Towards a virtual assay for aptamer-target interactions

Andrew Boardman

December 2018

1 Introduction

Aptamers are nucleic acid (either DNA or RNA) oligomers that bind to protein or small molecule targets (Tuerk & Gold 1990; Ellington & Szostak 1990). They can be synthesised extracellularly and used as probes, for example in super-resolution microscopy (Strauss et al. 2018) or expressed within living cells to control gene expression (Win & Smolke 2007).

Strong and specific interactions between biomolecules are key to the machinery of life (Rolland et al 2014), allowing functional modules to work reliably in the messy environment of the cell. Combining simple modules allows complex gene circuits to be built up (Kashtan & Alon, 2005); recently, scientists and engineers have begun to create new gene circuits in the emerging discipline of synthetic biology. Much has been achieved by simply reusing the molecules that nature has given us; in the long term, however, the challenge is to go beyond tinkering with these existing parts, to design molecular circuits performing desired functions from scratch.

Despite improvements in our understanding of the hierarchical nature of protein folding (Hocker 2012), the design of proteins with desired shape and custom interactions is still an extremely difficult problem. Nucleic acids, by contrast, can easily be engineered to take specific shapes and to form predictable interactions with other nucleic acids - they can be used to engineer simple circuits. Integrating aptamers into these circuits allows them to respond to changes in the concentration of the target.

Antibodies are used as ligands in many common protocols, in medical testing, and as therapeutics. Aptamers, unlike antibodies, are stable at room temperature, so offer benefits over antibodies for use in medical testing where refrigeration is not easily available. Aptamers are cheap to synthesize chemically and do not provoke an immune response; antibodies must be selected and synthesised in vivo and can be targeted by the immune system.

While the chemical space accessible to nucleic acids (with only 4 bases in the DNA and RNA alphabets) may not be as large or as diverse as that accessible to proteins (with 22 possible amino acids, ranging from charged to hydrophobic residues), there is reason to be hopeful about the ability of nucleic acid molecules to bind a wide variety of targets - high-throughput methods such as CLIP-Seq (Stork & Zheng 2016) have revealed that many more proteins bind to mRNAs than had been previously thought, many of which lack 'classical' RNA-binding domains (Hentze et al 2018). Nucleic acid ligands which are artificially synthesised can extend the 'chemical repertoire' of DNA and RNA by incorporating artificial nucleobases; for example, SOMAmers (Gold et al 2012) incorporate nucleic acids with protein-like side chains, allowing the aptamer to present hydrophobic surfaces. Backbone modifications are also possible; RNA aptamers are susceptible to degradation by RNAse, but backbone modifications (e.g. 2'-fluropyrimidine RNA) can improve this.

Most approaches to aptamer selection have been based on the SELEX protocol (Tuerk & Gold 1990), which has dominated the field of aptamer discovery for nearly 30 years now. This begins with a large random library of candidate sequences and selects progressively for the ligands with the highest affinity to the target by repeated rounds of separation and amplification. The technique has spawned many variants; for example, proteins, small molecules or even whole cells can be used as targets. However, the success of the technique depends on the size of the original pool, and the target - some targets, particularly negatively charged ones, may not bind well to any nucleic acid (Rohloff et al 2014). The process is also vulnerable to experimental bias; many confounding factors may influence the abundance of a sequence in the final round of sequencing. These include bias in the synthesis of the original library, non-specific binding to the experimental apparatus, and bias in the amplification of selected sequences. Therefore, the SELEX process will not necessarily converge to the sequence with optimal affinity for the target. (Hoinka et al. 2015)

However, by high-throughput sequencing of samples from each selection round (a process known as high-throughput SELEX, or HT-SELEX for short) it is possible to generate a huge amount of data which can reveal how the selection proceeds. Most computational approaches to the analysis of this data have focussed on clustering of the existing aptamer families. In this work we present machine learning algorithms which use the sequences obtained from a HT-SELEX dataset to predict whether a given sequence will bind to a target. This is done by dividing the sequences into 'active' and 'decoy' datasets and training a classifier which can distinguish the two sets. We use this classifier to predict new aptamers for verification by Markov chain-Monte Carlo sampling.

