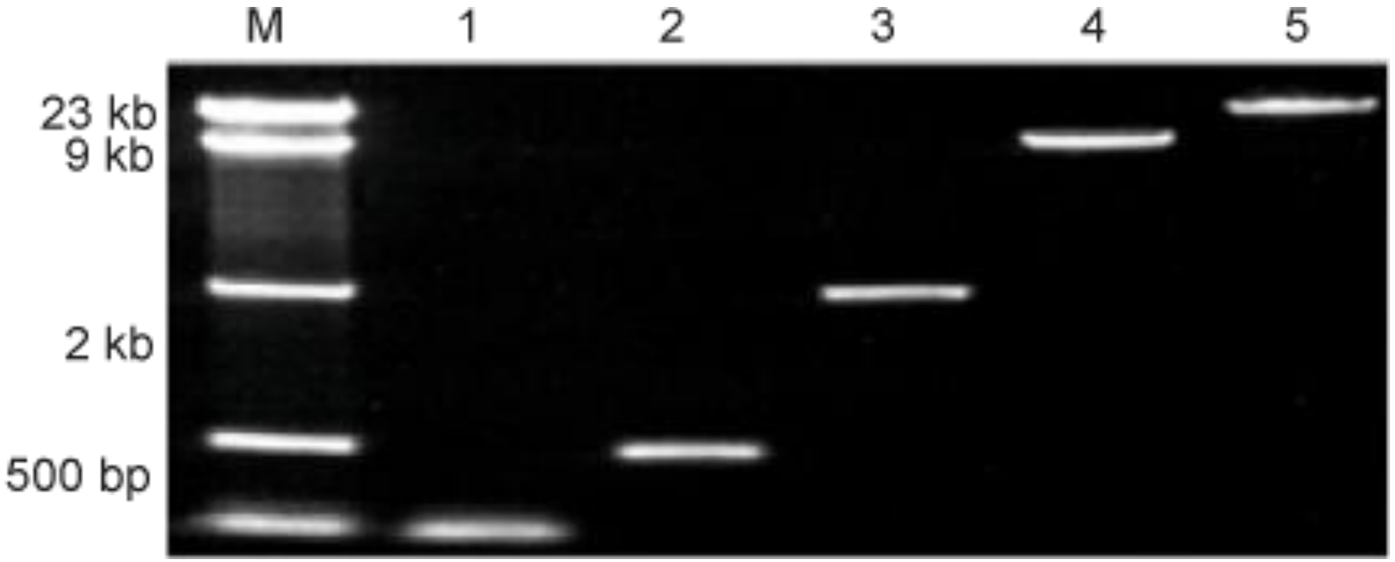
Quantitative Cellular and Molecular Biology Laboratory
Computational Biology Department
Comp Bio 02-261
Spring 2019

Lab 2 – Molecular Biology Computational Lab

January 25, 2019

Gel Electrophoresis

- Experimental method to determine distribution of DNA strand sizes in DNA sample.
- More details in the next lecture…

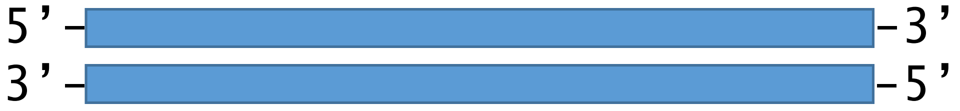


DNA Notation (primary and secondary structure)

ssDNA = 5’-ACTGCGATAGACGATGTCCGGATGACA-3’ ← Shows sequence

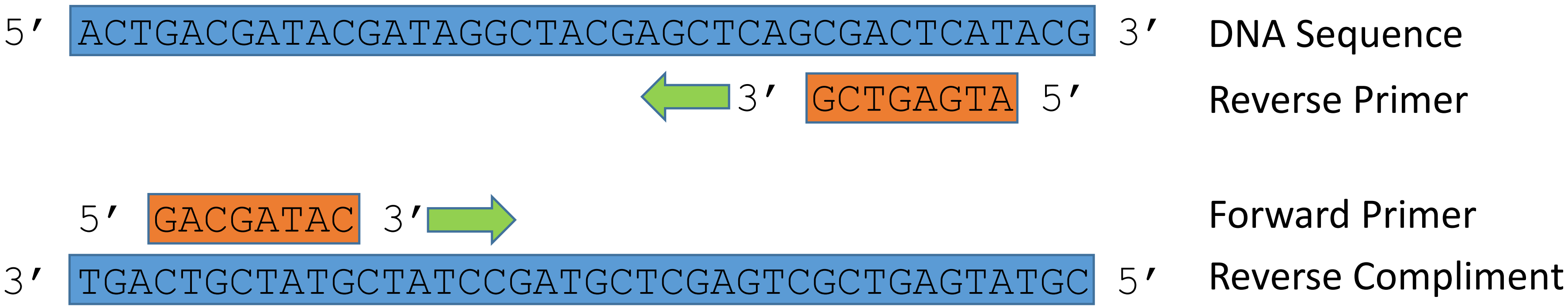
dsDNA = 5’-ACTGCGATAGACGATGTCCGGATGACA-3’
3’-TGACGCTATCTGCTACAGGCCTACTGT-5’ ← Shows sequence and pairing

