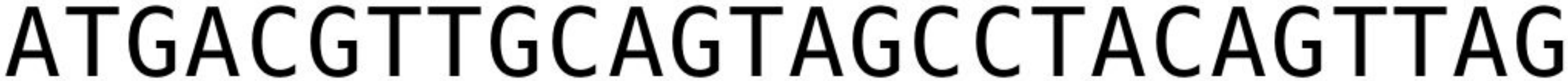


```
TGGGAATAAAACTGTCGCTAGCGCGGCTTTTTT
GAAAATAGCGGC---------------------
```

What’s wrong with this picture?

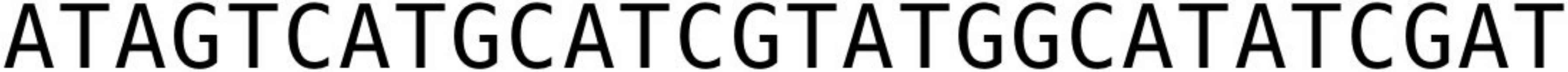
```
AATACCAACTCCRRTMYCTGCCCTTANTCCATGCTGCGACTATGG
AACAAGCCCCGTTGGCTTTGATGATAGTGACCATTATAGGGTCTT
TAAAAAAAGTGATCGCTTACTCGACTTGTAGTCAGTTGGGGTATA
ACTTAATGAACCATGCTTTTTTTAAGGCTTTATTATTCTTAAGCG
AAATGGGGGGGTTAATAAAGTCCATTCCCCTTACTTACACCATGG
GTTTCTATTCTAAAGATTTAATTTTAGAGTTGGCCTATGATCAAT
TAACAGCCTTTTATTCAATCCGATTGGTTTATTTAACTTTTATAA
GTTCTTGGAATTTAACCCTACCCTTGATATTATTAGCCTTGGGGA
```

What’s wrong with this picture?



ATGACGTTGCAGTAGCCTACAGTTAG

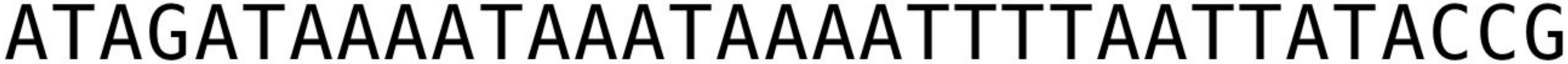
What’s wrong with this picture? (Assume that this is a full-length sequence.)



ATAGTCATGCATCGTATGGCATATCGAT

Hint: recall the error that Biopython gave some of you when you tried to translate your sequence objects last week!

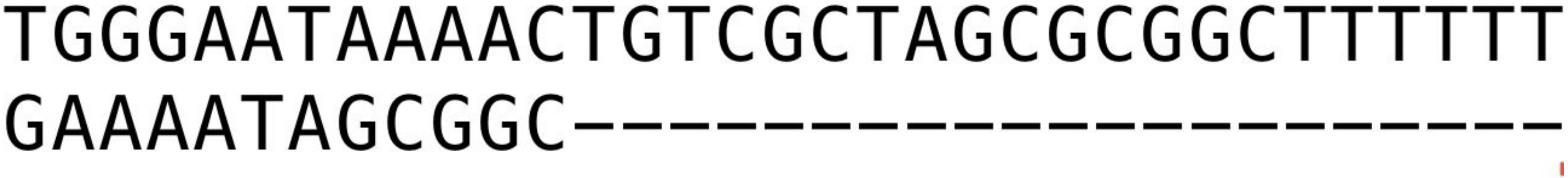
What’s wrong with this picture?



ATAGATAAAATAAATAAAATTTTAATTATACCG

the reveal...

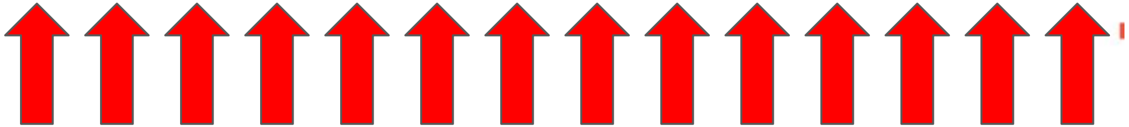
What’s wrong with this picture?



```
TGGGAATAAAACTGTCGCTAGCGCGGCTTTTTT
GAAAATAGCGGC---------------------
```

What’s wrong with this picture?

```
TGGGAATAAAACTGTCGCTAGCGCGGCTTTTTT
GAAAATAGCGGC---------------------
```



What’s wrong with this picture?

```
AATACCAACTCCRRTMYCTGCCCTTANTCCATGCTGCGACTATGG
AACAAGCCCCGTTGGCTTTGATGATAGTGACCATTATAGGGTCTT
TAAAAAAAGTGATCGCTTACTCGACTTGTAGTCAGTTGGGGTATA
ACTTAATGAACCATGCTTTTTTTAAGGCTTTATTATTCTTAAGCG
AAATGGGGGGGTTAATAAAGTCCATTCCCCTTACTTACACCATGG
GTTTCTATTCTAAAGATTTAATTTTAGAGTTGGCCTATGATCAAT
TAACAGCCTTTTATTCAATCCGATTGGTTTATTTAACTTTTATAA
GTTCTTGGAATTTAACCCTACCCTTGATATTATTAGCCTTGGGGA
```