2 Results

2.1 Selection of data for classifier training

We extracted aptamer sequences from the dataset provided by Hoinka et al. (2015), in which RNA aptamers were screened against murine Interleukin 4 in 5 progressive rounds of SELEX; samples of cDNA were taken from each of the last 4 rounds and sequenced. To date, this is the largest publicly available dataset from a HT-SELEX experiment. For this dataset, we treat the 2,146,796 distinct sequences present in the last round as our active set, and the 5,058,805 distinct sequences from the earliest sequenced round as our decoy set. Treating these as decoys is a conservative estimate of the background; the first sequenced aptamers will not in fact be completely non-binding - however they will be subject to experimental bias, and we have no negative control data to eliminate this.

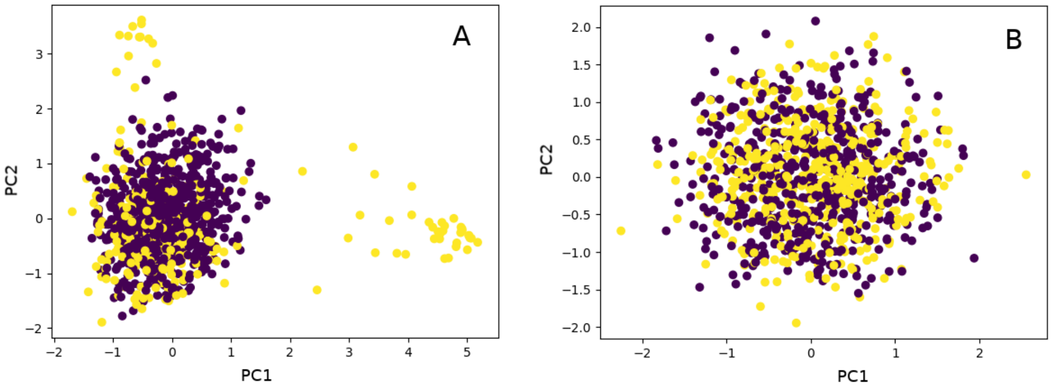

Figure 1: Plots of the first 1000 active & inactive sequences in principal component space, for the Ilk (A) and IgE (B) datasets. Active sequences are shown in yellow, inactive sequences in purple.

We also tested our models on data from a micro-free flow electrophoresis - SELEX experiment (W. Arter, unpublished data). DNA aptamers were screened against human immunoglobulin E (IgE), using a filter-free microfluidic approach. We obtained 106345 distinct sequences from the first SELEX round, and 118097 distinct sequences from a negative control experiment performed without target present.

Duplicate sequences were removed; as Hoinka et al. (2015) point out, the number of copies of a sequence present in the final round is subject to significant bias due to the sampling process and the PCR and cannot be taken as a reliable measure of binding affinity. Leaving duplicates in would provide the classifier with many repeated examples, biasing the classifier towards recognising these highly replicated sequences at the expense of performance across the final aptamer pool.

2.2 Exploratory data analysis reveals structure in aptamer datasets

Exploratory data analysis was performed to examine the two datasets for any obvious structure. For each dataset, samples were taken of active and decoy sequences; each sequence was encoded to form a binary vector of length 160. Principal component analysis was performed on each set of binary vectors. The IgE dataset shows no evidence of structure in principal component space (although this does not rule out the presence of non-linear structure); this suggests that the selection pressure in this experiment was not strong enough. This could be increased by performing more rounds of selection or by decreasing the protein concentration. The first two principal components from the Ilk data separate the samples into three clear clusters.

2.3 Classifier building and evaluation

We built models to distinguish sequences from the 'active' set from decoys. These return a score for each sequence; by putting a threshold on this score we can measure the accuracy of the resultant predictions. The models were evaluated by calculation of the AUROC (area under the receiver-operator characteristic) and the BEDROC (Boltzmann-enhanced discrimination of the ROC) scores. The AUROC is a classic measure of classifier performance, defined by the area under the curve of false-positive rate (FPR) against true-positive rate (TPR) - both of these depend on the threshold chosen. A classifier that assigns random scores should achieve an AUROC of 0.5. The BEDROC is a metric defined to solve the 'early recognition' problem: in virtual screening of potential drug candidates, it is far more costly to synthesise and test a false positive than it is to ignore a false negative. Therefore, the BEDROC weights performance at the low-false positive end of the curve exponentially more by a factor α. We used $\alpha = 20$ for all BEDROCs reported in this work.

