From Humans to Mice:
An Experiment in Transverse Evolution

Jason Saunders, Evan Hymanson {jsaund, hymanson}@stanford.edu

PREDICTING

Motivation:

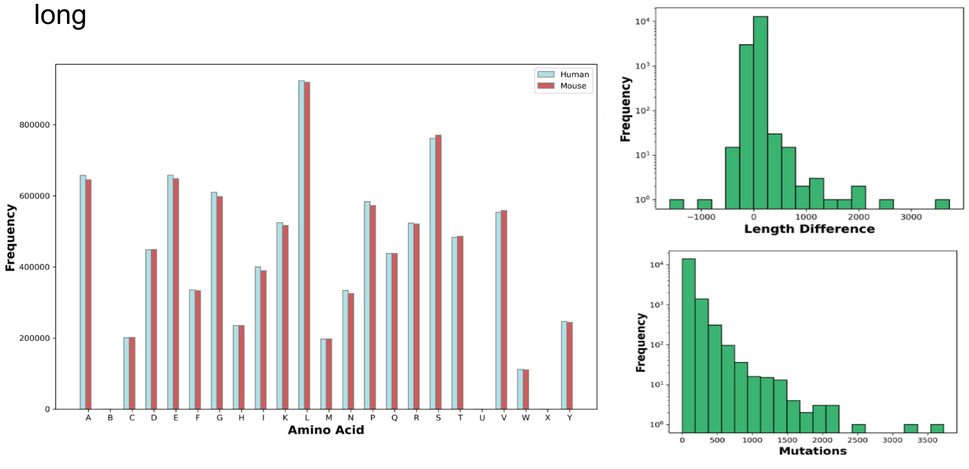
- Proteins are commonly used in biotech, but may require translating a human protein to function in different conditions, like different pH and temperatures
- What if there was a way to alter a protein in order to allow it to survive in a new set of conditions, but without deteriorating its function?

Problem:

- We aimed to predict mouse protein sequences from human proteins using a transformer encoder-decoder model.

DATA AND FEATURES

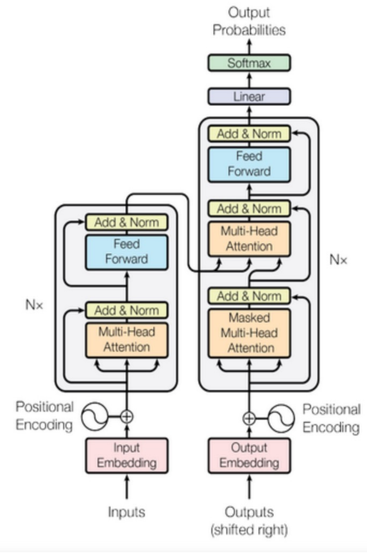
- We parsed through Swiss-Prot group at the Swiss Institute of protein lists and created a dataset of proteins which were shared between both the human and mouse proteomes, which was 15,945 proteins long



FUTURE WORK

- Improve the predictive capabilities of the model.
- Implement a cost function which more heavily weights predicting mutations rather than predicting shared amino acids.
- Attempt to express the protein sequences predicted by our model.

METHODS + EXPERIMENTS



- We used the general transformer model

Key Features:

1) Input/output embedding: 23 amino acid total tokens–we have 20 common amino acids and a few additional markers for uncommon, variable, or unknown amino acids)
2) Positional encoder adds a position-dependent sine term to each of the even dimensions of our encoder, and a position-dependent cosine term to each of the odd dimensions.
3) A look-ahead mask is used–looks like an upper-triangular matrix and masks all 'future' tokens
4) Cross entropy loss function: $L_{\mathrm{CE}} = -\sum_{i}^{n} t_i \log(p_i), \quad \text{for n classes,}$

where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

5) Hyper parameters: dembedding = 60, dhidden layer = 60, Nlayers = 4, Nattention heads = 4. And dropout probability of 0.2 in our model parameters.

RESULTS + DISCUSSION

Key Parameters:

- number of differences between our input human sequences and the true mouse sequences (Nit).
- number of differences between our predicted mouse sequences and the true mouse sequences (Npt).
- number of differences between our predicted mouse sequence and the input human sequence (Npi)

Epoch	Loss	N_{it}	N_{pt}	N_{pi}	ΔN
1	0.28	284455	460238	456739	-175783
4	0.22	284455	284638	1800	-183
8	0.22	284455	284420	535	35
12	0.22	284455	284395	539	60
16	0.22	284455	284395	512	60
20	0.22	284455	284400	538	55
Test	0.29	387088	386984	524	104

- If we were to just guess that the mouse sequence is identical to the human sequence, then we'd predict 93% of the amino acids correctly.
- Second, notice how Npi decreases as we continue training. That means the model's output is becoming more and more similar to the human sequence input, but it is not identical.
- The indicator ΔN shows us that our model's output is slightly closer to the true mouse sequence than the input. This changes most dramatically over the first 10 epochs, and then plateaus after that.
- Even at its best though, our model is only predicting .002% more amino acids correctly–this is a negligible percentage.