# Review of the litterature

Guillermo Martinez (matricule 20076854)
Dereck Piché (matricule 20177385)
Jonas Gabirot (matricule 20185863)

February 24, 2023

## Aptamers

Nucleic acids (DNA and RNA) carry the instructions on how an organism can grow, develop and replicate. Recent developments in our ability to generate large populations of degenerate oligonucleotides and isolate functional nucleic acid molecules that can bind to a specific target have led to the in vitro evolution of nucleic acid molecules for other than biological functions. These affinity reagents based on DNA or RNA are referred to as aptamers and posses two fundamental characteristics: their ability to fold into shapes with defined function (phenotypes) and the ability of their respective sequences (genotypes) to be replicated in vitro to produce progeny molecules with similar characteristics to their parent sequence. Aptamer-engineering consists in exploiting their unique phenotype-genotype connection to amplify individual molecules with desired phenotypes, and thus, optimizing their binding properties to the desired analyte. For example, through directed evolution procedures such as Systematic Evolution of Ligands by Exponential Enrichment [SELEX], one can apply recursively the principles of natural selection on a large population of nucleic acid sequences incubated with a target— i.e. select and amplify the molecules that bind to the target to create a new population of molecules that are more enriched in sequences that can perform the desired function, until the pool is dominated by the sequences that bind with high affinity to the desired target. These resulting sequences with high affinity towards a specific analyte are called aptasensors (Dunn et al., 2017. Their relatively simple chemical structure (Jeddi et al., 2017)— i.e. sequences composed of roughly of 20 to 100 nucleotides in length (Zhang et al., 2019)— is crucial for the insertion of electrochemical or fluorescent reporter molecules, as well as surface-binding agents. When binding with its target, the aptamer undergoes a conformation change which can be exploited to generate an analytical signal (Jeddi et al., 2017). Consequently, aptasensors have been found to be very useful in tracking the propagation of various molecules in their respective environments such as pathogens, toxins, antibiotics and pesticides in food, water and soil samples (Dunn et al., 2017), as well as adenosine triphosphate in cells ((Zhang et al., 2019). Although DNA

aptamers are more stable and more robust than their RNA counterparts, the large array of computational tools available for single stranded* RNA structure prediction are less pervasive for its DNA counterpart. In fact, the computational tools available for DNA were restricted to model only double-stranded DNA structures until 2017. Correctly predicting the 3-dimensional structure of single-stranded DNA hairpins and other more complex structures from 1 dimensional sequences has the potential to not only revolutionize aptamer-engineering process but also to bolster the range of application for aptasensors in more difficult environments (Jeddi et al., 2017) The most crucial difficulty in aptamer-analyte binding analysis is the pre-folding of the aptamer to the correct equilibrium structure. Each DNA sequence may adopt a variety of folded structures and brute force techniques such as naïve molecular dynamics search are exceedingly computationally expensive. End-to-End-DNA [E2EDNA]— an end-to-end aptamer-analyte binding pipeline for UTP complexing with simple hairpins—generates a set of structures a given aptamers is likely to adopt, as well as their respective probabilities given molecular dynamics simulations with appropriate force-fields. The pipeline finally implements 'NUPACK' and 'seqfold' Python packages to identify the minimum free energy structure (Kilgour et al., 2021) using nearest-neighbor empirical parameters of a given temperature and ionic strength specified by the user (Zadeh et al., 2011). The purpose of this research is therefore to train deep learning neural networks with randomly generated DNA sequences to predict the minimum free energy structure given by 'NUPACK'. The most stable aptamer sequences will be potential candidates to undergo the entire E2EDNA protocol and be tested on their binding affinity to a wide range of analyte of interest. The aptamers that are the most stable and possessing the highest binding affinity will be potential candidates to be synthetized and used to solve specific problems such as trace the oil molecules in the oceans after a spill. Given that shorter random-sequence libraries of doubly modified aptamers usually posses a higher binding affinity for a target than larger traditional sequences (30-mer VS 40-mer random regions), the length of our random sequences will be of 30-mer (Dunn et al., 2017).

