Machine Learning Challenges in DNA Sequencing

Tomáš Vinař

Computational Biology Research Group

Faculty of Mathematics, Physics and Informatics

Comenius University in Bratislava

http://compbio.fmph.uniba.sk/

slido `#mlmu`

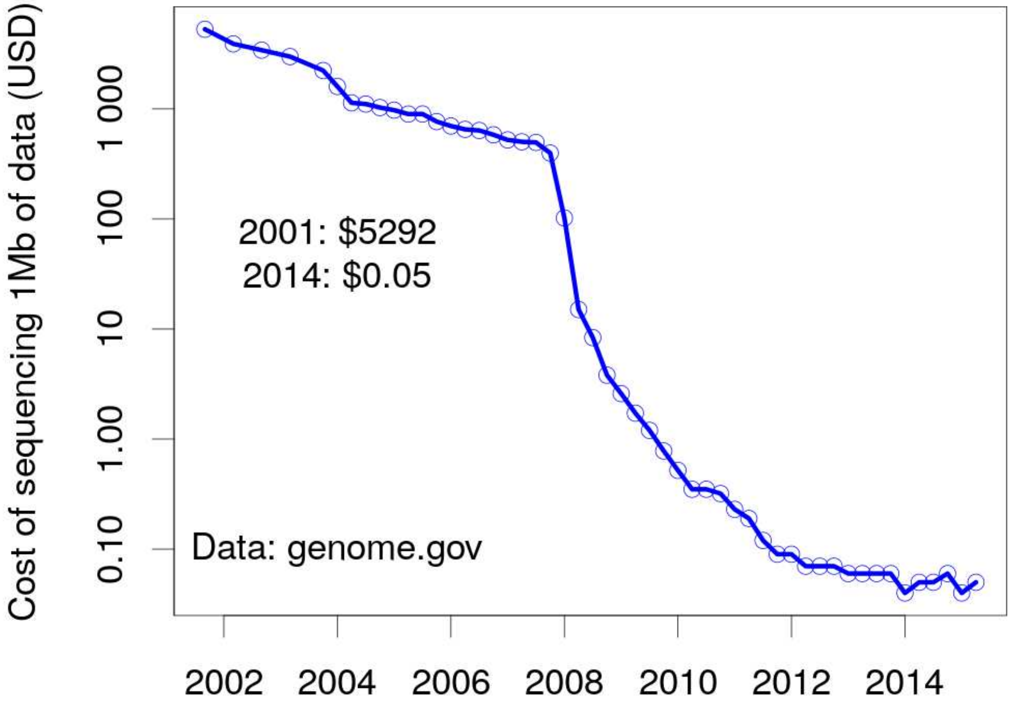
DNA Sequencing is Evolving Rapidly

ABI Sanger sequencing 2001: 115 kB per day

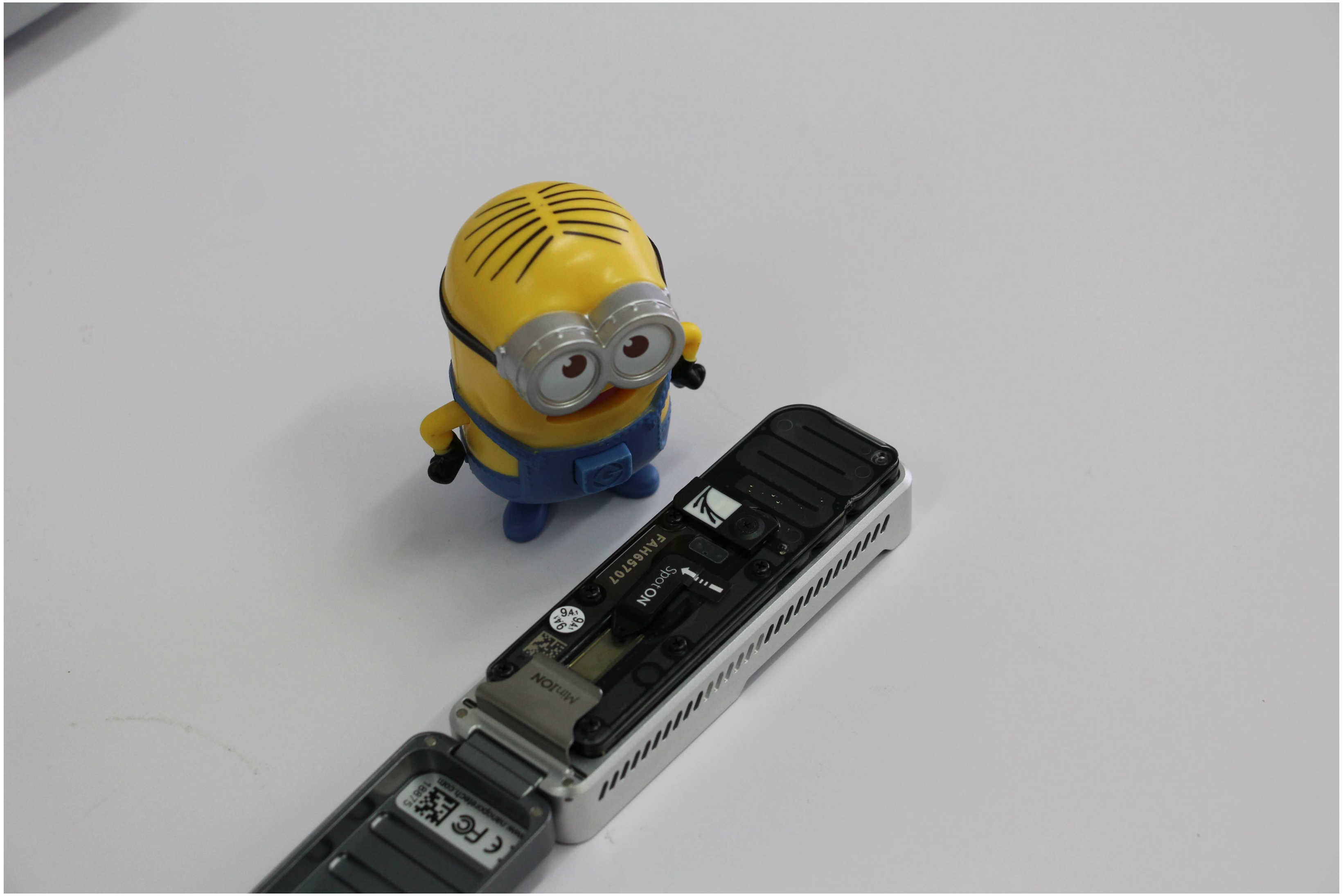
limited to large sequencing center, international consortia

