Supplementary Information for

Chemical Recognition and Binding Kinetics in a Functionalized Tunnel Junction

Shuai Chang[1,2], Shuo Huang[2,a], Hao Liu[3,2], Peiming Zhang[2], Feng Liang[2], Rena Akahori[2,b], Shengqin Li[2,c], Brett Gyarfas[2], John Shumway[1], Brian Ashcroft[2] Jin He[2,d] and Stuart Lindsay[1,2,3]

[1]Department of Physics, [2]Biodesign Institute, [3]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA

.

[a] Present address: Chemistry Research Laboratory, 12 Mansfield Rd., Oxford OX1 3TA, UK, shuohuangasu@gmail.com
[b] Present address: [4]Biosystems Research Department, Central Research Laboratory, Hitachi, Ltd, .1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601 Japan, rena.akahori.uo@hitachi.com
[c] Present address: Department of Chemistry, College of Science, Huazhong Agricultural University, Wuhan 430070, P.R. China, sqingli@mail.hzau.edu.cn
[d] Present address: Department of Physics, Florida International University, 11200 SW 8th Street, CP204, Miami, FL 33199, jinhe@fiu.edu

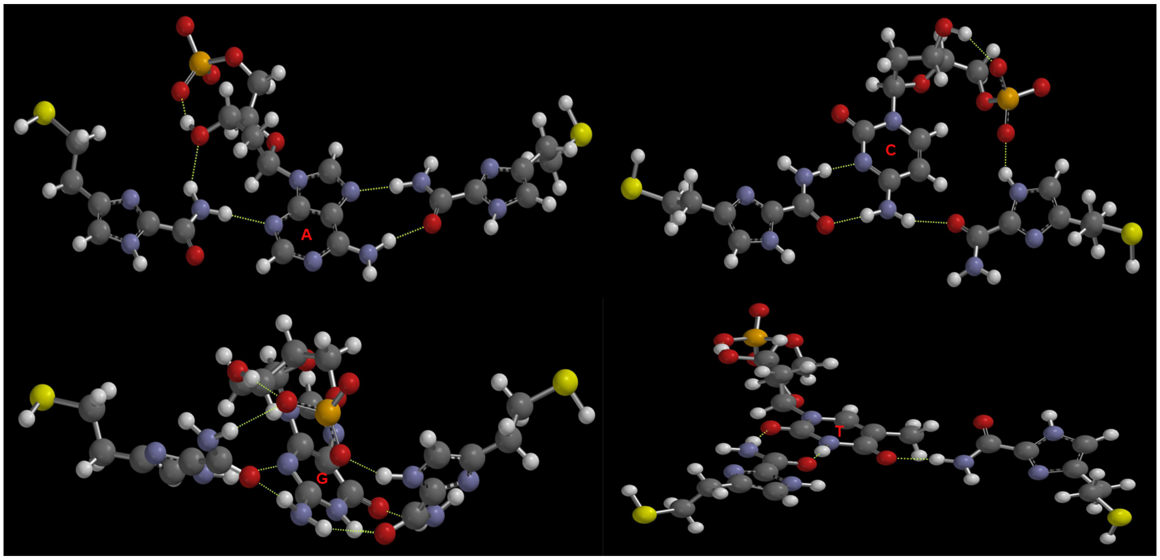


Figure S1: Binding motifs for all four bases. Structures are calculated as described in the caption for Figure 1.

