Application of Recognition Tunneling in Single Molecule Identification

by

Yanan Zhao

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2014 by the
Graduate Supervisory Committee:

Stuart Lindsay, Chair
Quan Qing
Robert Ros
Peiming Zhang
Robert Nemanich

ARIZONA STATE UNIVERSITY

August 2014

ABSTRACT

Single molecule identification is one essential application area of nanotechnology. The application areas including DNA sequencing, peptide sequencing, early disease detection and other industrial applications such as quantitative and quantitative analysis of impurities, etc. The recognition tunneling technique we have developed shows that after functionalization of the probe and substrate of a conventional Scanning Tunneling Microscope with recognition molecules ("tethered molecule-pair" configuration), analyte molecules trapped in the gap that is formed by probe and substrate will bond with the reagent molecules. The stochastic bond formation/breakage fluctuations give insight into the nature of the intermolecular bonding at a single molecule-pair level. The distinct time domain and frequency domain features of tunneling signals were extracted from raw signals of analytes such as amino acids and their enantiomers. The Support Vector Machine (a machine-learning method) was used to do classification and predication based on the signal features generated by analytes, giving over 90% accuracy of separation of up to seven analytes. This opens up a new interface between chemistry and electronics with immediate implications for rapid Peptide/DNA sequencing and molecule identification at single molecule level.

ACKNOWLEDGMENTS

I still remember the day five years ago, the day that I was accepted to Dr. Lindsay's group. Through the past years, there are good experiment results, bad experiment results, excitement and disappoint. Fortunately, I was guided through darkness by a lot of people and eventually sight the silver linings.

First of all, I want to thank my advisor Dr. Stuart Lindsay. His passion enlightens me to continue exploring the unknown world of science. His intuition about science left me a lifetime-lasting impression. One day, he came to my cabinet and said, "Hey, Yanan. Let's try if we can get Amino Acid signals using RT method." And, that sentence led to the most important discovery I have made so far, which I believe, have already open a door to do rapid peptide sequencing in the near future. He taught me not only how to conduct research but also how to bridge the experiment results with theory. I feel so lucky to have such a fantastic advisor and to join such a great group. I believe everything I have learned from here will continue guiding me through my future career.

Besides, I would like to thank Dr. Peiming Zhang. His expertise in organic synthesis, surface chemistry helped me to improve my experimental procedure and he is very kind to answer my questions with respect to chemistry. He also brought new idea to my research as Ibuprofen identification and disaccharide identification, etc.

Thirdly, I appreciate all the help I got from my colleagues. Prof. Jin He guided me how to synthesis Graphene when I first joined this lab. I still miss the excitement when I first got the graphene after numerous failure. Dr. Brett Gyarfas is a terrific colleague to work with. His expertise in Computer Science and Electronic Engineering is a treasury for our lab. Dr. Brian Ashcroft is the brain in charging of data analysis. When I was struggling to separate amino acids using old methods, he brought us the FFT analysis and Support Vector Machine method, and complete the RT method with robust analysis tools. Dr. Shuo Huang and Dr. Hao Liu taught me how to run STM measurements and helped me a lot both in and out of the lab. Besides, Dr. Pei Pang, Dr. Shuai Chang, Dr. Di Cao, Dr. Padmini Krishnakumar and Weisi Song, Suman Sen, JoneOne Im all helped a lot with my experiments and my PhD life.

I am so grateful that I have so many talented, helpful and dedicated people around me in these five year of memorable time.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

Methods of Single Molecule Conductance Measurement

When people use a nano-gap to conduct measurements of single molecule electrical properties, it is crucial to have an exactly constant gap distance every time because tunneling currents have exponentially decay relationship with gap distances. Even a small change in the gap size could change the current dramatically. Besides, people also need devices that could have variable gap size to match the length of analyte and to study the influence of gap distance on molecular conductance. After Scanning Tunneling Microscope (STM) was developed by Gerd Binning and Heinrich Rohrer at IBM in 1981, it has been widely used a tool to image surface at atomic level. Realizing STM satisfies requirement of controllable gap distance, people started using STM to examine the electrical properties of single molecule recently. In the following sections, I will give a brief introduction of methods of single molecule conductance measurement developed so far.

1.1 Nano Electronics

The semiconductor industry has made great progress since it formed around 1960s. Intel's co-founder Gordon Moore published his prediction that over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years (Figure 1.1), which later is recognized as Moore's law [69]. While the size of the chip features are becoming smaller and smaller, the size of semiconductor manufacturing process nodes has been decreased tremendously (Figure

1.2) [70]. While semiconductor industry is still marching on the roadmap, one important trap stone is at the scale of 10nm or less, quantum tunneling, especially through gaps, becomes a significant phenomenon. Sadly, electronics behaviors in the quantum realm are quite different from the conventional behaviors in macroscopic realm, and this open a new era of technology: nano-electronics.

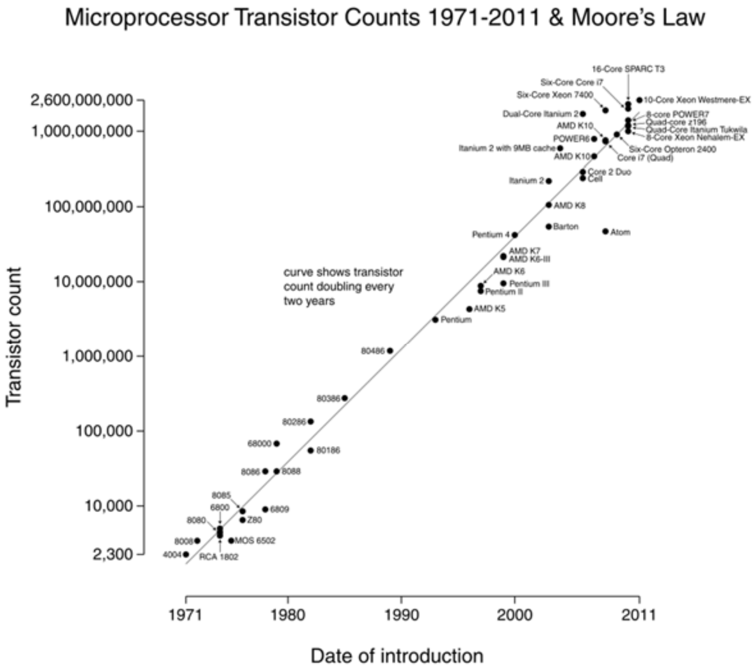


Figure 1.1 Plot of CPU transistor counts against dates of introduction.

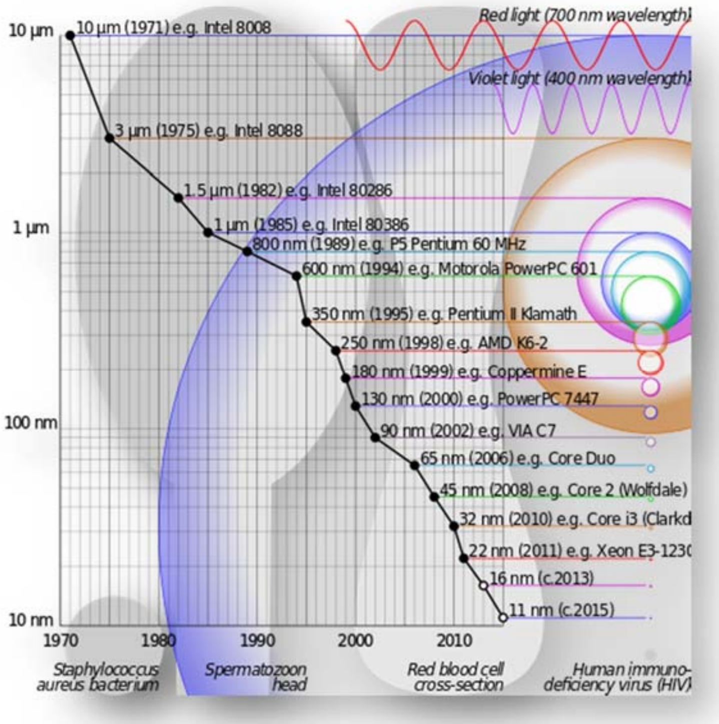


Figure 1.2 Process of miniaturisation, and comparison of sizes of semiconductor manufacturing process nodes with some microscopic objects and visible light wavelengths. http://en.wikipedia.org/wiki/Semiconductor_device_fabrication.

1.2 Quantum Tunneling

For classical electronics, the current I flow through a bulk material is given by *Ohm's law*,

$$I = V/R$$

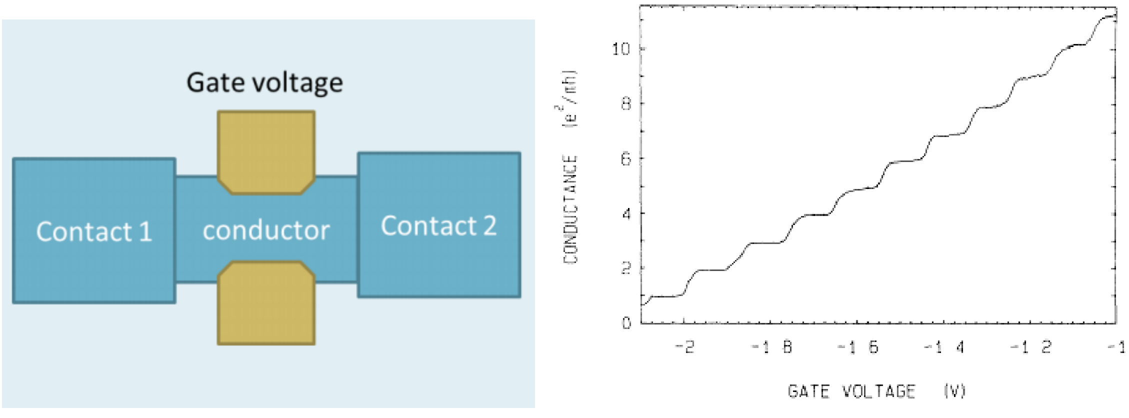
Where V is the voltage applied across the material and R is calculated by the following formula:

$$R = \rho \frac{l}{A}$$

l is the length of the material and A is the cross-section area. ρ is the electrical resistivity. Another derived physical property of material is called conductance, which defined as the G = 1/R. So,

$$G = \sigma \frac{A}{L}$$

and $\sigma = 1/\rho$, named electrical conductivity. According to this formula, *if L* goes to *0, then G* goes to ∞. However, when the size of the material is close to the mean free path and the phase coherence length of electrons, electron transportation through junction changes from the traditional diffusive to ballistic, and the conductance is becoming quantized (Figure 1.3). The first experimental results were obtained by B.J. van Wees in 1988 [2]. In this experiment, they used gate voltage to control the width of the gap.



(a) (b)

Figure 1.3 (a). Experiment setup. 2-dimensional electron gas at an AlGaAs-GaAs interface. (b). Point-contact conductance as a function of gate voltage

To illustrate what's going to happen in quantum physics realm, we start from Schrödinger equation and end up with the general formula to calculate tunneling current through a 1-D channel.

1.2.1 1-D Potential Well

Consider a potential well with length L, the Schrödinger equation for this 1-D circumstance is

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(x)\right]\Psi(x) = E\Psi(x) \qquad \text{(eq. 1.1)}$$

With boundary conditions:

$$\Psi\left(-\frac{L}{2}\right) = \Psi\left(-\frac{L}{2}\right) = 0$$

Inside the well, *V(x)* = 0. So,

$$\Longrightarrow\ E_n = \frac{\hbar^2\pi^2n^2}{2mL^2}$$

$$k_n = \frac{n\pi}{L}, with\, n\ \in\ \mathbb{N}$$

1.2.2 1-D Rectangular Potential Barrier

To step further, consider a similar condition but with a potential barrier (Figure 1-4):

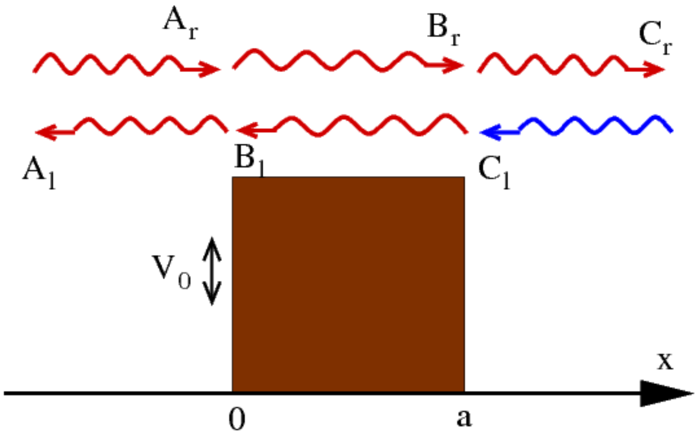


Figure 1.4 Scattering at a finite potential barrier

The corresponding Schrödinger equation is,

$$H\Psi(x) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(x)\right]\Psi(x) = E\Psi(x) \qquad \text{（eq. 1.2）}$$

$$V(x) = V_0[\Theta(x) - \Theta(x-a)]$$

V(x) is the barrier potential with height V_0 and $\Theta(x)$ is Heaviside step function. $\Theta(x) = 0, x < 0$; $\Theta(x) = 1, x > 0$. Let's divide the space into three regions. x < 0 as left (with subscription *L*), 0 < x < a as center (with subscription *C*) and x > a as right (with subscription *R*).

If $E \neq V_0$,

after solving Schrödinger equation separately, the wave function for each part could be written as:

$$\Psi_L(x) = A_r e^{ik_0x} + A_l e^{-ik_0x}, x < 0$$

$$\Psi_C(x) = B_r e^{ik_1x} + B_l e^{-ik_1x}, 0 < x < a$$

$$\Psi_R(x) = C_r e^{ik_0 x} + \; C_l e^{-ik_0 x}, x > a$$

Whereas the wave numbers are:

$$k_0 = \sqrt{2mE/\hbar^2}, x < 0 \; or \; x > a$$

$$k_1 = \sqrt{2m(E - V_0)/\hbar^2}, 0 < x < a$$

The boundary conditions of the wave function at $x = 0$ and $x = a$ require that the wave function and its first order derivative have to be continuous everywhere, so:

$$\Psi_L(0) = \; \Psi_C(0)$$

$$\Psi_C(a) = \; \Psi_R(a)$$

$$\frac{d}{dx}\Psi_L(0) = \; \frac{d}{dx}\Psi_C(0)$$

$$\frac{d}{dx}\Psi_C(a) = \; \frac{d}{dx}\Psi_R(a)$$

Thus, the coefficients have the restrictions:

$$A_r + A_l = \; B_r + \; B_l$$

$$ik_0(A_r - A_l) = \; ik_1(B_r - \; B_l)$$

$$B_r e^{iak_1} + B_l e^{-iak_1} = C_r e^{iak_0} + C_l e^{-iak_0}$$

$$ik_1\left(B_r e^{iak_1} - B_l e^{-iak_1}\right) = ik_0(C_r e^{iak_0} - C_l e^{-iak_0})$$

When *E* = *V₀*, the wave function of the center region becomes:

$$\Psi_C = B_1 + B_2 x, 0 < x < a$$

and the restrictions are:

$$A_r + A_l = B_1$$

$$ik_0(A_r - A_l) = B_2$$

$$B_1 + B_2 a = C_r e^{iak_0} + C_l e^{-iak_0}$$

$$B_2 = ik_0(C_r e^{iak_0} - C_l e^{-iak_0})$$

If we let $A_r = 1$ (incoming particle), $A_l = r$ (reflection), $C_l = 0$ (no incoming particle from the right), and $C_r = t$(transmission). After eliminate the coefficients B_l, B_r and solve for r, t. We have:

$$t = \frac{4k_0 k_1 e^{-a(k_0 - k_1)}}{(k_0 + k_1)^2 - e^{2iak_1}(k_0 - k_1)^2}$$

$$r = \frac{(k_0^2 - k_1^2)\sin(ak_1)}{2ik_0 k_1 \cos(ak_1) + (k_0^2 + k_1^2)\sin(ak_1)}$$

The transmission probability through the barrier is T= $|t|^2$. The most surprising result is even $E < V_0$, there is a non-zero probability that the particle is able to transpass the barrier, which in classical physics, is forbidden. This is a general picture how a STM works. For STM, the energy barrier is due to the vacuum or other medium between the probe and the substrate. Since the tunnel current depends exponentially decay on the barrier width, it is extremely sensitive to the surface topography of the sample.

1.2.3 The Landauer Formula

To study electron transportation at low dimension, for example, a 1-D system shown in Figure 1-5:

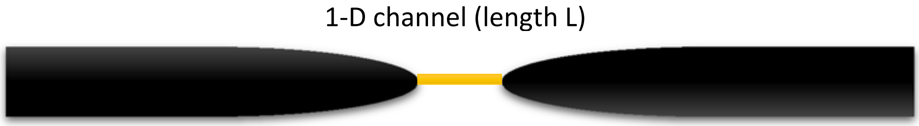


Figure 1.5 Sketch of one-dimensional system

Landauer formula gives the conductivity of this channel:

$$G(E) = \frac{2e^2}{h}\sum_{i,j} T_{ij}(E)$$

$T_{ij}(E)$ is corresponding to the probability that electron transmit from the ith mode at the left electrode to the jth mode at the right electrode. And the current when applying V on the sample is:

$$I = \int_0^{\infty} \frac{dE}{e}[f(E+eV) - f(E)]G(E)$$

Where, *f* is the Fermi function $f(E) = \frac{1}{e^{(E-E_F)/kT}+1}$, E_F is the Fermi level of the electrode.

If this 1-D channel is bridged by a molecule, it becomes the classical metal-molecule-metal sketch of device that people use to measure the conductivity of the molecules.

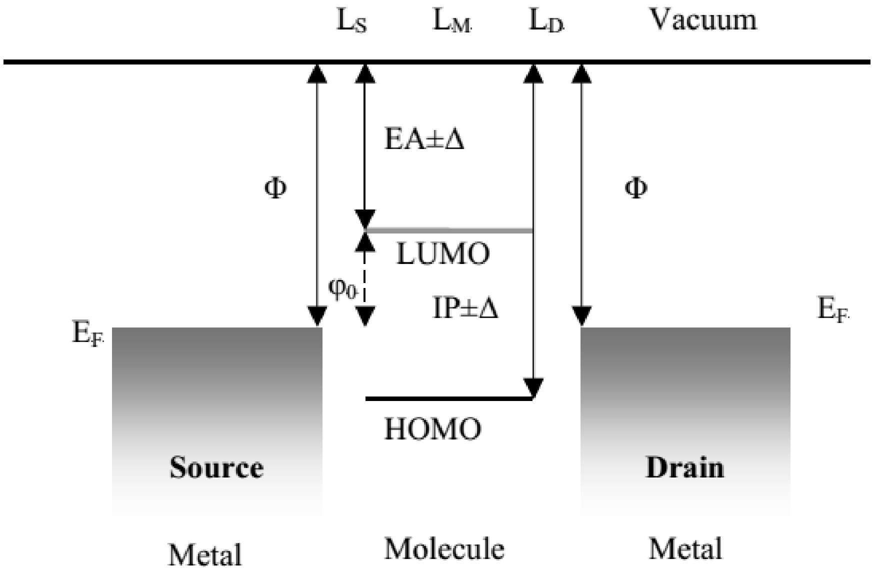


Figure 1.6 Schematic plot of energy (vertical distance) vs. distance (horizontal axis) for a metal-molecule junction. Fermi level EF is between lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO) [3].

As the calculations shown at the 1-D barrier, in general, the transmission probably through the metal-molecule-metal could be written as $e^{-\beta L}$. For tunneling through LUMO, $\beta = 2\sqrt{\frac{2m^*(E_{LUMO}-E)}{\hbar^2}}$, while for tunneling through HOMO, $\beta = 2\sqrt{\frac{2m^*(E-E_{HOMO})}{\hbar^2}}$, where E is the incident electron energy. It equals to the EF if there is no bias applied on the metal electrode. While after applied bias, it shifts the Fermi level and have an effect on the tunneling process.

At this point, it is clearly that by bridging the gap by molecule, we are able to increase the conductivity of the junction. Since different molecules might have different HOMO/LUMO level, length, geometry and structure, it is not only provides a method measuring molecule conductance but also a way to differentiate molecules theoretically.

1.3 STM and Other Nano-gap Devices

As mentioned before, the most challenging working to do the tunneling current measurement is to get a reliable and reproducible nano-gap. So far, three techniques have been employed to form a nano-gaps (Figure 1.7) [4].

1. Fixed electrodes
2. Mechanically formed molecular junction
3. Scanning probe techniques

1.3.1 Fixed Electrodes

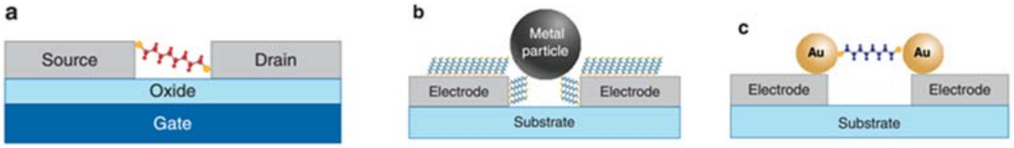


Figure 1.7 Sketch of fixed electrode setups. (a) A single molecule bridged between two electrodes with a molecular-scale separation prepared by electromigration, electrochemical etching or deposition, and other approaches.

(b) Formation of molecular junctions by bridging a relatively large gap between two electrodes using a metal particle.

(c) A dimer structure, consisting of two Au particles bridged with a molecule.

For Figure 1.7(a), molecules with anchoring groups on both ends are directly connected with the electrodes of the nano-gap. While forming a molecule size gap is the most changing work because it is beyond the capability of traditional fabrication facilities. New methods have been developed to achieve nano-gap such as electromigration [5] [6], electrochemical etching or deposition [7-10], and some other methods [11-14].

To ease the strict requirements of gap size, people came up with other approaches. Instead of creating molecular size gap, they fabricate a much large gap. Then the sample molecules are modified on the electrodes. A metal particle with the diameter larger than the gap is driven into the gap by dielectrophoresis [15] or a magnetic field [16] thereafter. Finally a double serial connected metal-molecule-metal junction is formed (Figure 1.7(b)). Another similar approach is by modifying the molecule with huge Au particles. And the dimer structure is then trapped in the gap (Figure 1.7(c)) [72].

1.3.2 Mechanically Formed Molecular Junction

The difficult part of fixed electrodes method is to form pre-required gap size precisely. One way to solve this problem is to make the gap size controllable during the fabrication process. Using a piezoelectric transducer (PZT) could help control the size with sub-angstrom precision. Generally, there are three kinds of techniques developed based on this concept - mechanically controllable break junction, conducting atomic-force-microscopy break junction, and scanning-tunneling-microscopy-based break junction.

1.3.2.1 Mechanically Controllable Break Junction (MCBJ)

The concept of mechanically formed molecular junction is to break a nano-wire into two facing electrodes. This breaking process is controlled by a PZT and stepping motor (Figure 1.8(a)).

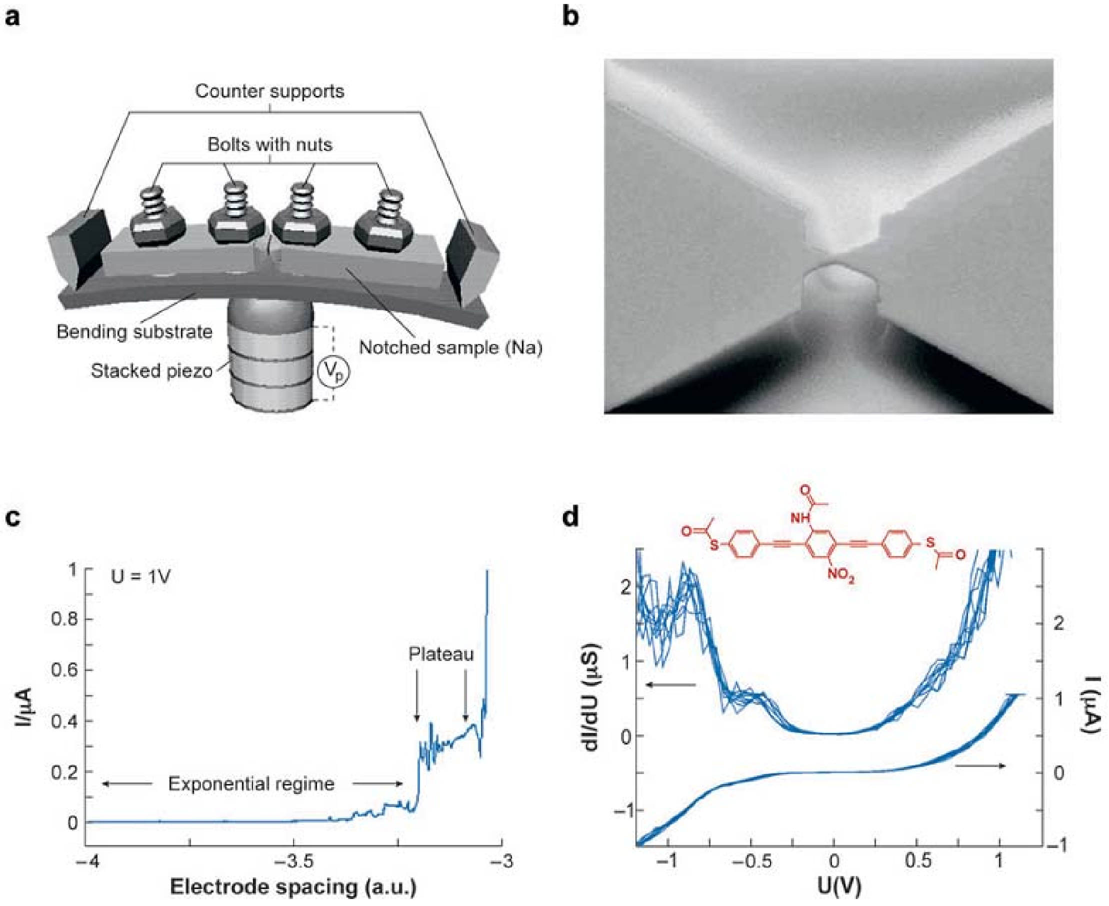


Figure 1.8 Mechanically formed molecular junction (a) Schematics of a MCBJ [17]. (b) SEM picture of lithographically fabricated break junction [18]. (c) Change of current vs the electrode spacing [19]. (d) I-V curve and dI/dV of a molecular junction [20].