Illumina HiSeq 4000 2015: 107 GB per day

can be operated in a small(ish) laboratory

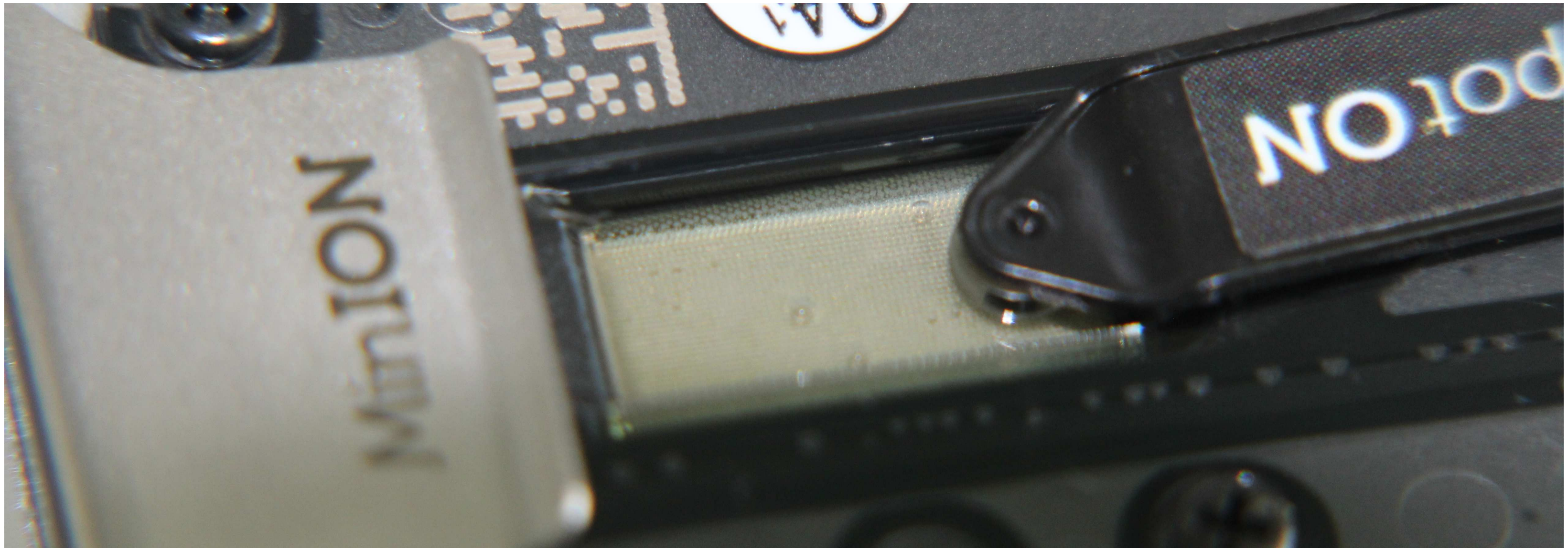


MinION: Sequencing of Anything, Anywhere, by Anyone



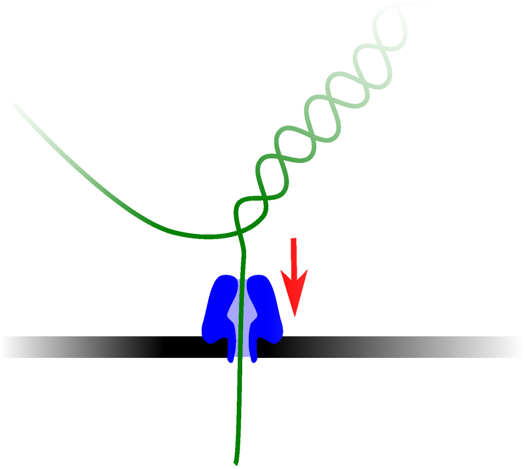
Sequencing in Our Group

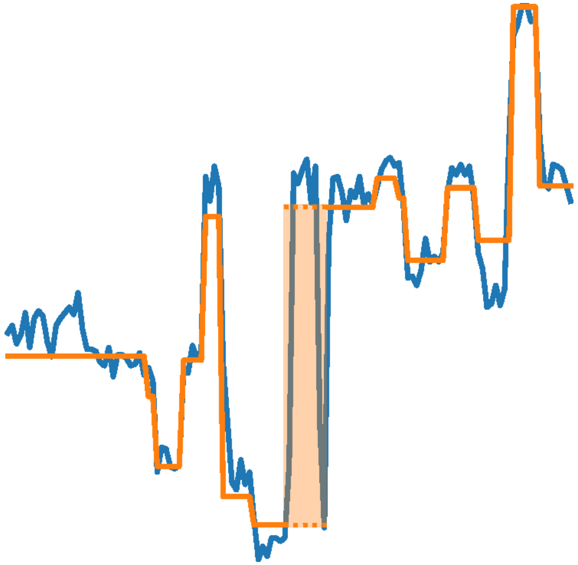
- DNA and RNA of non-conventional yeasts (*Magnusiomyces* clade)
- >10 GB of data in approx. 24 hour sequencing run
- About 0.10 EUR per 1 MB of data



