Supplementary information

for

Single-molecule Detection of Modified Amino Acid Regulating Transcriptional Activity

Y. Komoto*, T. Ohshiro, Y. Notsu & M. Taniguchi*

Correspondence to: taniguti@sanken.osaka-u.ac.jp

This PDF file includes:

SI1. Methods

Sample preparation

L-lysine ≥98% (Sigma Aldrich) and N_ε-acetyl-L-lysine ≥98% (Sigma Aldrich) and Glycine ≥ 99% (Sigma Aldrich) were dissolved in milliQ water to prepare a 1 μM solution.

Device fabrication

Schematic structure of MCBJ device is shown in Figure 1. Polyimide film was coated as an insulating layer onto thin-silicon substrate with spin coating. A gold nanowire was drawn on the silicon substrate by electron-beam lithography. Then, SiO_2 film was coated on the gold nanowire by chemical vapor deposition. Narrowest part of nanowire is several ten nm. Polyimide layer under gold nanowire were removed by dry-etching to form free-standing gold-wire.

Electrical measurements

Single-molecule conductance measurements were performed at the optimal gap distance of the nanogap electrodes as described previously in reference 17-18 in main text. A lithographically fabricated gold nanowire on a thin silicone substrate was broken by mechanically bending the substrate under ambient condition with application of a bias voltage of 100 mV, and the single detection part of the nanogap electrodes was formed. Throughout the junction breaking process, the junction conductance (G) was monitored using a picoammeter (Keithley 6487). A series of conductance jumps of the order of $G_0 = 2e^2/h$ (where e and h are the elementary charge and Planck's constant, respectively) was observed, and the final conductance was 1 G_0. Several seconds after reaching the 1 G_0 state, a gold atomic junction naturally ruptured in the nanowire, creating a nanogap. The gap size was controlled using the piezo bias voltage. The gap width was

0.56, 0.58, and 0.60 nm. The gap distance was estimated from the baseline tunnelling current as following section.

Estimation of gap distance

The gap distance is estimated using by following current equation of direct tunneling current

$$I = const\, \exp\left(-\frac{4\pi}{h}\sqrt{2mwl}\right).$$

Here, h, m,w, and l represents plank constant, electron mass, work function of gold electrode, gap distance. We used electron mass of 9.1 $\times 10^{-31}$ kg as m, and work function of Au(111) 5.3 eV as w. Effective mass and work function of gold nanogap not (111) surface should be used for accurate estimation. Furthermore, the inelastic gold gap broadening just after breaking atomic junction is not under consideration. Hence, the experimental gap length is larger than the target width of 0.56, 0.58, and 0.60 nm .

Theoretical Calculation

DFT calculations of lysine and acetyl lysine dG isolated molecules were calculated using Gaussian 09. The basis set was B3LYP/6-31G. Considering the charge at pH 7, charge was set to +1, amino groups were set to be protonated and the carboxyl group were set to be deprotonated for initial structures.

Machine Learning Identification

Random forest classifier in scikit-learn library was used to machine learning analysis. Default hyper parameters of random forest classifier was adopted. The analysis was performed using Python 3.10.9 with scikit-learn library version 1.2.1.

SI2. Estimation of classification accuracy for accumulated signals

The classification performance index (F-measure) is 0.72. This accuracy is not the accuracy determined by using multiple signals during application but only that for a single pulse. The classification accuracy can be improved by statistical analysis. In the method reported in this manuscript, each signal is classified one by one; the molecule is classified with majority vote of all signal classification results.

Here, we consider the relation between the classification accuracy and the number of signals n. The prediction ratio for a single pulse of the true molecule p is set to 0.72. Then, the probability of accurate prediction by the majority vote P is described using the following equation:

$$P = \sum_{k>n/2} \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

where k denotes the number of true signals. The relation between P and n is shown in Figure S1. The accuracy determined by the majority vote is over 90% for 9 signals, 99% for 25 signals.

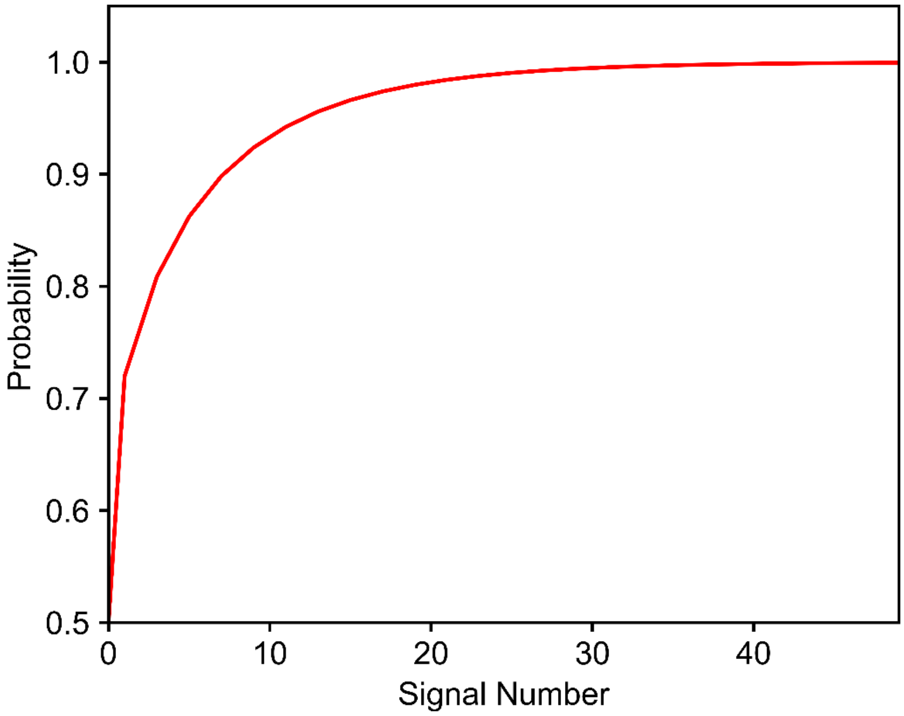

Figure S1. Relation between probability of accurate prediction by majority vote and number of signals for the classification with single-molecule discrimination accuracy of 0.72.