dsDNA =



← Shows pairing

Polymerase Chain Reaction



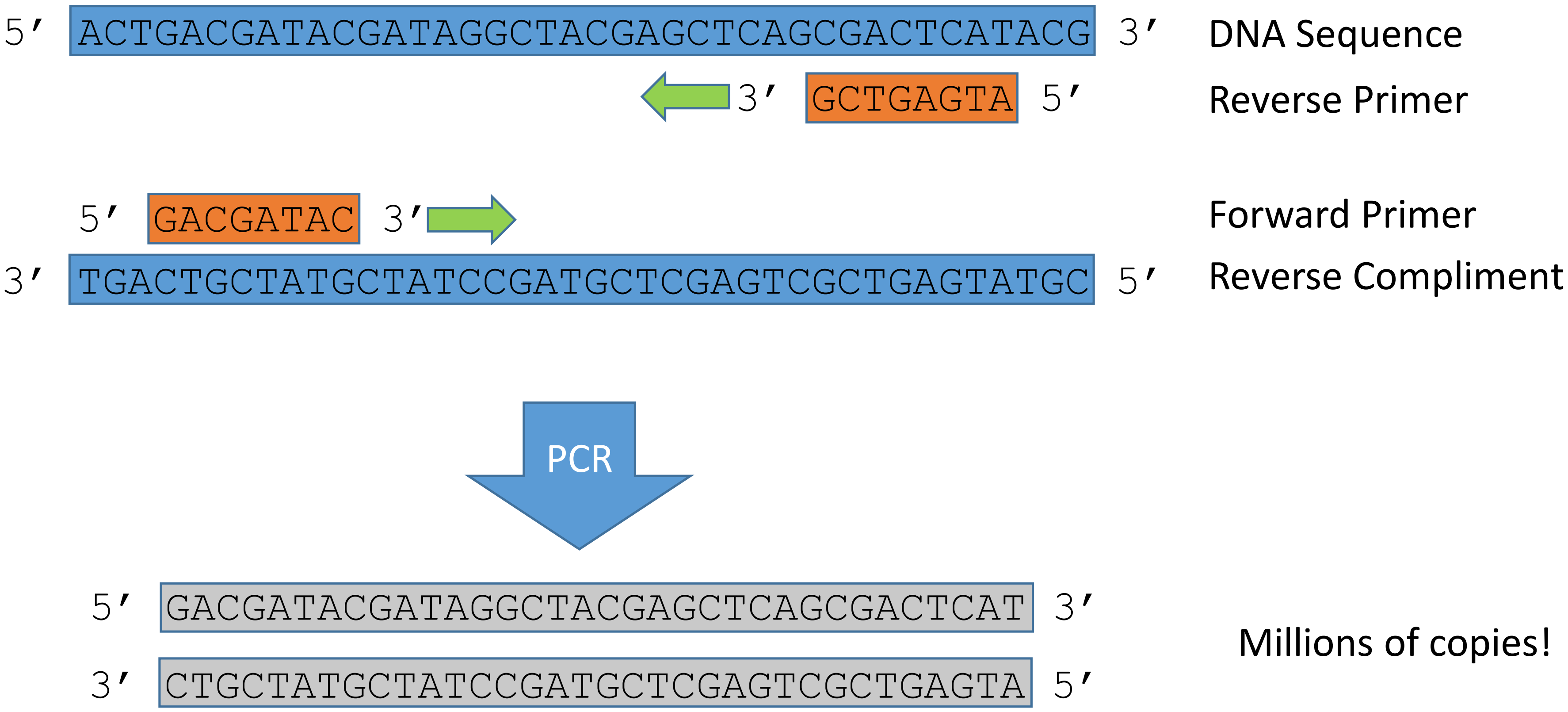
5′ ACTGACGATACGATAGGCTACGAGCTCAGCGACTCATACG 3′ DNA Sequence

