Deep seq2seq architecture for DNA sequence decoding from noisy data

Frederic Lechenault, Antoine Baker, Florent Krzakala
Laboratoire de Physique l'Ecole Normale Supérieure
PSL Research University, Sorbonne University, CNRS
24 rue Lhomond, 75005 Paris, France
frederic.lechenault@phys.ens.fr

Abstract

We investigate the use of a deep seq2seq architecture to solve a complex inverse problem in biophysics, that is to accurately decode DNA strands from noisy experimental data. The data corresponds to the thermal unbinding time of collections of oligonucleotides on unzipped DNA together with approximate hybridization location along the strand. Without noise, this data for all 64 possible triucleotides build from the four bases allows for DNA deciphering. However, experimental noise severely impairs this reconstruction. By modeling the various sources of noise, we train a deep neural network on synthetic data to solve this inverse problem. Various benchmarks against noise intensity demonstrate the relevance of this approach to real world DNA decoding.

1 Introduction

In a seminal series of work work by Pihlak et al. [2008] and Ding et al. [2012], an innovative technique for DNA decoding was put forward, based on mechanics. The idea is to unzip a DNA molecule by pulling each strand apart using magnetic tweezers, and then to dip it in a bath containing an oligonucleotide, a very short single stranded sequence of bases. This oligonucleotide will thus bind to the DNA strands typically where it finds a complementary sequence. The hairpin is then pulled out of the bath and the force keeping it open is slowly released. If the strands were pristine, the DNA strands would just zip back into the original molecule. The local presence of oligonucleotides momentarily slows down this zipping process, allowing for spatial identification of the binding sites along the chain. This information can be made more accurate through cycling, allowing for the accumulation of statistical information in the form of binding histograms. Collecting such data for all possible, say, trinucleotides, *i.e.* all 64 three-letter combinations of base pairs, allows in principle for the full sequence decoding. It turns out that this can actually be done, but with various sources of experimental noise which severly impair the final reconstruction. For example, pulling on the strands stretches them, introducing imprecision in the location of the binds.

Modelling the various sources of noise with realistic, experimentally driven parameters, we have devised a way to generate vast amounts of synthetic data in the form of approximate binding probabilities along the strands for all 64 trinucleotides. We have then trained a neural network to reconstruct the base sequence from this noisy data. The trained net was able to achieve significant decoding accuracy, beating previous efforts based on other algorithmic techniques, illustrating the power of neural networks to solve such a complicated inverse problem. This accuracy was aso found to be robust for values of the noise comparable to that found in the experiments.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