Inside MinION

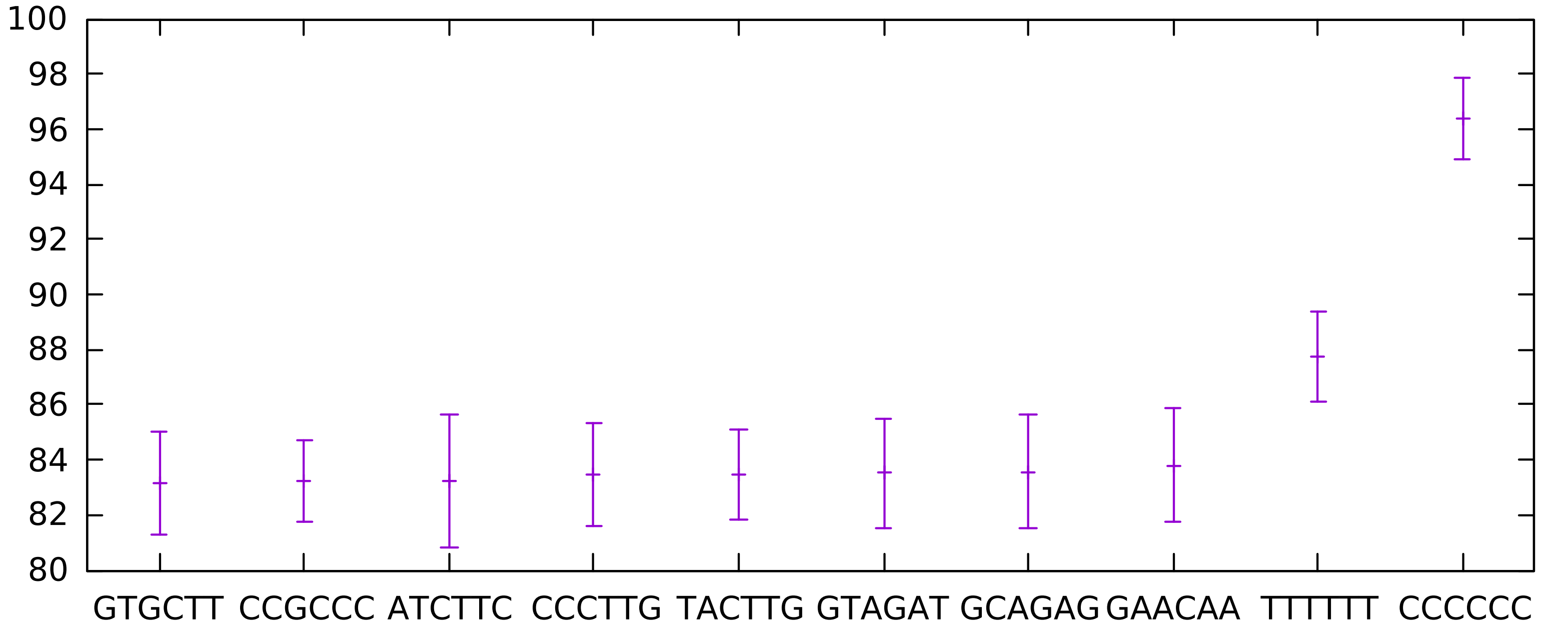
- DNA passing through a nanopore causes **changes of electrical** current based on the **context** of $\mathrm{k}(=6)$ bp
- 4000 measurements per second (approx. 10 measurements per context) $\Rightarrow$ **squiggles**
- Squiggles are translated to DNA sequences through **base calling**





Figures: Eduard Batmendijn

Base Calling for Nanopore Sequencing is Difficult



Boža, Brejová, Vinař (2017) PLoS One

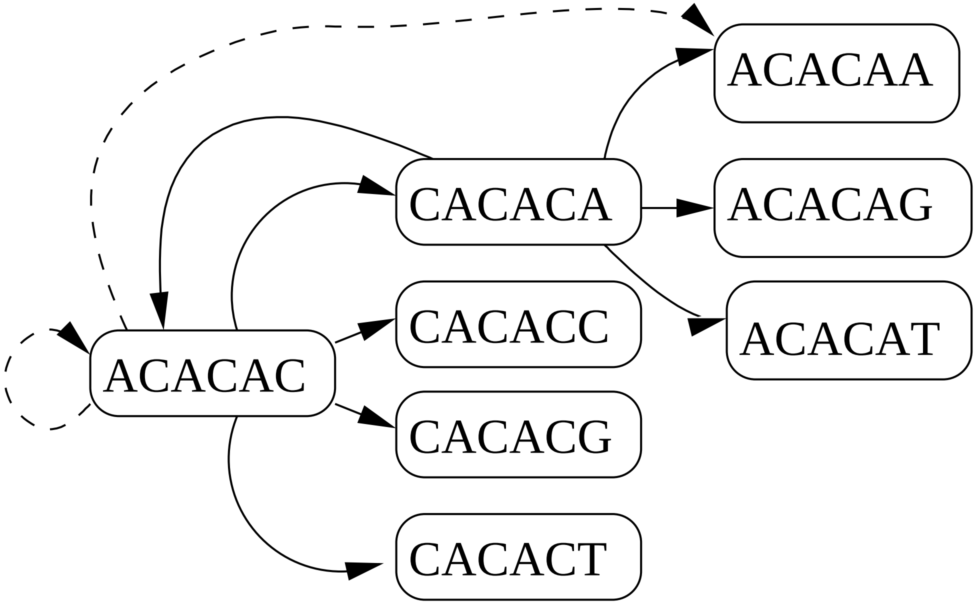
Base Calling Using hidden Markov Models (80%)