2.4 Biophysically inspired approaches recognise aptamer structural families

We assumed that the active sequences are drawn from some probability distribution, which we approximated; we then attempted to use the likelihood of a sequence under this model to classify actives from inactives. For simplicity, we first assumed that the base appearing at each position in an active sequence is independently chosen from a generalised Bernouilli distribution. This is equivalent to the construction of a position-specific weight matrix across the whole sequence. Although the assumption of independence is unjustified, we expected that in certain positions on the aptamer, having a specific base would be vital to the formation of hydrogen bonds during binding.

$$p(s) = \prod_{i=1}^{L} f_i(s_i) \implies log(p(s)) = \sum_{i=1}^{L} F_i(s_i) + c$$

This model is easy to fit by maximum likelihood; we calculate $f_i(b)$ as the fraction of aptamers with base b at position i To test the fit of our model, we split 20% of the active sequences from each dataset as the test set, and use the other 80% to fit our model. We then evaluate the log-likelihoods of the test sequences from the active set, and an equal number of test sequences from the decoy set and randomly generated sequences.

In our model building, we have not so far considered the important role that aptamer secondary structure plays in binding. It is generally advantageous for a binding molecule to be relatively inflexible in its unbound state, as this decreases the conformational entropy that will be inevitably be lost when it binds. Moreover, the formation of loops or hairpins by base-pairing may direct the bases to the correct positions to form direct interactions with the aptamer. Structural characterisation of aptamers has revealed a range of secondary structural motifs that presumably play a role in binding, such as G-quadruplexes and kissing hairpins. As secondary structure arises from the interaction between bases, we attempted to introduce structural constraints into our model by introducing a

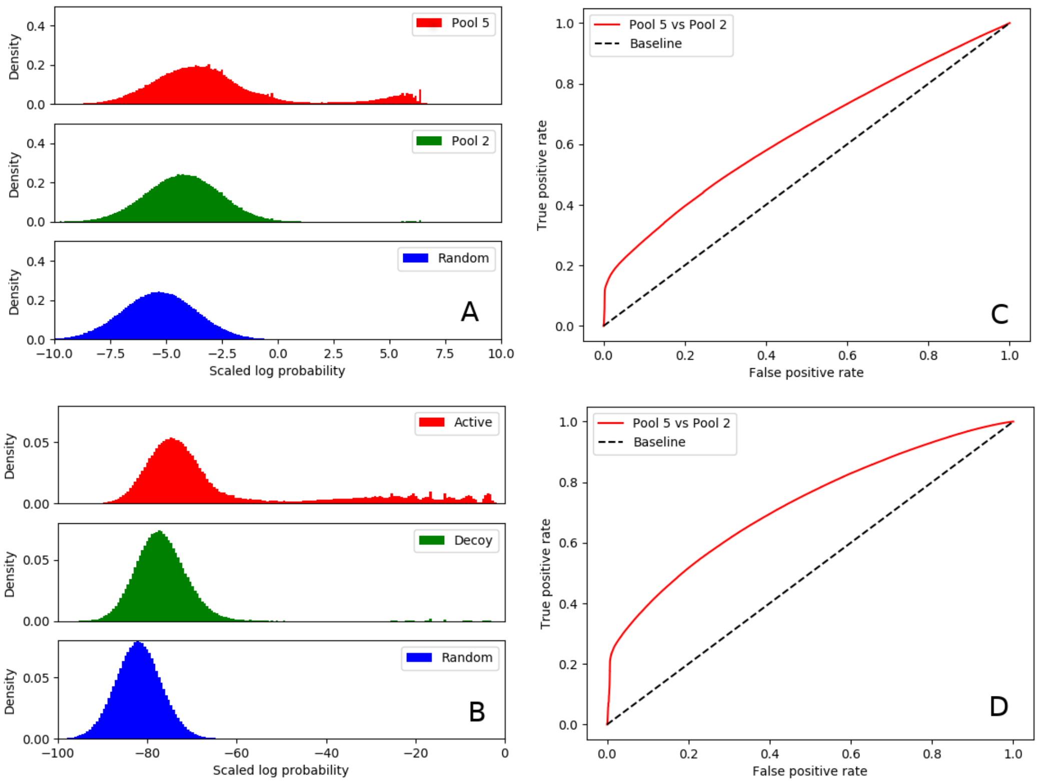

Figure 2: A & B: Histograms of the output of the generalised Bernouilli (A) and Ising (B) models, for test sets of active, decoy and random sequences. C & D: ROC curves for classification between actives and decoys for the generalised Bernouilli (C) and Ising (D) models