A nano-wire with fixed by notch (Figure 1.8(a)) or a nano-bridge on a substrate using electron-beam lithography (Figure 1.8(b)). Then stepping motor and PZT are pushing from the back to break the wire or the nano-bridge, after retrieving PZT slowly, a controllable nano-gap are formed. Reichert et al. [19] used MCBJ to investigate molecule conductance. A droplet of solvent was dropped into the gap after the gap had formed, then when retrieving PZT slowly, it first showed an exponential regime, following with a plateau barely change with gap size, which is supposed to be corresponding to the molecules are connecting the two electrodes (Figure 1.8(c)). A much clearly illustration is shown in Figure 1.8(d), after derive the first order of I/V, a constant conductance was obtained, suggesting that only one or very small number of molecules were bridging the gap.

1.3.2.2 Conducting Atomic-Force-Microscopy (AFM) Break Junction

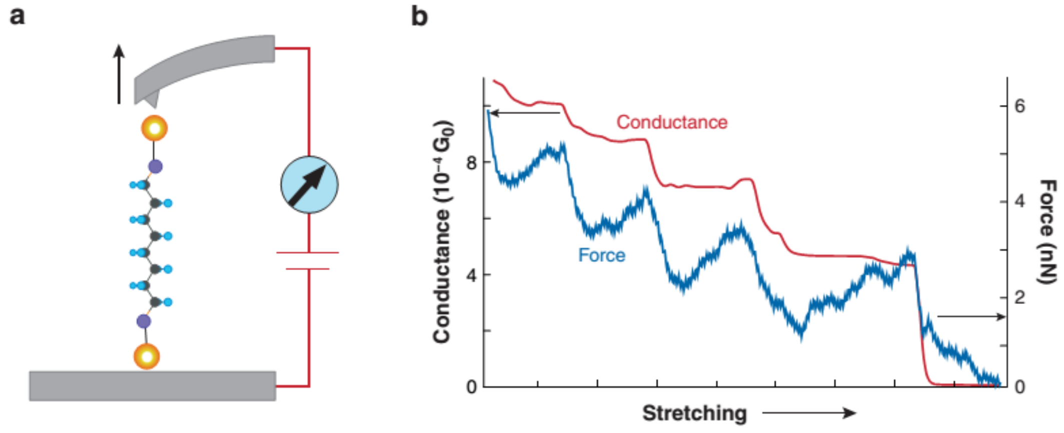


Figure 1.9 C-AFM break junction (a) Conducting AFM measuring force and conductance simultaneously. (b) Force and Conductance measurement during the stretching

Xu et al. [22] reported a method to measuring single molecule electromechanical properties using a conducting AFM (Figure 1.9). Analyte was dissolved in toluene and bonding with the tip and substrate covalently. When AFM tip was pulling away, they observed the force behaved saw-like structure while the conductance has plateau in compliance with the force (Figure 1.9(b)). They claimed at the end of each plateau, it was corresponding to one break of molecule or small number of molecules from the bridge, especially for the last breaking step. Thus by studying the last step they can measure the conductance of a single molecule. Results are shown in Figure 1.10

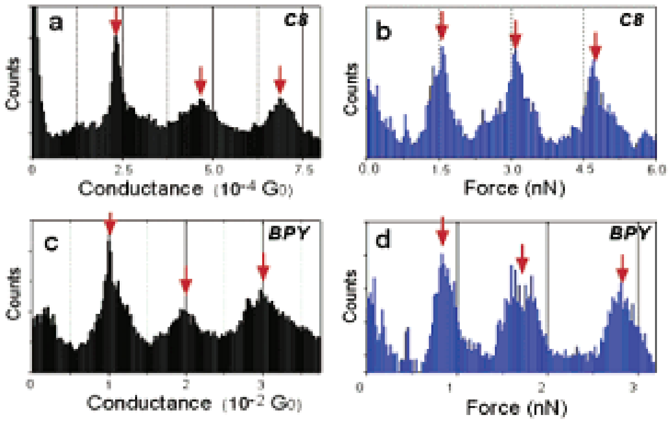


Figure 1.10 Histogram of conductance and force for 1,8'octanedithiol (C8) and 4,4'bipyridine (BPY)

1.3.2.3 Scanning-Tunneling-Microscopy-based Break Junction

This method was first invented by Xu & Tao [21]. It operates alike CAFM, but the measuring time is reduced tremendous. It is capable of collecting hundreds of current-distance curves in a minute.

The idea of conducting measurement is by repeatedly moving a scanning-tunneling-microscopy tip into and out of contact with the substrate electrode in the presence of sample molecules (Figure 1.11) [4].

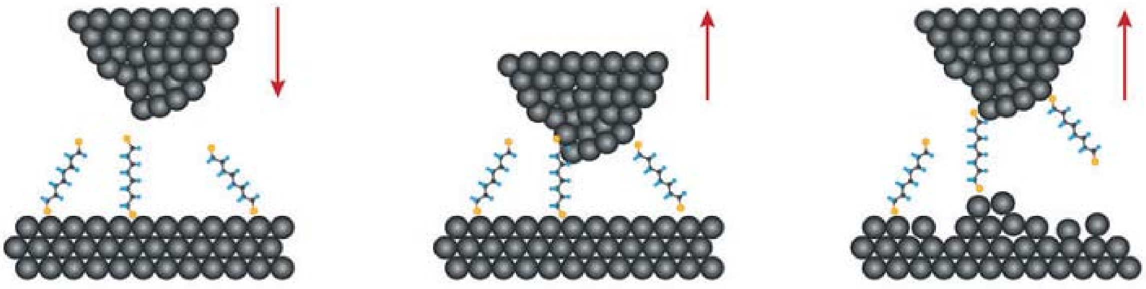


Figure 1.11 Measurement of single-molecule conductance using a scanning tunneling microscopy (STM) break junction.

I will give details of method in the next chapter by studying how the functional group influence the conductance of carotenoid molecule wire.

Compared with the previous fixed electrodes method, the advantages of MCBJ methods are that they have controllable gap size and could be varied during the measuring process to comply with the length the molecule.

1.3.3 Scanning Probe Techniques

Scanning tunneling microscopy (STM) are probably one of the most useful tools while studying molecular electronics. Not only it is used to image the molecules absorbed on a conductive substrate with submolecular resolution and manipulate molecules on surface [23], but also it can perform tunneling spectroscopy measurement [24-27]. In this method, analyte molecules absorb on the substrate and form a metal-molecule contact, while the tip moves above the surface [4].

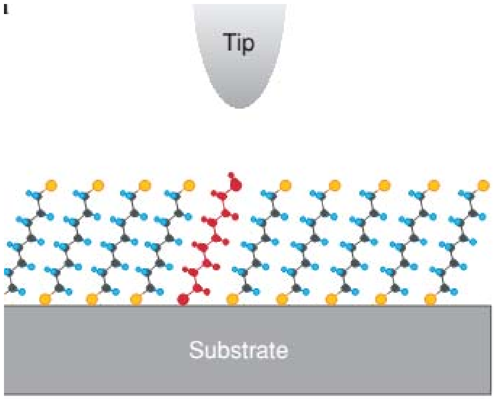


Figure 1.12 STM study of electron transport through a target molecule (red) inserted into an ordered array of reference molecules (cyan).

To reduce the consequence of multiple molecules connected with the tip at the same time, less conducting molecules (usually an insulating mono-layer as alkanethiol [27]) was previous modified before the analyte. So, single molecule conductance could be measured by using this method. An improved method that we called “Recognition tunneling” method will be illustrated in the following chapters combined with my work in later chapters.

Chapter 2

Conductance Measurement by STM-based Break Junctions

This chapter describes the molecule conductance measurement using STM-based break junction method. This work was published on Chemistry - A European Journal [28]. The contribution of the current author was conducting the STM measurements and data analysis. N.J Tao's lab at Biodesign Institute at ASU provided STM setup and programs to collect data and tools to perform data analysis.

2.1 Introduction

In molecular electronics, much of the attention has been forced on the device fabrication as diodes, transistors, etc. While to build electric circuits, it is equally important to fabricate circuits wiring devices. One of most promising organic molecular wires synthesized is carotenoid. Because of consisting of extensive π conjugation, it is supposed to have very high electric conductance [29]. He et al. [91] reported that the conductance of the carotenoid was exponentially decay with the length of the molecule. While there has no systematic study on the influence of side chain influence on the conductance yet. In this paper, we reported the conductance measurement of carotenoids with different polar aromatic substituents (OMe, Me, H and Br) to the conjugated polyene chain and also the combined effects of these polar aromatic substituents.

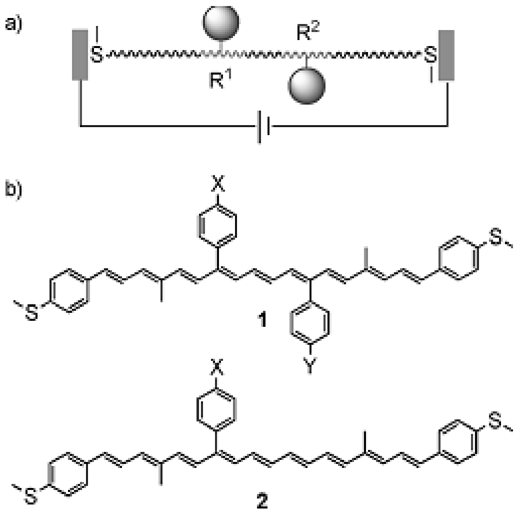


Figure 2.1 Conceptual design to measure carotenoid molecular wire. (a) Conceptual design of the circuit consisting of the carotenoid wires with polar substituents R[1] and R[2] as conductance of the carotenoid wires with polar modulators. (b) Carotenoid molecular wires 1 and 2 with the aromatic substituents(s) X of different electronic nature.

As shown in Figure 2.1, we changed the polar modulators of X and Y position with OMe, Me, H or Br and use the STM-based break junctions to study the electrical properties of the carotenoid molecule.

2.2 Experiment Procedure

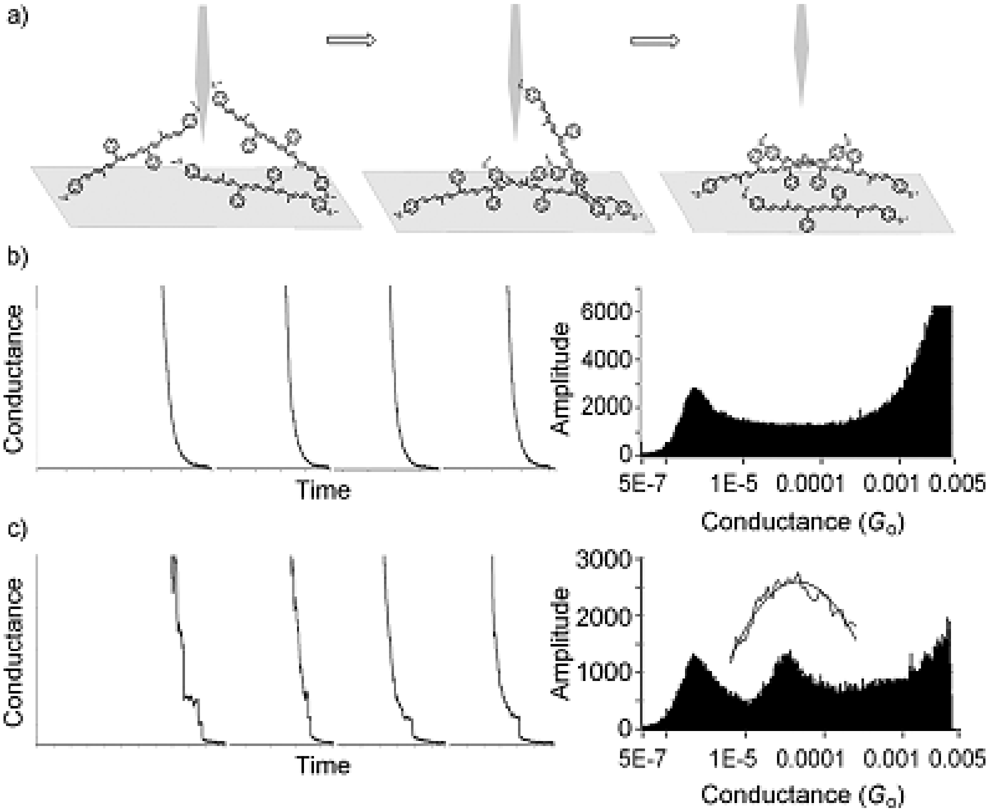


Figure 2.2 Conductance measurement of the carotenoid wire. (a) STM-based break junction. Carotenoid wires anchored to the gold substrate and anchored to the STM tip while the tip pushed very close to the surface and while the tip withdrawal, wires break. (b) Typical control conductance curve with only mesitylene solvent and conductance distribution (log scale). (c) Conductance curves after adding carotenoid wire. Plateaus in the curve indicate the break of one or a few molecules. A peak shows up in the histogram of conductance distribution, which gives the conductance of the single molecule conductance. Gaussian fit was applied to find the peak.

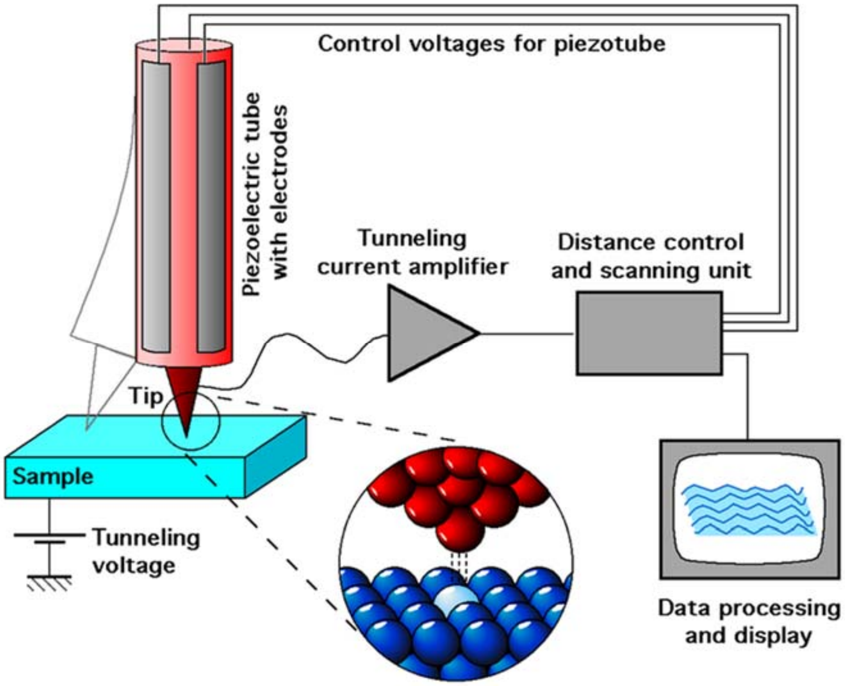

Control voltages for piezotube
Piezoelectric tube with electrodes
Tunneling current amplifier
Distance control and scanning unit
Tip
Sample
Tunneling voltage
Data processing and display

Figure 2.3 The schematic diagram of Scanning Tunneling Microcopy

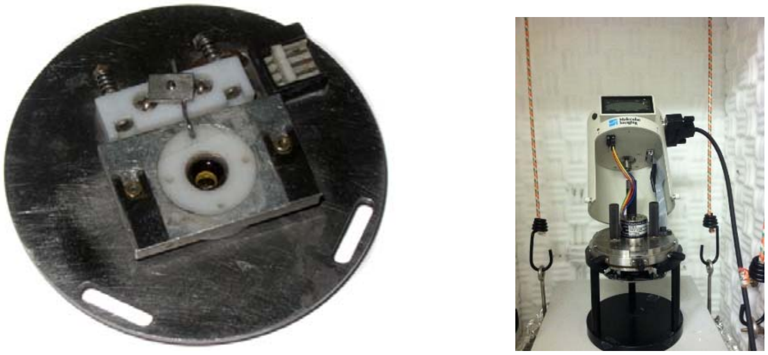
The concept of how STM-based junction works has been explained briefly in previous chapter. In Figure 2.3 [21], it shows how STM capture the topography of the sample. First a bias is applied on the sample and a set point current is pre-set according to the expected conductivity of the sample and other concerns. Then a stepper motor and PZT works alternatively to let the tip approach to the sample. Once the probe-sample gap is close enough, we can recall the transmission that we calculated in the previous chapter for a barrier. Tunneling currents start to be measureable. The stepper motor and PZT continue working until the tunneling current reaches to the pre-set value and stables eventually. Tip movements are controlled by a piezoelectric tube with a feedback loop

(PID control, usually just P Gain and I Gain is used). There are two modes for scanning: 1) servo on; 2) servo off. For servo on mode, the feedback loop makes the system keep a constant tunneling current, and topography of sample is mapped based on the movement of the tip in z-direction. For servo off mode, the feedback loop makes the system keep a constant tip-sample height, and uses the tunneling current to map the topography of the sample.

Back to the STM-based break junction, scanning the sample is usually not performed. This method only requires that the tip push into the surface (or close enough to the surface to pick up the analyte molecules) shown in Figure 2.2(a).

Most of STM-based break junction measurements run in aqueous solution or in organic solution [88-89], even it is possible to conduct measurements in ambient environment [31]. The use of solvent helps to reduce contaminations and it facilitates introduction of sample molecules. While the drawback of using solvent is that if solvent molecule has non-zero polarization such as water molecule, the ionic current would dominate the current and make the approaching to sample surface impossible. The solution to this problem is either using zero polarization solvent or coating the tip with wax or polymers to let only a tiny apex exposed so that ionic current could be reduced to negligible value (<1 pA). Details of coating tips will be given in the following chapter.

For the set up the STM-based break junction, PicoSTM (Agilent, Chandler) (Figure 2.4 (b)) with a fluid cell stage (Figure 2.4(a)) was used to perform measurements. We chose gold on mica as substrate (1.4cm X 1.1cm, Annealed gold-coated substrates with 1500 angstroms of Au (111) coving 1.0cm X 1.1cm). We used gold wire with diameter 0.25mm, 99.99% trace metals basis (Sigma-Aldrich) as the material to fabricate STM tip. Au wire was etched using chemistry method (Appendix A).



(a) (b)

Figure 2.4 (a) STM stage with a fluid cell. (b) STM (Agilent, Chandler)

Before using, Au substrates were hydrogen flamed for 30 seconds to remove contaminations and to reconstruct the surface to get bigger flat layer (Figure 2.5).

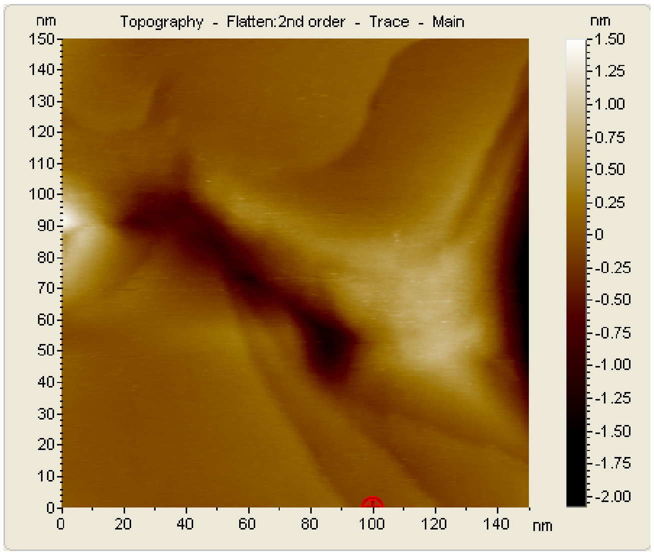


Figure 2.5 Gold layer on mica after hydrogen flaming. STM tip is Pt, with sample bias = -0.5V, set point = 10 pA, I gain = 1.5, P gain = 1.3

Au tip was also hydrogen flamed shortly (less than 1 sec), so the flame would not damage the apex of the tip.

Carotenoid wires were prepared in Dr. Sangho Koo's lab from Myong Ji Univeristy, Korea. The preparation process could be obtained online from the supporting information [28]. Nuclear magnetic resonance spectroscopy (NMR) for each carotenoid wire was also shown in the supporting information (Figure 2.6).

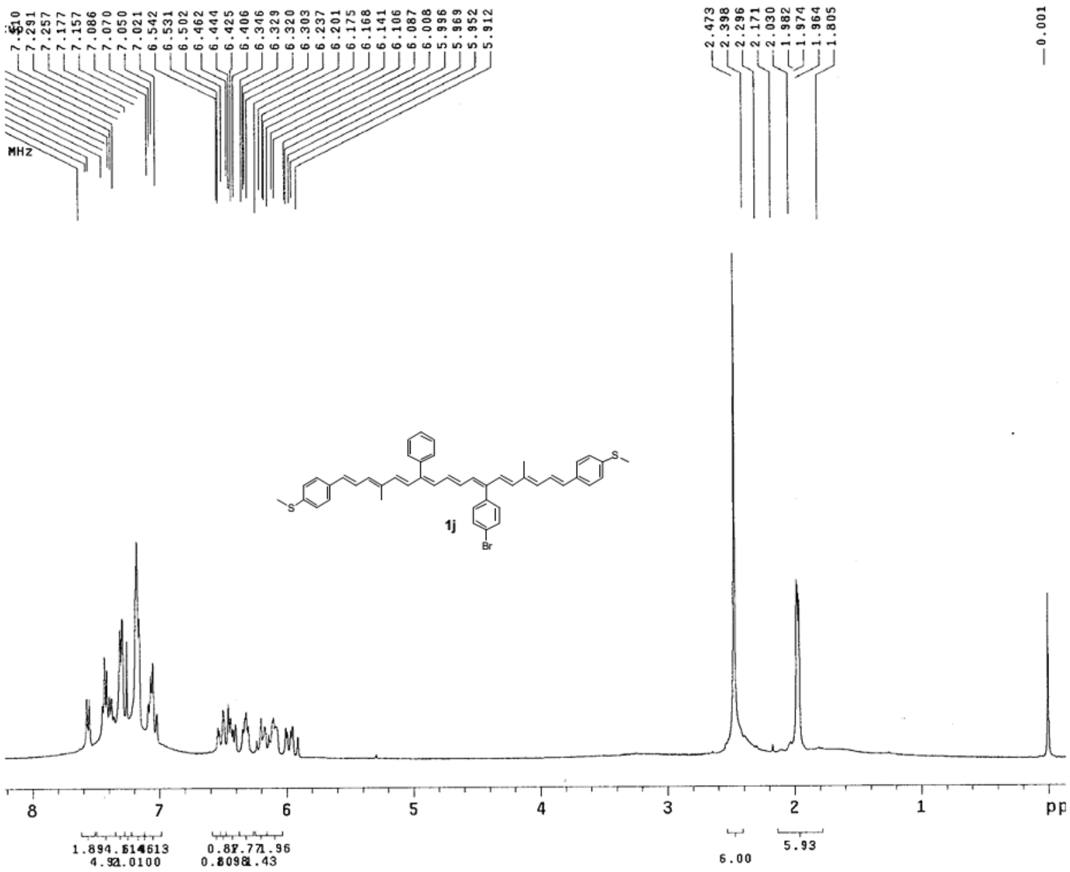


Figure 2.6 NMR of one of the carotenoid wires (X = H, Y = Br)

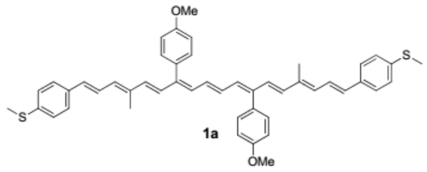
Mesitylene was used as solvent to dissolve carotenoid wires because 1). Carotenoid molecules have better solubility in organic solvents. 2). It has zero polarization. So there is no need to worry about ionic current during the measurement. 3). It has a relative high boiling point (164.7°C), so that solvent would not evaporate fast and we could collect thousands of conductance curves under the same conditions (such as concentration of the sample, etc.).

To collect the control data, only mesitylene solvent was introduced for the first step. STM tip was approached to the surface using the following settings: sample bias -0.3V, current set point 300pA, ramp rate 8Vs^{-1}; pre-amplification 10 nAV^{-1}. A custom LabView interface was developed in N.J Tao's lab, and these parameters were used to calculate the distance of the tip from the surface, conductance, etc. Once the tunneling current reached to the set point 300pA, we started tapping the tip. The LabView program had already implement this module, which continued pushing the tip very close to the surface until predefined maximum current reached and withdrawal it until the current went below predefined stop current. Thousands of conductance curves (only the withdrawal conductance curves are collected) obtained within a short period of time. Typical conductance curves of control experiment are shown on Figure 2.2(b). Initially the tip was very close the surface and very high conductance observed. As the tip pulled away from the surface at a constant speed, the tunneling current followed an exponentially decay. The histogram of the conductance distribution showed no peak except the initial part and end part, which were corresponding to the cutoff set point pre-set. Figure 2.2(c) shows conductance curves after adding one of the carotenoid wires

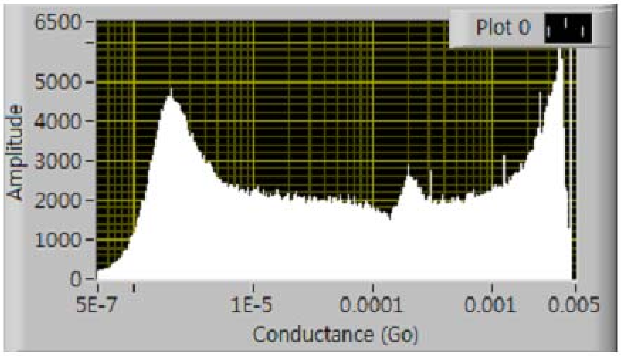
(X=Y = H). Compared with the control experiment conductance curve, plateaus started showing up, which were corresponding to the break of carotenoid wire. An obvious peak came up in the histogram of conductance histogram (Figure 2.7).