3′ GCTGAGTA 5′ Reverse Primer

5′ GACGATAC 3′ Forward Primer

3′ TGACTGCTATGCTATCCGATGCTCGAGTCGCTGAGTATGC 5′ Reverse Compliment

Polymerase Chain Reaction



5′ ACTGACGATACGATAGGCTACGAGCTCAGCGACTCATACG 3′ DNA Sequence

3′ GCTGAGTA 5′ Reverse Primer

5′ GACGATAC 3′ Forward Primer

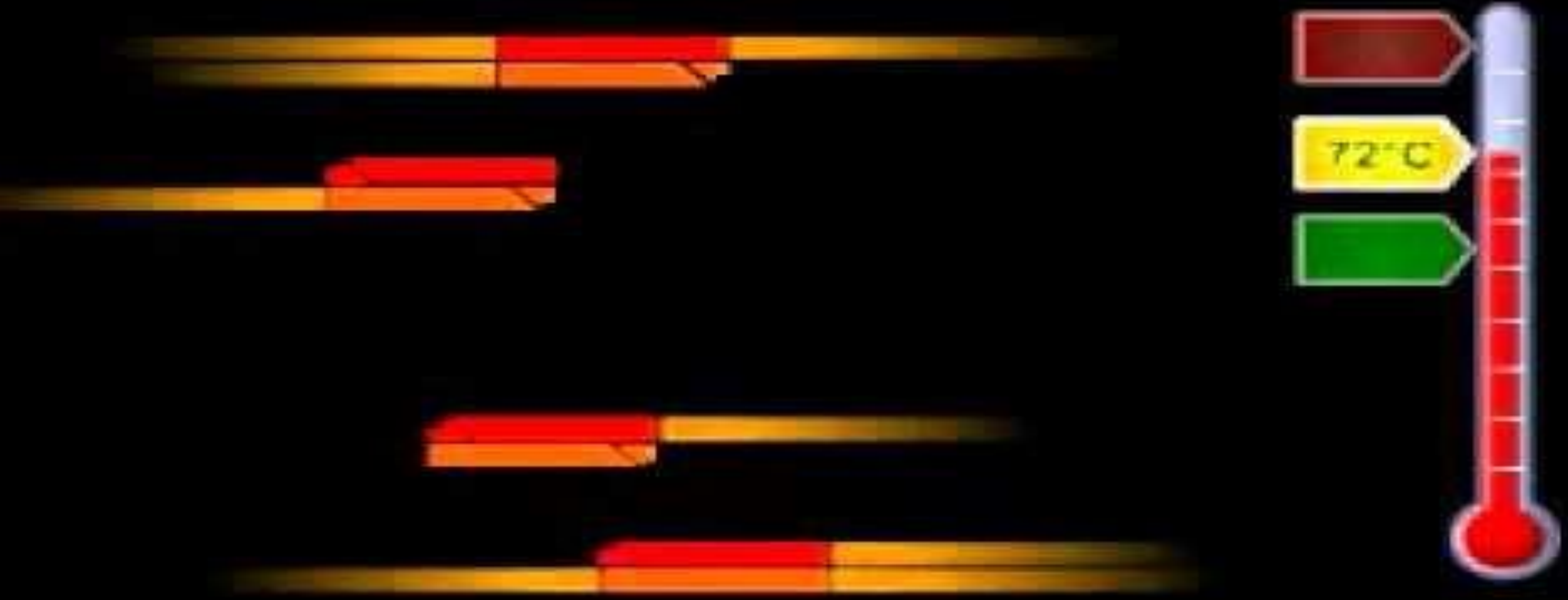
3′ TGACTGCTATGCTATCCGATGCTCGAGTCGCTGAGTATGC 5′ Reverse Compliment

PCR

5′ GACGATACGATAGGCTACGAGCTCAGCGACTCAT 3′

3′ CTGCTATGCTATCCGATGCTCGAGTCGCTGAGTA 5′

Millions of copies!