‘coupling’ term between positions in the sequence into our unsupervised model. This is equivalent to a Potts model, a family of graphical models which have proven effective in the the analysis and design of protein sequences within structural families. We have enough samples to estimate the couplings matrix using the mean-field expansion.

$$log(p(s)) = \sum_{i=1}^{L} F_i(s_i) + \sum_{i=1}^{L} \sum_{j=1}^{L} J_{ij}(s_i, s_j) + c$$

Note the small number of decoys with scores comparable to the highest binding sequences; this implies that some decoys may in fact be active ROC curve for classification of active sequences against decoys This model dramatically broadens the distribution of likelihoods in the active test set. This improves the AUROC to 0.72; we still see a kink in the AUC curve, but at a higher true-positive rate. The generalisation of this model is better than the previous (Bernouilli) one, but it still remains limited to the structural family which dominates the final pool. Examining the couplings matrix fit to the Ilk data reveals a mixture of short and long range couplings; this is as we might expect, if certain patterns are important but also long-range matching. In this heatmap we plot the sum of squares of magnitudes of all the couplings between each pair of positions. The ultimate failure of models which perform classification based on linear transformations of the sequence features is the failure to account for motif translation. It is well-known that transcription factors bind DNA by the recognition of motifs; it is reasonable to assume that a sequence of bases which strongly affects binding might have a similar effect wherever it is placed in the aptamer sequence. However, it was found in testing that small translations of a known motif could dramatically reduce the effectiveness of recognition.

2.5 Convolutional neural networks classify whole aptamer pools

To construct a model capable of motif recognition regardless of position in sequence, we constructed convolutional neural networks (Goodfellow et al., 2016). These are a class of model used in the recognition of ‘structured data’ such as sequences, images and audio. The network consists of a series of layers; each layer consists of a set of linear kernels, passed along an input, which act as ‘filters’ to detect local patterns. The kernels are of fixed length; the parameters which define which sequences will activate them are variable, and are optimised by using stochastic gradient descent (minimisation of a loss function calculated on random batches of the data). The first layer extracts patterns from a sequence, and the following layers combine the patterns in the preceding layers. The final layer takes a linear combination of the units in the preceding layer and transforms this to a class probability using a logistic sigmoid function. Adding more layers improves the model’s flexibility, but can lead to overfitting. Two network architectures were tested; one with two convolutional layers (‘shallow’), and one with four (‘deep’). The shallow network achieves an AUROC of 0.84 and a BEDROC of 0.91; the deeper network performs slightly better, with an AUROC of 0.87 and a BEDROC of 0.94. The convolutions can effectively recognise patterns, and the networks show better

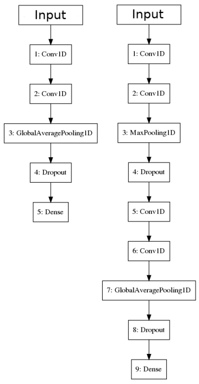

Figure 3: Network architectures used. Left: Shallow, Right: Deep

generalisation than the unsupervised models.

We also tested both architectures on the IgE dataset, revealing a small difference between the active and decoy sets; the test AUROC was 0.55 and the BEDROC was 0.54. This result was verified by training the same architecture on the same dataset with randomised class labels; this reduced the test set accuracy to the baseline value of 0.50, as expected. Using the deep network architecture rather than the shallow one did not improve the classification; due to the smaller dataset and the weaker signal, the advantages of a complex model are reduced.

2.6 Prediction of binding affinity is a more difficult task

Using our classifier, we wanted to see whether the class probabilities generated could be used to predict binding affinity. A double-filter binding assay was used to measure the dissociation constant of 20 radiolabelled RNA aptamers. The class probabilities and binding affinities of these sequences are shown. The deep network is much better at recognising these binding sequences as actives; however there is only weak correlation between the class probability and the dissociation constants. We would hope for there to be a strong negative correlation.