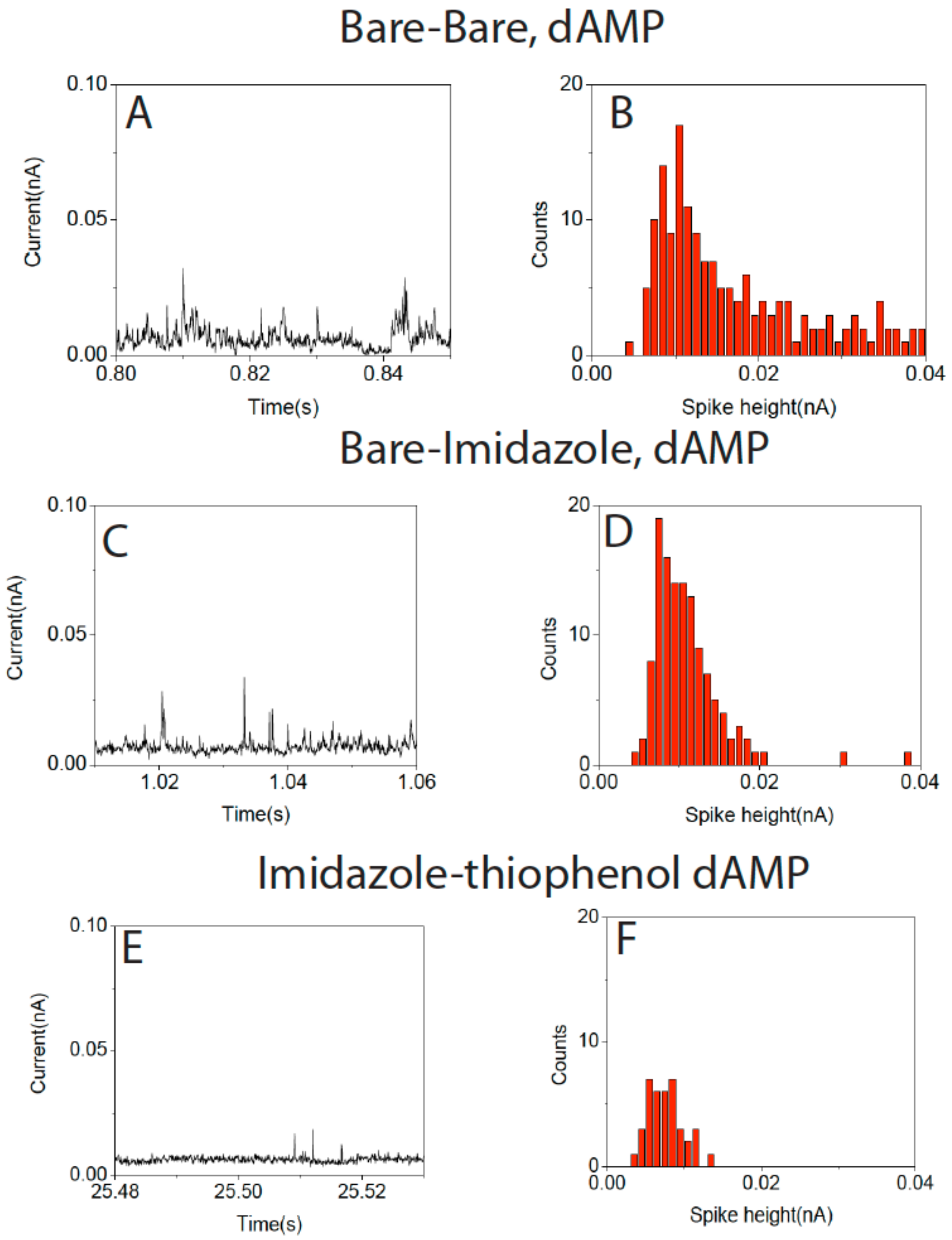


Figure S2: Typical current recordings (left) and spike height distributions for 10 μM dAMP in 1 mM phosphate buffer (pH 7.0) for (A,B) bare electrodes, (C,D) a bare probe and imidazole functionalized surface and (E,F) and imidazole functionalized surface and thiophenol functionalized probe.