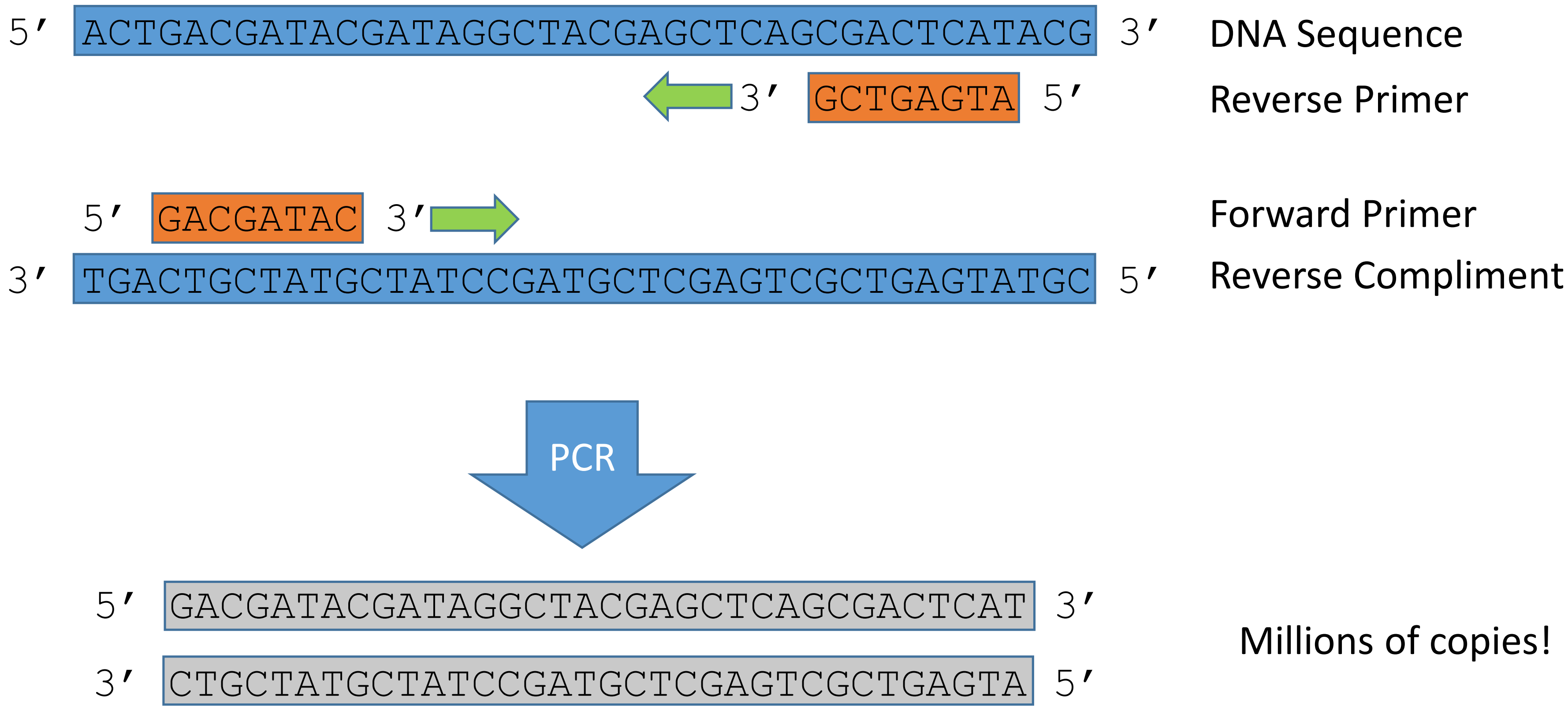


2nd Cycle

Video URL

https://www.youtube.com/watch?v=YJKYSlJREIc

Polymerase Chain Reaction



5′ ACTGACGATACGATAGGCTACGAGCTCAGCGACTCATACG 3′ DNA Sequence

3′ GCTGAGTA 5′ Reverse Primer

5′ GACGATAC 3′ Forward Primer

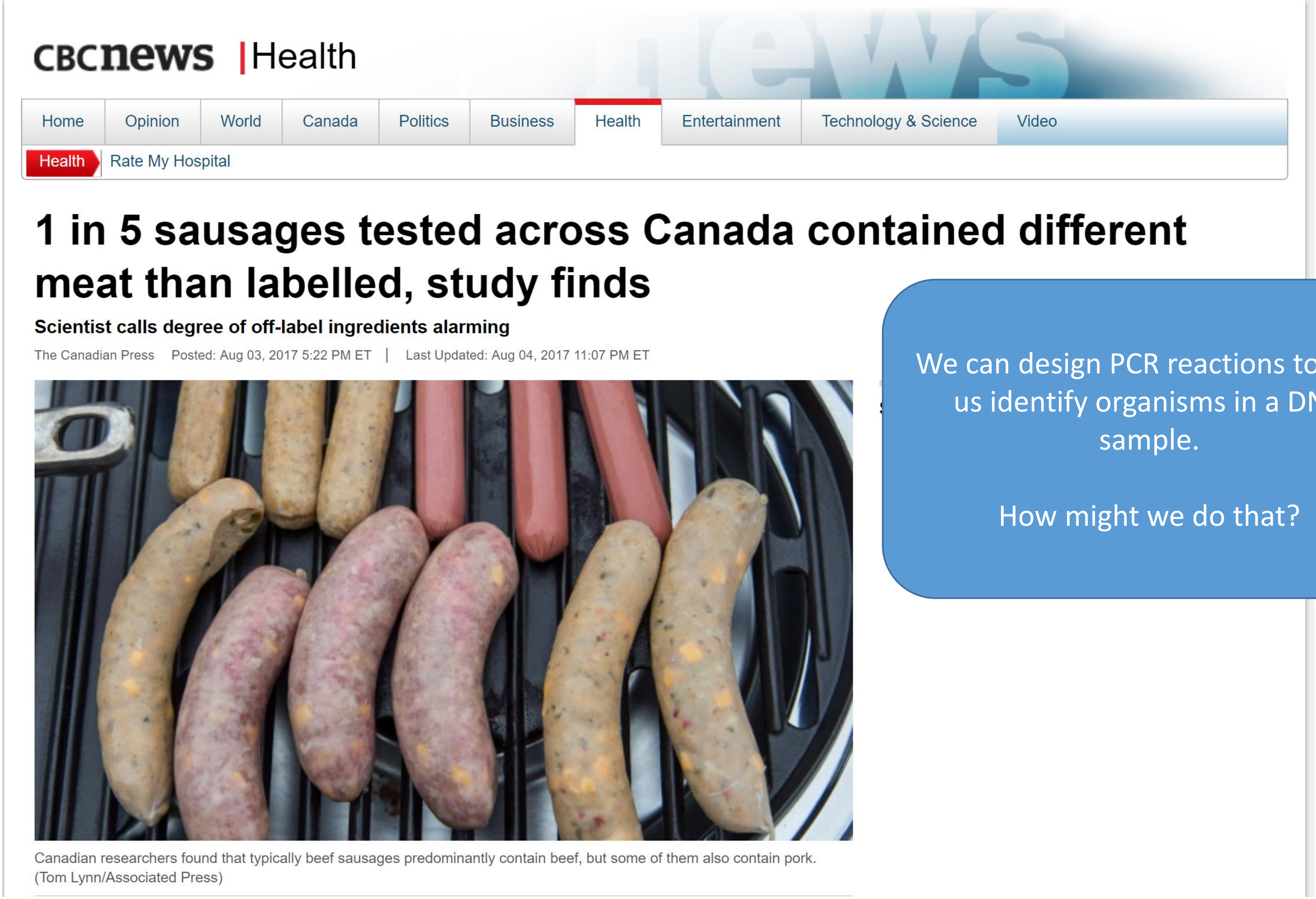
3′ TGACTGCTATGCTATCCGATGCTCGAGTCGCTGAGTATGC 5′ Reverse Compliment

