Generative Modelling of Insulin Proteins

Jonas Vestergaard Jensen, s162615 and Emma Ahrensbach Rørbeck, s173733

Introduction

Design of proteins with specific properties and functions is a complex problem and advances in this field could have a large impact in biology and medicine.
The joint probability $P(\mathrm{x})$ of a protein sequence, $\mathrm{x} = (\mathrm{c}, \mathrm{a})$ with properties $\mathrm{c} = (c_1, c_2, \ldots, c_{n_c})$ and amino acids $\mathrm{a} = (a_1, a_2, \ldots, a_{n_a})$ can be factorized using the chain rule of conditional probabilities

$$P(\mathrm{x}) = \prod_{i=1}^{n_c+n_a} p(x_i | x_{i-1}, x_{i-2}, \ldots, x_1) \tag{1}$$

and thus the problem of generating insulin proteins can be described as a next-token prediction problem [2]. During training we then minimize the negative log likelihood of of a dataset with N protein sequences

$$\mathcal{L} = -\sum_{n=1}^{N} \log P(\mathrm{x}^n) \tag{2}$$

In this project we have explored different popular deep learning architectures for modelling sequential data and applied them to the problem of generating insulin proteins

GRU & LSTM architectures

A recurrent neural network (RNN) is a neural network, which has been successful in modelling sequential data, like protein sequences. The RNN performs its computations in a cyclic manner, and can recall previous computations in a way, which can be interpreted as the memory of the RNN.

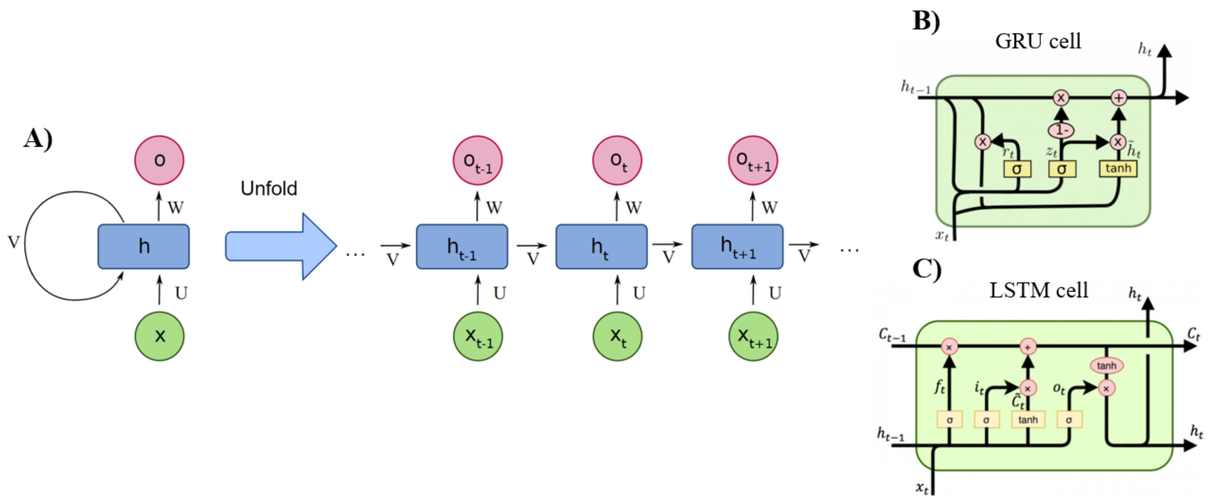

Figure 1: A) Representation of an RNN [3]. B) and C) The architectures of a Gated Recurrent Unit (GRU) and a Long-Short Term Memory (LSTM) cell, respectively [1].

Transformer

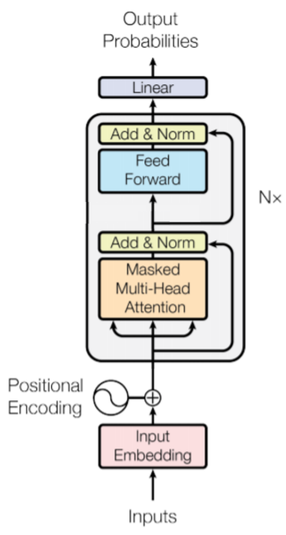

Figure 2: The model architecture of the Transformer Network [5].

The Transformer is a model architecture relying on an attention mechanism to draw dependencies between input and output. The attention mechanism is used to assign relevance to the context of the sequential inputs, and has an infinite reference window as memory.

$$Attention(Q, K, V) = softmax\left(\frac{mask(QK^T)}{\sqrt{d_k}}\right)V$$

Multi-headed attention is used, which allows the model to jointly attend to information from different representation subspaces at different positions. The multi-headed attention layer is masked so the prediction for position i does not depend on information about subsequent positions.

WaveNet

The WaveNet model was designed to generate raw audio signals. Here we apply the same architecture to generate insulin proteins. The main elements of the model are stacked dilated causal convolutions, skip connections and residual connections. WaveNet utilizes the following gated activation function

$$\mathrm{z} = \tanh\left(W_{f,k} * \mathrm{x} + V_{f,k}^T \mathrm{h}\right) \odot \sigma\left(W_{g,k} * \mathrm{x} + V_{g,k}^T \mathrm{h}\right) \tag{3}$$