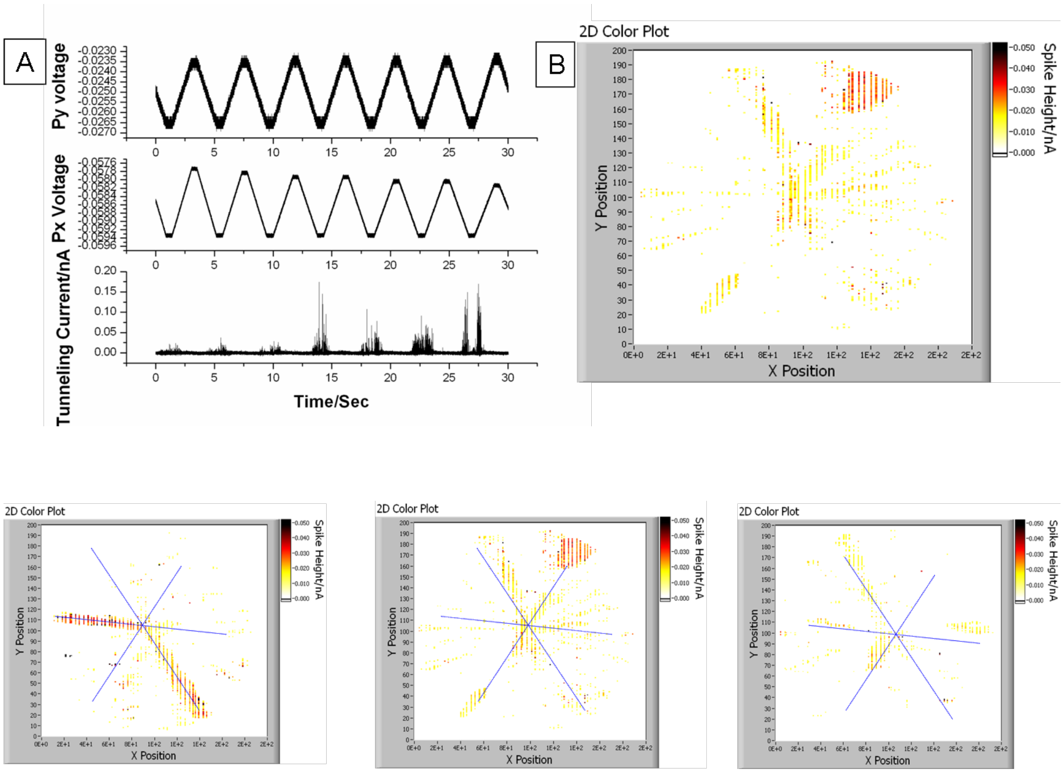


Figure S3: Clock scanning: A shows the voltages applied to the X and Y PZTs together with the recorded tunnel current showing bursts when the probe passes over a DNA oligomer. B shows the current distribution mapped onto the X,Y surface plane. Note that the signals tend to align along axes rotated by 60°. The panel below shows three separate scans taken at different places on the same Au(111) terrace, showing how the symmetry axes line up as a result of the constant orientation of the underlying gold lattice.

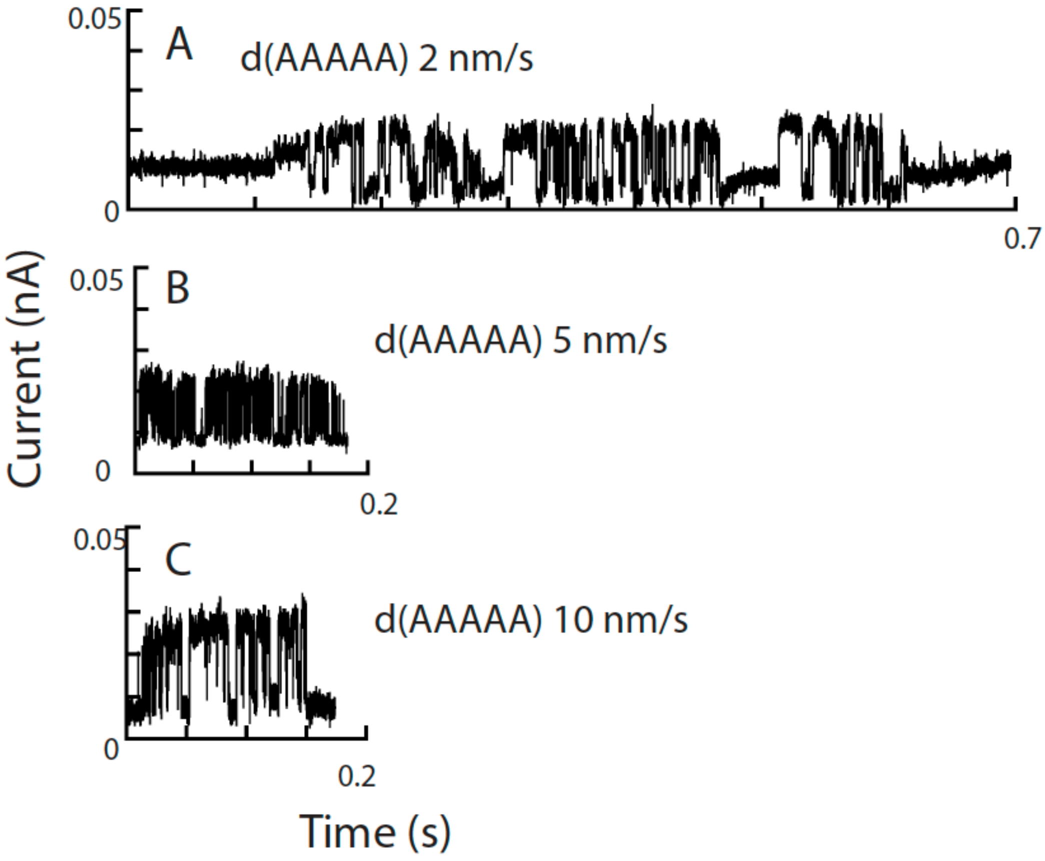


Figure S4: Periodic signal bursts from d(AAAAA) scanned at the speeds as marked (note that in C four bursts are shown, so the distance per burst is about 0.3 nm for all three traces shown here).