PCR

5′ GACGATACGATAGGCTACGAGCTCAGCGACTCAT 3′

3′ CTGCTATGCTATCCGATGCTCGAGTCGCTGAGTA 5′

Millions of copies!



1 in 5 sausages tested across Canada contained different meat than labelled, study finds

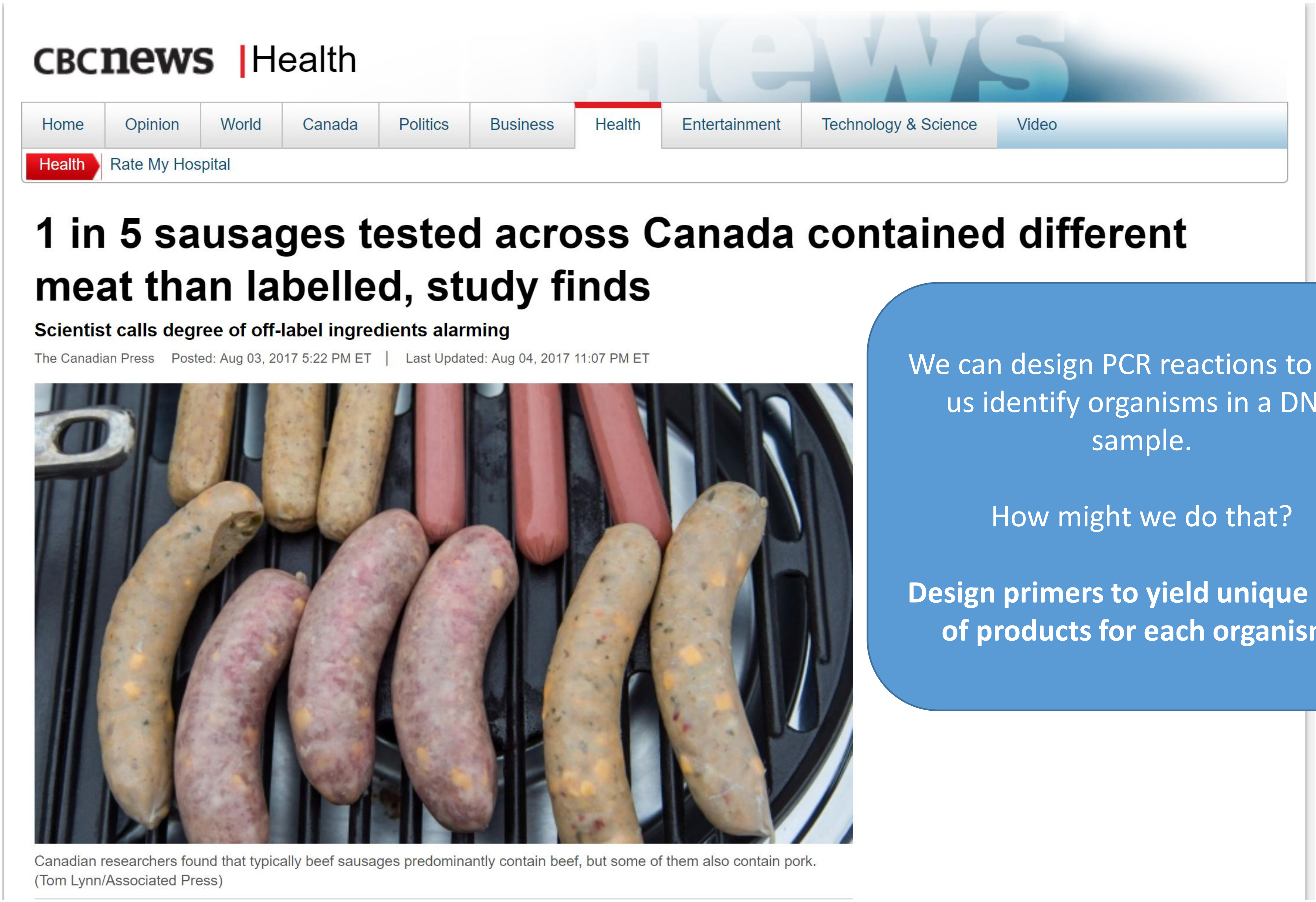
Scientist calls degree of off-label ingredients alarming