- Split the squiggle into **events** corresponding to approx. one context shift
- Events will usually **overlap** by $k-1$ basepairs
- This can be represented by hidden Markov models

  hidden states = k-mers of DNA

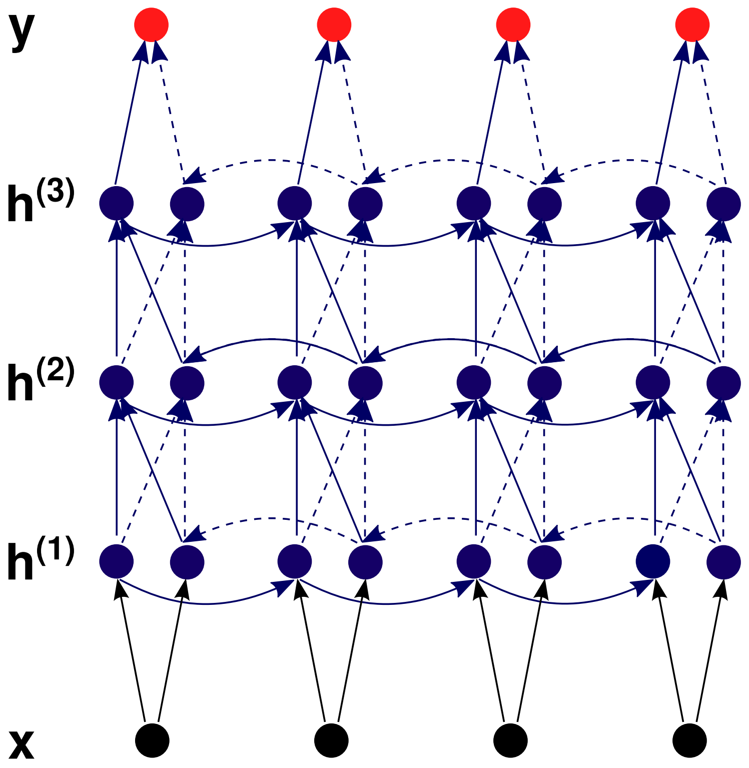
  emission probabilities = Gaussian distributions according to expected signal

  in the context of a k-mer



Discriminative Modeling Instead of Generative (85%)

- Recurrent neural networks
  (in our case: GRUs)
- Input vectors: for each event
  (mean, stdev, length)
- Output vectors: for each event
  0, 1 or 2 basepairs
- **Training problem:** Outputs are not uniquely aligned to events
- **Solution:** Start with some alignment
  Periodically realign based on newest predictions



Boža, Brejová, Vinař (2017) PLoS One

Connectionist Temporal Classification (90%)

- Softmax layer with one additional divider symbol |
  (output values interpreted as probabilities)
- Special CTC layer with the effect of summing up “similar” signal segmentations into one value

$$\left.\begin{array}{l} \texttt{HHHEEEEE|LLLL|LO} \\ \texttt{HEEEEELL|LLLLLLL|OOOOO} \\ \texttt{HHHHHHHHHHHEL|L|O} \\ \texttt{H|E|L|L|O} \end{array}\right\} \Rightarrow \texttt{HELLO}$$

- Works transparently with gradient descent training
- Heuristics for finding **most probable** sequence labeling
  (no fixed assignment of the outputs to inputs)

Graves et al. ICML 2006; Boža unpublished

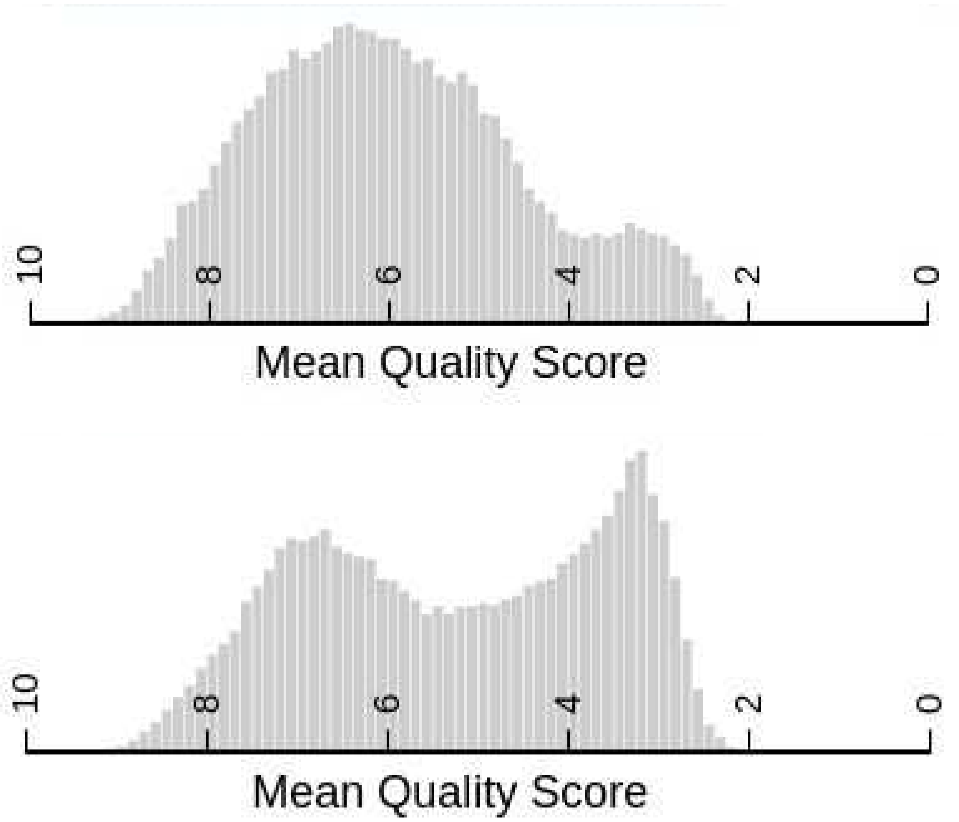
MinION Base Calling Summary

- MinION devices produce data reasonably fast and reasonably cheaply under **almost home** conditions
- They produce very long reads (compared to previous technologies) 200 B (Illumina) vs. 10 KB mean (MinION)
- Sequencing squiggles need to be base called: translated to DNA
- Lots of ambiguity $\Rightarrow$ need to use RNN with CTC
- Even best methods give **only 90% accuracy** (compared to 99.99% Illumina)
- Still, if we use **consensus of multiple overlapping reads**, we can get to 99% accuracy