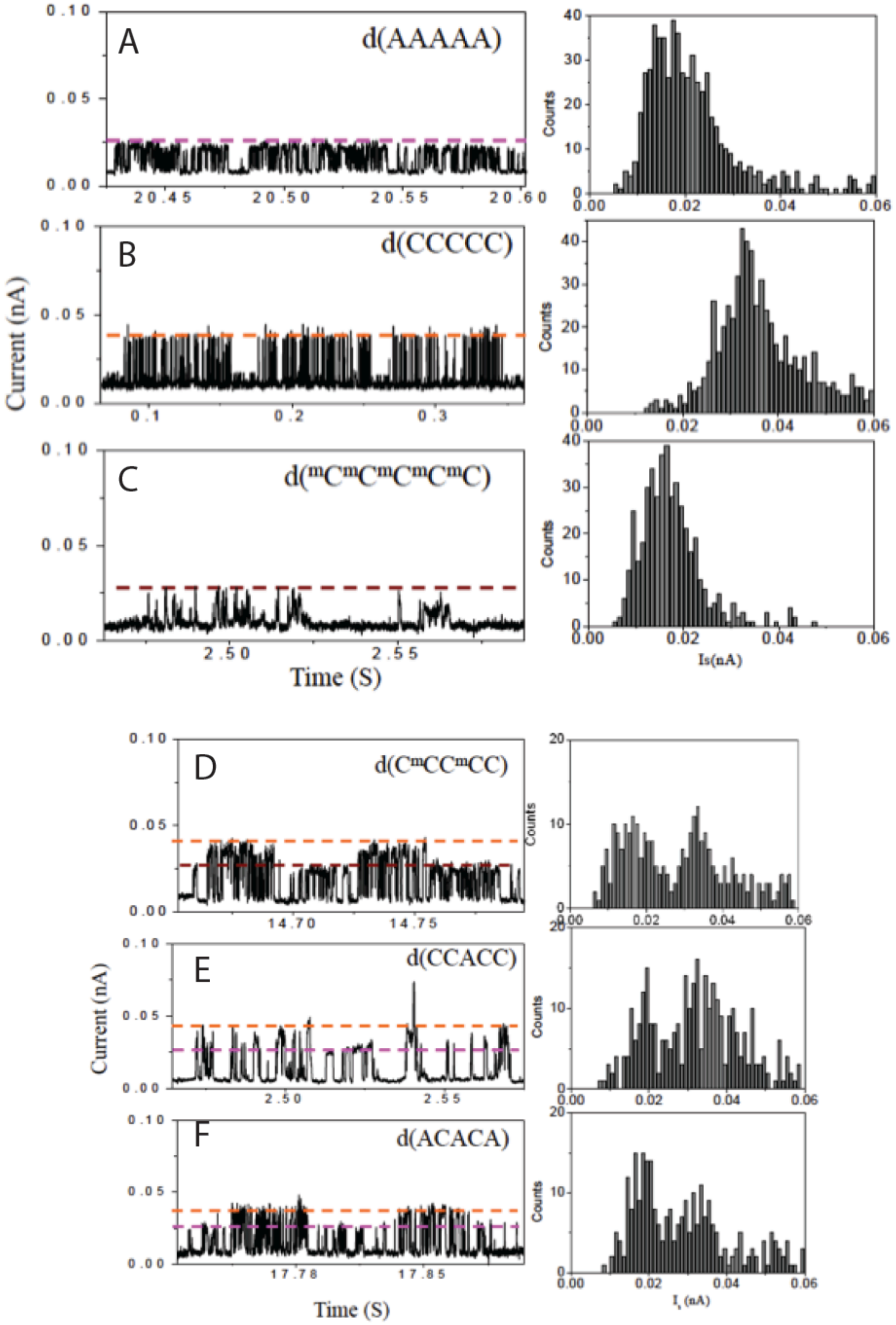


Figure S5: Clock-scans over homopolymers as marked (A,B,C) with spike height distributions to the right, and over heteropolymers as marked (D,E,F). Spike height distributions are clearly bimodal over heteropolymers.