In an attempt to improve the prediction of binding affinity, we selected actives from the Ilk dataset more stringently. For each distinct aptamer sequence, we calculated the relative abundance as $log(\frac{C+2}{N})$, and performed a linear fit for the relative abundance against the round of SELEX. The 1763019 sequences which had a positive slope were classed as 'active', and the 9728712 sequences with a negative slope were classed as 'decoys'. We hoped that, following Hoinka et al., there might be a stronger relationship between relative round-to-round enrichment and binding affinity than between abundance in the final round and binding affinity. Training a CNN with the deep architecture used above on this dataset produces a test AUROC of 0.80 and a BEDROC of 0.68. I suspect the decreased performance of the this dataset may be due to class imbalance; this could be improved by weighting the positive samples better. Unfortunately, the correlation between activity probability

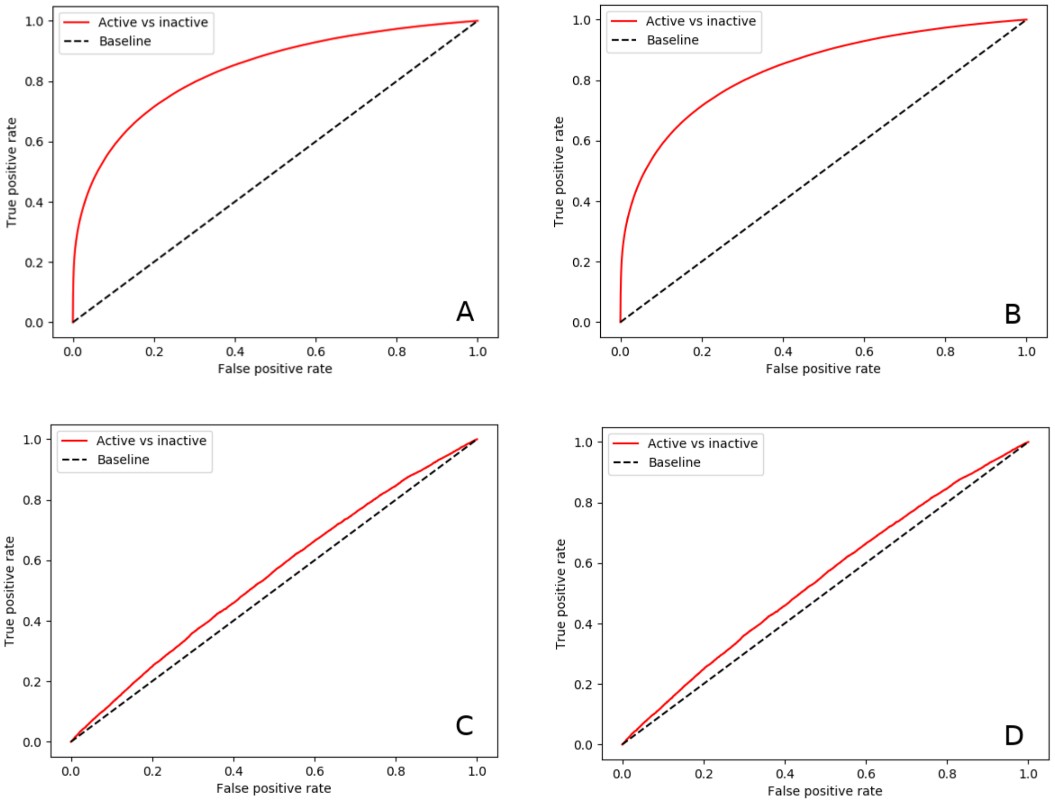

Figure 4: ROC curves for classification of actives from decoys. A (shallow) & B (deep) use the Ilk dataset; C (shallow) & C (deep) use the IgE dataset.

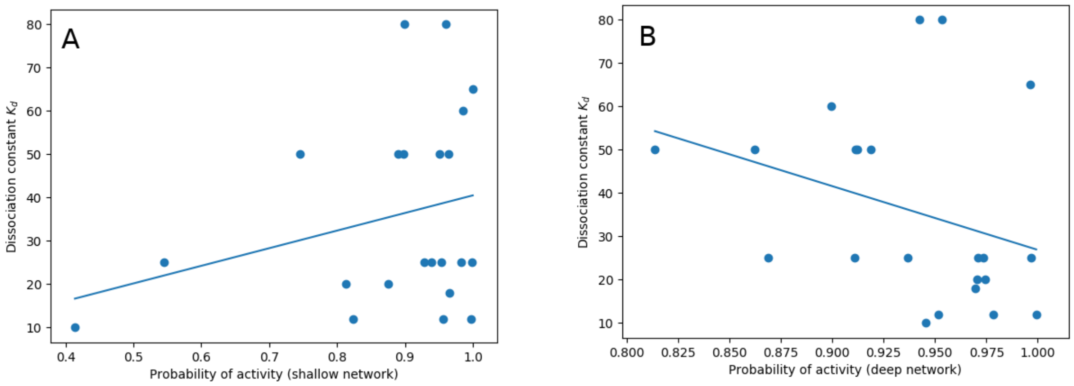

Figure 5: A: Predictions from a shallow network trained on the Ilk dataset plotted against measured binding affinity to Ilk. B: Predictions from a deep network trained on the Ilk dataset plotted against measured binding affinity to Ilk