2.3 Results and Discussion

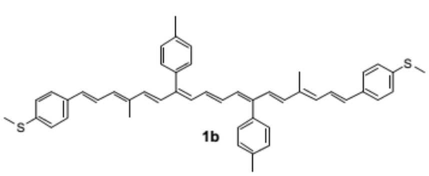
We measured ten carotenoid derivatives (Table 2.1). Five of the conductance histogram of carotenoid wires are shown in Figure 2.7. Complete results are available in the supplementary of the paper.



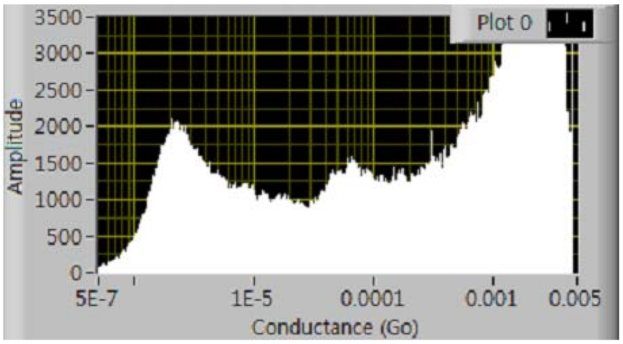
1a



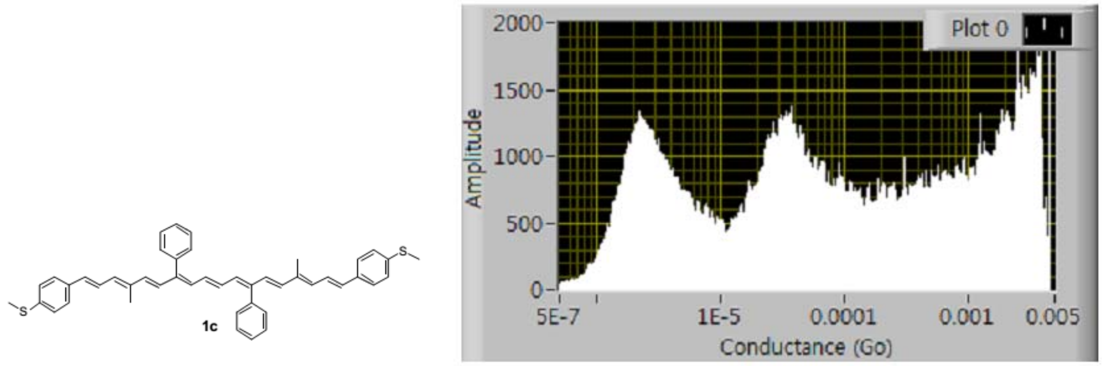
(a)



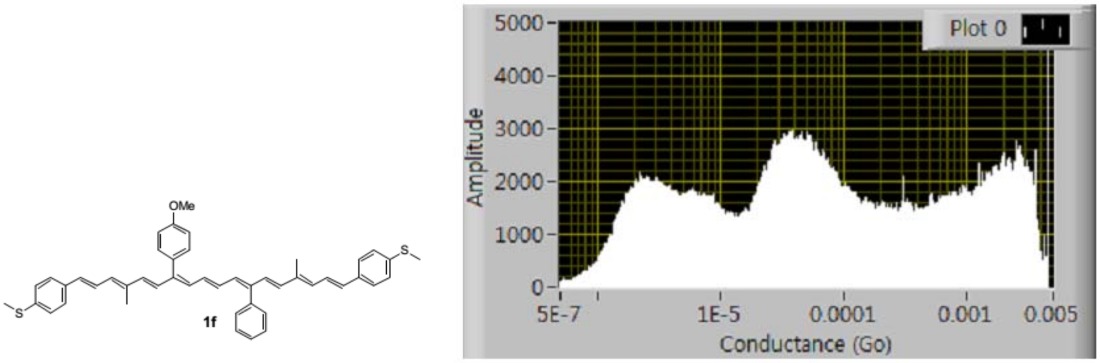
1b



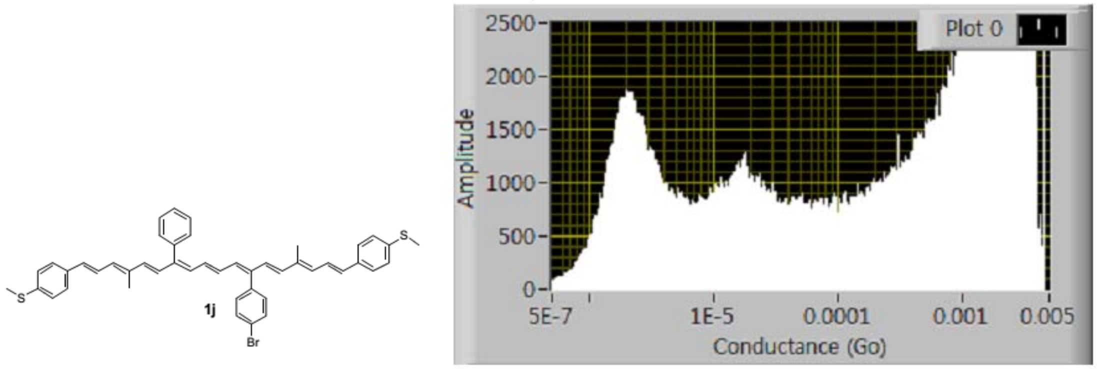
(b)


1c
Plot 0
Amplitude
2000
1500
1000
500
0
5E-7
1E-5
0.0001
0.001
0.005
Conductance (Go)

(c)


OMe
1f
Plot 0
Amplitude
5000
4000
3000
2000
1000
0
5E-7
1E-5
0.0001
0.001
0.005
Conductance (Go)

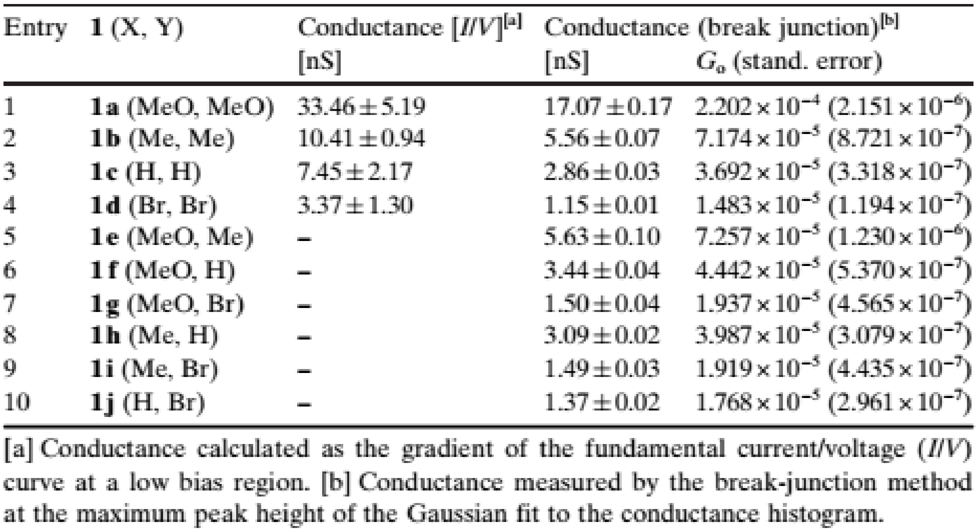
(d)


1j
Br
Plot 0
Amplitude
2500
2000
1500
1000
500
0
5E-7
1E-5
0.0001
0.001
0.005
Conductance (Go)

(e)

Figure 2.7 Conductance histograms of carotenoid wires. (a) (X,Y) = (MeO, MeO); (b) (X,Y) = (Me, Me); (c) (X,Y) = (H, H); (d) (X,Y) = (MeO, H); (e) (X,Y) = (H, Br)

Table 2.1 The conductance values of the carotenoid wires 1 with the substituents C_6H_4-X and C_6H_4-Y.



Entry	**1** (X, Y)	Conductance [*I*/*V*][a]	Conductance (break junction)[b]	
		[nS]	[nS]	G_o (stand. error)
1	**1a** (MeO, MeO)	33.46 ± 5.19	17.07 ± 0.17	2.202×10^{-4} (2.151×10^{-6})
2	**1b** (Me, Me)	10.41 ± 0.94	5.56 ± 0.07	7.174×10^{-5} (8.721×10^{-7})
3	**1c** (H, H)	7.45 ± 2.17	2.86 ± 0.03	3.692×10^{-5} (3.318×10^{-7})
4	**1d** (Br, Br)	3.37 ± 1.30	1.15 ± 0.01	1.483×10^{-5} (1.194×10^{-7})
5	**1e** (MeO, Me)	–	5.63 ± 0.10	7.257×10^{-5} (1.230×10^{-6})
6	**1f** (MeO, H)	–	3.44 ± 0.04	4.442×10^{-5} (5.370×10^{-7})
7	**1g** (MeO, Br)	–	1.50 ± 0.04	1.937×10^{-5} (4.565×10^{-7})
8	**1h** (Me, H)	–	3.09 ± 0.02	3.987×10^{-5} (3.079×10^{-7})
9	**1i** (Me, Br)	–	1.49 ± 0.03	1.919×10^{-5} (4.435×10^{-7})
10	**1j** (H, Br)	–	1.37 ± 0.02	1.768×10^{-5} (2.961×10^{-7})

[a] Conductance calculated as the gradient of the fundamental current/voltage (*I*/*V*) curve at a low bias region. [b] Conductance measured by the break-junction method at the maximum peak height of the Gaussian fit to the conductance histogram.

We used highly diluted carotenoid solutions (0.5 μM) when conducting measurement to make sure only single molecule was connecting tip and substrate each time. According to the histogram of conductance distribution, only one peak between the cutoff settings was observable. We believed the conductance we measured was indeed single molecule conductance, otherwise multiple peaks would be observable [92].

To calculate the single molecule conductance of each molecule, we used Gaussian fit to the conductance histogram at the maximum peak height (Figure2.2). The conductance of each molecule and error are summarized in Table 2.1.

We also used conducting AFM (cAFM) to measure conductance of each molecule wire (Table 2.1). The results measured by these two different methods were consistent.

We found some very interesting relationships between the conductance of the asymmetric wires 1e-j and that of the symmetric ones 1a-d. The conductance of wire 1e (MeO, Me) is 7.257G_0 and this value was very close to the wire 1b (Me, Me) with conductance 7.174G_0. The conductance values of asymmetric wires 1f (MeO, H) (conductance = 4.442 G_0) and 1h (Me, H) (conductance = 3.987 G_0) were almost same as symmetric wire 1c (H, H) (conductance = 3.692 G_0). The same trend was observed for the asymmetric wires 1g (MeO, Br) (conductance = 1.937 G_0), 1i (Me, Br) (conductance = 1.919 G_0), and 1j (H, Br) (conductance = 1.768 G_0) compared with symmetric wired 1d (Br, Br) (conductance = 1.483 G_0). The total conductance of the carotenoid wire containing two polar aromatic substituents was controlled mainly by the more electron-withdrawing substituent, whereas the more electron-donating one contributed to the fine tuning.

To explain the combined effect of the two different polar aromatic substituents on the conductance of the carotenoid wire, we believe that the electron-releasing anisyl substituent would increase the near-by π-electron density along the conjugated polyene chain, whereas the electron-withdrawing bromophenyl group would decrease it (Figure 2.8). The flow of π electrons along the carotenoid wire is controlled by the more electron-withdrawing group of the two acting like a "bottle-neck effect": the more resistive contributor controls the total conductance of the carotenoid wires. The overall effect after quantum interference of the two polar substituents was lowering the HOMO of the carotenoid relative to the gold Fermi level by the resistive modulator, thereby lowering the total conductance of the wire.

The conductance of the carotenoid wire containing the anisyl and the bromophenyl group is obviously controlled by the more resistive bromophenyl substituent.

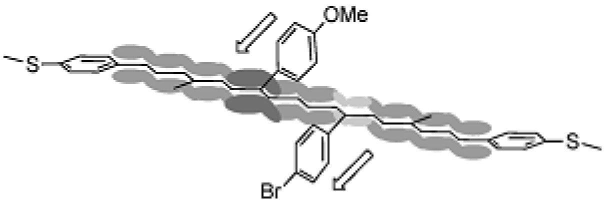


Figure 2.8 Control of the electronic flow in the carotenoid wire by the resistive substituent: a "bottle-neck" effect in the π-electron density by the electron-withdrawing substituent.

Chapter 3

Recognition Tunneling Method and Peptide Sequencing

Single molecule conductance measurement is challenging, while the outcome is quite rewarding. By studying unique electrical properties of molecules, not only people are exploring the revolutionary single molecule devices such as organic luminous diodes and transistors, but also by differentiating molecules based on their intrinsic properties, people could manufacture products and develop techniques applicable for industry, especially in production quality control (up to single molecule level) and sequencing research. In these areas, conductance is one of the unique intrinsic characters that molecules have [81, 82], while more sophisticate tools and methods are needed to achieve the high precision requirement for sequencing. In this chapter, I will illustrate the recognition tunneling (RT) method developed in our lab and its application in amino acids identification and peptide sequencing.

This work has been published on Nature Nanotechnology [73]. The current author carried out the measurements.

3.1 Recognition Tunneling

The principle of STM-based break junction method is to bonding and breaking the molecules from probe and substrate by changing gap size. Since STM is capable of forming adjustable molecular size gap by varying bias and current set point, it is capable of carrying electric properties measurement at a fixed gap. Wolfgang Haiss et al. [31][32]

observed that stochastic tunneling currents showed up by placing a gold STM tip above a bi-thiolated alkane chain functionalized gold surface at a fix gap, the time course of the tunneling current through metal/molecule/metal junction showed spontaneous switching between 2 levels (“telegraph noise”) (Figure 3.1) [31] [80].

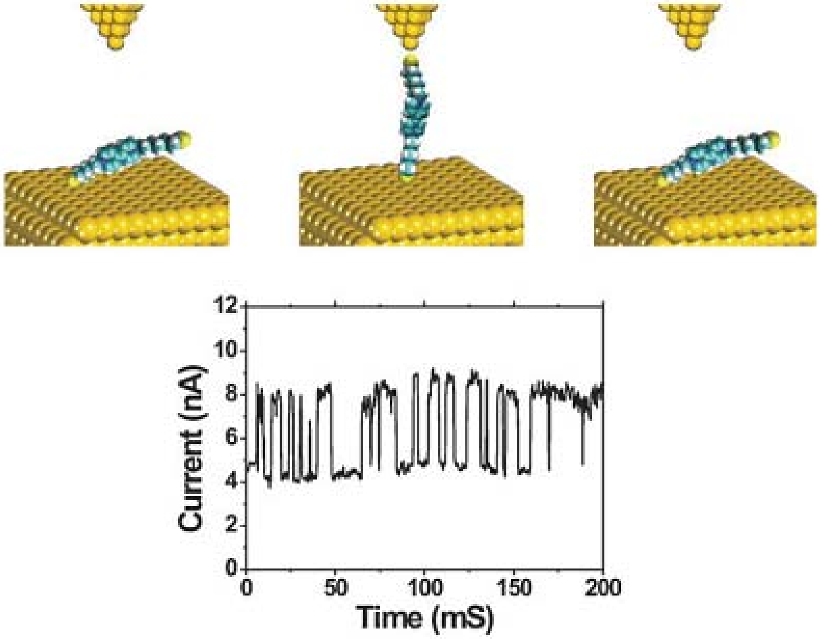


Figure 3.1 Schematic representation of the *I(t)* technique. Molecular junction stochastically form and break in a constant height gap between STM tip and metal substrate. The junction formation and cleavage process is characterized by current jumps.

The STM-based break junction method discussed in the previous chapter is usually referred as *I(s)* technique because tunneling current is measured in tip-substrate distance domain. While holding the tip steady at a constant height relative to the substrate and measuring the tunneling current in time domain is referred as *I(t)* technique.

Haiss's method provides another way conducting single molecule conductance measurement. While one limitation for this method is that the target molecule has to have anchor groups on both ends such as thiol group to bond with the tip and substrate. Haiss carried measurement in ambient environment. Contaminations could be a problem during measurement unless running in vacuum. Compared with *I(s)* method, *I(t)* method provides another important character about of molecule – the life time t_2 (Figure 3.2). The life time did not catch people's attention in their paper, but for molecule identification and sequencing, it is actually one of the most important parameters to evaluate the bond strength between molecules. I will discuss this in following chapters. Even counting in life time parameter, it is still almost impossible to apply this technique to peptide sequencing we are planning to do because we need to distinguish at least 20 amino acid and the distribution of these two parameters or joint-distributions would be so overlapped that the accuracy of differentiation would be very low.

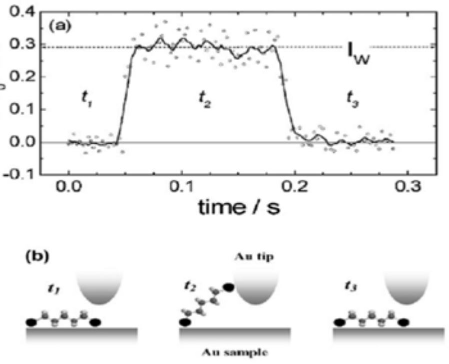


Figure 3.2 ON-OFF states switching analysis (a) A typical signal of spontaneous ON-OFF switching during an *I(t)* measurement. The ON-OFF region are marked by the label t_1 (OFF), t_2 (ON) and t_3 (OFF). (b) The corresponding schematic diagram of the bonding situation during the time period of t_1 (unbonded), t_2 (bonded) and t_3 (unbonded) [31].

To overcome the technique limitations and pioneer new methods to do molecule identification and Peptide/DNA sequencing, we advanced this method, which was named “Recognition Tunneling” (RT) method.

To solve anchor group problem, we came up with an idea, which was to modify the STM-tip and substrate with some “Reader” molecules. The requirements of the “Reader” molecule were: First, it should have anchor group to bond with probe and substrate of STM covalently such as thiol-group. Second, based on the functional groups on analytes, it should has special designed groups that could be bonded with target molecules. Third, it should has certain flexibility so that while bonding with target molecules, the bonding structure could be optimized. Fourth, its conductivity should be high enough so that tunneling currents are measureable when small bias applied. In result, once the “Reader” molecules are modified on the tip and substrate, there should be no requirement that target molecules having anchor groups because the connections for RT method becomes Metal-Reader-Analyte-Reader-Metal. In our lab, we have been focusing on the application of this method in DNA/Peptide sequencing. The “Reader” molecules that we have been designed all have a thiol-group to bond with STM tip and substrate (Au or Pd) and specified ammonium and carboxylate groups to form H-bonds with DNA bases or Amino Acids. The “Reader” molecules we have been developed so far as shown in Figure 3.3.

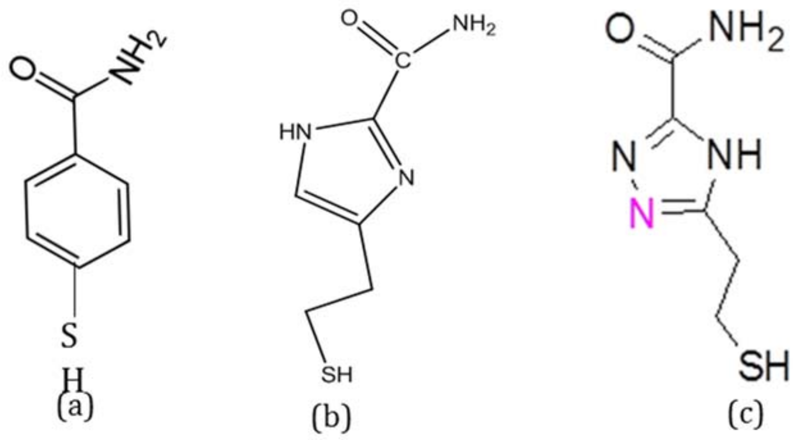


Figure 3.3 Reader molecules (a) 4-mercaptobenzamide (b) 4(5)-(2-mercaptoethyl)-1H-imidazole-2-carboxamide. (c) 5-(2-mercaptoethyl)-H-1,2,4-triazole-3-carboxamide

The method of modification and the quality will be discussed later in this chapter.

To solve the potential contamination problem if measurements were carried in ambient environment, we purposed to run the experiment in aqueous solution, phosphate buffer (PB) to be specific. Another advantage using PB is all analyte solutions have the same PH, so PH of solution would not be a variable influencing on results. Ionic current leakage problem need to be solved when measuring in aqueous solution, as mentioned in the STM-based break junction method. So, we built an auto coating machine to coat STM tips with polyethylene and only left a tiny apex exposed after coating. The coating process and images of the tips before and after coating will be discussed in section 3.2.

The problem that there were not enough parameters to identify multiple molecules were solved once we introduced "Reader" molecule. Because for "Metal-Reader-

Analyte-Reader-Metal" connection, not only the conductance was measured, lots of physical/chemical information from the intrinsic characters of the molecule and bonding of "Reader-Analyte-Reader" were reflected in the structure of "telegraph noise" as well. We named these intrinsic characters as "Fingerprint". The "Fingerprint" of different molecules were so different that we could separate 20 amino acid by an accuracy over 90%. I will discuss the details in section 3.3 and 3.4.

To give an accurate definition of RT method, it involves a traditional STM, coated tips with neglectable ionic current (less than 1 pA when applied 0.5 V bias) in aqueous solution, special designed "Reader" molecule and well modified on both tip and substrate, and last, algorithms to analysis the "Fingerprint" of the analytes.

3.2 Peptide Sequencing

3.2.1 Why Peptide Sequencing

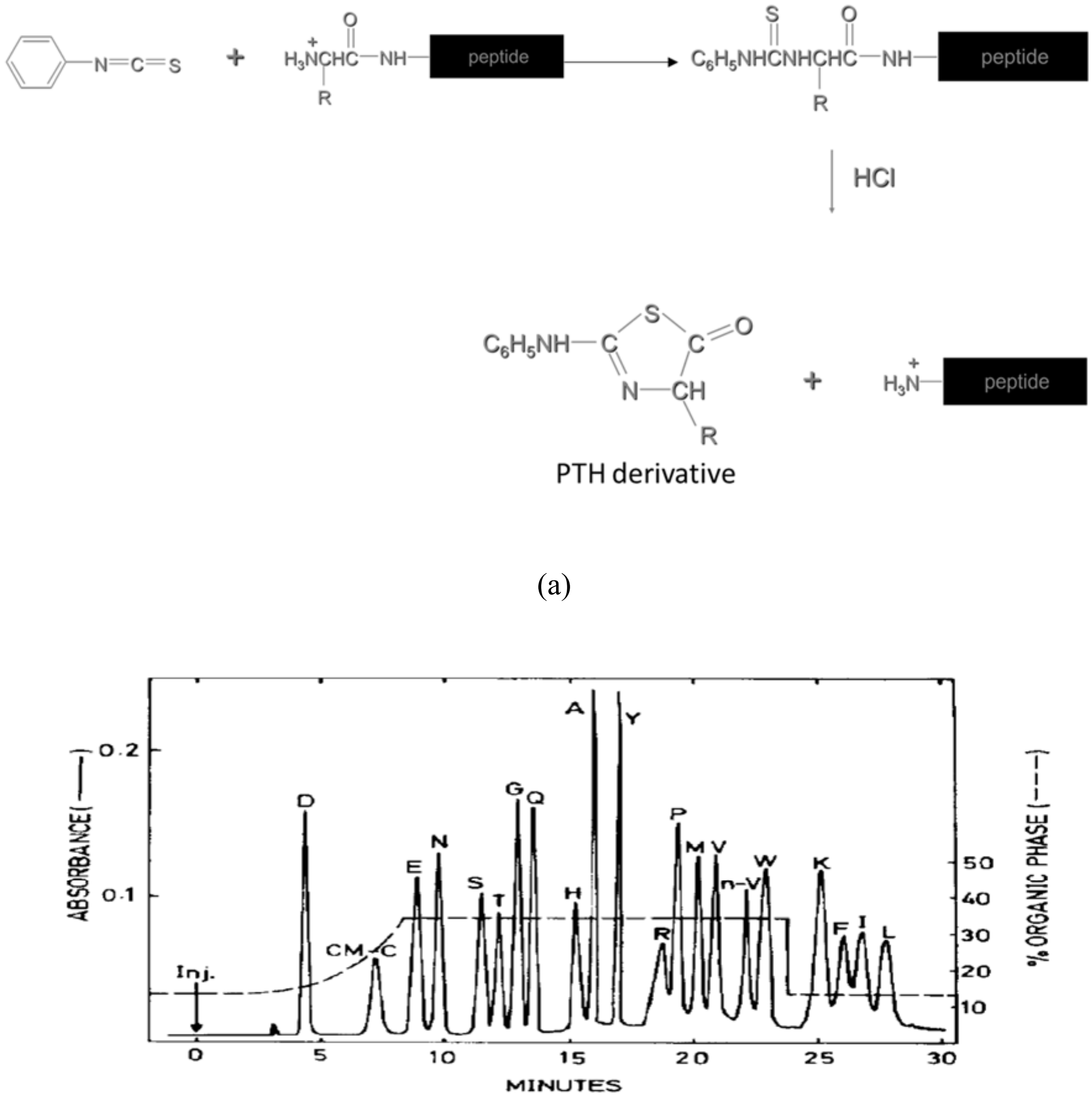
Peptide sequencing or protein sequencing has not been caught people's as much attention as DNA sequencing nowadays. The main reason, I think, is that people believe it is much more challenging to sequence billions of DNA bases. And since there are only four possibilities for each deoxyribonucleotide – deoxyadenylate (or dAMP), deoxyguanylate (or dGMP), thymidylate (or dTMP) and deoxycytidylate (or dCMP), it is easier to distinguish four molecules compared with 21 amino acids in eukaryotes. However, the challenging of Human Genome Project is not just identify the sequence of base of human DNA, but also to identify and to map all the genes of the human genome from both a physical and functional standpoint, which means how these genes express and synthesis of a functional gene product - proteins. One important fact that need to be mentioned is that 98% of human genome is noncoding DNA [32]. The functions of these noncoding DNA is still under studying [74][75], but from perspective of disease prediction and medical diagnosis, it is rather inefficient to sequence billions of bases just to figure out if there is a very short subsequence that would produce some specified protein. In other words, if we are able to detect the specified protein sequencing that are corresponding to the gene mutation, virus invading, or anything related to abnormal protein, we could use peptide sequencing to achieve the same purpose as DNA sequencing. Compared with DNA sequencing, it is more directly, efficient, fast and cheap.