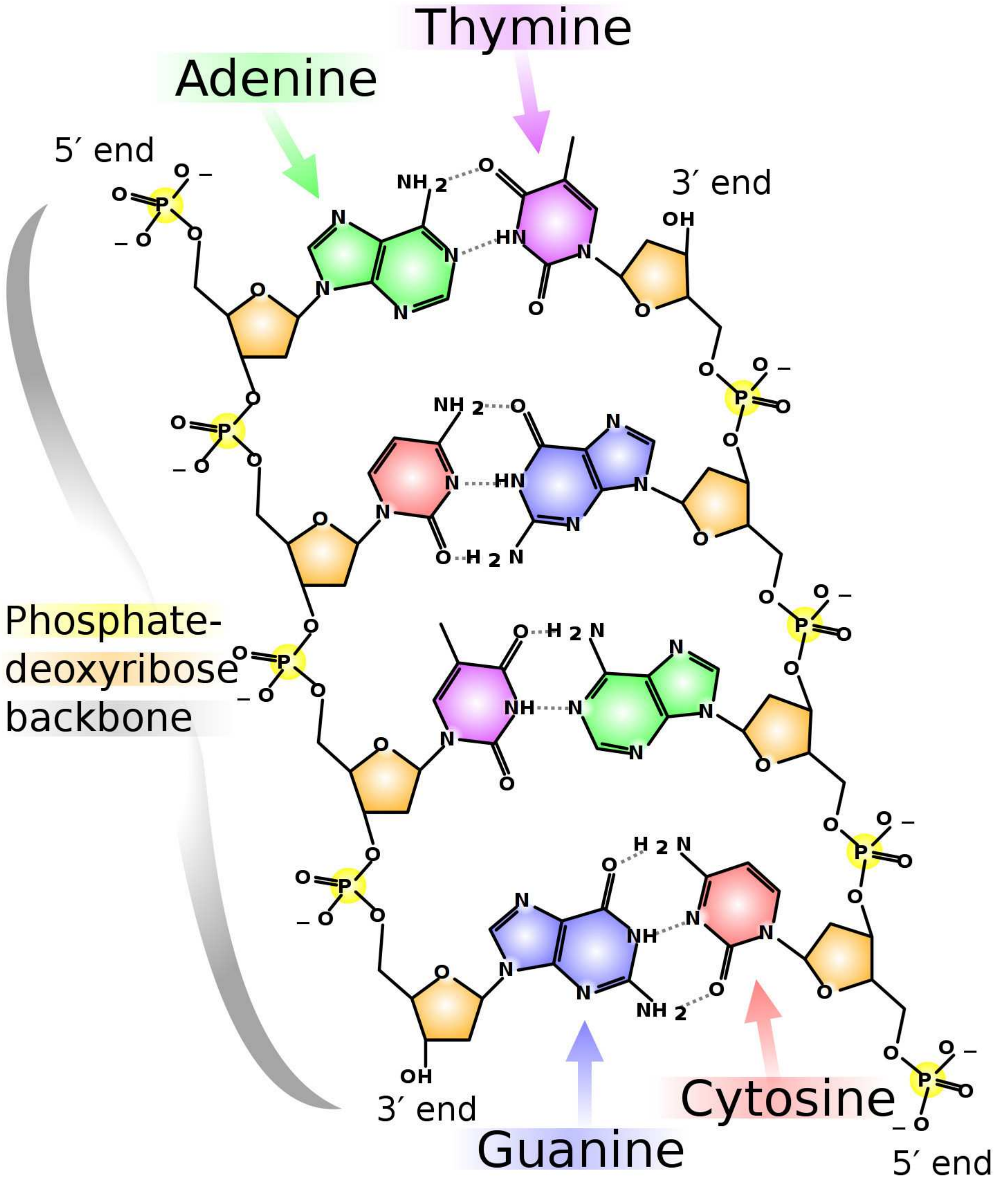
More Bad News About RNN Solutions

- The training requires large amounts of data

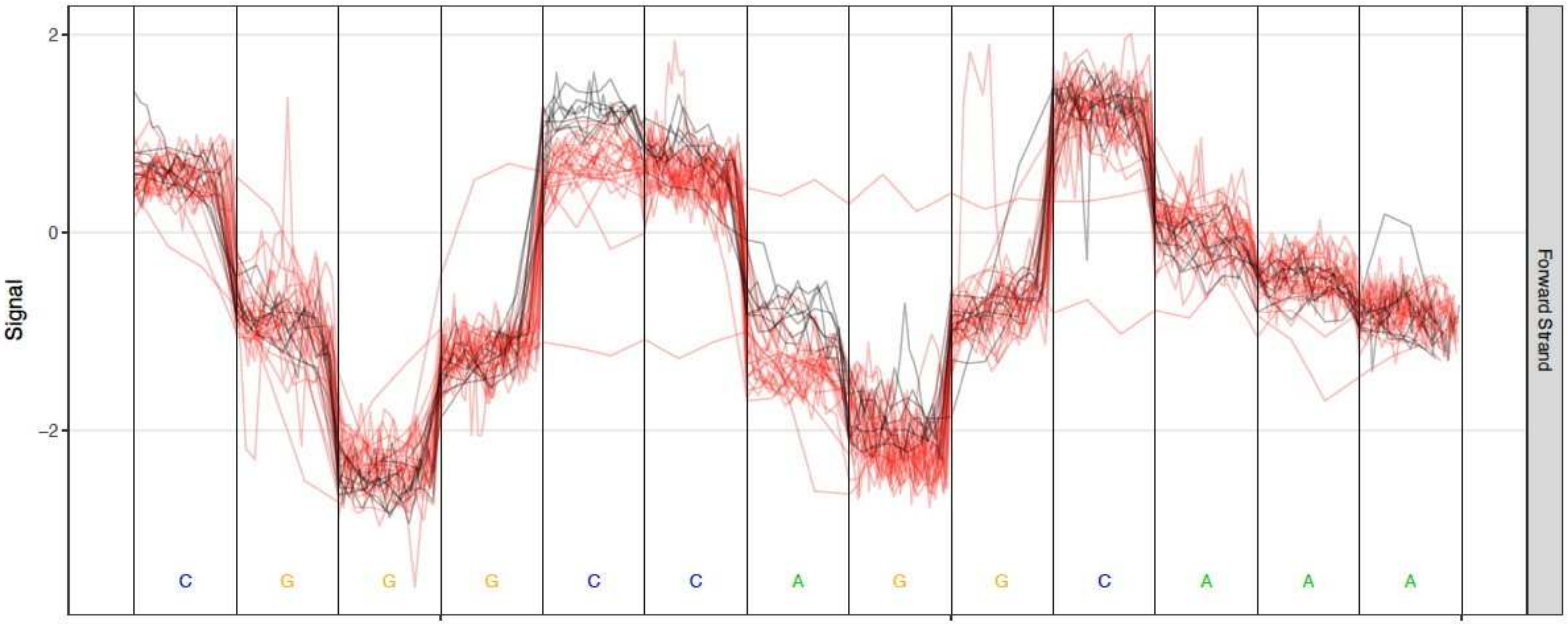
  . . . some of which is **very difficult or impossible** to get!