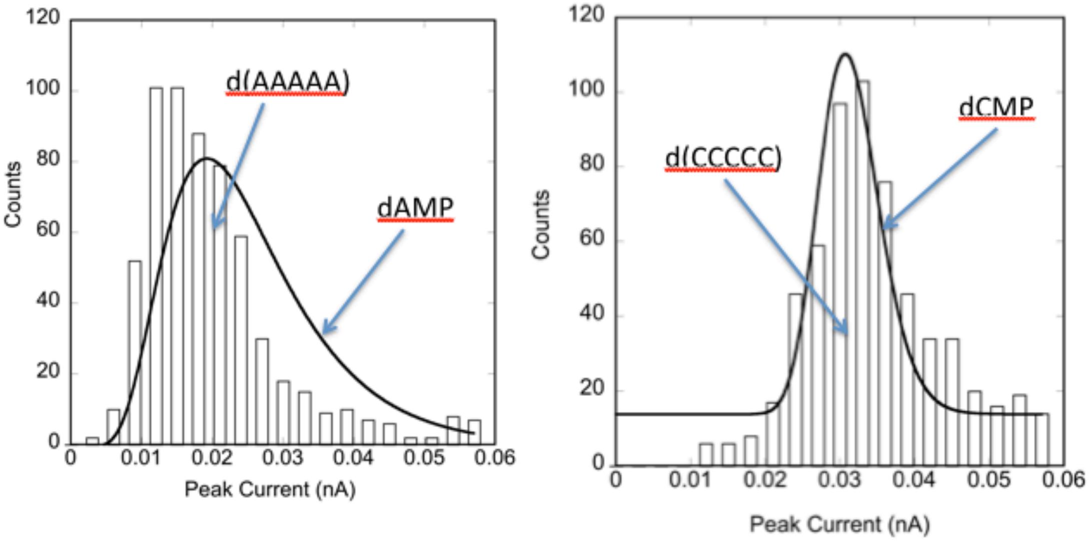
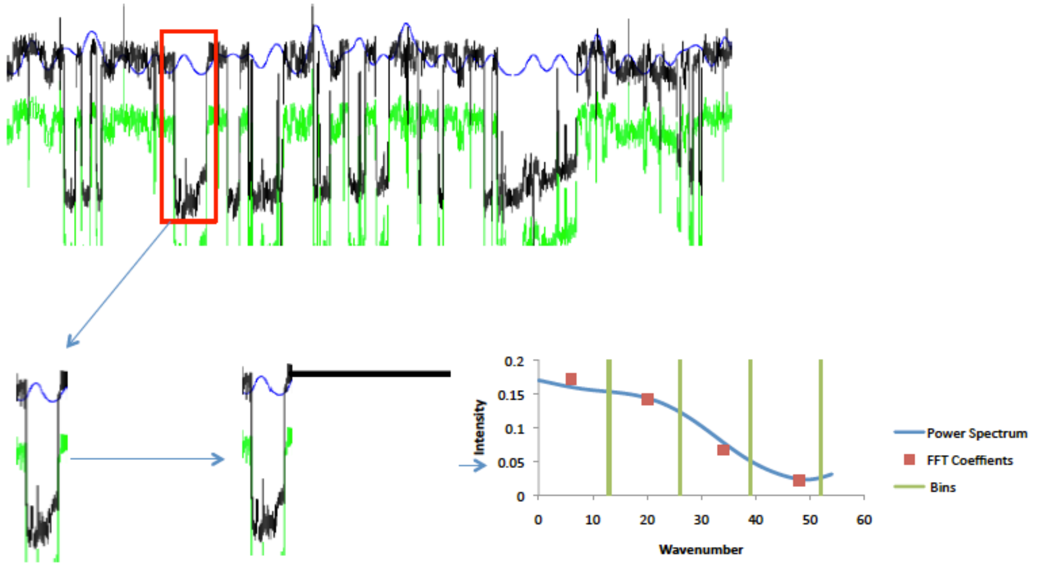


Figure S6: Spike height distributions from homoplymers (left d(AAAAA), right d(CCCCC)) shown as bars. The fitted distributions to the nucleotide data (Figure 4) are replotted here as solid lines.

FFT Analysis

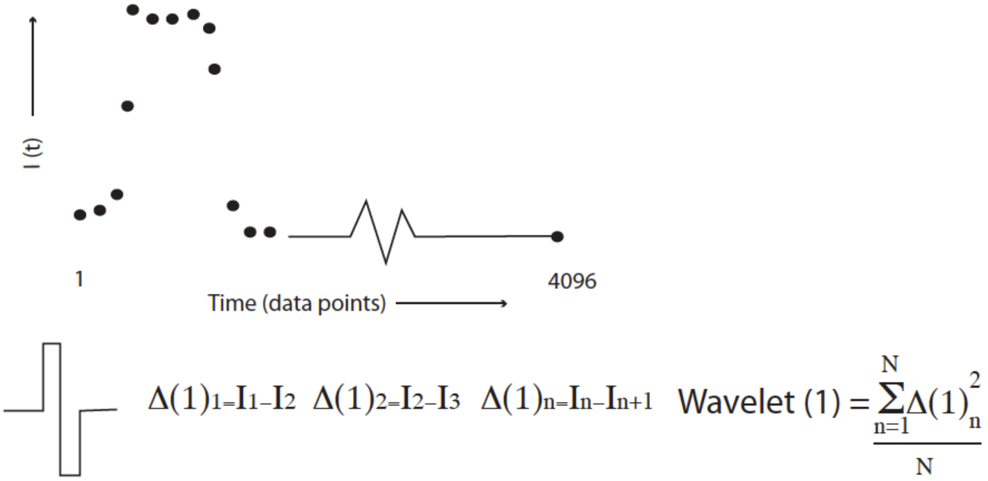


Clip Peak

Pad to standard size

Power Spectrum/ Binning

Figure S7: Fourier analysis of spikes. Spikes are first baseline subtracted and amplitude-normalized, then inserted into a 4096 point data array, taken to be periodic, for processing by an FFT. The power spectrum (going from 0Hz to the Nyquist limit of 25 kHz) is separated into four equal bins that are each averaged to produce four coefficients.