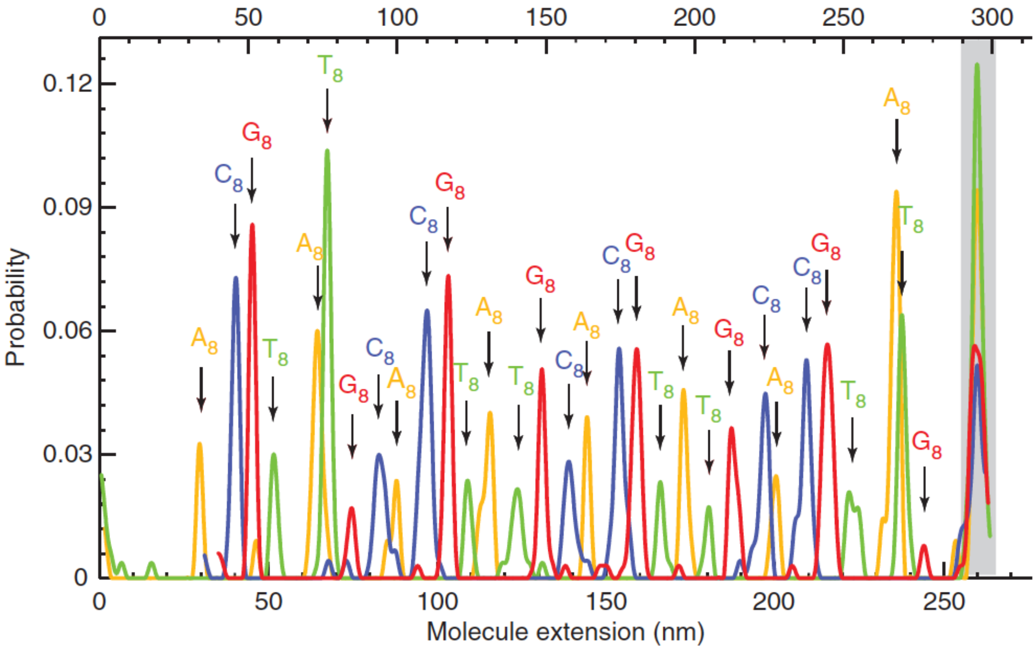

Figure 1: Example of experimental binding time histograms, in this case for sequences of 8 bases (from Ding et al. [2012])

.

2 Results

2.1 Training data

The data generation mimics the output of the experiment (see Fig 1) it picks a random DNA sequence and generates histograms of binding probabilities along the chain for all 64 trinucleids, introducing a tunable random stretch, blurring the correspondance between space and bases, fluctuations on the position of the binds and on the binding force. This input data is space based: we pick a physical length in nm and consider sequences of this given length. The target data is simply the one-hot encoding of the base sequence.

2.2 Network architecture

The input of the network consists of these 64 'temporal' sequences of the binding histograms, where time here is space along the DNA strand. The representation layer of the network is two series of 5 aggregated residual 1D convolutional layers as those introduced by Xie et al. [2017], each with 8 blocks of 8 filters with doubling dilation rate, borrowing from Dieleman et al. [2018]. All non-linearities are *leaky relu* units. This representation, followed by two LSTMs with 64 memory units, is the encoder of a sequence-to-sequence architecture implementing a classical attention mechanism, as described in Chorowski et al. [2015]. The decoder, a stack of two 64-unit LSTMs, is fed the internal state of the encoder. Encoder and decoder outputs are doted and thresholded to produce the attention matrix, which is then doted with the encoder output to form the context. This context is then concatenated with the decoder output and fed to two time-distributed dense layers, the first one with 64 units and a *tanh* activation, and the last one 5 units and a *softmax*, to produce one of 4 coding base and a start character. We have trained the net with output sequences of 50 bases, in batches of 32 sequences, with teacher forcing. The net was implemented in *keras* running with the *tensorflow* kernel. For each data point, it was trained for several days on a NVIDIA Titan Xp GPU with an adjusted learning rate schedule.