(changes in base call quality after the sample was stored for a few weeks)


Thymine
Adenine
5′ end
3′ end
OH
Phosphate-
deoxyribose
backbone
3′ end
OH
Cytosine
Guanine
5′ end
NH 2
O
HN
N
P
O –
– O
H 2 N
NH

Modified Base Detection with MinION

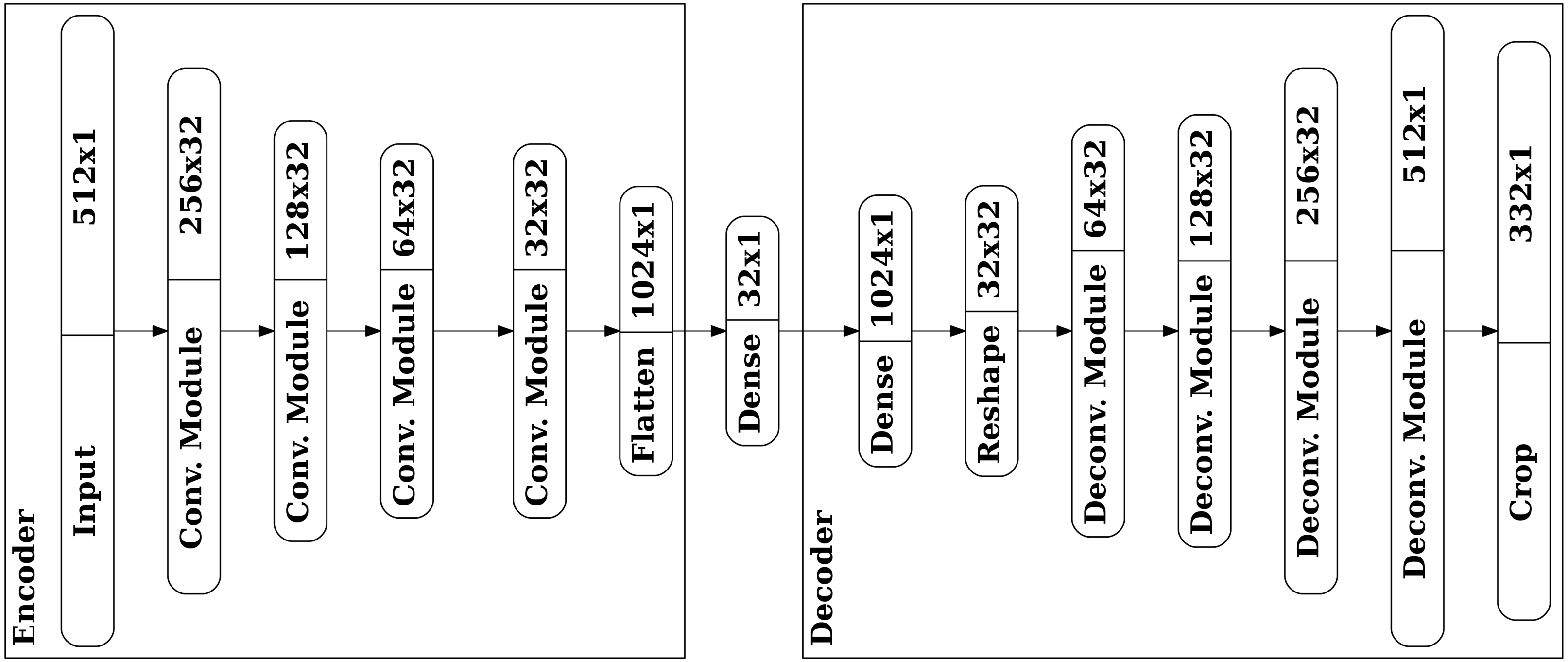


Oxford Nanopore, Tombo documentation

Learning “Clean” Signal Characteristics

Use autoencoder neural network to characterize a “typical” signal

(learn from **control sample** with modifications removed)