$$\Delta(1)_1 = I_1 - I_2 \quad \Delta(1)_2 = I_2 - I_3 \quad \Delta(1)_n = I_n - I_{n+1} \quad \text{Wavelet (1)} = \frac{\sum_{n=1}^{N} \Delta(1)_n^2}{N}$$

Figure S8: Illustrating how the Haar wavelet components are calculated. The first wavelet comes from convolution with 0,1,-1,0 to produce differences for all neighboring components. These differences are summed and averaged. The process is repeated for each successive (N^{th}) wavelet in which the filter is increased to 0, +2^N points, -2^N points, 0.

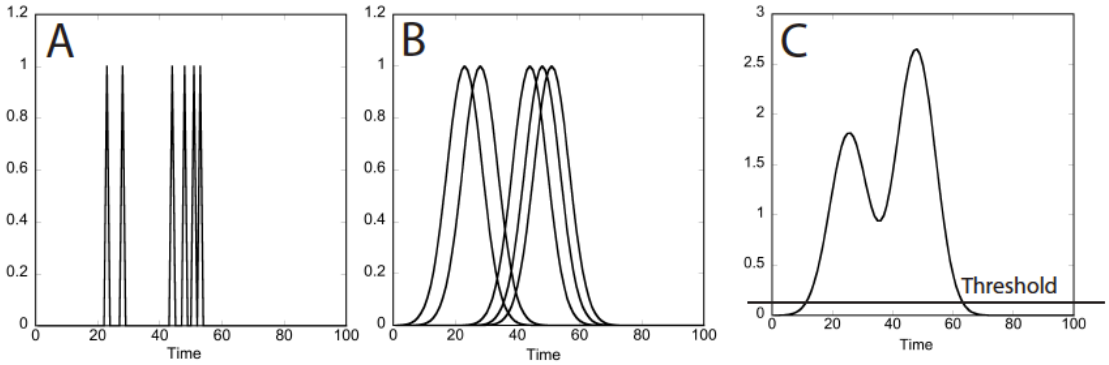


Figure S9: Cluster location algorithm. Each spike is replaced with a unit delta function centered at the middle of the spike (A) and a unit Gaussian of 4000 points FWHH placed on each spike (B). These Gaussians are summed and the cluster duration defined by the period over which this sum exceeds a threshold.

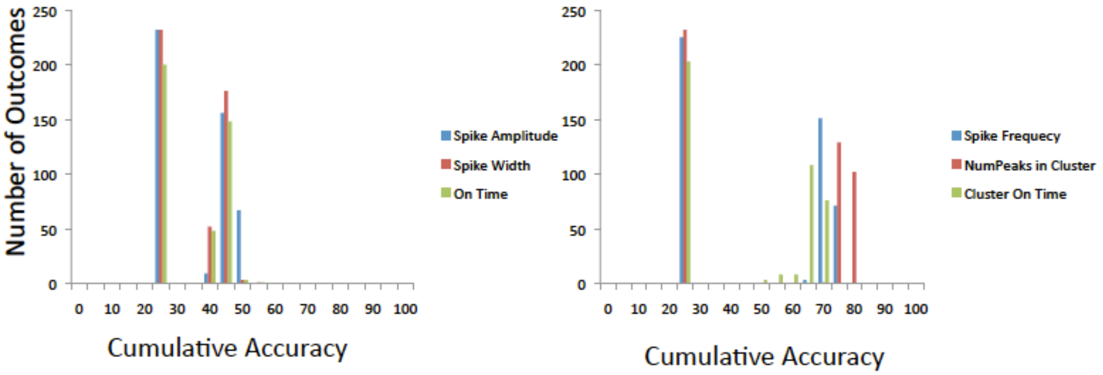


Figure S10: Distribution of base-calling accuracies with single-spike parameters (as marked) left, and (right) with cluster parameters. Cluster parameters allow calls of up to 80% accuracy for the best parameter combinations whereas single-spike parameters permit up to 50% accuracy.