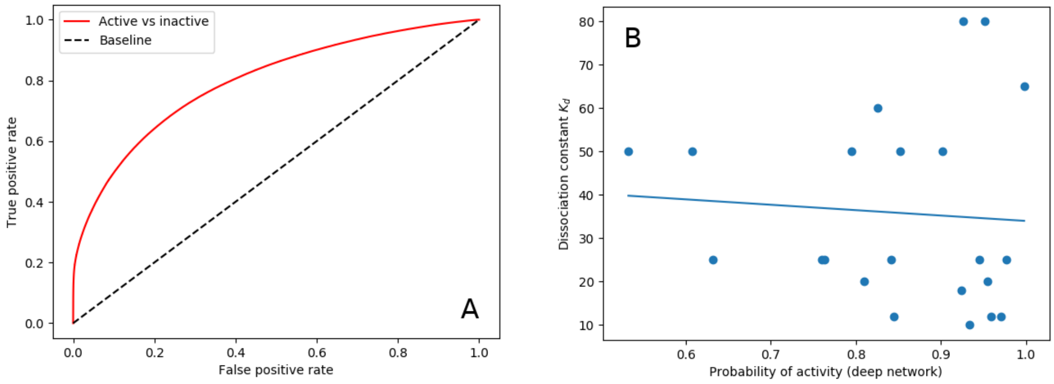

Figure 6: A: ROC curve for classification using the deep CNN on the Ilk dataset with a reduced active set. B: Plot of prediction versus affinity for this dataset.

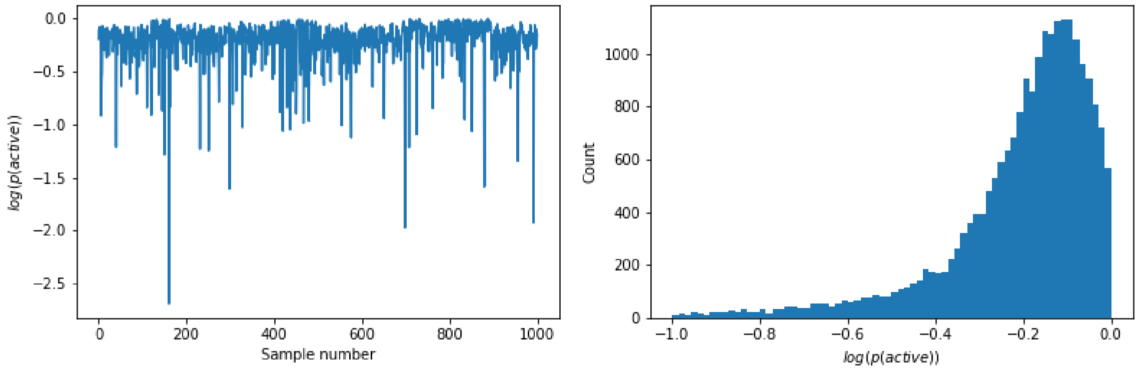

Figure 7: A: A biased random walk through sequence space. B: The histogram of log probabilities of the generated sequences

and binding probability is worse than for the deep model trained on the previous dataset.

2.7 Monte Carlo sampling generates new aptamer sequences

To validate our classifier, we wanted to make testable predictions of sequences which should bind to our target. These could be validated by measuring their Kd. We performed Markov chain-Monte Carlo sampling to explore sequence space, with $-log(p(active))$as the 'energy' to be minimised. Starting with the sequences which were tested by Hoinka et al, we made a random change to each sequence, calculated the change in log probability under the model and accepted the change with a probability 1 if the log probability increased, and probability $exp(-\delta log(p))$if it decreased. Running this for 100000 steps and taking samples every 100 steps generates 357 distinct sequences with p(active) $>$ 0.99.