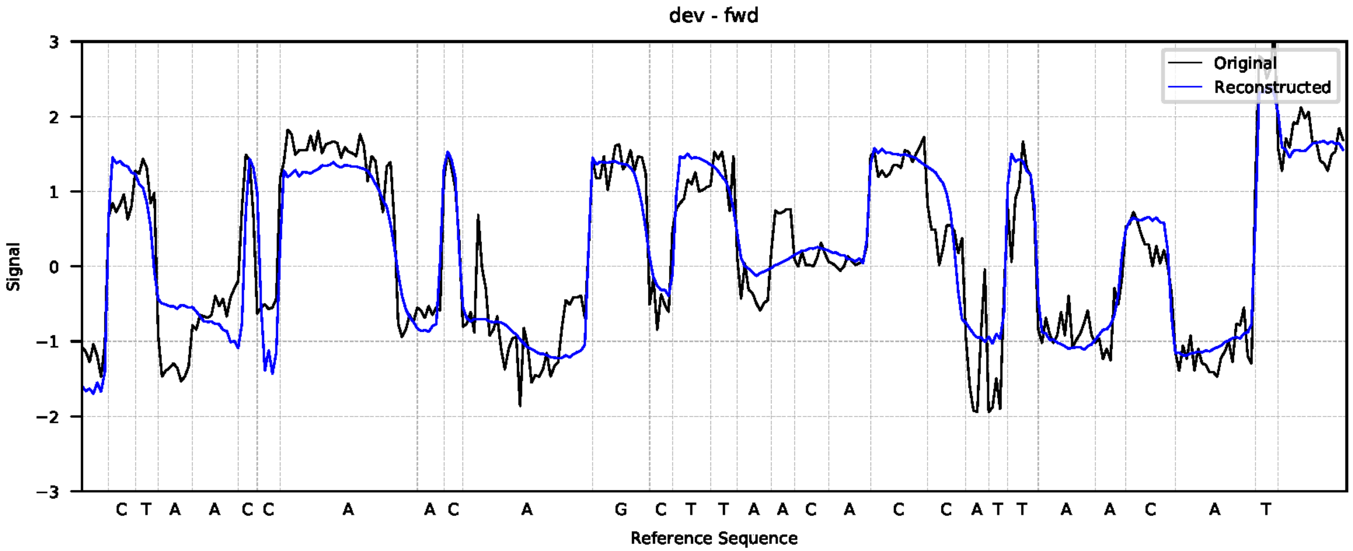


Rabatin et al. 2018, unpublished

Learning “Clean” Signal Characteristics (cont.)

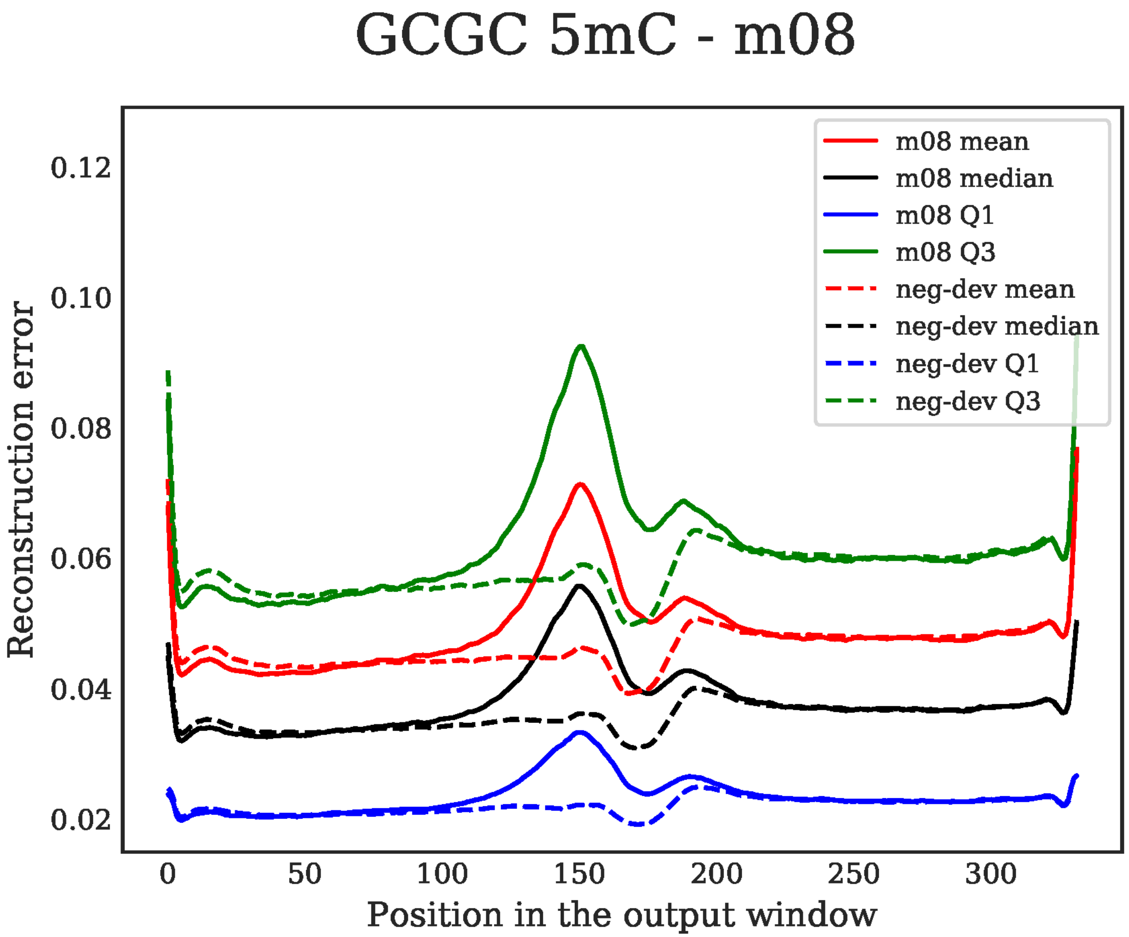
Signal can be reconstructed from the “bottleneck” layer