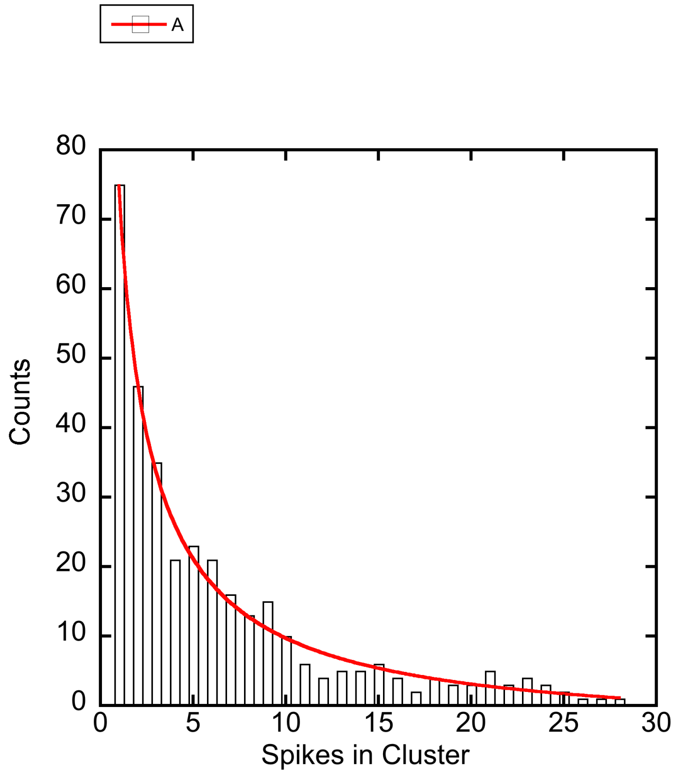


Figure S11: Distribution of the number of spikes in a cluster for dAMP scanned at 5 nm/s. The red line is a heavily damped log normal distribution centered on 13 spikes per cluster (this function fits better than an exponential).

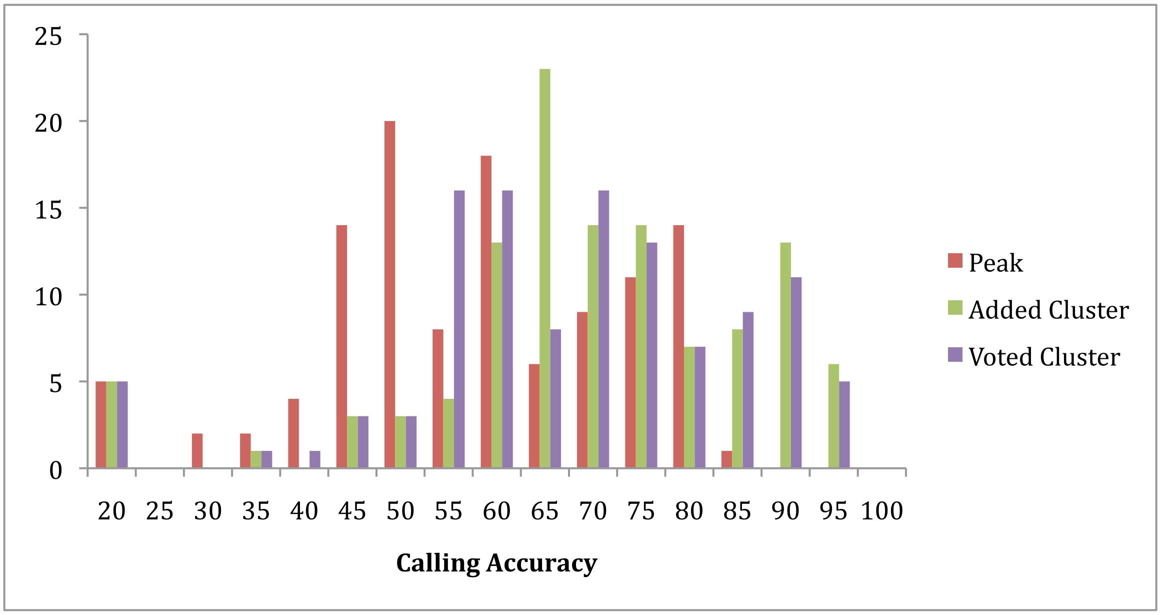


Figure S12: The red bars are the distribution of calling accuracies based on various combinations of peak characteristics (including cluster information). The best parameter combinations yield a calling accuracy of a little over 80%. By voting within a cluster, either by simple majority (purple bars) or adding probabilities returned by the SVM (green bars) a significant number of parameter combinations yield >95% accuracy.