The Canadian Press Posted: Aug 03, 2017 5:22 PM ET | Last Updated: Aug 04, 2017 11:07 PM ET

Canadian researchers found that typically beef sausages predominantly contain beef, but some of them also contain pork. (Tom Lynn/Associated Press)

We can design PCR reactions to help us identify organisms in a DNA sample.

How might we do that?



CBCnews | Health

Home | Opinion | World | Canada | Politics | Business | Health | Entertainment | Technology & Science | Video

Health | Rate My Hospital

1 in 5 sausages tested across Canada contained different meat than labelled, study finds

Scientist calls degree of off-label ingredients alarming

The Canadian Press Posted: Aug 03, 2017 5:22 PM ET | Last Updated: Aug 04, 2017 11:07 PM ET

Canadian researchers found that typically beef sausages predominantly contain beef, but some of them also contain pork. (Tom Lynn/Associated Press)

We can design PCR reactions to help us identify organisms in a DNA sample.

How might we do that?

Design primers to yield unique sizes of products for each organism.

Primer Design Example

		Sample DNA		
		Beef	Chicken	Pork
Primers	Beef Fwd + Beef Rev	100 bp	None	None
	Chicken Fwd + Chicken Rev	None	200 bp	None
	Pork Fwd + Pork Rev	None	None	300 bp

Primer Design Example

		Sample DNA		
		Beef	Chicken	Pork
Primers	Beef Fwd + Beef Rev	100 bp	None	None
	Chicken Fwd + Chicken Rev	None	200 bp	None
	Pork Fwd + Pork Rev	None	None	300 bp

One or both pork primers have no binding sites on **Beef** DNA.

One or both pork primers have no binding sites on **Chicken** DNA.

One or both pork primers have a single binding site on **Pork** DNA.

Tasks for Computational Lab

1. Generate features to allow prediction of primer melting points
2. Implement function for predicting PCR products
3. Design primers for PCR reaction to identify three types of DNA

Task 1 – Primer Melting Point Prediction

Features: Numerical descriptors of an object

Design ***features*** to help predict the melting point for a primer. Implement your feature calculation methods.

Assess with N-fold cross-validation using a RandomForest regressor model for generating predictions.

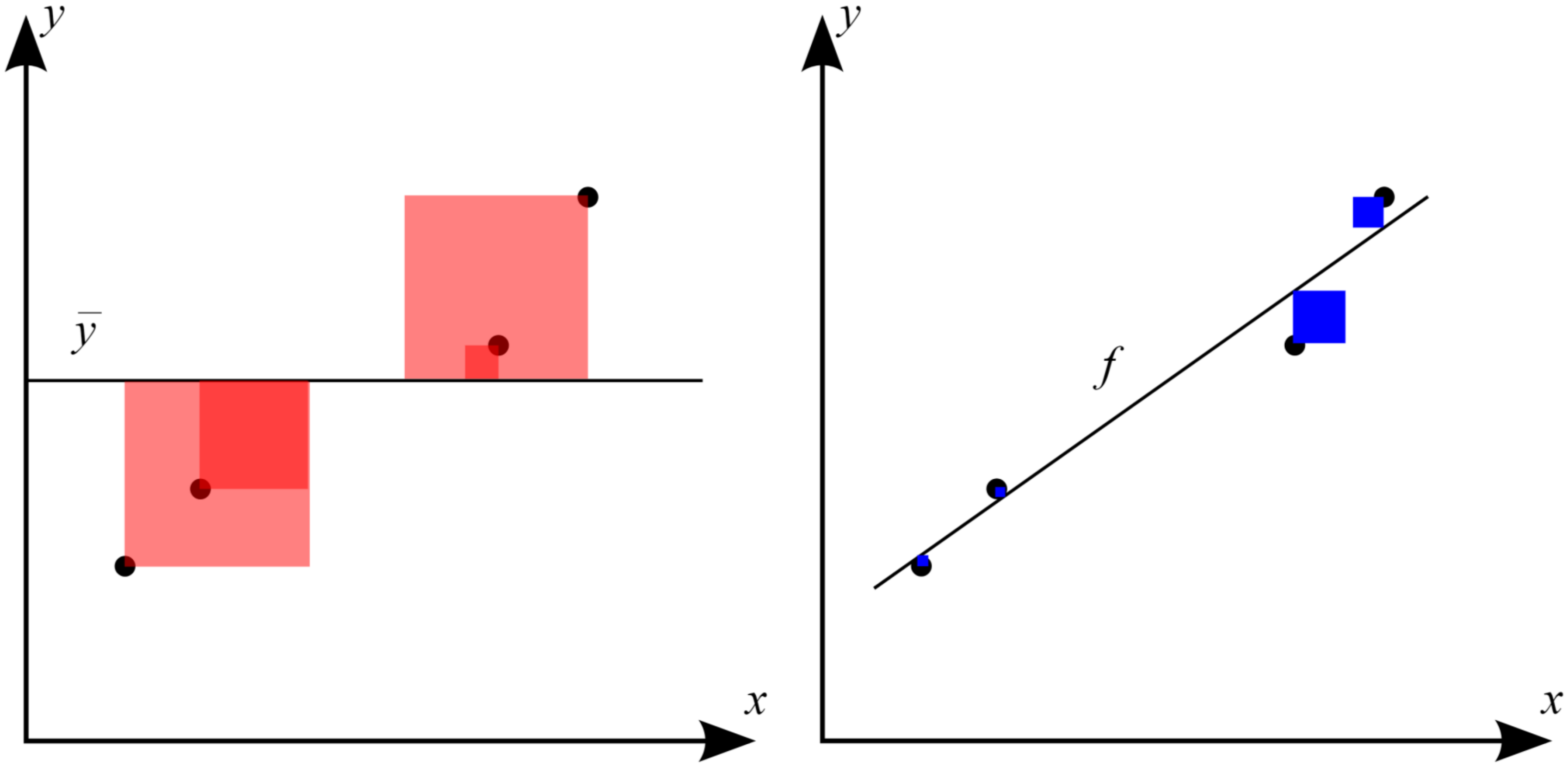
Regressor/regression – predict continuous value