Large reconstruction error indicates unusual patterns



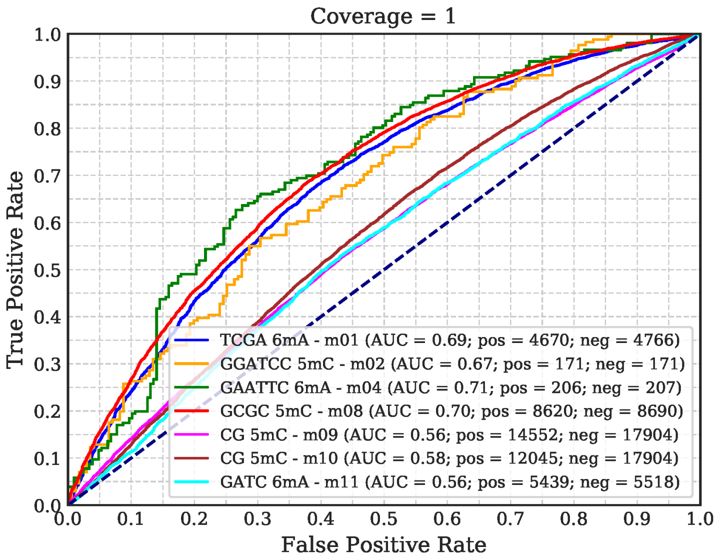
Rabatin et al. 2018, unpublished

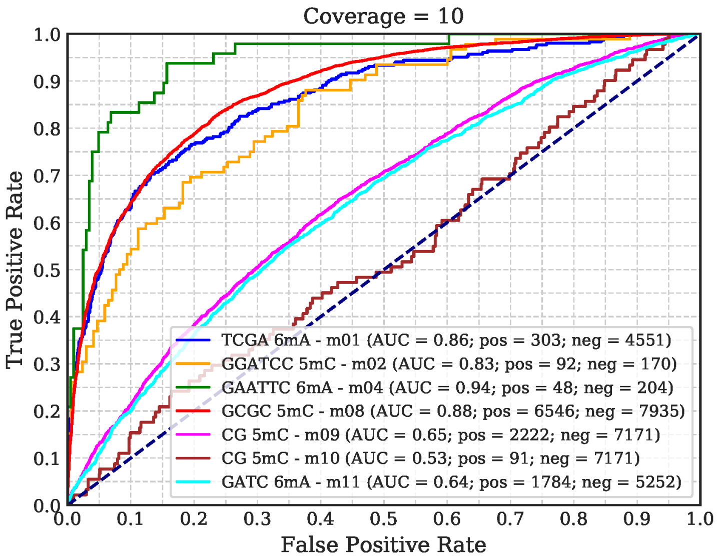
Looking for Large Reconstruction Errors



Rabatin et al. 2018, unpublished

Evaluation on Data Sets with Known Methylation Status





can detect modified bases

even without knowing anything about their characteristics!

(using also 11bp sequence context as input to decoder)

Rabatin et al. 2018, unpublished

Selection Using ReadUntil

- We look at the squiggle on-line (take 1s beginning)
- If signal corresponds to **wanted** sequence, continue reading
- Otherwise, **terminate the read**

Problems:

- 500 pores in parallel x 4000 measurements per second

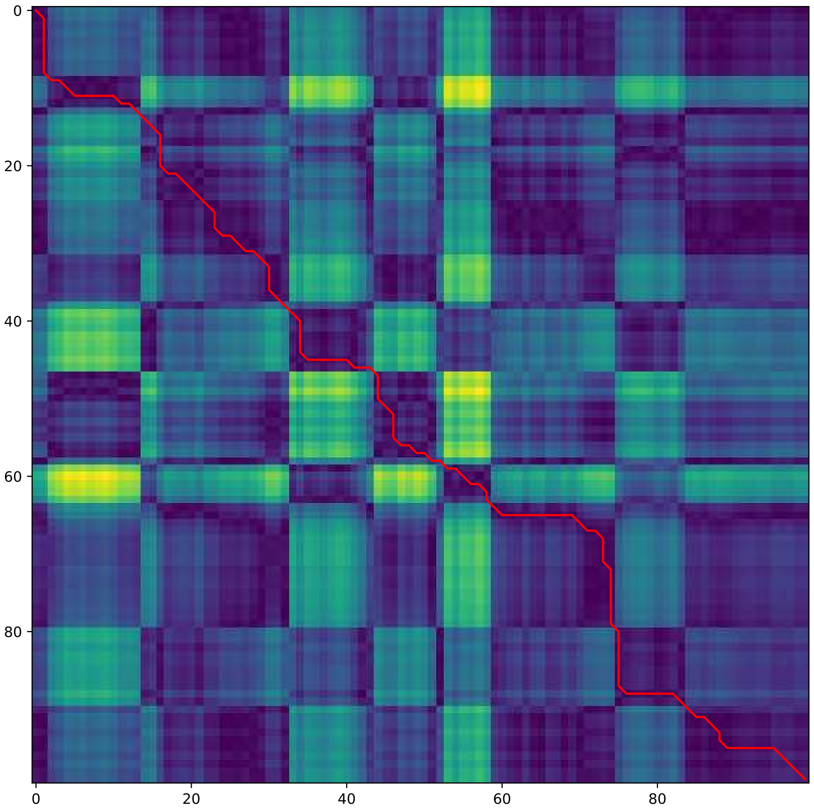
  (cca 10 measurements per base)
- base calling in real time requires high-performance server (24 CPUs)

Possible solutions:

- special hardware accelerators
- **working with squiggles at the signal level**

Loose et al. (2016) Nature Methods

Dynamic Time Warping from Speech Processing



Sakoe and Chiba 1978; figure: David Barbora

Summary

- MinION reads contain information about methylation patterns
- Autoencoders can pick up irregularities in squiggles indicating methylation patterns
- Working directly with squiggles seems to be the way to go in many applications (future work)

Bioinformatics Research Group, http://compbio.fmph.uniba.sk

Facuty of Mathematics, Physics and Informatics, Bratislava

#NGSchool2018

Nanopore sequencing and Personalised Medicine
16-23 September 2018, Lublin, Poland

People from all over the world!

MinION hackathon

©Hand Drawn Goods

Personalised medicine

Apply by June 30 at https://ngschool.eu/2018

Visegrad countries receive priority

Want to work with us?

- Post-doc positions (applications by June 30, 2018)
- PhD studies (applications by April 30, 2019)
- Maybe we will be looking for programmers (September 2018)
- “Bioinformatics and Machine Learning” option in master’s CS program
- Bioinformatics bachelor’s program

Courses at the Faculty of Mathematics, Physics and Informatics

- Methods in Bioinformatics (Fall 2018 - T. Vinař, B. Brejová)
- Machine Learning (Fall 2018 - Vlado Boža)
- Graphical Models in Machine Learning (Spring 2019 - Tomáš Vinař)
- Modern Topics in Machine Learning (Spring 2020 - Vlado Boža)
- Bioinformatics Seminar (all the time)

Tomáš Vinař

Computational Biology Research Group

http://compbio.fmph.uniba.sk/

Funding: APVV-14-0253 (JN), VEGA 1/0684/16 (BB), VEGA 1/0458/18 (TV)