Spikes in cluster	P(A)	P(C)	P(G)	P(T)	P(meC)
3.0000	0.67495	0.11754	0.13033	0.010480	0.066702
10.000	0.64892	4.8600e-06	5.2500e-05	3.6500e-06	0.35102
10.000	0.91724	0.0084734	0.0051912	0.052177	0.016914
10.000	0.30575	0.15436	0.075989	0.16042	0.30348
10.000	0.99999	5.0600e-08	4.4600e-07	1.1600e-05	7.0100e-07
10.000	0.99971	1.6800e-07	0.00012862	0.00013496	2.1500e-05
10.000	0.99852	9.5000e-06	0.0014556	3.2200e-06	1.3000e-05
10.000	0.94007	6.4500e-06	7.6000e-05	0.00012719	0.059722
10.000	0.65244	2.2400e-06	0.00049840	1.7000e-06	0.34706
1.0000	0.61618	0.054975	0.11677	0.027437	0.18463
2.0000	0.97333	0.00014965	0.0099610	0.00020670	0.016358
2.0000	0.28232	0.28171	0.068545	0.090825	0.27659
1.0000	0.48941	0.17596	0.23492	0.025562	0.074144
1.0000	0.87122	0.046533	0.063004	0.0062350	0.013007

Table S1: Some examples of base calling accuracy from a selection of different clusters of various size for a sample of dAMP. The first column lists the number of spikes in a cluster while subsequent columns list the accuracy of the call as returned by the SVM as dAMP, dCMP, dGMP, dTMP and dmeCMP. Some of the clusters with 10 spikes in them can be called to better than 99% accuracy.

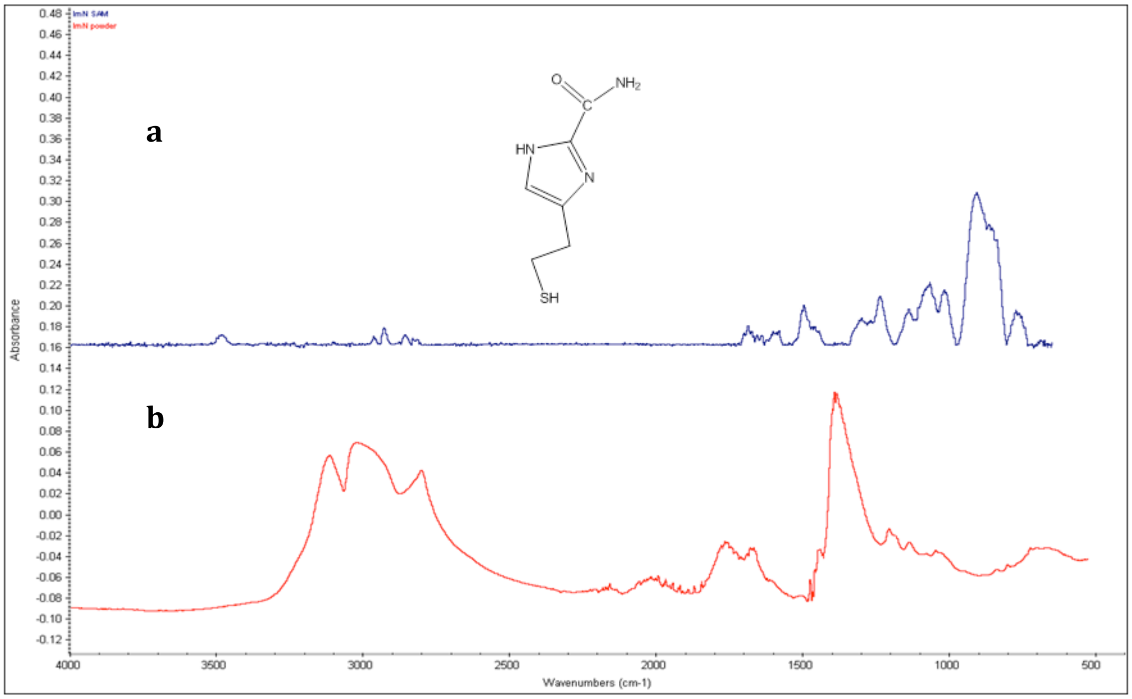


Figure S13. FTIR of 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide (a) in a monolayer; (b) in a powder. A gold substrate was cleaned by hydrogen flaming, and then immersed in a 0.2 mM ethanolic solution of 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide for 24 h. The substrate was copiously washed with ethanol and dried by gently blowing a stream of dry nitrogen on the surface. Thickness of the monolayer was measured as 8.46 ± 0.23 Å by ellipsometry (the distance between O of the amide and S of the thiol is 8.44 Å and the bond length of Au – S is about 2.45 Å [1]). The XPS data show that the monolayer contains C, N, O, and S atoms. FTIR spectrum of 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide shows a similarity with that in a powder.

1. Majumder, C.; Briere, T. M.; Mizuseki, H.; Kawazoe, Y., Structural investigation of thiophene thiol adsorption on Au nanoclusters: Influence of back bonds. *J. Chem. Phys.* **2002,** *117* (6), 2819-2822.