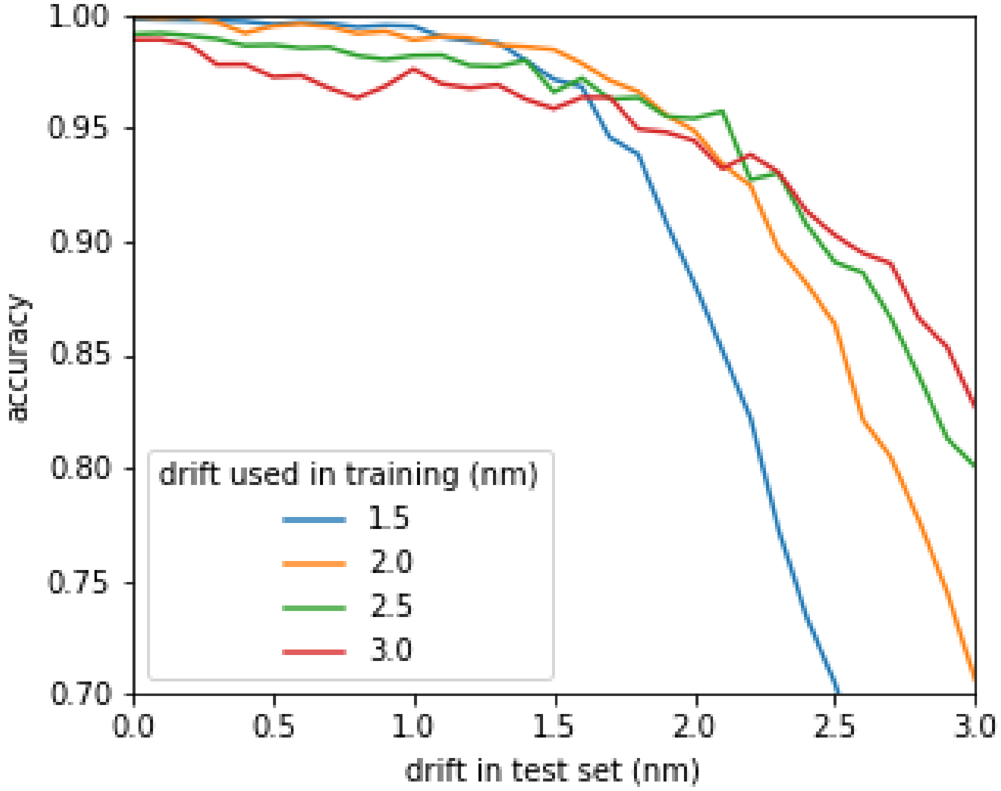

Figure 2: Benchmark of accuracy versus levels of drift noise in test set. Each line represents a model trained assuming a specific amount of drift noise

.

2.3 Benchmarks

If the peaks in the binding time histograms were perfectly aligned with the positions of the underlying trinucleotides, the decoding problem would be almost trivial. Indeed if we knew the exact positions of the trinucleotides, we could easily reconstruct the full DNA sequence from the overlaps. Unfortunately, in actual experimental conditions, the peak position fluctuates around the actual position of the trinucleotide by a random shift. The typical shift (as measured by the standard deviation) is called the *drift noise* and constitutes the main hindrance to perfect reconstruction. Experimentally the levels of drift noise range from 1.5 nm to 3 nm.

We generated several test sets with varying levels of drift noise. For each test set, we compared the actual and predicted DNA sequences and simply computed the average accuracy to quantify the model perfomance. In Fig 2, we benchmark several models each trained with a different amount of drift noise. As expected, for all models the performance degrades with the amount of drift noise in the test set. For the model trained assuming a drift = 1.5 nm (blue line) the accuracy is very good at small amounts of drift but sharply decreases for drifts higher than 1.5nm. When we consider models trained with higher amounts of drift (orange, green and red lines), the accuracy can decrease but still remains quite good for low values of drift noise, while the model becomes much more robust to higher values of drift. Overall, for the typical amount of drift noise obtained experimentally (up to 3nm), the trained models can achieve compelling decoding accuracy (for instance always greater than 80$\%$ for the green and red models).

3 Conclusion

This paper proposes a new method for inferring DNA sequence from binding time histograms. We exploit the fact that the forward process (from the DNA sequence to the binding histograms) can be easily simulated to generate vast amount of synthetic data. Given the intrinsic difficulties in the direct writting of the inverse model, this makes a deep learning approach extremely appealing.

Using these synthetic data, we thus trained a neural network to solve the inverse problem (predict the DNA sequence from the binding histograms). The accuracy obtained on synthetic test sets, with typical amount of noise observed experimentally, is encouraging (from 80% decoding accuracy to nearly perfect reconstruction). The predictive performance depends crucially on the level of drift noise that was indeed expected to be the main hindrance to perfect reconstruction.

References

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

Sander Dieleman, Aaron van den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. In *Advances in Neural Information Processing Systems*, pages 7989–7999, 2018.

Fangyuan Ding, Maria Manosas, Michelle M Spiering, Stephen J Benkovic, David Bensimon, Jean-Francois Allemand, and Vincent Croquette. Single-molecule mechanical identification and sequencing. *Nature Methods*, 9:367–372, 03 2012.

Arno Pihlak, Göran Baurén, Ellef Hersoug, Peter Lönnerberg, Ats Metsis, and Sten Linnarsson. Rapid genome sequencing with short universal tiling probes. *Nature biotechnology*, 26(6):676, 2008.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.