Electronic Supplementary Information (ESI)

High-bandwidth nanopore analysis by using a modified hidden Markov model

Jianhua Zhang,*[a] Xiuling Liu,[a] Yi-Lun Ying,*[b] Zhen Gu,[b] Fu-na Meng[b] and Yi-Tao Long[b]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, P.R. China

[b] Key Lab for Advanced Materials and Department of Chemistry, East China University of Science and Technology, Shanghai 200237, P.R. China

*E-mail: zhangjh@ecust.edu.cn.
*E-mail: yilunling@ecust.edu.cn.
Tel.: +86-21-64253808

Hidden Markov Model

A standard HMM contains the following parameters[1-3]:

1. The observation sequence $O_1, O_2, \L, O_T$ denoting a set of samples with the length of *T*.

2. The set of hidden states $S = \{S_1, S_2, \L, S_N\}$. Each observed sample $O_t, t = 1, 2, \L, T$ belongs to S with certain probability.

3. The $N \times N$ state transition probability matrix $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, $1 \le i, j \le N$ denotes the state of the HMM at time t.

4. The initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i), 1 \le i \le N$ is *N*-dimensional column vector.

5. The observation probability distribution matrix in the state S_j: $B = \{b_j(O_t)\}$, where O_t is the observation at time t, and $b_j(O_t) = P(O_t \mid q_t = S_j), 1 \le j \le N$.

The HMM is usually used to solve the following three typical problems:

Problem 1. Given the model $\lambda = (\pi, A, B)$, determine the occurrence probability $P(O \mid \lambda)$ of observation sequence $O_1, O_2, \L, O_T$. The typical method for this problem is Forward and Backward algorithm.

Problem 2. Given the model $\lambda = (\pi, A, B)$ and observation sequence $O_1, O_2, \L, O_T$, find the optimal state sequence $q_1, q_2, \L, q_T$ such that the probability $P(O, S \mid \lambda)$ is maximized. The typical method for this problem is Viterbi algorithm, which will be briefly introduced later on.

Problem 3. Adjust the parameters in the model $\lambda = (\pi, A, B)$ such that the probability $P(O \mid \lambda)$ is maximized. There are two typical methods to optimize the HMM parameters: the Viterbi training algorithm (aka. segmental k-means in some literature)[4] and Baum-Welch algorithm

The Viterbi algorithm

To find the optimum state sequence $q_1, q_2, \L, q_t$ of observation sequence $O_1, O_2, \L, O_t$, we define the maximum probability along a single path at time t which accounts for the first t observations by the hidden state S_i as:

$$\delta_t(i) = \max_{q_1, q_2, \L, q_{t-1}} P(q_1 q_2 \L\ q_t = S_i, O_1 O_2 \L\ O_t \mid \lambda) \tag{9}$$

then we have

$$\delta_{t+1}(j) = \max_{1 \le i \le N}[\delta_t(i) a_{ij}] b_j(O_{t+1}) \tag{10}$$

Moreover, we use $\psi_t(j)$ to indicate the state that maximizes $\delta_t(j)$. The procedure of finding the optimal state sequence consists of the steps as follows.

Step 1 - Initialization:

$$\begin{aligned} &\delta_1(i) = \pi_i b_i(O_1), 1 \le i \le N \\ &\psi_1(i) = 0 \end{aligned} \tag{3}$$

Step 2 - Recursion:

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t),\ 2 \le t \le T, 1 \le j \le N$$
$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}],\quad 2 \le t \le T, 1 \le j \le N \tag{4}$$

Step 3 - Termination:

$$P^* = \max_{1\le i\le N}[\delta_T(i)]$$
$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)] \tag{5}$$

Step 4 - Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\quad t = T-1, T-2, \text{L}\ 1 \tag{6}$$

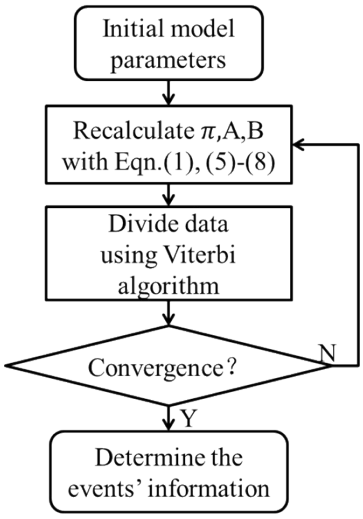

Fig. S1. Flowchart of Viterbi training algorithm that alternately executes parameter re-estimation and Viterbi algorithm until convergence is reached.

Table S1. Comparison of the duration of the events estimated by MHMM, FWHM and DBC algorithms.

Method	Duration of the detected event (μs)				
	I	II	III	IV	V
MHMM	70	80	70	80	140
FWHM	170	140	110	150	160
DBC	110	110	80	100	120

*MHMM was used to process the unfiltered experimental data. † FWHM and DBC were used to analysis the filtered experimental data. A poly$(dA)_{30}$ traversing through an α-hemolysin nanopore under 100 mV.

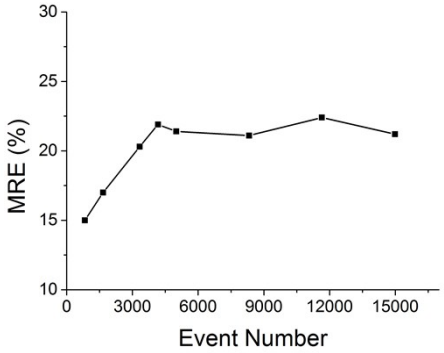

Fig. S2. The change of MRE index as a function of the number of events. The MRE of the MHMM method for 70 μs-long simulated events with σ = 0.5 pA. The simulated events are generated in random, whose number is varied from 1000 to 15000.

References

(1) L R. Rabiner, *Proc. IEEE*., **1989**, *77*, 257-286.

(2) R. Dugad, U. B. Desai, *Bombay Technical Report No.: SPANN-96.1*., **1996**, 1-16.

(3) B. H. Juang, L. R. Rabiner, *Technometrics*., **1991**, *33*, 251-272.

(4) T. K. Bhowmik, J. P. van Oosten, L. Schomaker, Presented at *Int. Conf. on Pattern Recognition and Machine Intelligence*, June 27 - July 01, 2011, Moscow, Russia.