Problem 1

1A Dragonn: Training convolutional networks

- **1-layer training**
  - At the first epoch, the training auPRC was .511, and the validation auPRC was .526.
  - At the final epoch, the training auPRC was .577 and the validation auPRC was .516.
- **2-layer training**
  - At the first epoch, the training auPRC was .529, and the validation auPRC was .500.
  - At the final epoch, the training auPRC was .928 and the validation auPRC was .716.

1B: Using and interpreting trained convolutional networks

- **Input and output dimensions of the first Convolution2D layer** The input dimensions, 4 and 500, correspond to the 4 character, 500 base length input sequences. The output dimensions, 15 and 486, correspond to the 15 Channels of the convolution outputs. Position domains begin with (1,15) and continue to (486,500).
- **Input and output dimensions of the Dense layer**. The input layer dimension, 195, arises from the number of nodes in all the previous ”flatten” layer, 15 by 13. The output dimension of 1 corresponds to the NN's estimate of the probability of a positive condition given the input.
- **Test Confusion Matrix**

2-Layer Test Results	
positive_test_0	P(bound)=0.897705674171
positive_test_1	P(bound)=0.925364196301
positive_test_2	P(bound)=0.806158542633
positive_test_3	P(bound)=0.909259021282
positive_test_4	P(bound)=0.490884274244
negative_test_0	P(bound)=0.708925485611
negative_test_1	P(bound)=0.161382764578
negative_test_2	P(bound)=0.941750407219
negative_test_3	P(bound)=0.781042456627
negative_test_4	P(bound)=0.526986956596

True Positives	4
False Positives	4
True Negatives	1
False Negatives	1

- **Filters vs Final Motifs** The filters do not obviously resemble the final motifs. This is not surprising. Filters are different objects, for example they ave both positive and negative values. The are optimized to be used in conjunction with each other in the next neural net layer. Finally, they can describe interactions between nearby positions, while PWMs assume strict independence regarding positions.
- **DeepLIFT scores** The DeepLIFT scores vividly highlight occurrences of motif2.png, NFKB_known1. The motif from motif1.png, IRF_known1 is not found, or at least not highlighted by the algorithm. The sequence diagrams are not exact, but they are much closer to the motif than to any of the trained filters.

1C: Implementing a forward pass - Programming assignment: forward_pass.py

Problem 2

2A: Estimating relative abundance

$$\begin{aligned}
\hat{f}_1^{unique} &= \frac{\frac{620}{40}}{\frac{620}{40}+\frac{405}{270}+\frac{5500}{1000}+\frac{40}{120}+\frac{180}{100}} = 0.629\\
\hat{f}_2^{unique} &= 0.0609\\
\hat{f}_3^{unique} &= 0.223\\
\hat{f}_4^{unique} &= 0.0135\\
\hat{f}_X^{unique} &= 0.0731
\end{aligned}$$

$$\begin{aligned}
c_1^{rescue} &= 620 + 390\frac{0.629}{0.629+0.223} &&= 908\\
c_2^{rescue} &= 405 + 4100\frac{0.0609}{0.0609+0.0731} &&= 3760\\
c_3^{rescue} &= 5500 + 390\frac{0.223}{0.629+0.223} &&= 5600\\
c_4^{rescue} &= 40 + 4100\frac{0.0731}{0.0609+0.0731} &&= 785\\
c_X^{rescue} & &&= 180
\end{aligned}$$

$$\begin{aligned}
\hat{f}_1^{rescue} &= \frac{\frac{908}{40}}{\frac{908}{40}+\frac{3760}{270}+\frac{5600}{1000}+\frac{785}{120}+\frac{180}{100}} = 0.449\\
\hat{f}_2^{rescue} &= 0.275\\
\hat{f}_3^{rescue} &= 0.111\\
\hat{f}_4^{rescue} &= 0.129\\
\hat{f}_X^{rescue} &= 0.0356
\end{aligned}$$

2B: Estimating absolute abundance

$$\begin{aligned}
N_X & &&= 1000\\
N_1 &= 1000\frac{0.449}{.0356} &&= 12600\\
N_2 & &&= 7740\\
N_3 & &&= 3110\\
N_4 & &&= 3640
\end{aligned}$$

Problem 3

Gaussian processes for time series data

I propose first creating, for each gene, two GPs, one for the normal transient and one for the heat shock. I would do this with the GaussianProcessRegressor function in the sklearn package. After calling the fit and predict functions, y_pred and sigma vectors are returned for a linspace of interest.

Bayesian Equality Test

I am following the approach in Benavoli, Alessio, and Francesca Mangili. ”Gaussian Processes for Bayesian hypothesis tests on regression functions.” Artificial Intelligence and Statistics. 2015.

Our next task is to map these two GPs to a p-value for the null hypothesis: the gene is not differentially expressed in the two transients. They first note that the difference between two GPs is itself a GP. If the resulting is GP is ”near” the x-axis, the probability of the data being sampled from the null hypothesis is relatively high. To explicitly perform this mapping, we must scale the ”credible region” by choosing the smallest number of SDs such that the zero vector is completely contained. We then map this Z score to a p-value with a t-test conversion.

Multiple Hypothesis Correction

Finally, we must account for the number of genes tested. We should choose an acceptable FDR, and decide after examining the distribution on an appropriate correction method. Then choosing an appropriate q-value, deliver a list of possible DEGs for further inspection.