3.2.2 Traditional Peptide Sequencing Methods

One of the challenges as mentioned is that there are too many amino acids. Figure 3.6 shows the 21 amino acids founded in eukaryotes. Another challenge is that samples of DNA, RNA or Protein are frequently present in minute concentrations. For DNA and RNA, polymerase chain reaction (PCR) method is usually applied to amplify the low concentration, while, there is no such method to amplify protein sequence. Thus, for traditional protein sequencing, it is also quite different from DNA sequencing. The mostly used traditional methods used to sequence protein are Edman degradation reaction [34] and mass spectrometry [33].

3.2.2.1 Edman Degradation

Edman degradation was developed by Pehr Edman [35]. The idea of this method is to cleave the N-terminal amino acid one by one without disrupting the peptide bonds between other amino acid residues. The chopped N-terminal acid is labeled and analyzed (Figure 3.4 (a)).



(a)

(b)

Figure 3.4 Edman degradation process (a) and Ion chromatography for amino acid analysis (b)

The processes of the Edman degradation are as following: First, phenylisothiocyanate is reacted with an uncharged terminal amino group under slightly alkaline conditions. After the reaction, a cyclical phenylthiocarbamoyl derivative is formed. Second, this derivative is treated under acidic condition and heat, so the derivative of the terminal amino acid is cleaved as a thiazolinone derivate. Third, under the same condition, the unstable thiazolinone derivate becomes a stable phenylthiohydantoin (PTH) – amino acid derivative. At last, this PTH-derivate could be identified by using chromatography or electrophoresis. While the next amino acid at N-terminal could be analyzed the in same way. The first major limitation of this technique is that the length of the peptide is confined under 50-60 (in practice, under 30) because the cyclical derivatization is not always going to completion. Longer peptide has to be cleaved into shorter peptides before using Edman degradation. Another limitation of this technique is that Edman degradation would not work if the N-terminal amino acid has been chemically modified or if it is concealed within the body of the protein.

Ion Chromatography is mostly used method for the last step of Edman degradation to identify amino acid molecules. The separation of ions and polar molecules is based on their affinity to the ion exchanger. A sample is introduced into a sample loop of known volume. Buffered aqueous solution (mobile phase) carries the sample from the loop onto a column that contains stationary phase material, which is typically a resin or gel matrix consisting of agarose or cellulose beads with covalently bonded charged functional groups. The target analytes (anions or cations) are retained on the stationary phase but can be eluted by increasing the concentration of a similarly charged species that will displace the analyte ions from the stationary phase. The analytes of interest then could be detected by other methods as conductivity or UV/Visible light absorbance. Take the separation of amino acid for example. The procedure of running Ion chromatography is as following:

1. Buffered aqueous solution (pH = 2.2) carries the analytes are introduced into a column with cationic resin. Because in this pH level, PTH-amino acids are all carrying positive charges, they will be absorbed on the resin, while the bonding varies for different molecules because of the molecule's size, polarization, etc.
2. Increasing the pH, the positive charges amino acids carries starts to decrease and finally eluted from resin. The eluting order is based on the bond strength.
3. Ninhydrin reaction is used to detect the PTH-amino acid eluted. The reaction product has a purple color, which could be do quantitatively analysis using 440nm or 570nm light.
4. Based on the retention time to identify the PTH-amino acid as shown in Figure 3.4 (b).

3.2.2.2 Mass Spectrometry

Another major direct method to sequence a protein is Tandem Mass Spectrometry. There are lots of methods to deliver peptides to the spectrometer. One of the most popular ones is by electrospray ionization. Initially, the protein is digested by an endoprotease, and the resulting solution is passed through a high-pressure liquid chromatography column. The solution is sprayed out of a narrow nozzle charged to high positive potential into the mass spectrometer at the end this column. Charges on the droplets break apart the droplets until only single ions remain. The peptides are then fragmented and the mass-to-charge ratios of the fragments measured. 1 peak from selected from the spectrum of peptide mixture and pass collision cell to fragment peptide. After that, Mass Analyzer is selected to measure the spectrum and compared against a database of previously sequenced proteins to determine the sequences of the fragments. The process is then repeated with a different peak from the spectrum of peptide mixture until all the peaks are analyzed (Figure 3.5).

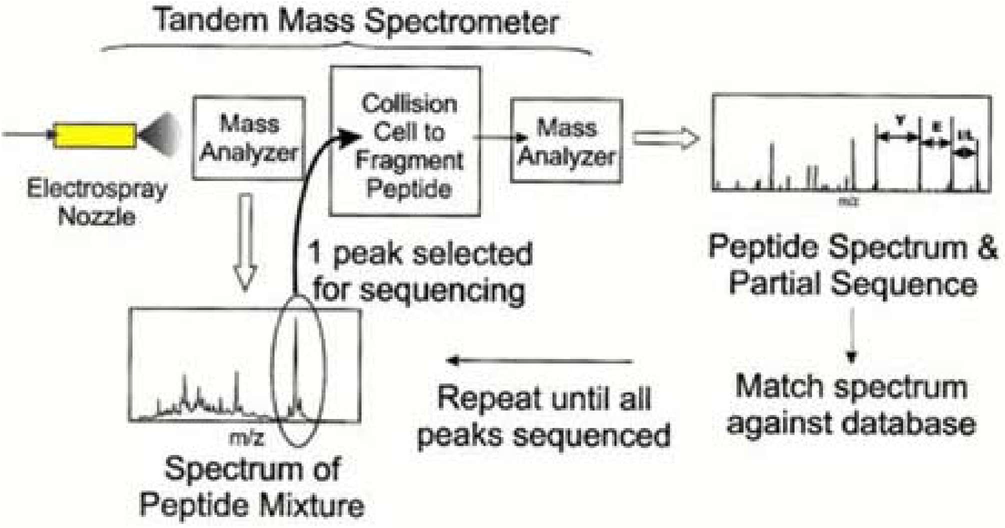

Tandem Mass Spectrometer
Electrospray Nozzle
Mass Analyzer
Collision Cell to Fragment Peptide
Mass Analyzer
m/z
Peptide Spectrum & Partial Sequence
1 peak selected for sequencing
m/z
Spectrum of Peptide Mixture
Repeat until all peaks sequenced
Match spectrum against database

Figure 3.5 Tandem Mass Spectrometer

While all the traditional methods have their own drawbacks, as for Edman degradation, it requires mg weight scale analytes and Tandem MS has limitations on the length of peptide. Is it possible to sequence a peptide in single molecule basis? So, the first step we tried is to use the recognition tunneling mentioned earlier to see if it was possible to identify the 21 amino acids found in eukaryotes (Figure 3.6). 1H-imidazole-2-carboxamid (ICA) (Figure 3.3(b)) is the reader molecule used in our measurements.

3.3 Methods and Experiment

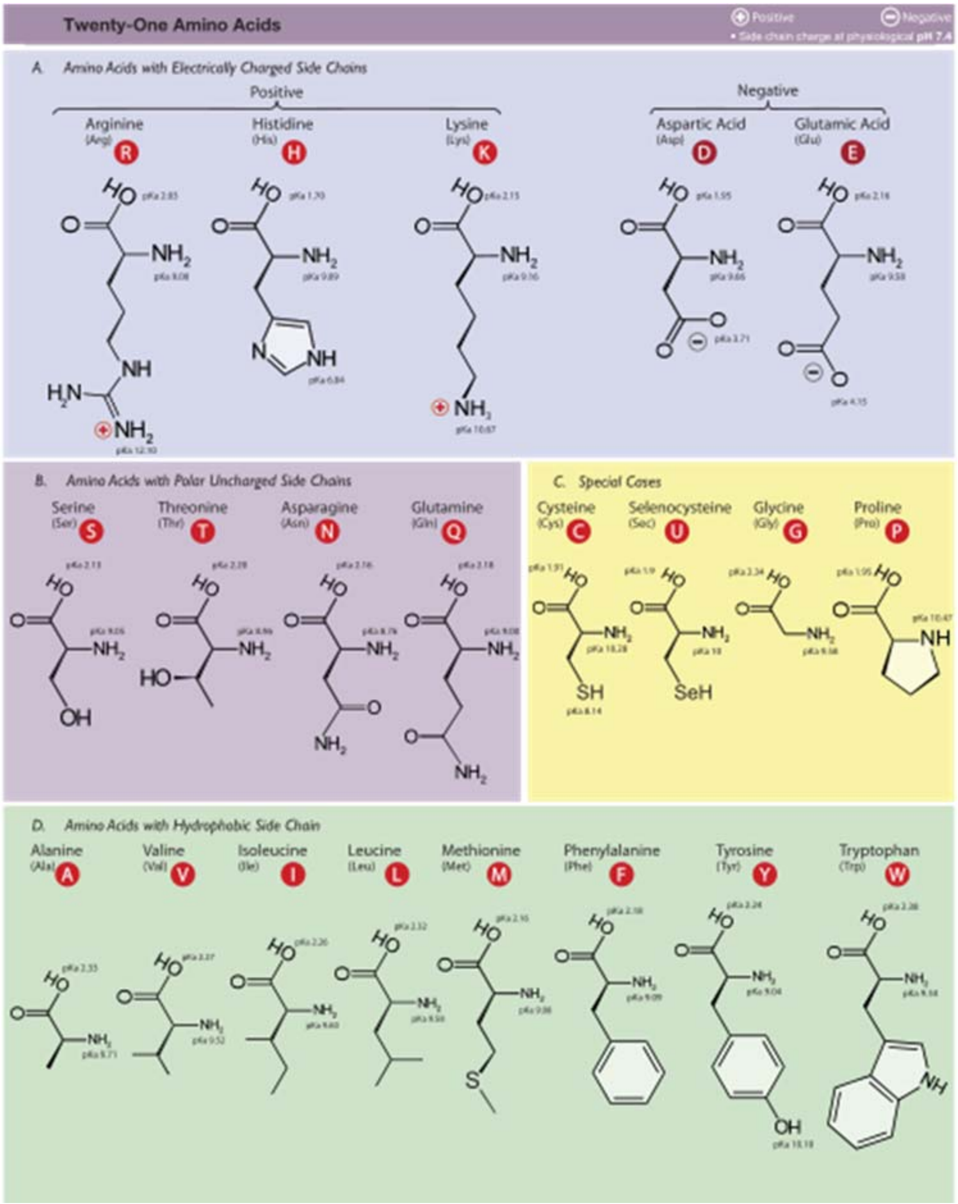


Figure 3.6 The 21 amino acids found in eukaryotes (From Dan Cojocari, Department of Medical Biophysics, University of Toronto)

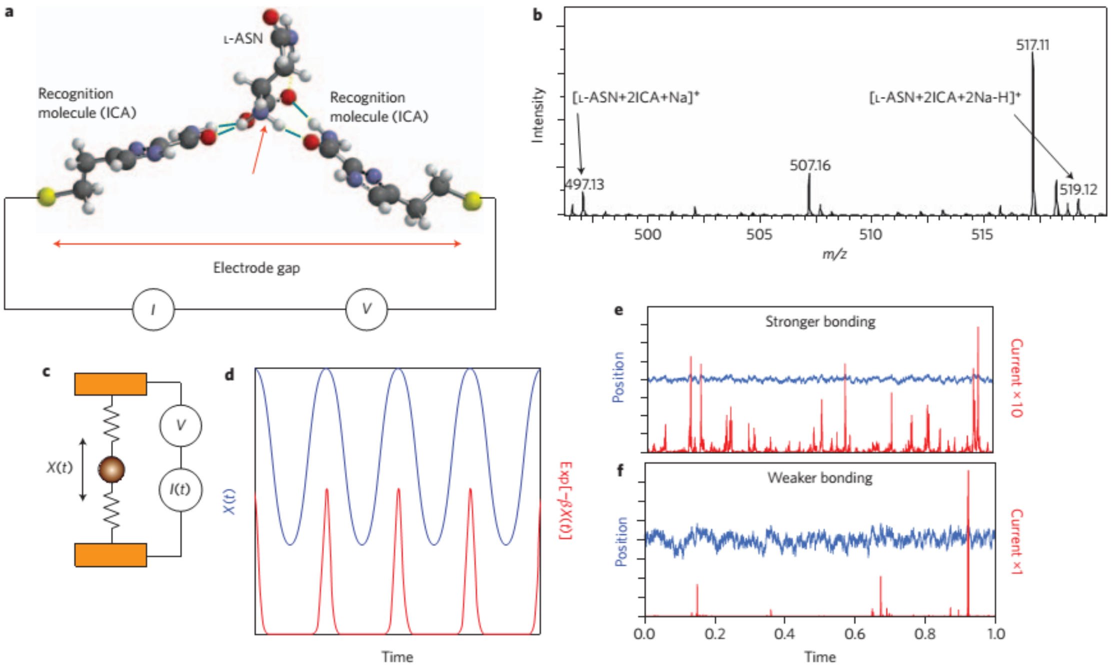


Figure 3.7 Recognition tunneling (RT). a, Recognition molecules (1H-imidazole-2-carboxamid, ICA) are strongly attached to a pair of closely spaced electrodes, displacing contamination and forming a chemically well-defined surface. An analyte (here shown as L-ASN) is captured by non-covalent interaction (blue bars show hydrogen bonds) with the recognition molecules. The bonding pattern is specific to the analyte. The red arrow shows the orientation of the molecular dipole for L-ASN. This orientation is different when D-ASN is captured. b, ESIMS shows the stoichiometric adducts form between reader molecule, here illustrated for 2:1 complexes of ICA and L-ASN. c, Generation of RT signals. Picturing the analyte as a mass (sphere) trapped by a pair of springs that represent the non-covalent bonds, the extend of analyte motion, X(t), depends on the strength of the springs. d, A simple sinusoidal motion of the analyte (blue trace) produces a series of sharp current spikes (red trace) because of the exponential dependence of tunnel current on position. e,f, Simulations for random thermal excitation of a strongly (e) and more weakly (f) bonded analyte, showing how the current fluctuations are much bigger when the bonding is weaker (red traces).

3.3.1 Preparation of Analytical Solutions

The 21 amino acids found in eukaryotes (Figure 3.6) were obtained from Sigma-Aldrich (98% purity) and dissolved in 1mM phosphate buffer (PH = 7.4) made using water from a Milli-Q system with specific resistance of 18 MΩ cm and total organic carbon contamination below 5 ppb.

3.3.2 Self-Assembled Monolayer and its Characterization

There are three widely used materials (Au/Pt/Pd) for substrate and STM tips (Table 3.1) [36]. Au is the most widely used because it can form a relative large flat area after annealing (Figure 3.5), but the problem for Au is that it is not compatible with CMOS fabrication facilities [37]. So we use palladium as the substrate and STM tip not only because it is compatible with CMOS fabrication but also theoretical calculation shows that it has higher conductance than gold after modified with ICA molecules and it is highly reactive toward thiol groups, which make it easier to modify the “Reader” molecules on the substrate and tip. Another advantage for Pd is that it resists oxidation in air up to 800°C, so there is no need to worry about surface oxidation during modification of “Reader” molecules.

Table 3.1 Au/Pt/Pd physical properties comparison

	Au	Pt	Pd
Shear Modulus	27 GPa	61 GPa	44 GPa
Reduction Potential	$Au^{3+} + 3e^{-} \Leftrightarrow Au$: +1.52V	$Pt^{2+} + 2e^{-} \Leftrightarrow Pt$: +1.188V	$Pd^{2+} + 2e^{-} \Leftrightarrow Pd$: +0.915V
Melting point	1064.18 °C (1947.52 °F)	1768.3 °C (3214.9 °F)	1554.9°C (2831°F)
Density at 20°C	19.30g/cm^3	21.45 g/cm^3	12.02g/cm^3
boiling point	2856 °C	3825°C	2963°C
Atomic radius	1.35Å	1.35Å	1.37Å

Polycrystalline Pd substrate were deposited on a 750 μm silicon wafer using electron-beam evaporation of 100 nm Pd onto a 10 nm Ti adhesion layer. Then the wafer was cut into small pieces of 1cm*1cm size. An STM image of the Pd substrate was shown in Figure 4.8. The Pd substrate was then rinsed by ethanol and water for 2 min each in sequence. After that, substrate was blow-dry by N_2. Traditional ultrasonic cleaning or H_2 flaming was not used because these procedures tend to either break the grid of the Pd or oxidize it [36].

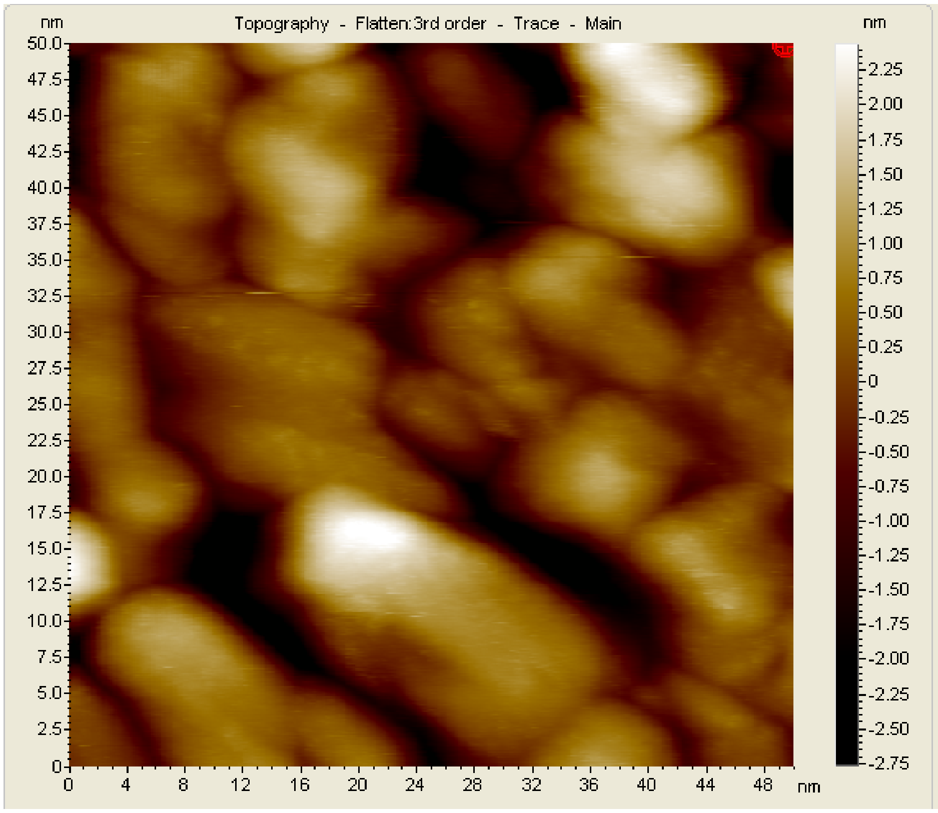


Figure 3.8 STM image of Pd surface (50nm*50nm). PicoSPM (Agilent Technologies) with I Gain = 1, P Gain = 1, Setpoint = 0.01 nA, Sample Bias = -0.7 V

To modify the Pd substrate with a Self-assembled monolayer (SAM) of ICA, thiol groups on ICA were bonded with Pd covalently [78]. HPLC grade ethanol was degased by Argon (Ar) for 2~3 minutes to remove O_2 dissolved in the ethanol. ICA molecule is relative stable. But if the molecule is exposed in O_2 too long, there could be some reaction (thiol oxidation). ICA was dissolved in the pre-degased ethanol to get a 0.5 mM solution. The pre-cleaned Pd substrate was immersed in this solution for 16 hours, after which a SAM was formed on Pd. X-ray photoelectron spectroscopy (XPS), ellipsometry, Fourier transform infrared spectroscopy (FT-IR) and STM imaging were employed to characterize the SAM.

XPS data were obtained by VG ESCALAB 200i-XL photoelectron spectrometer. The source of the spectrometer was 1500 eV Al-Kα radiation at 6×10^{-10} mbar base pressure. Wide scan spectra were obtained at a passing energy of 150 eV. Core level high resolution spectra as C1s, Pd3d, N1s, S2p were collected with a passing energy of 20 eV. CasaXPS package was used for the curve fitting of S2p spectra and atomic concentration calculation [36].

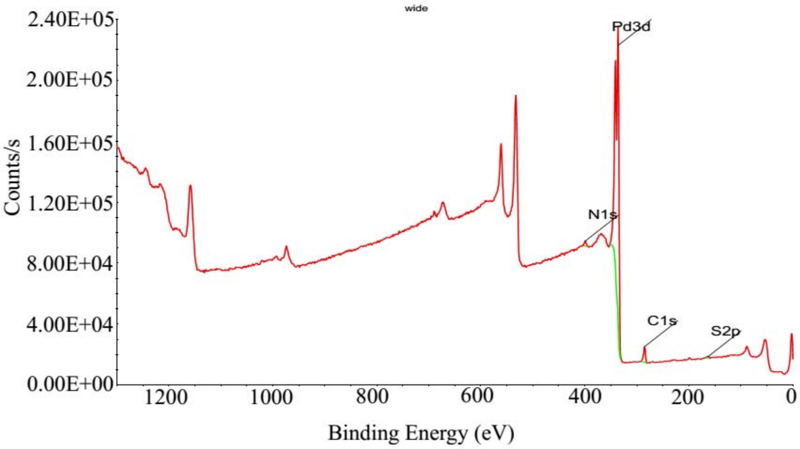


Figure 3.9 XPS spectrum of ICA on Pd substrate

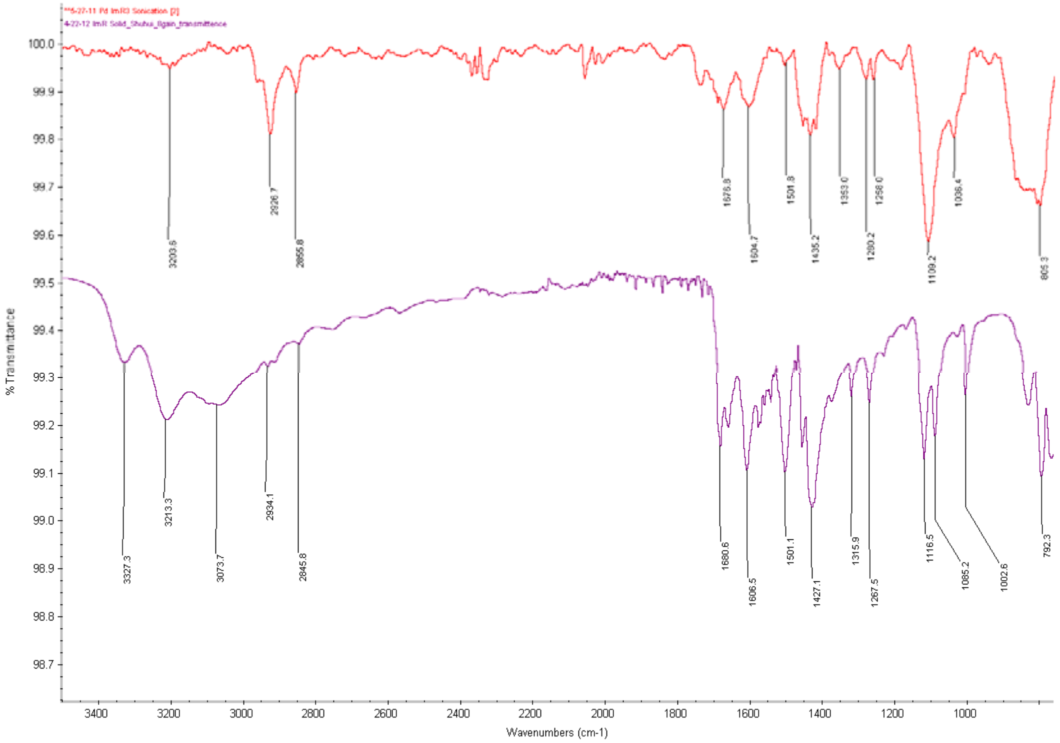


Figure 3.10 FTIR spectrum of modified Pd substrate by ICA

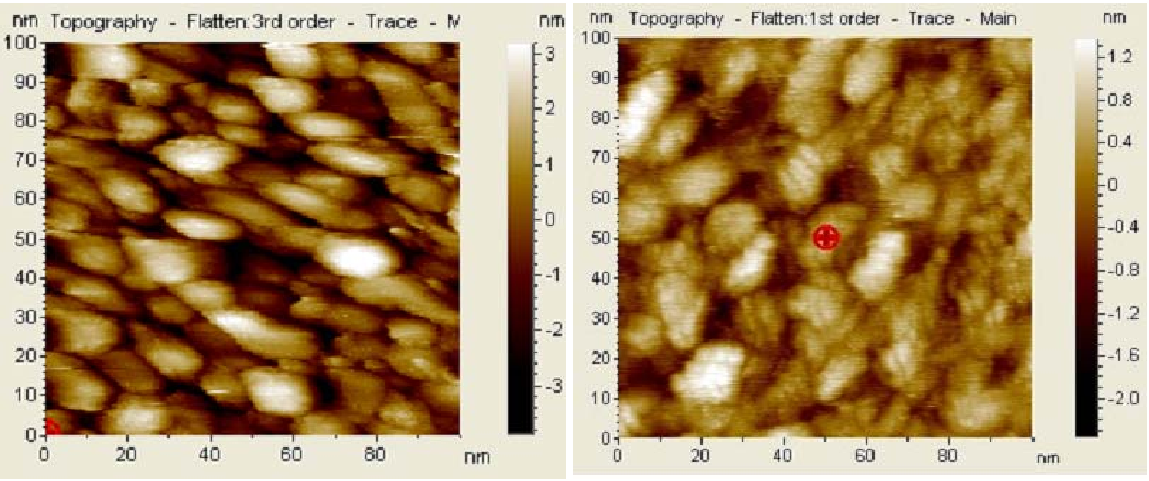
Besides XPS, we also used FTIR to verify the modification by checking the particular set of functional groups of ICA. A FTIR spectrum [38] of the powder sample of 4(5)-(2-mercaptoethyl)-1H-imidazole-2-carboxamide were recorded on a attenuated total reflection (ATR) accessory (Smart Orbit, Thermo Electron Corporation) at 4 cm-1 resolution with 128 scans. By comparing these two FTIR spectra, the monolayer spectrum has the similar characteristics as one of the powder sample in the region of 1700 to 800 cm-1. The broad band from 3400 to 2800 cm-1 in the powder spectrum implies the intermolecular hydrogen bonding interactions among the imidazole molecules. The formation of monolayer prevents the intermolecular interactions, resulting in a disappearance of the band in a polarized FTIR spectroscopy. The bands at 2927 and 2856 cm-1 are assigned to anti-symmetric and symmetric stretching of the CH2 groups of the flexible carbon chain (Figure 3.10).