3 Discussion

We have demonstrated that it is possible to use convolutional neural networks to differentiate high-affinity sequences observed in the later rounds of SELEX experiments from non-binding sequences. To establish this as an effective 'virtual assay' for aptamer-target interactions, we would need to successfully predict the affinity of sequences from a different dataset; predictions have been made, but there was not sufficient time to validate them. It seems that SELEX data are not sufficient to predict binding affinity. Further refinements of the active set selection may be able to improve this, for example by choosing the active set from the clusters which are enriched from round to round. If a larger dataset of aptamers with measured dissocation constants were available, we could train a model to simultaneously classify the SELEX data and perform regression on the dissociation constant-labelled data by multitask learning; two neural networks would be created which share all layers except the output layer. It may be possible to improve the performance of the model by computing DNA/RNA secondary structure and adding this as an input. It is known that RNA secondary structure is an important determinant of aptamer binding, and pre-computing it may improve this. However, we must be aware that RNA secondary structure may be variable and alter on binding to protein - this may affect the degree to which secondary structure is a useful predictor. With enough data and a deep enough neural network, pre-computation of secondary structures should not be necessary (as the network should learn to recognise any relevant aspects of the secondary structure); however it is not clear that we are in this regime. If a virtual assay can be validated, it should be possible to predict binding interactions between a target and any arbitrary sequence. This could be used to design aptamers which interact with multiple targets in a defined way. A virtual screening approach would allow us to specify more complex criteria; for example, it would be possible to modulate GC-content for efficient expression, or to design for selectivity against more than 2 targets. The NUPACK software (Wolfe et al. 2017) currently allows the *in silico* design and checking of DNA and RNA circuits; the ultimate goal of such efforts is to build a 'molecular compiler' which can create nucleic acid nanotechnology to perform a desired task. To design nucleic acid systems which interact with non-nucleic acid components, it is necessary for only the desired parts of the circuit to interact with these components, to prevent 'cross-talk'. Convolutional neural networks based on SELEX datasets could provide a way to perform automated checking that this is the case.

4 Materials and Methods

4.1 Data acquisition and pre-processing

DNA aptamers (300nM) were incubated with human IgE (28.5 uM); the bound protein-DNA complexes were separated from unbound strands by micro-free flow electrophoresis and purified. Sequencing was performed by GenXPro. The reads for screening against Ilk was obtained from the European Nucleotide Archive, study accession number PRJNA315881. The methods used to obtain it are described in Hoinka et al. 2015. Preprocessing was performed for both datasets using AptaPLEX, a dedicated demultiplexer for aptamer data created by Jan Hoinka (Hoinka & Przytycka, 2016). Adapters and barcodes were trimmed and sequences not of length 40, or containing bases of quality below 20 were removed. When reads contained the reverse complement adapters, the reverse complements of the reads were taken to give the true aptamer sequence. The FASTA files containing aptamer sequences from each pool were de-duplicated using the FASTX toolkit, labelled using a custom script in Biopython and shuffled together using the seq-shuf program in perl. Exploratory data analysis (PCA) was performed using the sklearn library in python.

4.2 Model building and testing

The models were written using the numpy, sklearn and keras libraries in Python. Datasets that did not fit into RAM were handled using the h5py library. The datasets were split randomly into 90% training and 10% test data. Optimal model parameters were chosen by varying each parameter and evaluating loss on the training set; the test set was then used to compare the models.

4.3 Sequence prediction

The Markov chain-Monte Carlo sampler was coded in Python using the numpy library and run for 100000 steps with an inverse temperature of 20, taking samples every 100 steps.

5 Bibliography

Tuerk & Gold 1990
https://science-sciencemag-org.ezp.lib.cam.ac.uk/content/249/4968/505.long
Ellington & Szostak 1990
https://www-nature-com.ezp.lib.cam.ac.uk/articles/346818a0
Strauss et al. 2018 - Super-res microscopy using SOMAmer labels
https://www-nature-com.ezp.lib.cam.ac.uk/articles/s41592-018-0105-0
Win & Smolke 2007 - using aptamers for synthetic biology
https://www-pnas-org.ezp.lib.cam.ac.uk/content/104/36/14283
Hentze et al. 2018 - a review of RNA-binding proteins
https://www-nature-com.ezp.lib.cam.ac.uk/articles/nrm.2017.130
Rohloff et al 2014 - SOMAmers
https://www-sciencedirect-com.ezp.lib.cam.ac.uk/science/article/pii/S2162253116303365?via