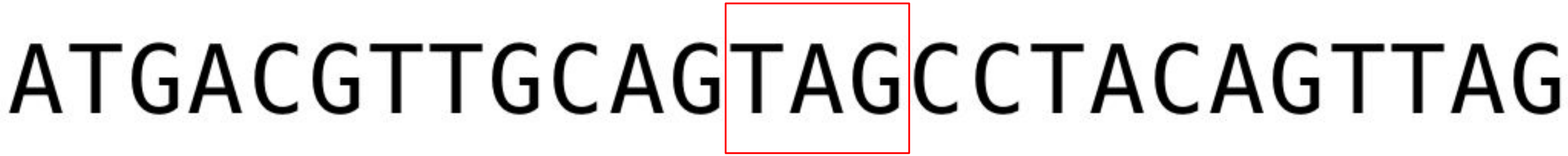
What’s wrong with this picture?

```
AATACCAACTCCRRTMYCTGCCCTTANTCCATGCTGCGACTATGG
AACAAGCCCCGTTGGCTTTGATGATAGTGACCATTATAGGGTCTT
TAAAAAAAGTGATCGCTTACTCGACTTGTAGTCAGTTGGGGTATA
ACTTAATGAACCATGCTTTTTTTAAGGCTTTATTATTCTTAAGCG
AAATGGGGGGTTAATAAAGTCCATTCCCCTTACTTACACCATGG
GTTTCTATTCTAAAGATTTAATTTTAGAGTTGGCCTATGATCAAT
TAACAGCCTTTTATTCAATCCGATTGGTTTATTTAACTTTTATAA
GTTCTTGGAATTTAACCCTACCCTTGATATTATTAGCCTTGGGGA
```

What’s wrong with this picture?

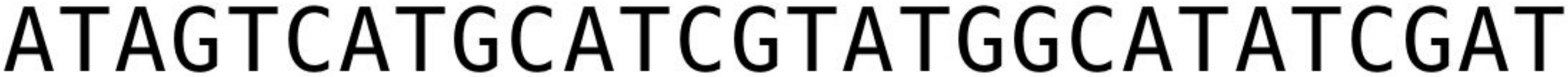
ATGACGTTGCAGTAGCCTACAGTTAG

What’s wrong with this picture?



ATGACGTTGCAGTAGCCTACAGTTAG

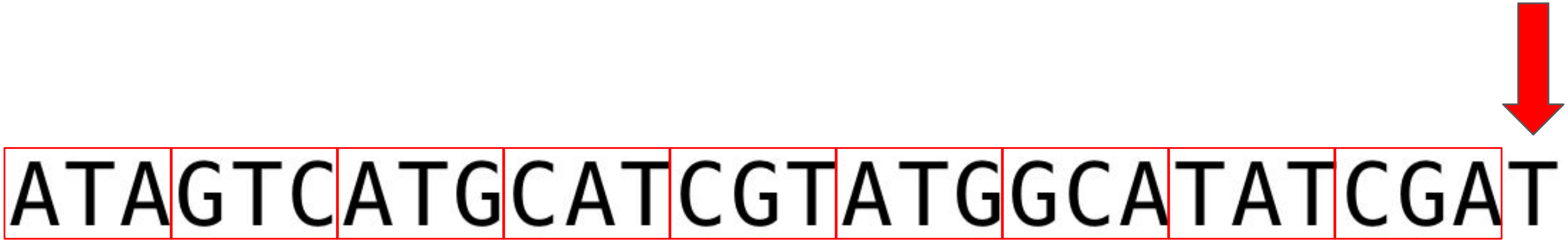
What’s wrong with this picture? (Assume that this is a full-length sequence.)



ATAGTCATGCATCGTATGGCATATCGAT

Hint: recall the error that Biopython gave some of you when you tried to translate your sequence objects last week!

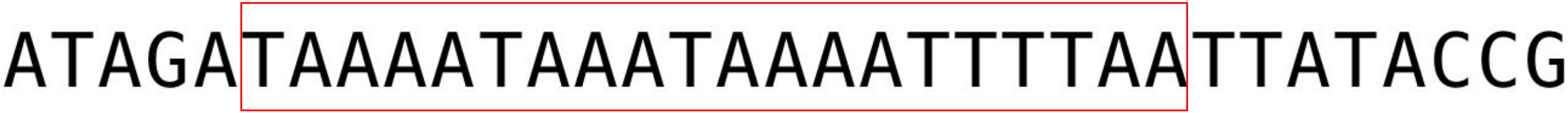
What’s wrong with this picture? (Assume that this is a full-length sequence.)



ATA GTC ATG CAT CGT ATG GCA TAT CGA T

Hint: recall the error that Biopython gave some of you when you tried to translate your sequence objects last week!

What’s wrong with this picture?



ATAGATAAAATAAATAAAATTTTAATTATACCG

Our findings:

- Missing nucleotide bases
- Bizarre IUPAC codes
- Stop codon in middle of sequence
- Sequence length not a multiple of 3
- Low complexity region

Our solution:

- As Biopython experts, we’ll be designing fixes to these problems.

- And so begins the **data cleaning** phase of our research!