Ellipsometry was used to measure the thickness of layer. The estimated length of the ICA molecules is around 8.7 Å [36]. In Table 3.2, we measured the thickness of layers on the substrates with different modification time. The thickness data showed valid indication of SAM was formed on Pd substrate.

Table 3.2 Time Course Ellipsometry analysis of SAM thickness

12h	24h	48h
10.3±0.6 Å	12.2±1.4 Å	13.2±1.1 Å

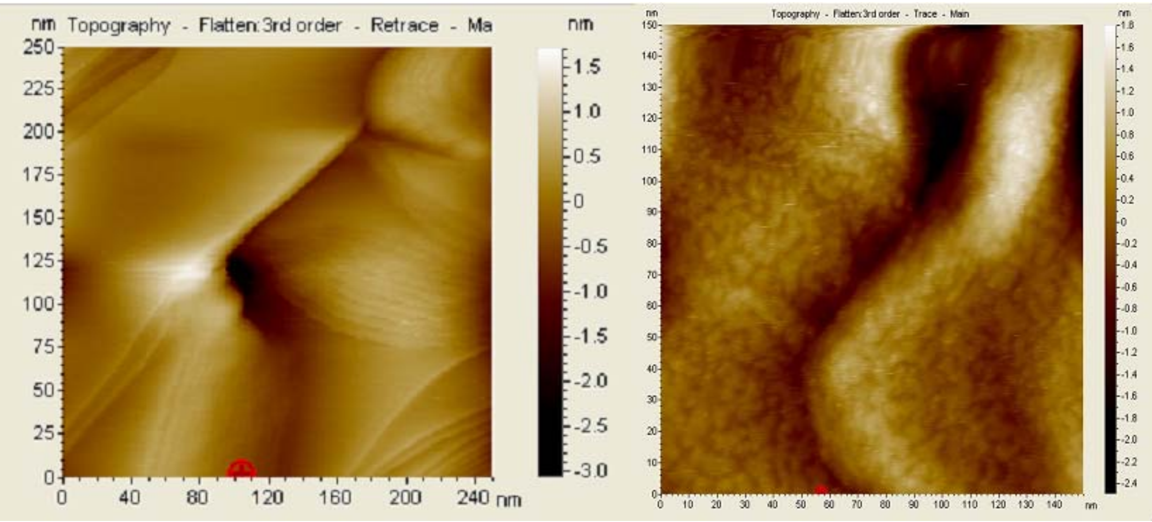
STM imaging was also used to check the topography of the sample.



(a) (b)

Figure 3.11 (a) STM image of Pd without modification. (b) STM image of Pd after modification.

For STM imaging, Chemical etched Pt-Ir (90:10) tips were used as the PicoSPM's tip. SAM was not quite obvious on the Pd, possibly because of the rough grid of Pd surface. We also tried to modify ICA on Au substrate, the modification procedure was the same as Pd's. A clearly modified layer showed on Au substrate (Figure 3.12).



(a) (b)

Figure 3.12 Au substrate before (a) and after modified with ICA (b).

3.3.3 Fabrication of STM Tips

Probes were etched from 0.25mm Pd wire (California Fine Wires). Then a chemical etching process was used to get a traditional functional STM tip. The etching process was almost the same etching Au tips shown in APPENDIX A. Since experiments were run in aqueous solution, ionic current would dominate and get tunneling current was impossible. We developed an auto coating machine and a standard procedure to coat the tips with high density polyethylene (Sigma-Aldrich). And only a tiny apex were exposed after the coating. Leakage current was checked by STM after coating. Only the tips with leakage currents less than 1pA (the minimum current STM could detect) under 0.5V bias were kept for further steps.

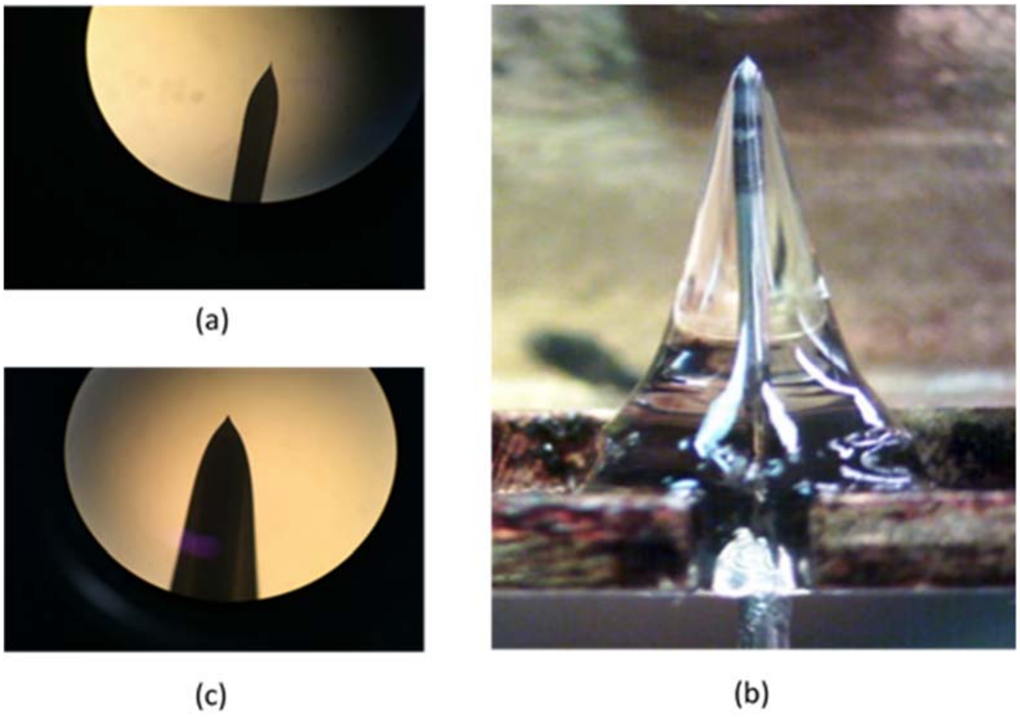


Figure 3.13 Pd tips before (a), during (b) and after (c) coating

Figure 3.13 shows the optical image of Pd tip before coating, in the middle of coating and after coating. The diameter of the Pd wire is 0.25mm.

After etching, tips were rinsed by water and ethanol in sequence to wash away etching solution residues left on the tip. Then, tips were dried gently with nitrogen. Then a home-made auto-coating machine was used to process the coating. High density polyethylene was heated to 225 °C and after that tip was slowly pushed through the polyethylene under the stage at the speed of 0.1 mm/s. The coating length of the tip was around 8 mm. After the first step coating finishes, machine withdrawal the tip along the same path at the same speed until the tip apex was under the stage. Finally, a second step coating followed the same process at the first one. The moment that the coating stopped, you should blow some cool air gently to the tip to cool polyethylene rapidly and remove the tip from the stage and check the structure of the tip under optical microscope to make sure the exposure was barely able to see. Leakage checking using STM was applied afterwards.

If the tip survived the previous checks, it was ready to be modified. We used the same ICA solution as the one used to modify the substrate. Both the tip and substrate were modified for 16 hours. After the monolayer were formed, both tips and substrates were rinsed with ethanol, gently dried with nitrogen and used immediately.

Glass vials for making solution and the liquid cell of the STM stage were rigorously cleaned by Piranha solution (Sulfuric acid: Hydrogen peroxide = 3:1) and then ultrasonic cleaned in water (Milli-Q system) (twice, each last 5 minutes), ethanol (twice, each last 5 minutes) in sequence.

3.3.4 Tunneling Measurement

We used two different PicoSPMs (Agilent Technologies) equipped with custom LabView interfaces for data acquisition. The standard produce running tunneling current measurement is in APPENDIX B. 1mM PB was added to the liquid cell to collect control data before adding amino acid solution. The current set point was set to 4pA with 0.5V bias applied (probe positive, as this results in less leakage). During the approach, I Gain and P Gain was set to 1.0. Once the tunneling current setting 4pA arrived, a gap size around 2.4 nm was obtained [39]. Tunneling current was sampled at 50 kHz. The -3 dB bandwidth of the current-to-voltage converter was 7 kHz, but useful signals were obtained out to the Nyquist limit of 25 kHz after correction for the instrumental response. Once the approach finished, a scanning was first applied to make sure the geometry of the tip was ideal to run experiment. A typical STM image of the substrate is shown in Figure 3.14. If the tip passed the test, it was withdrawn from the surface for 20 µm and sample bias was set to -0.1V for 2 hours to stabilize the tip and substrate. After 2 hours, sample bias was set back to -0.5V, and re-approached to the surface and I Gain and P Gain was set to 0.1 so that the system's response would be slow enough to collect signals at a relative constant gap. Control experiment data was collected. After the control experiment, tips was again withdrew from the surface with I Gain and P Gain set to 1 for 100 µm. Stage was taken out to add amino acid solution in the liquid cell and tip was approached to the surface again. Another 1 hour stabilization was applied to reduce the drift of the stage or vibration of the tip. The same procedure as collecting control data was used to collect amino acid signals.

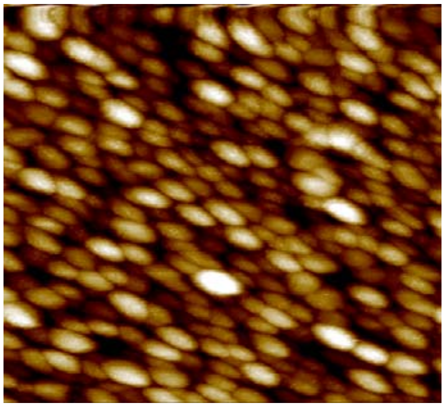
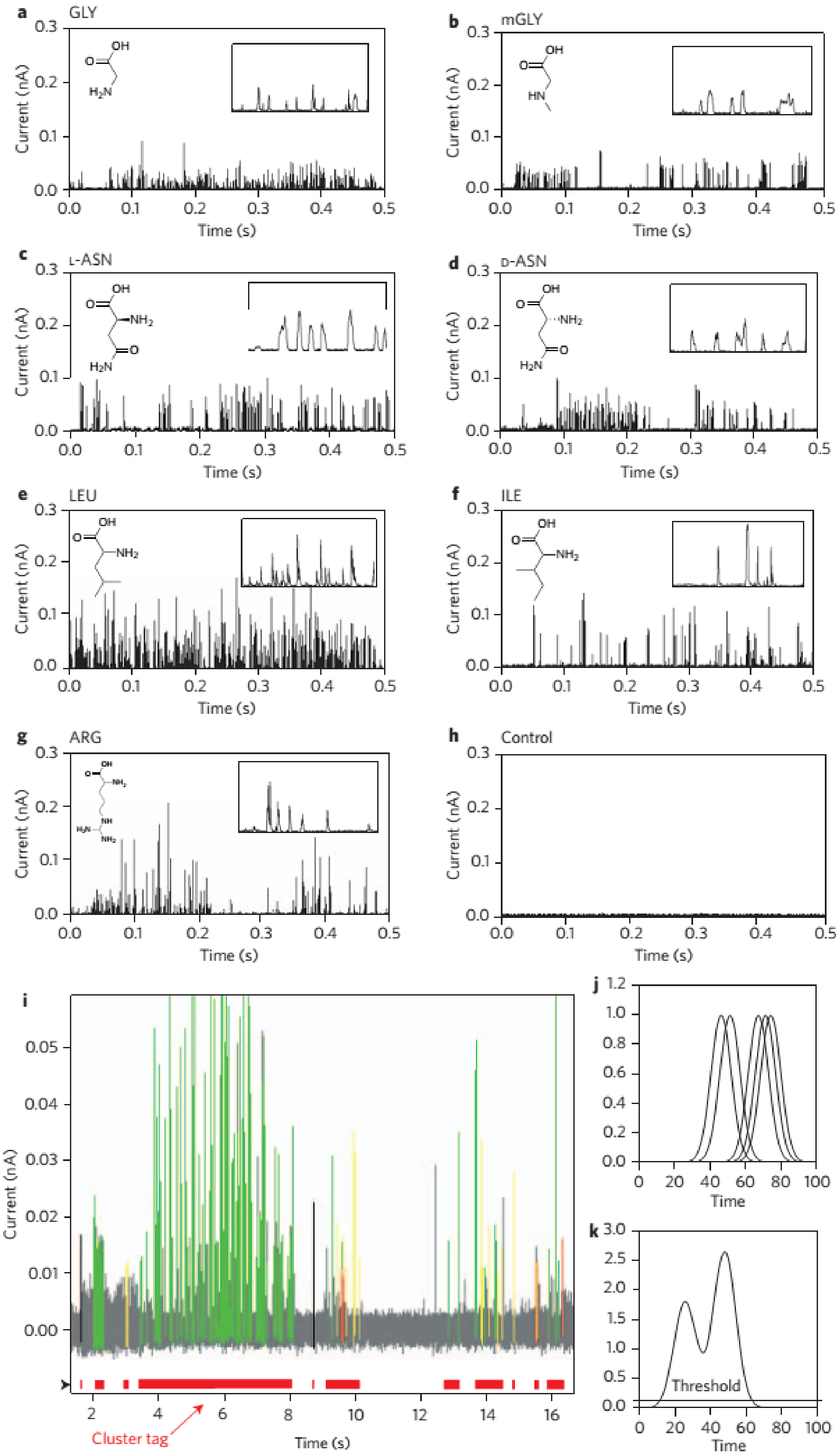


Figure 3.14 Grain structure of the Pd substrate as imaged with an insulated, functionalized Pd probe. Set point is 4 pA at 0.5V bias. Image size is 200 nm x 200 nm and height scale is 0-2 nm.

3.4 Results and Discussion

3.4.1 Bonding in the RT Junction

Several combinations of bias and Setpoint were tested and we found that the best condition to do the measurement was (Sample bias, Setpoint) = (-0.5V, 4pA). Under this setting, the trace of control experiment was very clean with rare noise signals, which might be corresponding to electron noise or vibration of the probe caused by random sound from the environment (Figure 3.15 h). While after adding in the amino acids, frequent spikes showed up in bursts (Figure 3.15 (a)-(g)).

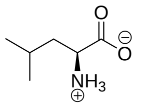
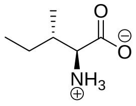
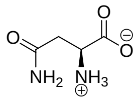
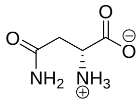
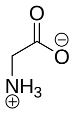


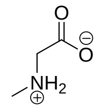
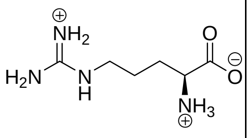
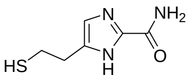
a
GLY
b
mGLY
c
L-ASN
d
D-ASN
e
LEU
f
ILE
g
ARG
h
Control
Current (nA)
Time (s)
0.0
0.1
0.2
0.3
0.4
0.5
OH
O
H2N
HN
NH2
i
Current (nA)
0.00
0.01
0.02
0.03
0.04
0.05
2
4
6
8
10
12
14
16
Cluster tag
Time (s)
j
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Time
k
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Threshold
Time

Figure 3.15 Examples of RT signals from amino acids. (a) GLY and its N-methylated modification, sarcosine (mGLY) (b). Enantiomers L-ASN (c) and D-ASN (d). Isobaric isomers LEU (e) and ILE (f). (g) Data for the charge amino acid, ARG. (h) Control data from buffer solution alone. Insets: expanded traces displaying the complex peak shapes that are important features in the analysis of these data. (i) Signal trace for ARG, colour-coded according to the peak assignments made by a machine learning algorithm (green, correct; red, wrong call; black, 'water peak'; yellow, common to all amino acids). (j),(k) Automatic cluster identification was carried out by placing Gaussians of unit height and full width of 4,096 data points (1 data point = 20 μs) at the location of each spike (j), summing them (k), and assigning a cluster to regions where this sum exceeds 0.05. This choice picks out obvious single-molecule events well.

Our models explained what happened after adding amino acid solutions. Take L-Asparagine (L-ASN) for example, density functional theory was used to calculate the optimized structure (Figure 3.7(a)). It showed that amino acid can be captured by hydrogen-bonding to ammonium and carboxylate groups of their zwitterionic centers. To study the interactions of ICA with amino acids, we used electrospray ionization (ESI) mass spectrometry. ESI is able to transfer weakly bound complexes to the gas phase for mass analysis. It has been used to study noncovalent interactions [40-45]. ESI should be able to give us what kind of species in the solution. First, we measured the ESI mass spectra of seven individual amino acid molecules and ICA molecule dissolved in PB buffer (Table 3.3). The observed m/z shows that both amino acids and ICA can form self-associated adduct including dimers and trimers besides monomer structure. After that, we mixed ICA with seven amino acids respectively in 1:1 and 2:1 ration and measured ESI mass spectra again. Figure 4.16 shows an example of ES-MS spectra of pure Leucine solution, ICA solution and ICA+Leucine (2:1) ratio. Clearly, there were new m/z peaks corresponding to 1:1 and 2:1 adducts in the spectra. We confirmed the complexes by tandem mass spectrometry (ESIMS/MS) (Figure 3.17). The ESIMS/MS data of ICA-amino acid solutions are given in Table 4.4. In summary, ESIMS/MS studies prove that ICA can form 1:1 and 2:1 complexes with amino acids in aqueous solutions. The procedure of running the experiment could be obtained from the supporting information of this paper.

Table 3.3 Structure information and MS data of Individual Amino Acids and ICA

Analyte	**Calculated Monoisotopic Mass**	**Solution pH**	**Molecular form**	[1] **Observed *m/z***
L-Leu	131.0946	8.1		154.04, [M+Na]$^+$, (82) 176.03, [M+2Na-H]$^+$, (85) 285.16, [2M+Na]$^+$, (100)
L-Ile	131.0946	8.0		154.04, [M+Na]$^+$, (65) 176.03, [M+2Na-H]$^+$, (100) 285.16, [2M+Na]$^+$, (50)
L-Asn	132.0535	8.1		155.00, [M+Na]$^+$, (100) 176.99, [M+2Na-H]$^+$, (98) 287.09, [2M+Na]$^+$, (23) 485.10, [3M+4Na-3H]$^+$, (81)
D-Asn	132.0535	8.1		155.00, [M+Na]$^+$, (34) 176.99, [M+2Na-H]$^+$, (66) 287.07, [2M+Na]$^+$, (11) 485.09, [3M+4Na-3H]$^+$, (100)
L-Gly	75.0320	7.8		97.96, [M+Na]$^+$, (39) 119.95, [M+2Na-H]$^+$, (100) 173.02, [2M+Na]$^+$, (70) 314.02, [3M+4Na-3H]$^+$, (53)

***N*-MeGly**	89.0477	7.9		111.98, $[M+Na]^{+}$, (100) 133.97, $[M+2Na-H]^{+}$, (66) 201.06, $[2M+Na]^{+}$, (45) 356.07, $[3M+4Na-3H]^{+}$, (16)
L-Arg	174.1117	8.1		175.08, $[M+H]^{+}$, (100) 197.07, $[M+Na]^{+}$, (37) 219.05, $[M+2Na-H]^{+}$, (23) 371.20, $[2M+Na]^{+}$, (57)
ICA	171.0466	8.3		172.02, $[M+H]^{+}$, (8) 194.01, $[M+Na]^{+}$, (76) 365.07, $[2M+Na]^{+}$, (25) 363.06, $[ICA'+Na]^{+}$

1. The relative Intensity (%) value of observed ions are given in parentheses next to each complex ion. The most intense peaks in single stage MS spectra are defined as 100.

Table 3.4 Characteristic ESIMS of ICA-amino acids 1:1 & 2:1 mixture and their MS/MS products

Analyte			**Observed *m/z***	**MS/MS Product Ion**
	Ratio	pH	Mass, adduct ion, (Intensity, S/N)	Mass, molecular ion, (Intensity)
ICA+L-Leu	1:1	7.8	325.12, $[ICA+Leu+Na]^{+}$, (15.3, 1703)	194.00, $[ICA+Na]^{+}$, (100)
	2:1	7.9	518.16, $[2ICA+Leu+2Na-H]^{+}$, (0.2, 80)	176.03, $[Leu+2Na-H]^{+}$, (100)
ICA+L-Ile	1:1	7.8	325.12, $[ICA+Ile+Na]^{+}$, (13.5, 1494)	194.00, $[ICA+Na]^{+}$, (100)
	2:1	7.9	496.18, $[2ICA+Ile+Na]^{+}$, (0.1, 42)	194.01, $[ICA+Na]^{+}$, (100)
			518.16, $[2ICA+Ile+2Na-H]^{+}$, (0.2, 60)	176.03, $[Ile+2Na-H]^{+}$, (100)
ICA+L-Asn	1:1	7.9	326.08, $[ICA+L-Asn+Na]^{+}$, (6.1, 800)	155.00, $[L-Asn+Na]^{+}$, (100) 194.00, $[ICA+Na]^{+}$, (5)
	2:1	8.0	497.13, $[2ICA+L-Asn+Na]^{+}$, (0.5, 60)	365.06, $[2ICA+Na]^{+}$, (100) 155.00, $[L-Asn+Na]^{+}$, (48)
			519.12, $[2ICA+L-Asn+2Na-H]^{+}$, (0.4, 42)	176.99, $[L-Asn+2Na-H]^{+}$, (100)

ICA+D-Asn	1:1	7.9	326.08, [ICA+D-Asn+Na]$^+$, (3.9, 691)	155.00, [D-Asn+Na]$^+$, (100) 194.01, [ICA+Na]$^+$, (5)
	2:1	8.0	497.13, [2ICA+D-Asn+Na]$^+$, (0.4, 67)	365.06, [2ICA+Na]$^+$, (100) 155.00, [D-Asn+Na]$^+$, (74)
ICA+L-Gly	1:1	8.0	269.05, [ICA+Gly+Na]$^+$, (0.2, 53)	194.01, [ICA+Na]$^+$, (100) 172.02, [ICA+H]$^+$, (24)
			291.03, [ICA+Gly+2Na-H]$^+$, (0.1, 30)	119.95, [Gly+2Na-H]$^+$, (100)
	2:1	8.1	462.10, [2ICA+Gly+2Na-H]$^+$, (0.1, 24)	119.95, [Gly+2Na-H]$^+$, (100)
ICA+N-MeGly	1:1	8.0	261.09, [ICA+N-MeGly+H]$^+$, (0.2, 23)	172.02, [ICA+H]$^+$, (100)
			283.07, [ICA+N-MeGly+Na]$^+$, (0.4, 35)	194.00, [ICA+Na]$^+$, (100)
	2:1	8.1	476.11, [2ICA+N-MeGly+2Na-H]$^+$, (0.1, 11)	133.97, [N-MeGly+2Na-H]$^+$, (100)
ICA+L-Arg	1:1	7.8	346.16, [ICA+Arg+H]$^+$, (0.2, 81)	175.09, [Arg+H]$^+$, (100)
	2:1	7.9	517.21, [2ICA+Arg+H]$^+$, (0.2, 59)	175.09, [Arg+H]$^+$, (100)
			539.19, [2ICA+Arg+Na]$^+$, (0.3, 92)	197.07, [Arg+Na]$^+$, (100)

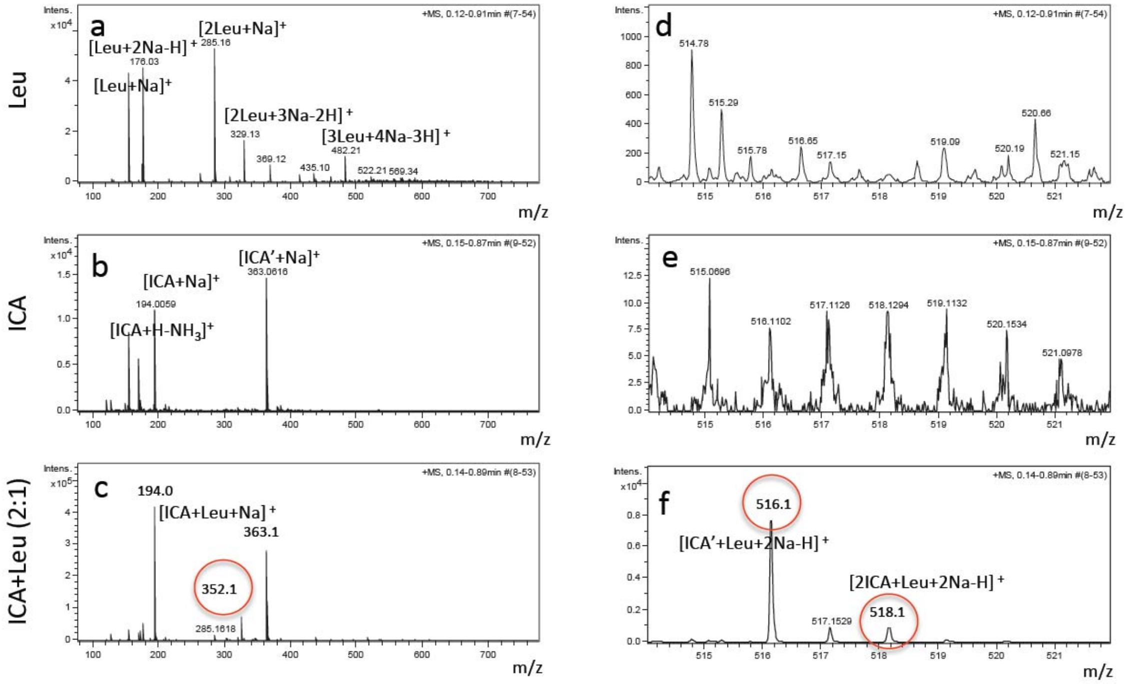


Figure 3.16 Example of ES-MS spectra of pure compounds and complexes. (a) Leucine. (b) ICA, (c) ICA+Leucine at 2:1 ratio. (d), (e), (f) show spectra at higher resolution.