SI3. Limit of detection of single-molecule measurement.

We prepared lysine solutions at concentrations of 10^{-6}, 10^{-9}, 10^{-10}, and 10^{-11} mol/L to establish the limit of detection (LOD) for single-molecule measurements. The criteria for signal detection are detailed in the Methods section. We measured the frequency of these signals at each concentration, and the results are presented in the figure. Since the blank data has background noise of around 120/min, it can be determined that signals below 120/min are not detected. As shown in the figure, the LOD is approximately 10^{-10} M.

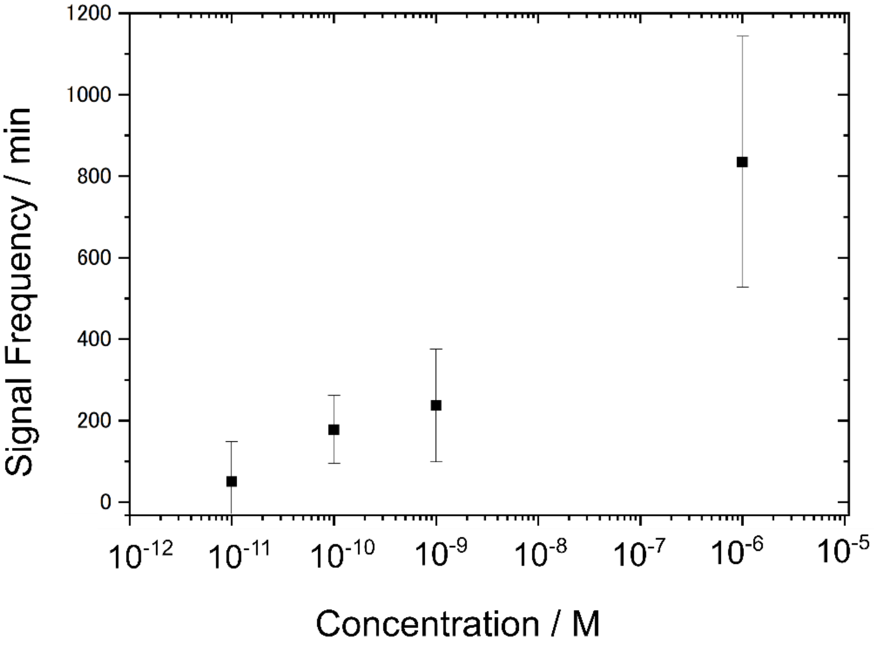

Figure S2 Concentration dependence of signal frequency.

SI4. Algorithm dependence

In main manuscript, we choose random forest classifier for AcLys-Lys classification. To exhibit the dependence of the classification algorithm, we also carried out with neural network and XGBoost classifier. Neural network is performed with Pytorch version 1.12.1. The model of neural network consists of 4 fully-conected layers with 13, 256, 128, 32 nodes and Rectified Linear Unit (ReLU) for activation function, and an output layer with softmax function. *F*-measure of the classification is 0.71 and 0.70 for neural networks and XGBoost, respectively. No significant differences by algorithm were observed for the data in this study.

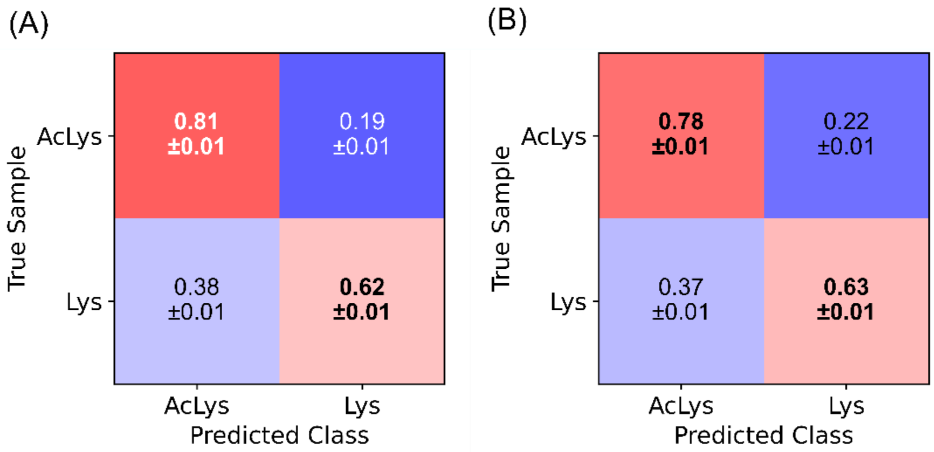

Figure S3 Confusion matrices of classification between Lys and AcLys by (A) neural network and (B) XGBoost.