# Machine learning algorithms

## Multilayer Perceptrons

This is the deep architecture that started it all and the one we shall use to create our baseline since it does not posess any bias towards the positionning of the different tokens. The use of MLPs is made possible by the fact that we are only dealing with 4 different possible elements as our tokens (A,C,G,T).

## Reccurent Neural Networks

The broad definition of a recurrent neural network is that it is the subclass of neural networks [1] which have cycles in the layers. RNNs (especially LTSM

RNNs) were the default deep learning models used for sequences of tokens for a while, until the recent rise of transformers These RNNs processed multiple tokens one at a time by storing information in the so called hidden state of the network. Since you have to process each token one at a time, there is no parallelization. One advantage of recurrent neural network is that they grow linearly with respect to the amount of tokens as input, which means that their really isn't a practical limit to the number of tokens that it can process with reasonable time complexity.

### Transformers

According to our assumptions, the transformer architecture [2] is by far the most appropriate for our task. Transformers use a multi-headed attention mechanism and self-attention. Let be $t_1, \ldots, t_n$ be a sequence of input tokens. Then a single head will create for each input token $t_i$ an output $y_i$ which is a linear combination of the other tokens given as input. It's complexity is $O(kn^2)$, where k are factors independant of input size. This process is repeated for multiple heads with different parameters (which are learned). Their outputs are concatened and fed into a fully connected network that combines their ouputs. The architecture is complex and impractical to spell out in greater detail here.

## Currently used classical algorithms (State of The Art)

There is currently little research and writing on learning aptamer properties with deep learning algorithms. Instead, biology-specific algorithms biology-specific algorithms are favoured, as well as clustering algorithms. clustering algorithms. For example, this article from January 2023 uses an original algorithm that combines clustering methods to find an optimal an optimal aptamer from a selection. https://pubs.acs.org/doi/pdf/10.1021/acssynbio.2c00462. However, some recent papers use deep learning. "Machine learning guided aptamer refinement and discovery" (https://www.nature.com/articles/s41467-021-22555-9) uses a standard MLP neural network to find the most compatible (high affinity) aptamers with target molecules. The estimation of free energy is a sub-step of the affinity calculation. It performs a truncation step to minimise the length of the aptamer without altering its properties. Another deep learning model with aptamers is AptaNet (https://www.nature.com/articles/s41598-021-85629-0). This model uses an MLP and a CNN to learn the relationship between aptamers and target proteins proteins (Aptamere-protein relations or API). The MLP works best, with a test accuracy of 91.38algorithms such as SVM, KNN and random forests. However, this model uses a very detailed database containing numerous auxiliary variables measured in the laboratory for each individual, but with only 1000 individuals. No published aptamer model uses transformers or RNNs to predict free energy, so the predict free energy, so our method would be original in this field.

1. Dunn, M., Jimenez, R. and Chaput, J. Analysis of aptamer discovery and technology. Nat Rev Chem 1, 0076 (2017). https://doi.org/10.1038/s41570-017-0076
2. Jeddi, I., and Saiz, L. (2017). Three-dimensional modeling of single stranded DNA hairpins for aptamer-based biosensors. Scientific reports, 7(1), 1178. https://doi.org/10.1038/s41598-017-01348-5
3. Kilgour, M., Liu, T., Walker, B. D., Ren, P., and Simine, L. (2021). E2EDNA: Simulation Protocol for DNA Aptamers with Ligands. Journal of chemical information and modeling, 61(9), 4139–4144. https://doi.org/10.1021/acs.jcim.1c00696
4. Zhang, Y., Lai, B. S., and Juhas, M. (2019). Recent Advances in Aptamer Discovery and Applications. Molecules (Basel, Switzerland), 24(5), 941. https://doi.org/10.3390/molecules24050941
5. Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011). NUPACK: Analysis and design of nucleic acid systems. Journal of computational chemistry, 32(1), 170–173. https://doi.org/10.1002/jcc.21596

# References

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9.

[2] Ashish Vaswani et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017.