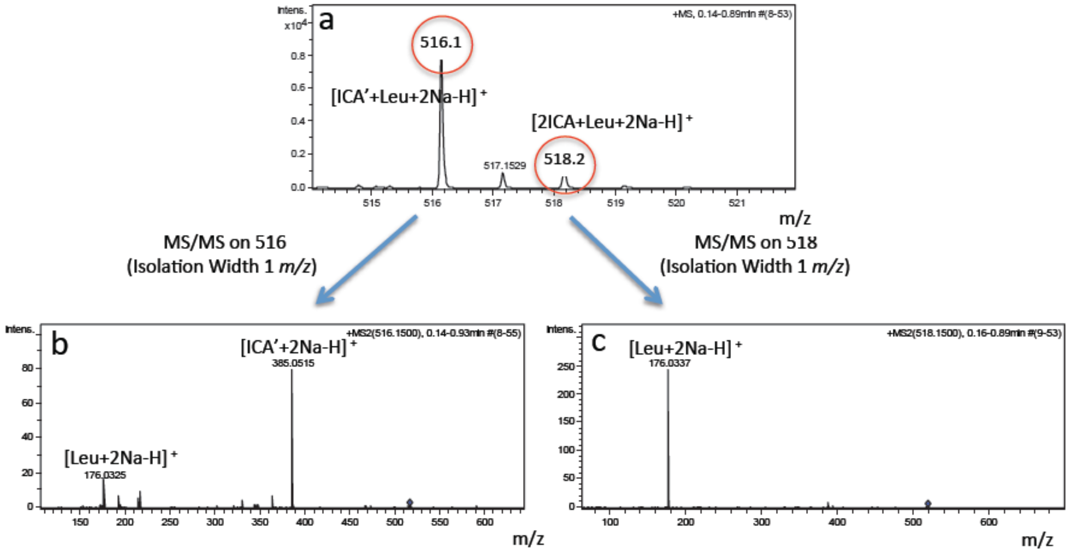


Figure 3.17 Example of MS-MS spectra. Two peaks are found in 2:1 mixtures of ICA with Leucine, circled in (a). MS-MS shows that the peak at 516 Dalton is a complex of an oxidized ICA (Labeled ICA’) in which two ICA molecules are joined by a disulfide linkage (b). The peak at 518 Daltons is shown (c) to consist of two non-oxidized ICA molecules with one Leucine.

The influence of strength of bond on the signals are simulated in Huang’s paper [46], as shown in Figure 3.7 (c-f).

3.4.2 Identifying Amino Acids

Figure 3.18-Figure 3.20 shows the raw signals of 16 amino acids with control experiment present at the same experiment condition (Sample bias, Setpoint) = (-0.5V, 4pA). While we could not get any signals from tyrosine (Tyr) and Tryptophan (Trp). Setpoint for Tyr was increased to 6pA and 10 pA for Trp to get much narrower gaps to capture signals for these two molecules (Figure 3.21).

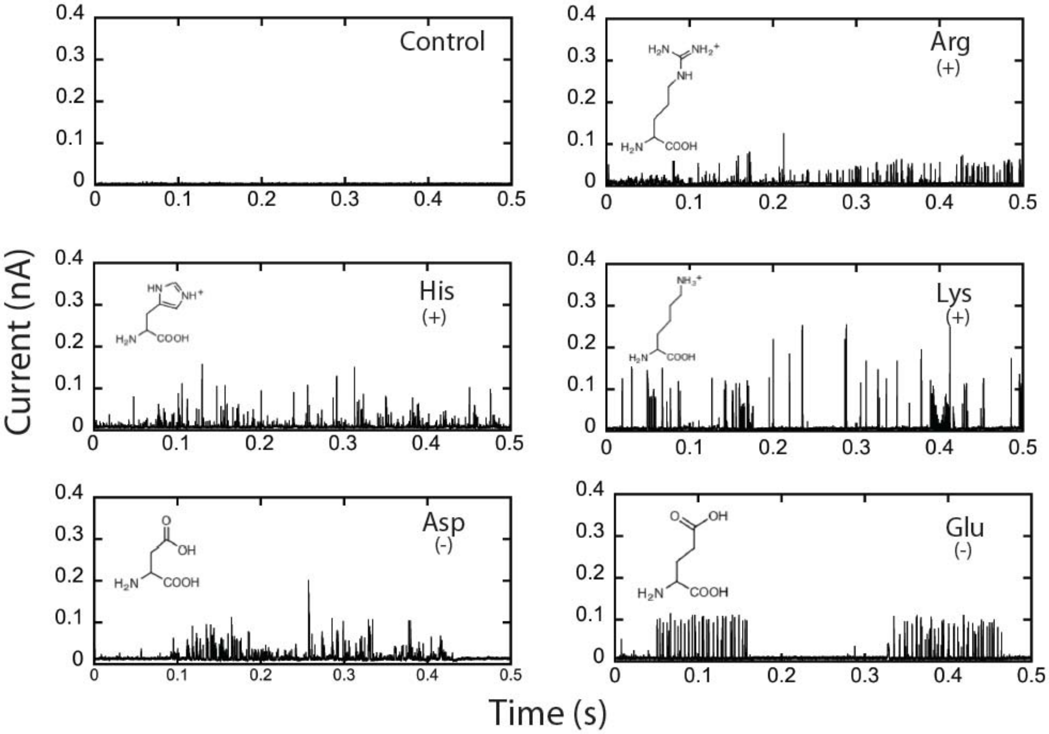


Figure 3.18 RT current traces for the charged amino acids, tunnel gap set to 4 pA at 0.5V bias with 100µM solutions in 1 mM phosphate buffer, pH 7.4. A trace for buffer alone is shown as the control in the upper left.

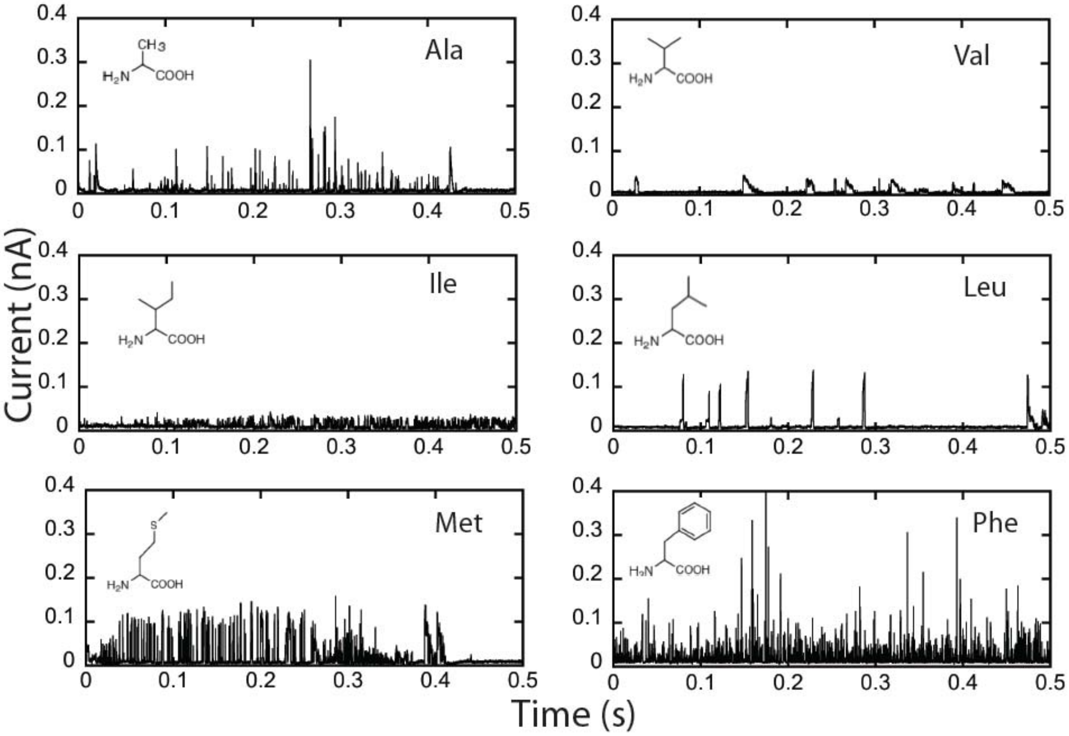


Figure 3.19 RT current traces for the hydrophobic amino acids (excluding tyrosine and tryptophan). Tunnel gap set to 4 pA at 0.5V bias with 100µM solutions in 1 mM phosphate buffer.

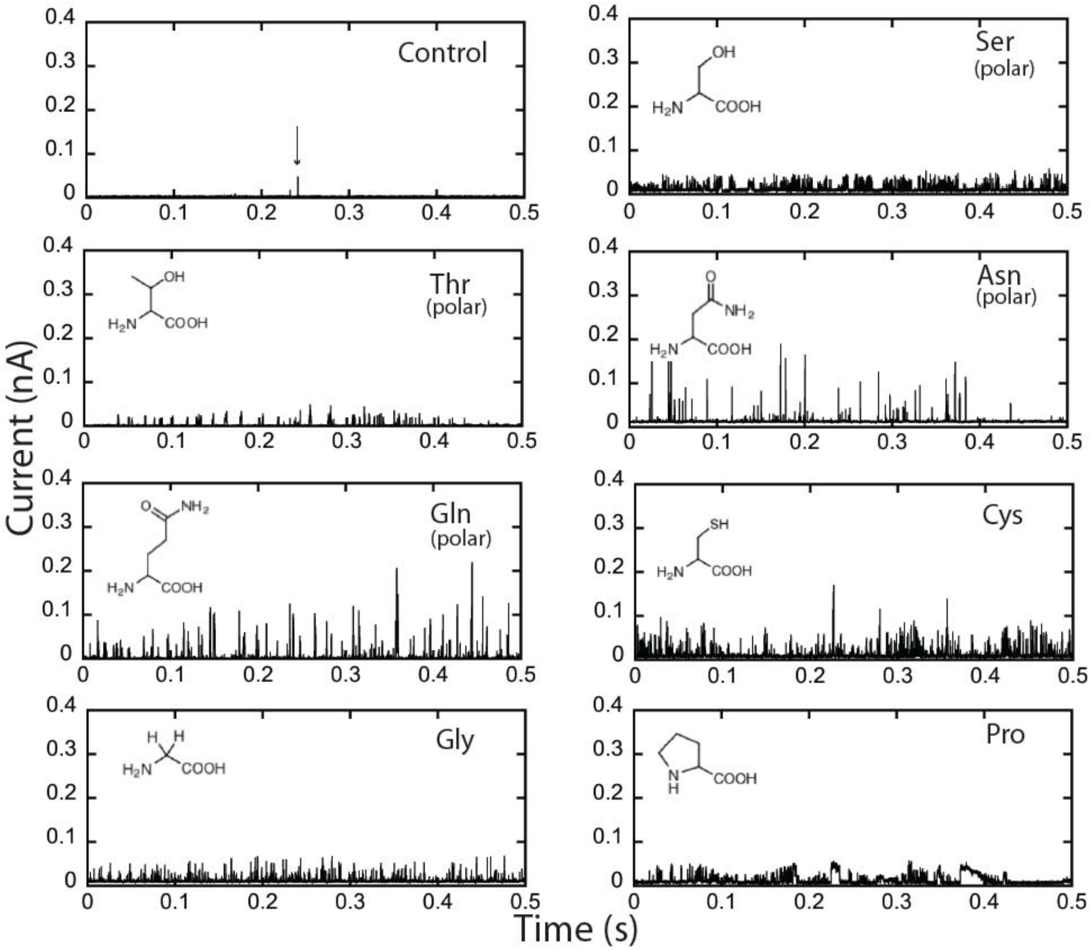


Figure 3.20 RT current traces for the remaining amino acids (excluding tyrosine and tryptophan). Tunnel gap set to 4 pA at 0.5V bias with 100μM solutions in 1 mM phosphate buffer. A trace for buffer alone is shown as the control in the upper left. The arrow points to a "water" spike.

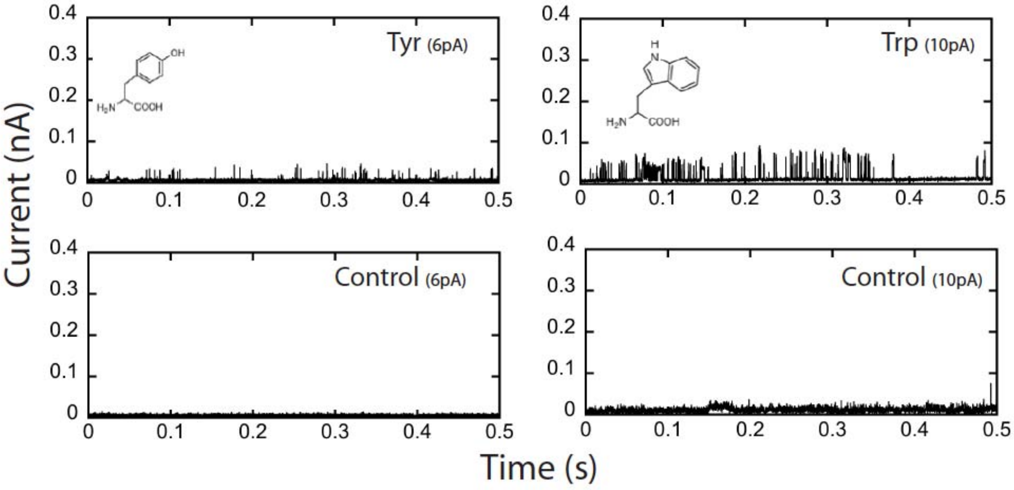


Figure 3.21 RT current traces for tyrosine (left) and tryptophan (right). Tunnel gap was set to 6 pA at 0.5V bias (for tyrosine) and 10 pA at 0.5V bias (for tryptophan). Control scans in these two tunneling conditions are shown below. Data for100μM solutions in 1 mM phosphate buffer.

Signals of different molecules were quite different, while traditional properties as conductance of molecule, the life-time and frequency of spikes could provide some certain separation. Considering there are 20 amino acids in total, the distributions or joint-distributions of these parameters would be so overlapped that the accuracy of separation could be quite low. To solve this problem, we kept traditional parameters as conductance, life-time, etc. and applied Fast Fourier transform (FFT) on the raw cluster data to transform the signal from time-domain to frequency-domain. A lot more information about the signals were extracted from the raw data and finally we ended up with 161 starting features about the raw signals (Table 3.5).

Table 3.5: 161 starting features used in the signal analysis. Details of their calculation are given in Chang et al [47].

Feature Number	Feature Name	Feature Description
1	Max Amplitude	Maximum current at the peak
2	Average Amplitude	Average of all the spike
3	Top Average Figure 4d	average of peak above half maximum
4	Spike Width	full width at half maximum
5	Roughness	standard deviation of the spike above half maximum height
6	Total FFT Power	Square root of the sum of power spectrum
7	FFT L	Average of three points within the first frequency band (0.9, 1.8, 2.7 kHz)
8	FFT M	Average of three points within the middle frequency band (9.3, 10.2, 11.1 kHz)
9	FFT H	Average of three points within the highest frequency band (23.2, 24.1, 25 kHz)
10	High Low Ratio	Ratio of FFT amplitude in the 22.3-25 kHz band to that in the 0- 2.7 kHz band
11	Spike Frequency	Number of peaks per millisecond over a window of 4096 samples
12	Odd FFT Components	Sum of all odd frequencies from the non downsampled FFT
13	Even FFT Components	Sum of all even frequencies from the non downsampled FFT
14	Odd Even Ratio	Ratio of the odd to the even FFT sums
15-23	Spike FFT Components (1-9)	Downsampled FFT spectrum
24	Spikes In Cluster	Number of peaks in the cluster
25	Cluster Peak Frequency	Number of peaks in cluster divided by ms length of cluster
26	Cluster Average Amplitude	Average amplitude of all cluster peaks
27	Cluster Top Amplitude	Average amplitude of all peaks above half maximum
28	Cluster Width	Cluster time in ms
29	Cluster Roughness	std deviation of whole cluster signal
30	Cluster Max Amplitude	average of the max of all the spikes in cluster
31	Cluster Total FFT Power	square root of the sum of the power spectrum
32	Cluster FFT Low	Average of three points within the first frequency band (0.136, 0.273, 0.410 kHz)

33	Cluster FFT Medium	Average of three points within the middle frequency band (12.710, 12.847, 12.983 kHz)
34	Cluster FFT High	Average of three points within the highest frequency band (24.726, 24.863, 25 kHz)
35-95	Cluster FFT Components (1-61)	Downsampled FFT spectrum of cluster
96-99	Cluster Frequency Location of Maximum Peaks (1-4)	Frequency of the 4 dominant peaks in the spectrum, ordered by the height of the peaks
100-161	Cluster Cepstrum (1-61)	Spectrum of the power spectrum of the cluster, downsampled to 61 points

Not all the parameters would be used because most of them were highly correlated. Based on the characters of signals of analytes, a correlation matrix was usually calculated and then the highly correlated parameters were reduced.

To demonstrate the applications of RT, we chose to distinguish a modified amino acid, sarcosine (or N-methylglycine (mGLY)), which is a potential cancer marker [48], from glycine (GLY), two enantiomers (L- and D-asparagine, ASN) and two isobaric amino acids (Leucine (LEU)) and isoleucine (ILE)). In addition, we were going to demonstrate the identification of all seven analytes, which was an essential step to develop peptide sequencing technique.

As shown in Figure 3.18-Figure 3.21, signals were shown up in cluster. We believed each cluster was corresponding one single molecule binding for three reasons: first the duration time of each cluster was on the order of 0.2 s (Figure 3.22), which was comparable to the time for which hydrogen-bonded complexes remain bound in a nanogap [49-51]. Second, signals within clusters were much more strongly correlated than signals from different clusters. And the third reason was from a mixture solution measurement, which would be shown in section 3.4.4. The colour-coded spikes shows each cluster represent one analyte or the other.

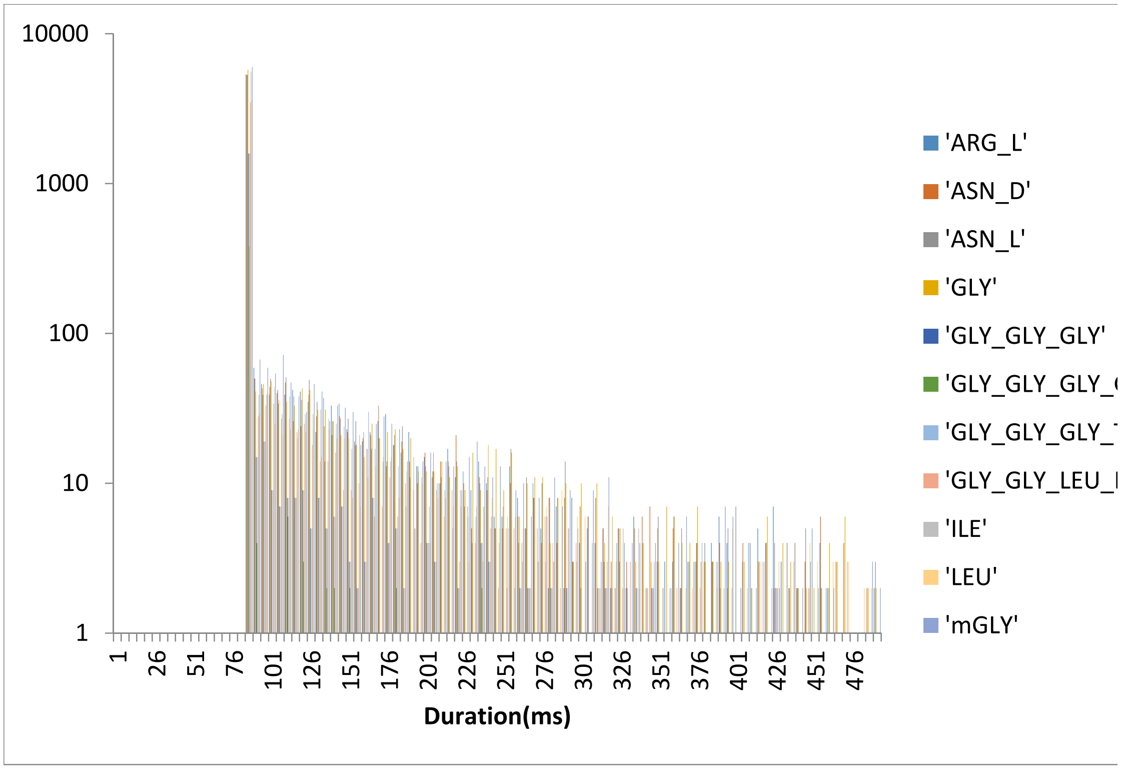


Figure 3.22: Distribution of cluster durations in ms for all 7 amino acids and four peptides. The 1/e time for the distributions is about 200 ms. The spike near 80 mS is an artifact of the data analysis which placed all short clusters into this first bin.

Figure 3.23 shows the data analysis results for LEU and mGLY. As shown in Figure 3.23(a), the traditional parameter average cluster amplitude for these two molecules have a big overlap. After running FFT analysis on the clusters, two signal features' distribution Cluster FFT (8.6-9 kHz) and Cluster FFT (22.6-23 kHz) shows some kind of separation. After plotting the two features' distribution in 2-D plot, a correct call over 95% can be achieved.

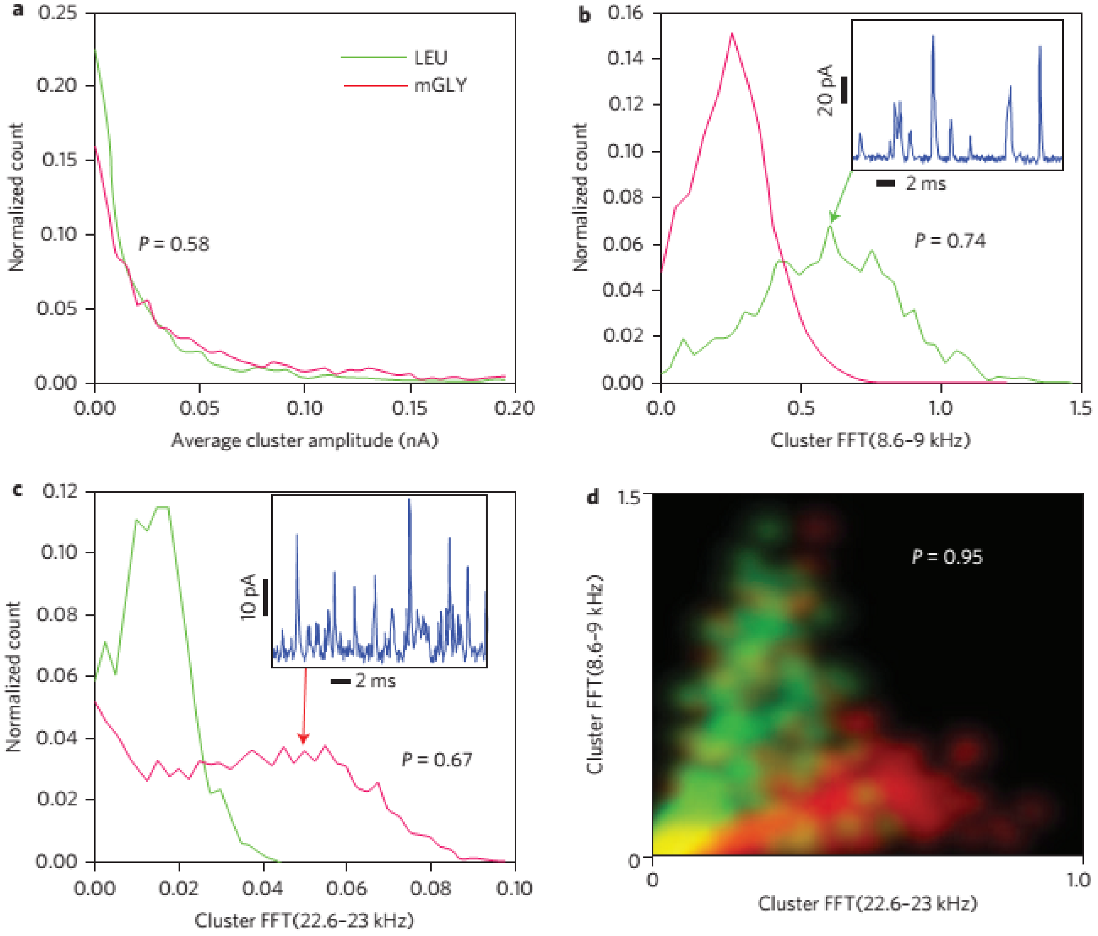


Figure 3.23 Signal feature identify analytes (LEU and mGLY). (a) Peak amplitude distributions for LEU and mGLY showing exponential decay with mostly overlapped. (b)&(c) Particular Fourier component distributions, which shows some discrimination. (d) Two-dimensional plot of probability density as a function of two FFT feature values. The colour scale shows mGLY data points as red and LEU points as green. Yellow regions are ambiguous calls. Based on this separation, we have a correct call 95% of the time.

We applied the same analysis method on the three chemically similar pairs of analytes (Figure 3.24 (a)-(i)). Even though the six features are quite overlapped for each pair (Figure 3.24 (a) (b) (d) (e) (g) (h)), the joint plot of the density plot of features gave a rather high accuracy to distinguish these pairs. It is worth mention that the separation of stereoisomers. We suggest the local-chiral adsorption geometry on surfaces caused different characters of their signal [52]. Our modeling shows a different bonding pattern for L-ASN (Figure 3.7(a)) and D-ASN (Figure 3.25) as well.