Classifier – predict discrete class

How to design features for predicting melting point?

http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html

Assessing Accuracy of Predicitons



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 values closer to 1.0 are better.

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - Binds to upper strand
    - Reverse compliment of binding site on upper strand
    - T_m ~ 60^oC
    - 18-25 bases long
  - Forward Primer
    - Binds to lower strand within 1000 bases upstream of the reverse primer binding location
    - Reverse compliment of binding site on lower strand
    - T_m ~ 60^oC (within 5^oC of reverse primer melting point)
    - 18-25 bases long

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - Binds to upper strand
    - Reverse compliment of binding site on upper strand
    - **T_m ~ 60^{o}C**
    - 18-25 bases long
  - Forward Primer
    - Binds to lower strand within 1000 bases upstream of the rever n
    - Reverse compliment of binding site on lower strand
    - **T_m ~ 60^{o}C (within 5^{o}C of reverse primer melting point)**
    - 18-25 bases long

Use Task 1!

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - Binds to upper strand
    - **Reverse compliment of binding site on upper strand**
    - T_m ~ 60^{o}C
    - 18-25 bases long
  - Forward Primer
    - Binds to lower strand within 1000 bases upstream of the reverse pr
    - **Reverse compliment of binding site on lower strand**
    - T_m ~ 60^{o}C (within 5^{o}C of reverse primer melting point)
    - 18-25 bases long

Reverse Complement:
ACTG -> CAGT

Complementary Base Pairs:
A <-> T
G <-> C

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - **Binds to upper strand**
    - Reverse compliment of binding site on upper strand
    - T_m ~ $60^{o}C$
    - 18-25 bases long
  - Forward Primer
    - **Binds to lower strand within 1000 bases upstream of the reverse primer binding location**
    - Reverse compliment of binding site on lower strand
    - T_m ~ $60^{o}C$ (within $5^{o}C$ of reverse primer melting point)
    - 18-25 bases long

How do we determine binding?

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - **Binds to upper strand**
    - Reverse compliment of binding site on upper strand
    - T_m ~ 60°C
    - 18-25 bases long
  - Forward Primer
    - **Binds to lower strand within 1000 bases upstream of the reverse primer binding location**
    - Reverse compliment of binding site on lower strand
    - T_m ~ 60°C (within 5°C of reverse primer melting point)
    - 18-25 bases long

How do we determine binding?

(Local) Sequence Alignment!