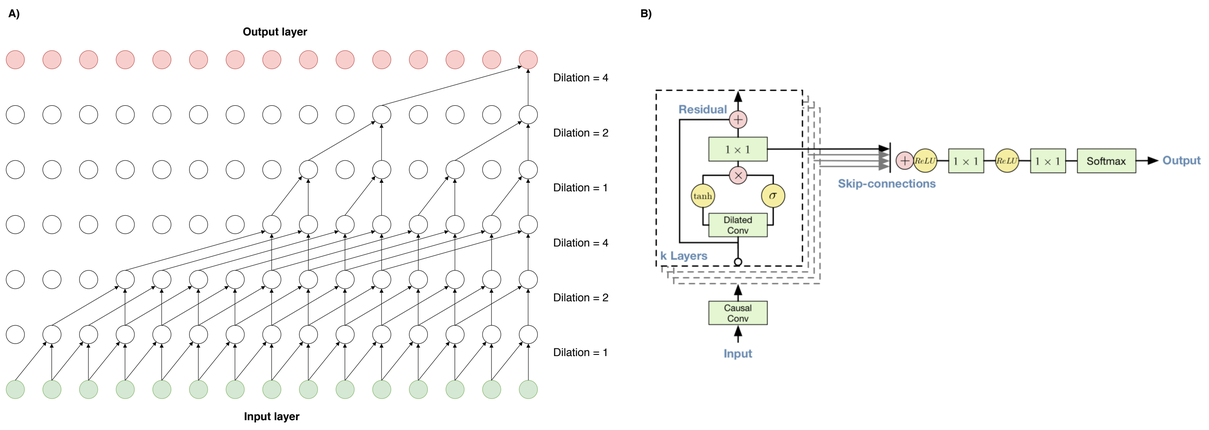

Figure 3: A) An example of stacked dilated causal convolutions. Only connections to the right-most output are shown. B) The full architecture of the WaveNet model [4].

Data

- 3272 insulin proteins, 2617 non-insulin proteins
- 5 protein keyword categories; Organism, Biological process, Cellular location, Molecular function, and Insulin
- Preprocessing procedures
  - Include all combinations of keywords for each protein
  - Sample one from each category for each protein

Model Performances

	LSTM	GRU	Transformer	WaveNet
Data set structure	S	S, I	C, I	C, I
Number of parameters	31,340,560	22,930,788	548,868	3,656,602
Avg. test perplexity	4.1	3.5	4.6	4.6

Table 1: Comparison of the different models. Data sets structure: Sample (S), Combination (C), Only Insulin proteins (I)

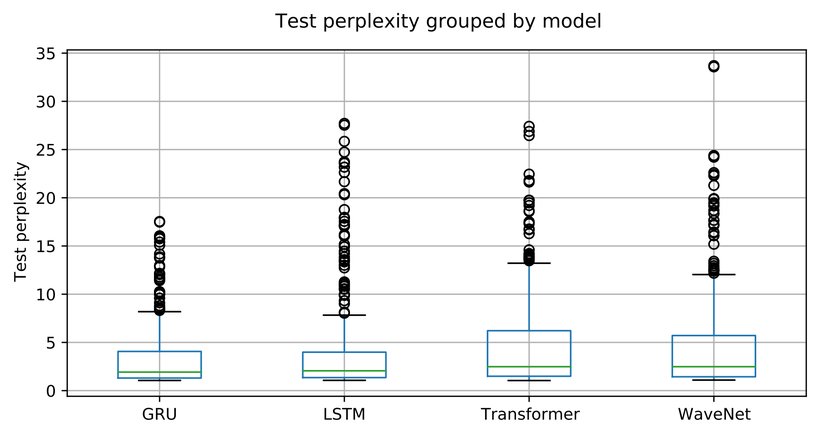

Figure 4: Box plot of the test perplexities for the most optimal version of the four models

Sequence Embedding

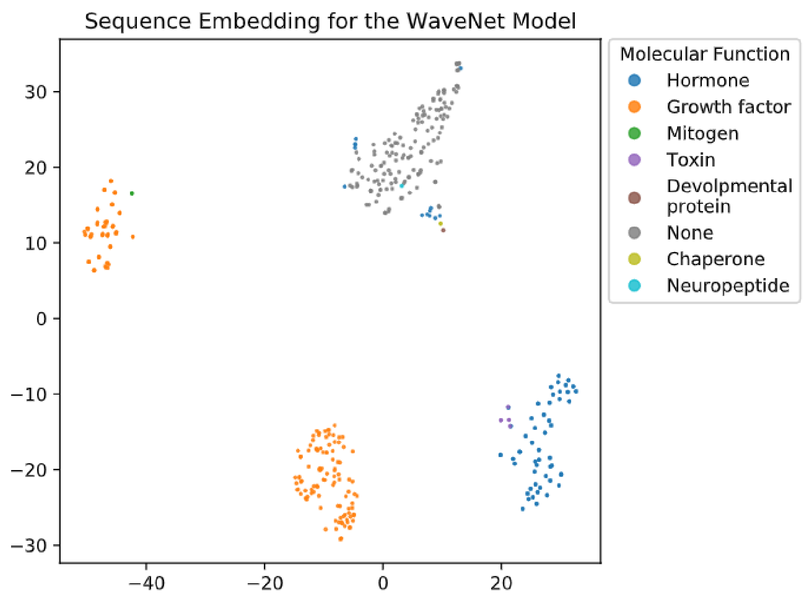

Figure 5: t-SNE representation of mean pooling of WaveNet's token embeddings for all test proteins.

Protein generation

Global alignments using the BLOSUM62 matrix were made between the test proteins and proteins generated by the models. These global alignments are compared to global alignments between the test proteins and a 50% mutation baseline, where half of the amino acids in the test proteins were randomly mutated.

Figure 6: Across all context lengths, sequence similarity has been calculated for each protein in the test set with the generated protein. This is compared to a 50% mutation baseline

References

[1] RNN, LSTM & GRU, 2019. Available at: `http://dprogrammer.org/rnn-lstm-gru`.

[2] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. Progen: Language modeling for protein generation, 2020.

[3] P. T. Perez. Deep learning: Recurrent neural networks, 2018. Available at: `https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010`.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.