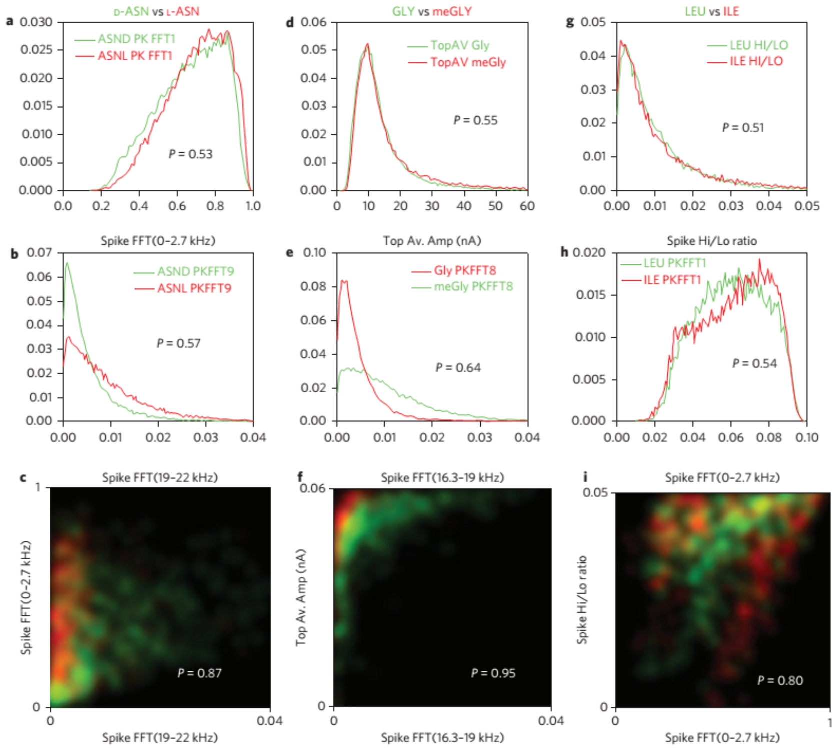


Figure 3.24 Closely related pairs of analytes can be significantly separated (>80%) using just two signal features together. (a)-(c) L-ASN and D-ASN. (d)-(f) GLY and mGLY. (g)-(i) LEU and ILE

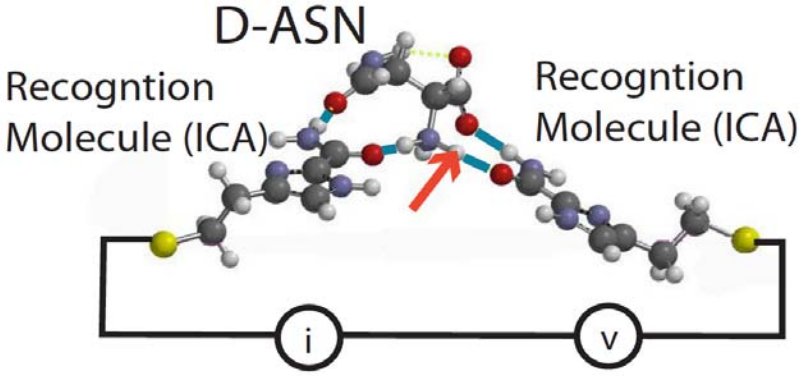


Figure 3.25 Binding of D-ASN in a RT junction. The arrow shows the molecular dipole, tilted relative to that for L-ASN. The structures shown here and in Figure 3.7(a) were calculated using DFT (B3LTY/6-311++G (2df, 2p), Spartan '10 Software) with the distance between the two sulfur atoms constrained to be 2 nm.

So far, we have not taken advantage of the most of the features list in Table 3.5. Because tens of features would be included in the analysis even after reduction of highly correlated parameters. A machine-learning algorithm named Support Vector Machine (SVM) to analysis multitudinous parameters and identify molecules is used. I will give detail description of SVM in the last chapter. Here is a brief introduction of mechanism of how SVM works.

Suppose there are two molecules signals, each has thousands of clusters. After FFT, each spike carries N feature values. A small pool for each molecules' features is chose randomly from whole data set with their corresponding label (the molecule that each spike corresponding to). A training algorithm is selected to construct a hyperplane (N-1 dimension) to classify data sets. So the N-dimension hyperspace is separated into two hyperspaces, each of which is corresponding to a specific molecule. The data points lay on the hyperplane are called support vectors. The remainder of the data is tested to which hyperspace it belongs to and compared with what it should belong to. So, an accuracy of correct call is build. In our experiment, we want to identify seven molecules, so the hyperspace was divided into seven sub-hyperspaces. And each signal spike is assigned to the amino acid corresponding to the highest confidence level.

3.4.3 Reproducibility of the SVM Analysis

A total of 30,000 data spikes (corresponding to 3,000 cluster) were collected for each of the seven analytes. After FFT analysis, each spike and the cluster that contains it were characterized by values of 161 signal features (Table 3.5). A correlation matrix was built and 40 features were removed because of high correlation with other parameters. The remaining 121 parameters were reduced to 106 after removing the parameters that vary from run to run, which were caused by slightly change of geometry of tip, STM device or other experimental artefacts. Noise spike (1-15% of the total data, varying from run to run) were eliminated by training the SVM to find signals common to all seven analytes. The last stage of noise filtering were adjusted by varying the soft margin (broadening of the partition boundaries) of the SVM parameters. Increasing the soft margin improved accuracy at the cost of rejecting more signals. A small subset of data were chosen randomly to train SVM and the rest of the data were tested to calculate the accuracy. Table 3.6 shows the correct call accuracy of each analyte, even based on single spike, SVM gave an accuracy over 95% percent for each analyte. The accuracy could be further improved if multiple spikes (3 or 5) were taken from different clusters. Details of SVM analysis are given in the last chapter and the supporting material of this paper.

Table 3.6 Accuracy with which any one of seven pure analytes is identified from the total pool of data taken from all seven pure samples using 52 signal features together*.

Number of spikes	ARG	ASN_$_D$	ASN_$_L$	GLY	ILE	LEU	mGLY
1	99.14	94.99	96.99	97.24	96.87	94.36	96.45
3	98.77	99.62	99.99	99.62	99.99	99.55	99.99
5	99.99	99.99	99.99	99.99	99.99	99.99	99.99

* Results in the first row are based on a single spike. The subsequent rows are based on a majority vote using three and five spikes taken from different signal clusters. These results were obtained with the noise filter soft margin set to reject ~70% of the data spikes

3.4.4 Analyzing Mixture of Analytes

I have shown how the pure L-ASN and D-ASN molecules could be identified by SVM previous. One interesting question is if we can tell them apart in mixture solution. Three different molecule ratio solutions were prepared: L-ASN:D-ASN = 1:1, 2:1, 3:1. We used pure L/D-ASN training data and applied support vectors analysis on mixtures' signal to assign each cluster with which molecules' signal it came from. Figure 3.26 (a) shows a typical trace of 1:1 ratio L/D-ASN mixture. The spikes colored purple were assigned to D-ASN and yellow to L-ASN by SVM analysis. Some black spikes were not recognized. From the red cluster tag marked under Figure 3.26 (a), we could tell for all the spikes in one cluster was only assigned to one molecule. This was summarized for 556 clusters in Figure 3.26 (b). Red dots were corresponding to raw cluster and we could tell less than 1% of the cluster contains more than one type of spike. After a common noise filter was applied, only 400 pure clusters remains (blue point). And all the spikes in the remaining clusters were only assigned to one type of spike D- or L-ASN. When increasing molecule ratios of L/D-ASN, although we could not get the exact same molecule ratio either by counting peaks or by counting clusters (Figure 4.26 (c)) because of molecules' different binding strength with ICA, a trend was obvious. The relationship could be used to conduct quantitative analysis after adjusted by the slope.

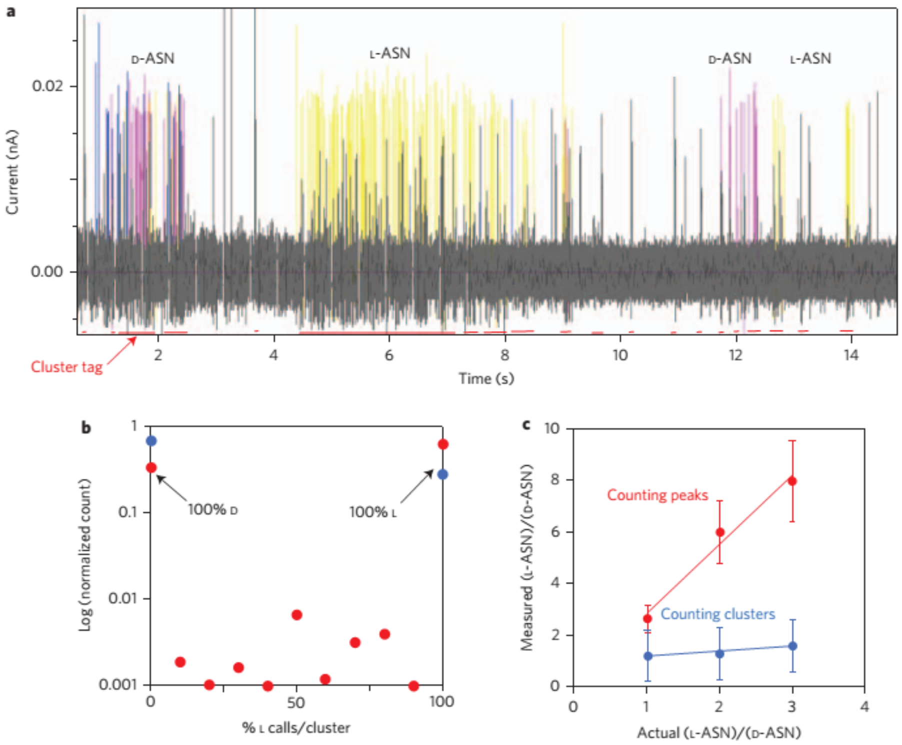


Figure 3.26 A mixture produces alternating cluster signals as different molecules diffuse into and out of the gap. (a) Signals obtained with a 1:1 mixture of L-ASN and D-ASN. The SVM assignment are coded purple (D-ASN) and yellow (L-ASN). Black spikes are unassigned. (b) Each cluster (red tags) contains only one type of signal, as shown statistically. (c) Quantification of the L/D ratio using trained on pure samples.

3.4.5 Signals for Peptides

To investigate if we can also get RT signals from peptide, two short peptides Glycine-Glycine-Glycine-Glycine (GGGG) and Glycine-Glycine-Leucine-Leucine (GGLL) were tested under the same condition as the amino acids using. Typical signals are shown in Figure 3.27. We also trained SVM by pure Glycine and Leucine data to see if it can recognize the signals from these short peptides. While SVM could not find any signals from GGGG or GGLL from either Glycine or Leucine. This result was not a surprise for us because to form the peptide, the amine group and carboxylic acid group react and form a peptide bond. Since the groups that bond with ICA are not present anymore for each amino acid component in peptide any longer, it is reasonable that SVM could not recognize these peptides.

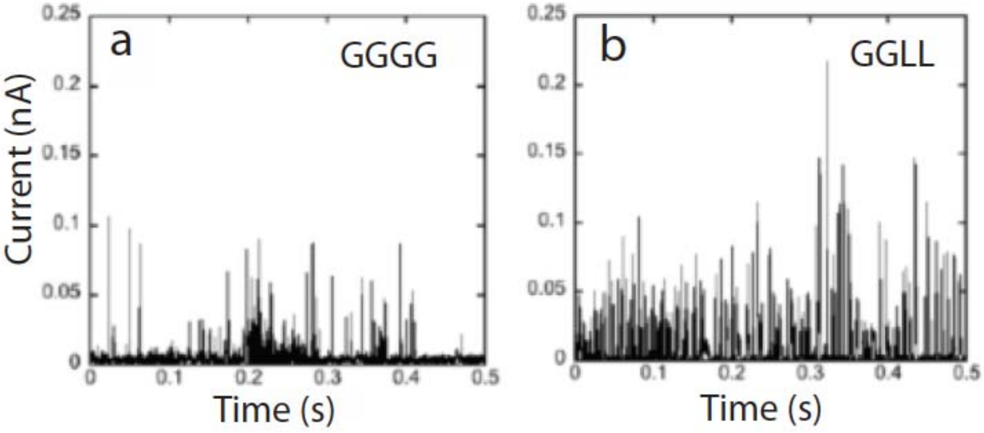


Figure 3.27 Representative RT signals for (a) GGGG and (b) GGLL

Thus, to do peptide sequencing using RT method, a similar method as Edman degradation should be applied on peptide to cliff amino acids first and then RT method could be applied to identify the components individually. To demonstrate this protocol, we developed a simple flow-through measurement. The liquid cell on the stage were drilled two holes and connected with a tube. One syringe was used to inject fluid into the cell, while the other one would pump solution out of the cell. Initially the cell was filled with PB buffer, then we injected Glycine (Gly) solution into it. After a short period of data collection, we injected PB buffer to clean the cell and followed by Phenylalanine (Phe) solution. The signal trace is shown in Figure 3.29. Obvious different types of spikes were observed and SVM gave a 96% true positive rate on the training.

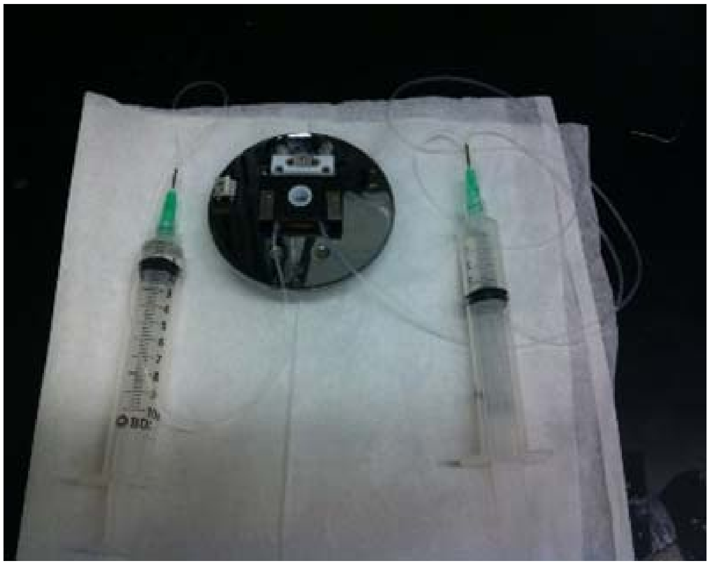


Figure 3.28 A flow-through STM stage

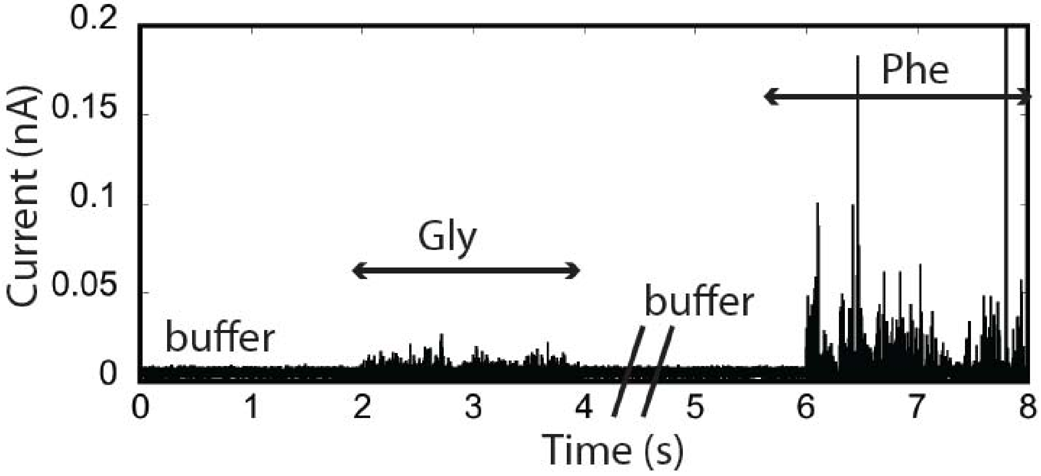


Figure 3.29 Reading analytes sequentially. An SVM trained on this data run separated the two analytes correctly at a 96% true positive rate.

The rest of amino acids were also analyzed in the same way.

In summary, we were able to do amino acids identification, but to do peptide sequencing, peptides have to be break apart into amino acids and amino acids have to pass the RT gap sequentially. In the next chapter, I will illustrate the design of several commercializable peptide sequencing devices.

Chapter 4

Commercializable Peptide Sequencing Devices

In this chapter, I am going to describe an invention of apparatuses that could run peptide sequencing based on the previous results. This invention has been filed a patent to World Intellectual Property Organization (WIPO) and entitled, "SYSTEMS, APPARATUSES AND METHODS FOR READING AN AMINO ACID SEQUENCE" Stuart Lindsay, Peiming Zhang, Yanan Zhao, WO2013116509 A1, Aug 8, 2013.

4.1 The First Apparatus

Let's start from Figure 4.1, which shows a tunneling gap for detecting analytes that bind transiently to recognition molecules (ICA showed in the figure), and generating current spike (I) when a bias (V) is applied across the gap. As illustrated, two facing electrodes 1,2 are separated by a gap 3 about 2.5 nm, but which could be as small as about 1 nm and as large as about 4 nm. Each electrode are functionalize with reagent 4 (like ICA molecule) that is chemically-bonded to the electrodes, to form non-covalent bonds with the target molecule. The materials of electrodes could be platinum, palladium, gold or silver, all of which could be modified by the regent molecules. In some embodiments, the entire system is immersed in an aqueous electrolyte 5.

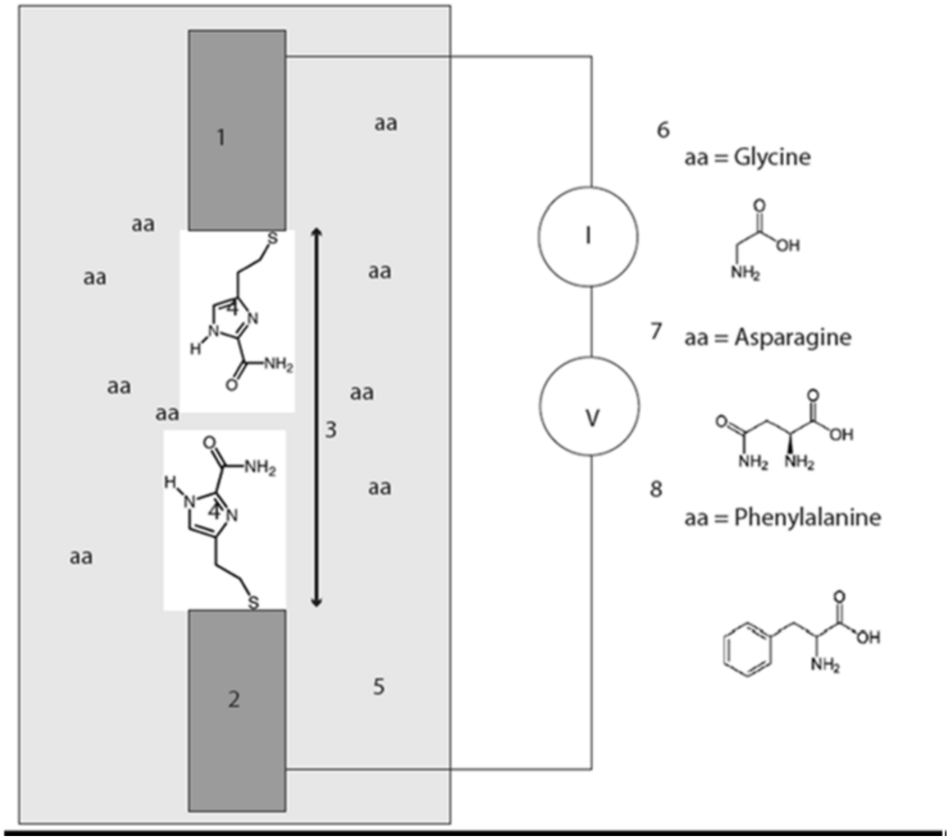


Figure 4.1 A tunneling gap apparatus for detecting analytes.

4.2 The Second Apparatus

As shown in the last section of previous chapter, the first apparatus might be able to readout only amino acids instead of a whole chain of peptide. Another drawback of the first apparatus is we cannot control the motion of target molecules. So we designed a second apparatus that is much more sophisticated shown in Figure 4.2.

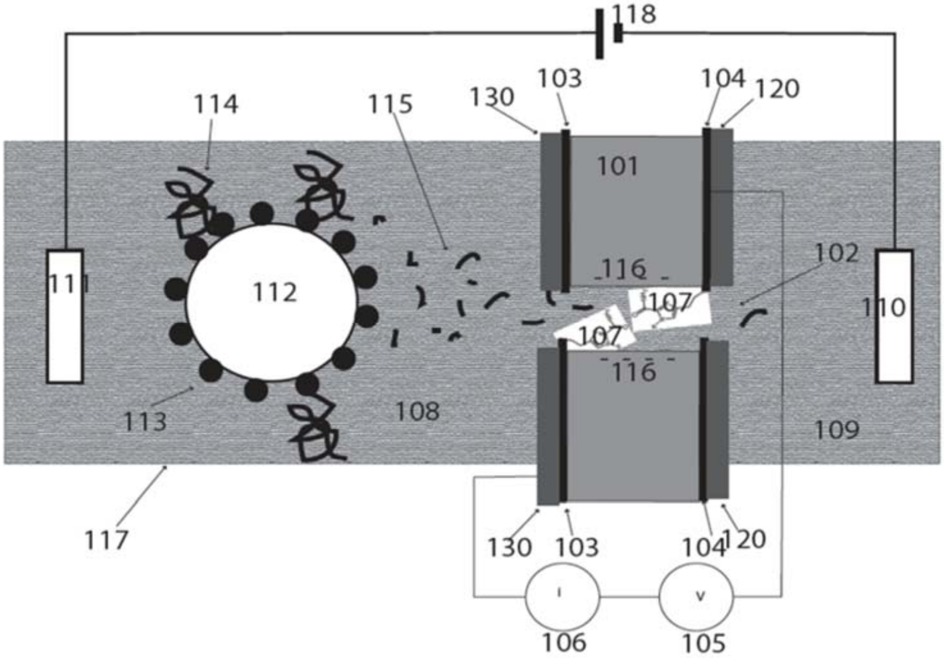


Figure 4.2 An apparatus for protein, peptide, amino acid, and/or modified amino acid identification and/or sequencing

As shown in Figure 4.2, a silicon or silicon nitride membrane 130 spans a channel 117 containing an electrolyte solution 108, 109. The membrane, typically about 10 to about 100 nm thickness, divides the channel into two chambers, a cis chamber 108 and a trans chamber 109, in fluid communication with each other only by means of a nanopore 102 that may be drilled through the membrane (and all the other layers 103, 104, 101, 120) by means of, for example, an electron beam. Ideally, the nanopore may be between about 1 and about 5nm in diameter at the point where it passes through the electrodes. The pore passes through planar electrodes 103, 104 separated by the dielectric layer 101 (for example). The dielectric layer may be between about 1 nm and about 4 nm thickness between the electrodes and may be made by depositing silicon nitride, silicon dioxide or hafnium oxide. The electrodes may be made by depositing between about 1 nm to about 5 nm of silver, gold, palladium or platinum onto the top of the first membrane 130 (for example). The dielectric layer 101 may then be deposited, followed by deposition of a second metallic layer, 104. This “top” electrode may then be covered by a second layer of dielectric 120. Fabrication of the assembly according to some embodiments is completed by drilling the nanopore. One or more of the electrodes (in two electrode embodiments, e.g., both; in plurality of electrode embodiments, two or more) may then be functionalized by immersing the assembly in a solution of the recognition molecules 107.

Each chamber of the apparatus may be filled with an electrolyte such as KCl or $NaClO_3$ (for example), in concentrations that range from about 1 mM to about 1M, and to which may be added Mg ions and/or ATP as required to activate enzymes 113 attached to one or more beads 112 which are fixed in turn to the walls of the channel 117 in close proximity to the nanopore 102. Alternatively, the enzyme is directly attached to the channel in close proximity to the nanopore. The enzymes may then be any one of the well-known proteases which include carboxypeptidase, aminopeptidase, trypsin, chymotrypsin, pepsin, papain or elastase (for example). Since each of these proteases is somewhat selective in their hydrolysis of peptide bonds, the beads may ideally contain a mixture of these enzymes. In some embodiments, the bead may be functionalized with proteosomes, assemblies that sequentially degrade proteins into their component amino acids. The isolated protein or peptide 114 to be identified/sequenced may also bound to the bead. In some embodiments, digestion of the protein may be initiated by binding the protein to the bead in the absence of Mg or other chemicals needed to initiate digestion, the bead placed in the channel, and then digestion is initiated by the addition of Mg or other chemicals (needed to initiate digestion). The resulting small fragments, small peptides, or optimally amino acids 115, may be released into the solution. A bias 118 may be applied between the cis and trans chambers by means of the reference electrodes 111, 110, for example. In the figure, this is shown with the negative electrode in the trans chamber 109. Accordingly, this draws positively charged amino acids and small peptides through the nanopore where they bind transiently to the recognition molecules 107 generating current spikes recorded by a transconductance amplifier 106. A bias 105 of between about 0.1V and about 1V is preferably applied between the electrodes 103, 104. The walls of an oxide layer (such as the surfaces of silicon, silicon nitride or silicon dioxide or hafnium oxide) in aqueous electrolyte are negatively charged 116 owing to the accumulation of OH groups on the surface. This negative charge causes an electro-osmotic flow of water towards the negative electrode 115. Thus, neutral amino acids that diffuse into the vicinity of the nanopore may be swept through it by the electroosmotic flow of the water. All the neutral and positively charged amino acid residues can be read by the same nanopore apparatus, according to some embodiments. Accordingly, the amino acid mixture will generate a characteristic set of tunneling signals that will allow the protein or peptide to be identified. To the extent that the digestion of the target proteins is sequential and synchronized, the train of signals could also be used to deduce protein sequence, while the sequence of small peptides may be read directly from the time series of signals generated as each amino acid in the chain passes through the tunnel gap.

