Nanopore Basecalling and Modification Detection Using Deep Learning

Rojin Safavi[1], Andrew Bailey[2], Miten Jain[2], Benedict Paten[2], Mark Akeson[2]

Department of Molecular, Cell, and Developmental Biology: University of California, Santa Cruz

Department of Biomolecular Engineering: University of California, Santa Cruz

Introduction

Figure: Oxford nanopore device

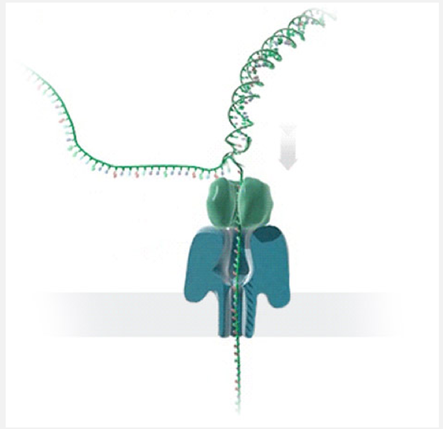

Figure: nanopore

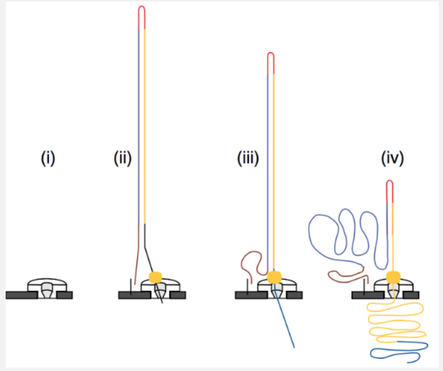

Figure: DNA trans-location

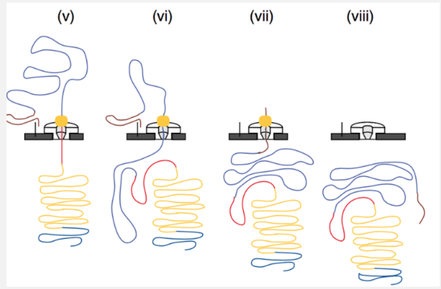

Figure: DNA trans-location

	mean	move	stdv	length	mp_state
0	60.72320...	1	0.791058...	0.0015	GCCAC
1	64.94183...	2	1.593811...	0.002250...	CACTA
2	64.71323...	0	1.619748...	0.004500...	CACTA

Figure: Event table, the Nanopore output

SignalAlign

MinION signal-level alignment and methylation detection using hidden Markov Models with hierarchical Dirichlet process kmer learning.

Event Index	Reference Position	Base	Posterior Probability
1403	837835	E	0.901382
1403	837835	C	0.087196

Figure: SignalAlign output

Analysis Workflow

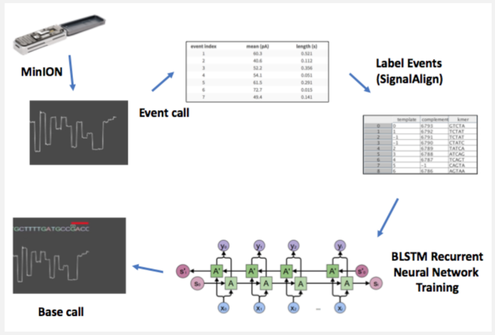

Figure: Analysis Workflow

Recurrent Neural Network

- RNN is used when we want to make prediction in sequential data.
- RNN has memory that influences future predictions.
- As an example, in voice recognition you may want to make prediction based on the previous.

Example of RNN implementation for canonical basecalling:

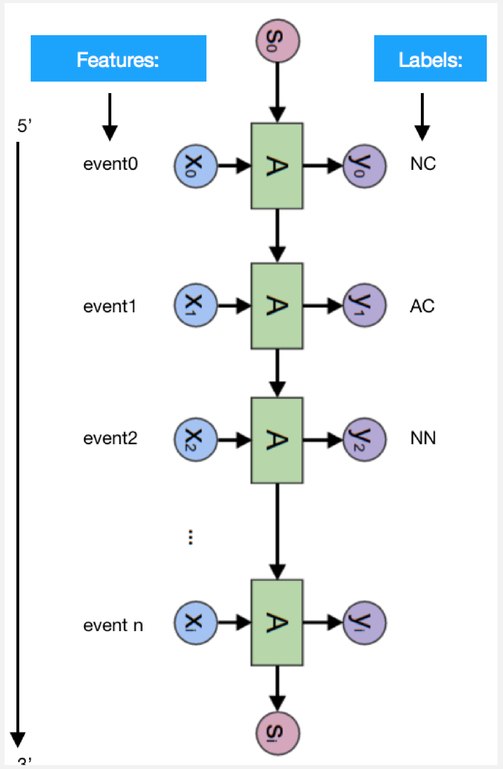

Figure: RNN implementation for canonical basecalling

- The RNN Network model is trained on a training set that has known labels for each event feature.
- The model prediction accuracy is then measured on a test set with known labels.
- **Accuracy for canonical basecalling:** 67%

Conclusion

- The accuracy is lower compare to other nanopore basecaller that use deep learning(deepnano = 77.9 for template strand)

Future directions

- Using convolutional neural network in addition to bidirectional neural network to boost the accuracy.
- Using signalAlign methylation information (posterior probabilities) to train our model for modifications
- Using raw signal information to generate features instead of using event table

References

1. http://colah.github.io
2. https://github.com/ArtRand/signalAlign

Contact Information

- Email: rsafavi@ucsc.edu