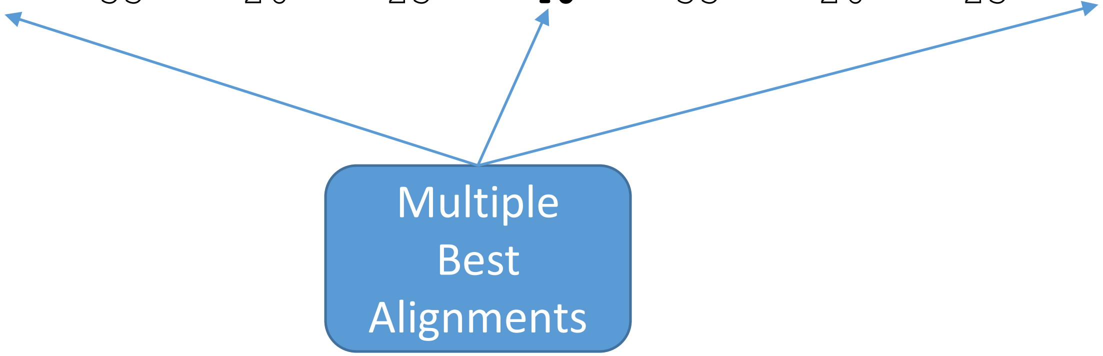
Task 2 – Predict PCR Products (Alignment)

```
> alignment.local_align("ACTG", "ACTG", print_output = True)
Scoring: match = 10; mismatch = -5; gap_start = 0; gap_extend = -7
A matrix =
             *      A      C      T      G
      *      0      0      0      0      0
      A      0     10      3      0      0
      C      0      3     20     13      6
      T      0      0     13     30     23
      G      0      0      6     23     40
Optimal Score = 40
Max location in matrix = (4, 4)
```

Best Alignment:
ACTG
ACTG

Task 2 – Predict PCR Products (Alignment)

```
> alignment.local_align("ACTGACTGACTG", "ACTG", print_output = True)
Scoring: match = 10; mismatch = -5; gap_start = 0; gap_extend = -7
A matrix =
        *    A    C    T    G    A    C    T    G    A    C    T    G
   *    0    0    0    0    0    0    0    0    0    0    0    0    0
   A    0   10    3    0    0   10    3    0    0   10    3    0    0
   C    0    3   20   13    6    3   20   13    6    3   20   13    6
   T    0    0   13   30   23   16   13   30   23   16   13   30   23
   G    0    0    6   23   40   33   26   23   40   33   26   23   40
Optimal Score = 40
Max location in matrix = (12, 4)
```



Multiple Best Alignments

Task 2 – Predict PCR Products (Alignment)

```
> alignment.local_align("AGTCACTGGCTT", "ACTG", print_output = True)
Scoring: match = 10; mismatch = -5; gap_start = 0; gap_extend = -7
A matrix =
            *     A     G     T     C     A     C     T     G     G     C     T     T
      *     0     0     0     0     0     0     0     0     0     0     0     0     0
      A     0    10     3     0     0    10     3     0     0     0     0     0     0
      C     0     3     5     0    10     3    20    13     6     0    10     3     0
      T     0     0     0    15     8     5    13    30    23    16     9    20    13
      G     0     0    10     8    10     3     6    23    40    33    26    19    15
Optimal Score = 40
Max location in matrix = (8, 4)
```

Position in String 1 of the last character in optimal alignment

Position in String 2 of the last character in optimal alignment

Best Alignment:
----ACTG----
ACTG

Best Score:
40/40
Best score possible for alignment of 4 characters.

Binding defined by 90%+ alignment.

Local Alignment Function

def local_align(x, y, score=ScoreParam(10, -5, -7), print_output = False):

x = sequence 1

y = sequence 2

score = Score Parameter (match = +10, mismatch = -5, gap = -7)
(optional)

print_output = binary indicating whether or not you want pretty output printed from alignment

(optional)

Task 2 – Predict PCR Products

- Given a sequence and a pair of primers, predict whether or not there will be a product. If so, predict the resulting product.
- Important Primer Pair Characteristics:
  - Reverse Primer
    - **Binds to upper strand (90%+ alignment)**
    - Reverse compliment of binding site on upper strand
    - T_m ~ 60^oC
    - 18-25 bases long
  - Forward Primer
    - **Binds to lower strand within 1000 bases upstream of the reverse primer binding location (90%+ alignment)**
    - Reverse compliment of binding site on lower strand
    - T_m ~ 60^oC (within 5^oC of reverse primer melting point)
    - 18-25 bases long

How do we determine binding?

(Local) Sequence Alignment!

Task 3

- Given sequences of genes from beef, chicken, and pork, design primers to identify each of them in a DNA sample.
- How do we do this?
  - Each primer pair must only generate a product in one source DNA sample.
  - Products from each source DNA sample must be of different lengths (50+bp different)

1 in 5 sausages tested across Canada contained different meat than labelled, study finds

Scientist calls degree of off-label ingredients alarming

The Canadian Press Posted: Aug 03, 2017 5:22 PM ET | Last Updated: Aug 04, 2017 11:07 PM ET

Canadian researchers found that typically beef sausages predominantly contain beef, but some of them also contain pork. (Tom Lynn/Associated Press)

What to turn in?

As a group:

- Code:
  - Your modified version of tm_prediction2.py (renamed usefully)
- Documentation:
  - Describe your approaches for Task 1 and Task 3 only.