# LSTM

## Reasons

* Can be a standardised specified space for any length kmers, to a degree
* Might be faster than DNA2VEC or other counting methods
* Generalised method for embedding not specific for a task

<https://arxiv.org/pdf/1906.03087.pdf>

this work uses an encoder and decoder lstm to train the model, once adequate the encoder lstm is used to represent the sequence in a fixed size. As the model can obviously decode it back into tits original state, the model must be able to encode it to a state where it can be represented in a fixed size

# DNA2VECs

Simple conversion using DNA2vec or similar, then give the output as the inputs to knn and svm

# Suggested one

Another simple conversion and then inputs for knn and svm

# Comparions

* Knn
* SVM
* Random Forest

# Data

Data sizes will be abit off sequence wise as they are currently fastq files. Will likely use the sequences with human reads and the bacterium removed.

* Use only csd\_17 data, 673 files, 379 GB
  + Not the worst still quite large though and will take a while for training
* Use only csd\_16 data, 392 files, 814 GB
  + Not this one too big
* Split be region, e.g. Europe
  + Don’t know if it will work on worldwide generalised scale. Could just work for Europe
* Use few thousand samples from each data file
  + Would still be quite large due to sheer number of files
* Use few thousand samples from just CSD 17 data
  + Probably the best option, wont be too large
  + Will still be global
  + Can use left over data for testing if required.

# Model

* <https://arxiv.org/pdf/1906.03087.pdf>
  + Can be based similarity of this paper
* Ideas for training the system
  + Use lstm autoencoder - would be useful for generally compressing a sequence
  + Use LSTM then directly to knn, ect. – would be more specific for the task.

RNN for certains tasks instead of LSTM

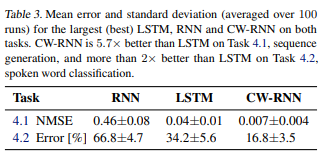
<http://proceedings.mlr.press/v37/jozefowicz15.pdf>

Comparisons of LSTM mdoels

<https://arxiv.org/pdf/1503.04069.pdf>

clockwork rnn looks promising

<https://arxiv.org/pdf/1402.3511.pdf>



Learning the Language of the Genome using RNN - <https://cs224d.stanford.edu/reports/jessesz.pdf>