Another improvement made to apparatus Figure 4.2 is to add another nanogap in the system (Figure 4.3). Different sign of bias are applied to different nanogap, so the system is able to readout amino acid no matter what kind of charge it carries. An illustration of the system is given in Figure 4.4

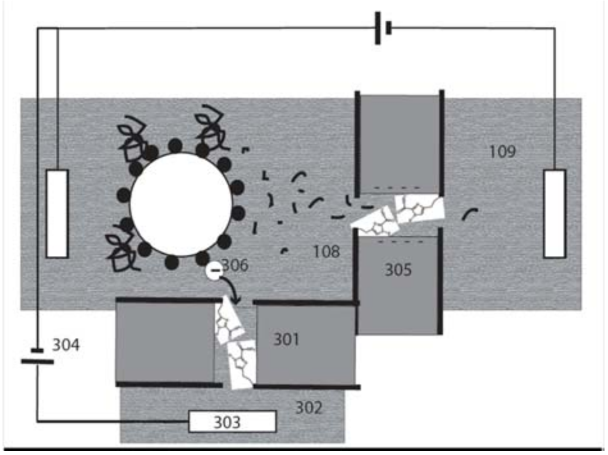


Figure 4.3 An improved peptide sequencing device

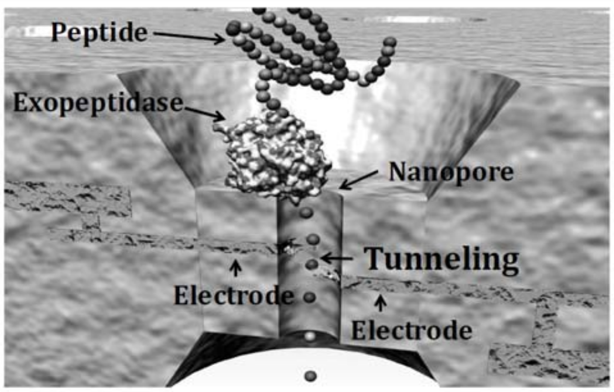
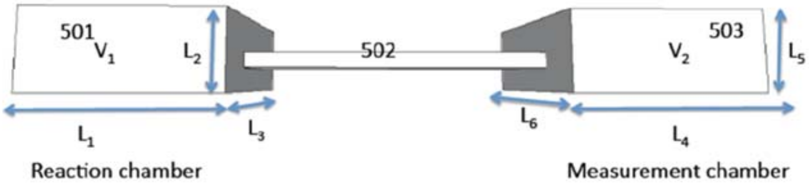
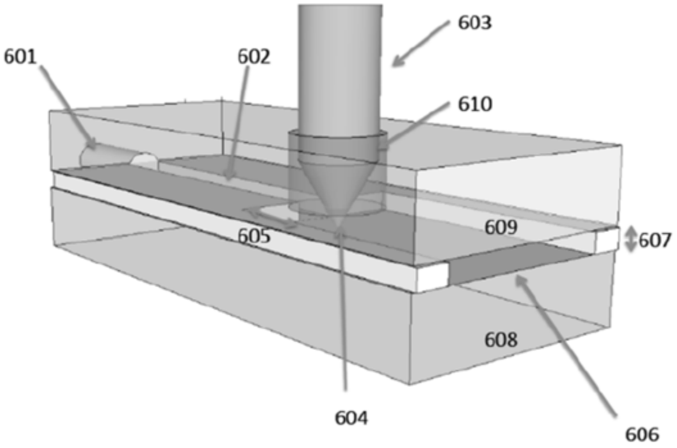


Figure 4.4 An illustration of peptide sequencing by enzymatically cleaving amino acids from a terminal of protein.

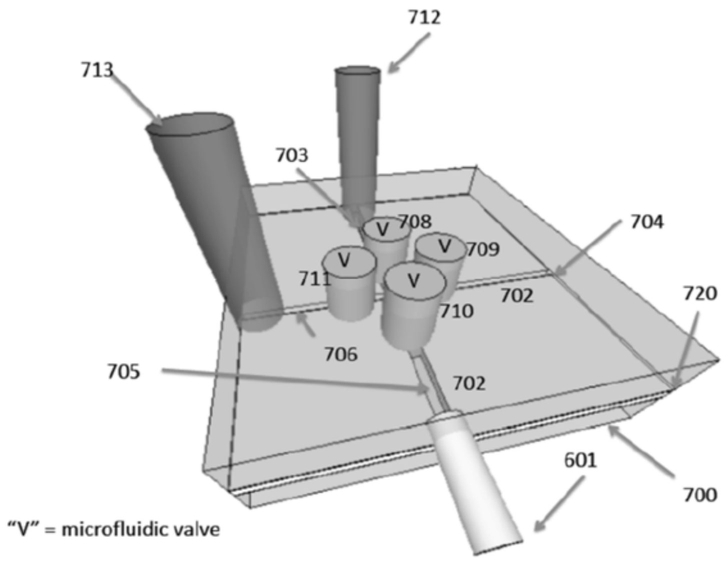
4.3 An Apparatus Combined with a 1-D STM.



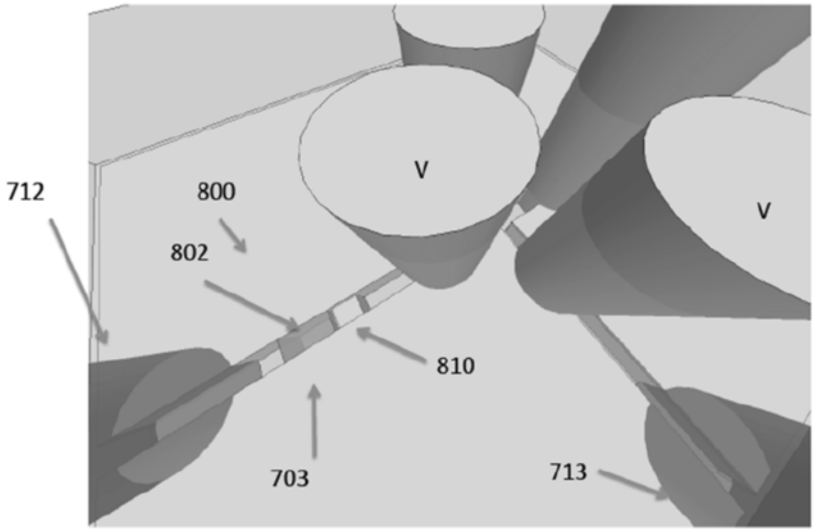
(a)



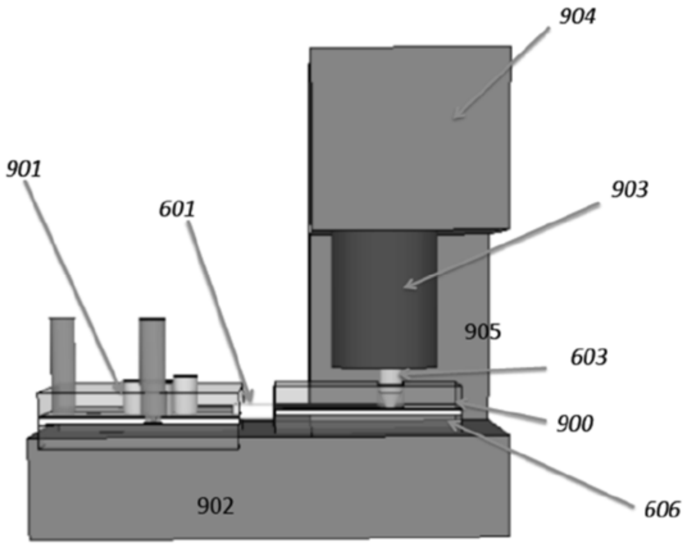
(b)


712
713
703
V
708
V
709
V
711
V
704
702
710
720
706
705
702
601
700
"V" = microfluidic valve

(c)


V
V
712
800
802
810
703
713

(d)



(e)

Figure 4.5 A 1-D STM embed sequencing apparatus

As shown in Figure 4.5, an example of a measurement cell according to some embodiments, with nano-liter volume, is shown and described herein. Accordingly, the aliquot from the reaction chamber is injected via a microfluidic coupling channel 601 into a channel 602 that has a width in the region of 100 microns and a similar height, set by the etching depth 607 into the glass substrate 608. The lower surface 606 is coated with a film of suitable metal such as Pd (for example), which is contacted via an electrical connection (not shown). A metal probe 603, made of a similar metal (usually Pd) having a sharp exposed apex 604 and otherwise covered with insulation (as described by Tuchband et al. [68]) is positioned through a hole 610 in an upper plate 609 that is bonded over the lower assembly to seal the channel 602. The probe apex 604 is placed within tunneling distance of the substrate 606 by means of a scanning tunneling microscope controller as is well known in the art. The apex of the probe 604 is placed a short distance 605 away from the end of the channel 602. If this distance is about 100 microns or less, then a nano-liter of sample passed down the channel 602 floods the tunnel junction between the probe apex 604 and the substrate 606. Providing that the hydrostatic pushing pressure is removed once the aliquot is delivered, it will remain in the vicinity of the junction for at least 30 seconds (as described earlier). Please note, the dimensions described above are merely exemplary, as the scale is readily reducible from about 100 micron distances down to about 20 micron distances or less.

In operation (for example), an aliquot to be measured is passed to the tunneling junction, measured by means of its characteristic tunneling signal, and then flushed out by passing clean buffer into the measurement cell via the channel 601/602. The cycle is then repeated as needed.

Another advantage of using a microfluidic channel with a scanning-tunneling microscopy (STM) apparatus, according to some embodiments, is that an adjustable tunnel gap can be included for added versatility and ease of manufacture. Such embodiments also may allow identification of species that require a different tunneling gap. For example, in the case of tyrosine and tryptophan, a “null” read in the standard tunneling conditions (0.5V, 4 pA) can be followed rapidly by a sample taken at a smaller gap (6 or 10 pA current) to see if the aliquot of sample that produced no signal was tyrosine or tryptophan. In that way, all 20 amino acids can be identified.

A view of an exemplary nano-liter scale reaction system, according to some embodiments, is shown in Figure 4.5(c). Two microfluidic channels 702 are etched in the form of a cross in a ction chamber. Reagents are pushed into this chamber by the pump/dispenser 712. Reaction products are released by opening the valves 708 and 710, where they pass into the output channel 705 to the coupler to the measurement cell 601. For rinse cycles of the reaction cell, valves 708 and 709 are opened, and buffer dispensed by the pump 712. For rinse cycles of the measurement cell, valves 711 and 710 are opened and clean buffer flowed from the dispenser 713 out via the coupler 601.

Figure 4.5(d) illustrates a view down into the reaction chamber (703 in Figure 4.5(e)) according to some embodiments. A weir 810 is formed in the reaction channel 800 by masking part of the channel during etching leaving a space of a few microns at the top of the channel to pass fluid and small molecules but trapping reaction column beads. A nano-liter or so of column material (such as C12 [76]) is flowed into the channel where it piles up against the weir to form the reaction cell 802. A heater (not shown) is placed underneath this reaction cell so that the standard cycle of base, acid and heat can be applied to complete the Edman reaction.

The overall assembly, according to some embodiments, is shown in Figure 4.5(e). The measurement chamber 900 and reaction chamber 901 are coupled via the coupler 601 and may sit on a common base 902 that contains controls for the reaction chip and its heater. The probe 603 is held above the substrate 606 by an actuator 903. This consists of a coarse mechanical approach based on a stepper motor in series with a fine adjustment that uses a piezoelectric material as is well known in the art. This may be coupled to the base 902 via a rigid support 905 and a top housing 904 that holds the actuator and control electronics for the tunnel gap.

Prior to sequencing, both the probe 603 and the substrate 606 may be functionalized with an adaptor molecule as previously described. Buffered electrolyte may then be pushed from a reservoir on the microfluidic chip 901 to fill the reading channel 602 and the probe advanced to the desired tunneling current by means of the actuator 903, and subsequently controlled by the servo circuit. A set point at a bias of 0.5V between the probe and substrate is a current of 4 pA (for example).

The first Edman degradation may then be carried out, and control valves on the microfluidic chip may be set so as to release the first amino acid aliquot released from the peptide into the reading channel up to the point where it is preferably centered on the junction region between the probe apex 604 and the substrate 606. RT signals may then be acquired for a period of between about 0.1 to about 1s and recorded for subsequent analysis. Valves on the microfluidic chip may then be set to flush the sample out of the junction area. Then a next cycle of Edman chemistry may be carried out, releasing the next amino acid to the junction for the next read. This cycle may be repeated out to the limit of reliable cleaving by Edman chemistry, which may be (for example) up to about 50 amino acid residues. In some embodiments, mixtures of amino acids can be analyzed by this technique with each read producing hundreds or thousands of signal peaks, each one of which can be assigned to a particular amino acid, enabling analysis of the data well into a number of amino acids beyond 50 amino acid residues, based on the identity of the last reliably called residue. This is because contaminants from residues that may fail to have cleaved in earlier reactions will have been previously recorded, so their presence in a subsequent reaction can be recognized as an artifact of the chemistry.

Chapter 5

DNA Sequencing by RT Method

In addition to protein sequencing, another attractive application area of RT method is for DNA sequencing [77] [87] [90]. Unlike so many amino acids consisting protein, there are only four nucleobase (guanine, adenine, thymine and cytosine) consisting DNA. Based on our previous result, RT method is quite promising for DNA sequencing [79] [83 - 86]. In this chapter, I will give a brief summary about what we have done so far for DNA sequencing. This result has been published on ***Nature Nanotechnology*** [46]. I helped run some experiments. The details of experiment and results could be obtained online or through my colleagues' theses [58] [59].

Figure 5.1 shows the energy levels calculated for four different DNA bases and DNA base paring by DFT method [61][62]. The calculation showed four nucleobase having significantly different LUMO levels. Recall the Landau formula described in Chapter 1. Different nucleobase should give different conductance if measured by RT method. A similar setup as the one used for identifying Amino Acids was used. One exception was that "Reader" molecule was benzamide (Figure 3.3 (a)) instead of ICA. And substrate and probe was Au instead of Pd for peptide sequencing.

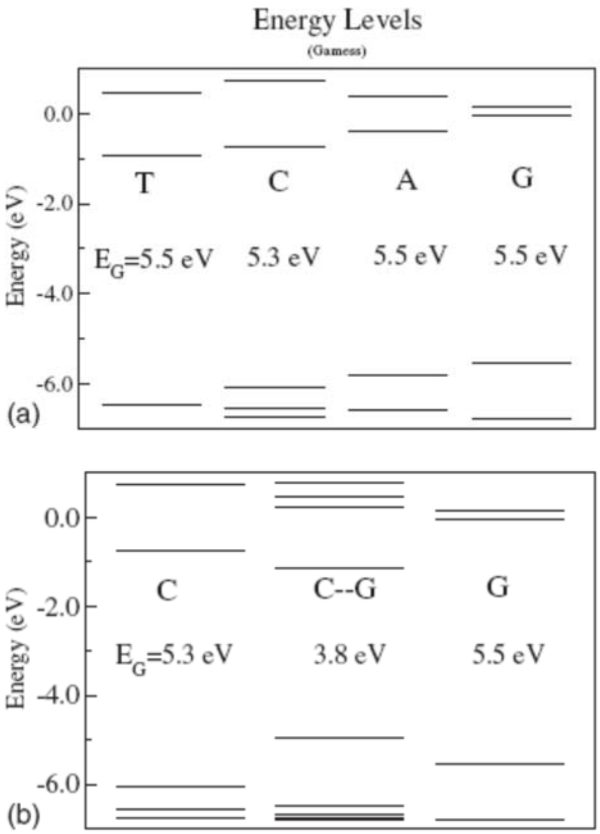


Figure 5.1 The energy level calculated for four different DNA bases and DNA base pairing by DFT method.

Figure 5.2 (a) shows our proposed dNMP bonding with reagent molecules. Figure 5.2 (c)-(f) shows typical signal trace for dAMP, dCMP, d^{m}CMP and dGMP. We could not get signals from dTMP. After switched to ICA as the "Reader" molecule, we got signals for dTMP as well [60].

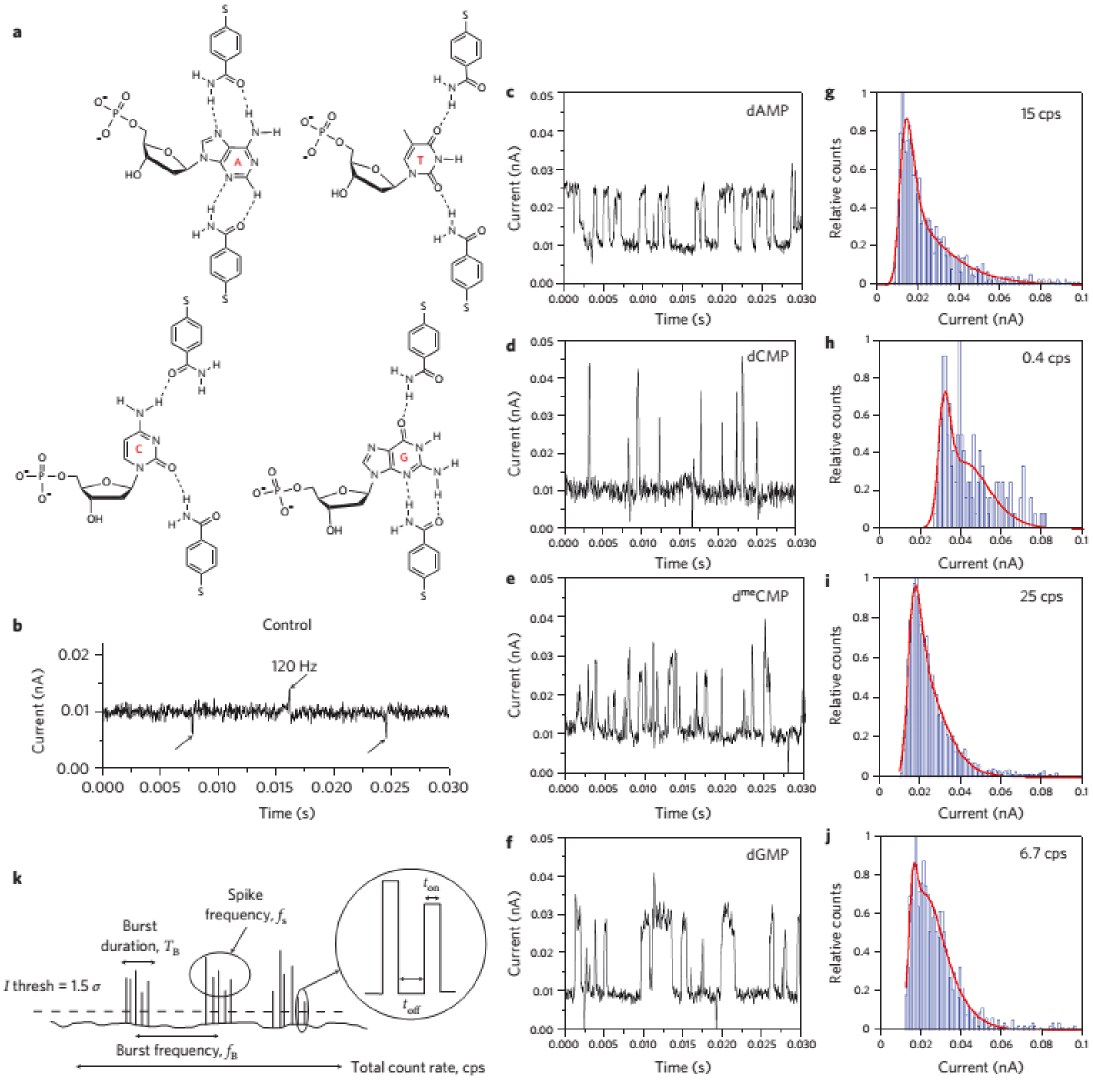
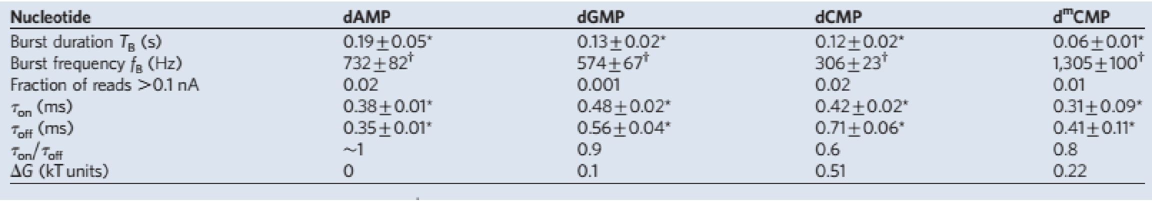


Figure 5.2 Tunneling signals from nucleotides trapped in a functionalized tunnel gap. (a), Proposed hydrogen-bonding modes for all bases. (b), In PB alone, a 20 pS gap (i=10 pA, V = +0.5V) gave a signal free of features, except for some a.c. coupled line noise. (c)-(f), Characteristic current spikes produced when nucleotides dAMP, dCMP, d^{m}CMP and dGMP were introduced. (g)-(j), Corresponding distribution of pulse heights. k, Definition of the parameters used to characterize the tunneling signals.

The tunneling current distribution for each molecule are shown from Figure 5.2 (g)-(j). Other parameters describing the signal character are summarized in Table 5.1. The definition of burst duration, burst frequency, fraction of reads, on time of spikes and off time are given in Figure 5.2 (k).

Table 5.1 Nucleotide tunneling noise characteristic



Nucleotide	dAMP	dGMP	dCMP	d^mCMP
Burst duration T_B (s)	$0.19 \pm 0.05^*$	$0.13 \pm 0.02^*$	$0.12 \pm 0.02^*$	$0.06 \pm 0.01^*$
Burst frequency f_B (Hz)	$732 \pm 82^\dagger$	$574 \pm 67^\dagger$	$306 \pm 23^\dagger$	$1{,}305 \pm 100^\dagger$
Fraction of reads >0.1 nA	0.02	0.001	0.02	0.01
τ_{on} (ms)	$0.38 \pm 0.01^*$	$0.48 \pm 0.02^*$	$0.42 \pm 0.02^*$	$0.31 \pm 0.09^*$
τ_{off} (ms)	$0.35 \pm 0.01^*$	$0.56 \pm 0.04^*$	$0.71 \pm 0.06^*$	$0.41 \pm 0.11^*$
τ_{on}/τ_{off}	~1	0.9	0.6	0.8
ΔG (kT units)	0	0.1	0.51	0.22

Based on the distributions, we found that dAMP signals are well separated from dCMP signals, and dmCMP signals were well separated from dCMP signals in spike amplitude and in life-time distribution of their signals. So, we tested some oligomers (Figure 5.3). One of the amazing results is shown in Figure 5.3(g) and Figure 5.3(9), both of which show two level signals and current distributions also proved that the presents of two nucleotides.

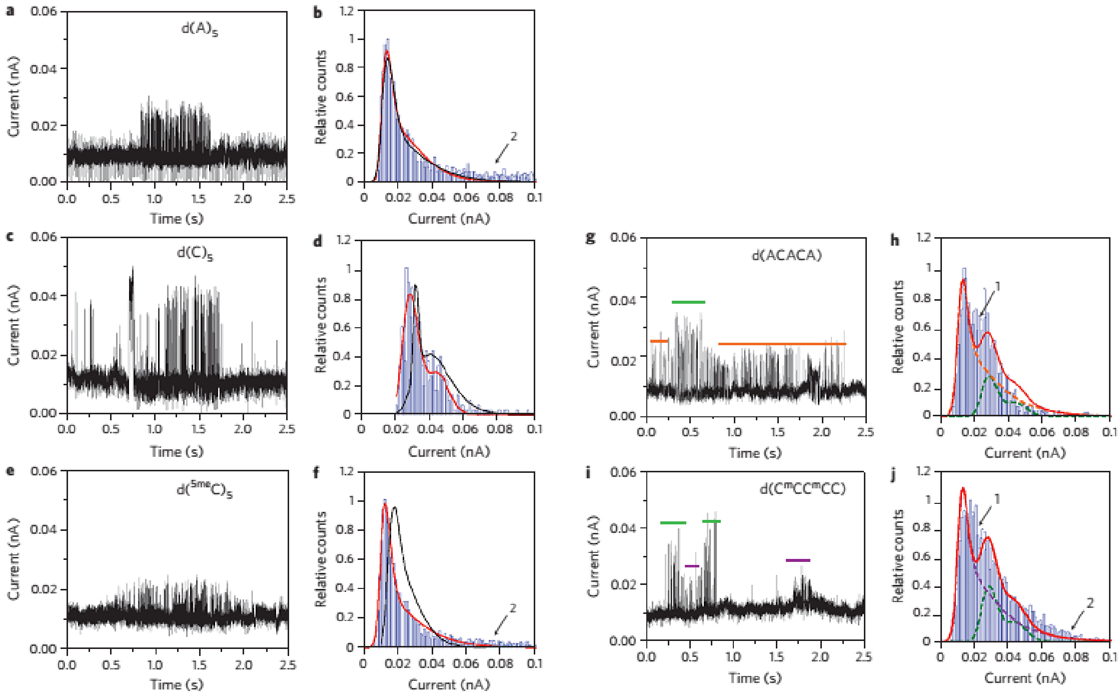


Figure 5.3 Tunneling signal distributions from Oligomers resemble those of the constituent nucleotides. (a)-(f), Current traces for d(A)$_5$ (a), d(c)$_5$ (c) and d(^{m}C)$_5$ (e), with the corresponding distributions shown in (b), (d) and (f). (g)-(j), Current traces from mixed oligomers d(ACACA) (g) and d(C^{m}CC^{m}CC) (i) with corresponding current distributions.

Does replacing the regent molecule to ICA and applying SVM analysis could result in real-time DNA sequencing? So far, the answer is possible. In the following chapters, I will give a brief introduction of a revolutionary device that we are still working on based on our current promising STM setup. And we will get much closer to the answer of the question.

Chapter 6

Application of RT Method in Peptide Phosphorylation Detection, Monosaccharide, Disaccharide and Ibuprofen Identification

In this chapter, I will present my presently undergoing projects including peptide phosphorylation detection, disaccharide and ibuprofen identification using RT method. Some of the results are preliminary data and further experiments (such as ESIMS/MS) are need to confirm the bonding of reagent molecule (ICA) and analytes.

6.1 Peptide Phosphorylation Detection

Lots of cellular processes as signal transduction, gene expression, cycle cytoskeletal regulation and apoptosis are controlled by the reversible phosphorylation of proteins [53-55]. Among all the possible phosphorylation on a peptide, the most common and important sites occur on serine, threonine and tyrosine residues. In this section I will show we could use previous introduced peptide sequencing method to detect if there exist phosphorylated amino acids and thus monitor cellular processes. Another approach is that if we are able to detect some unique signals from phosphorylated peptide, then sequencing a whole peptide would not be necessary. So, in our project, we select two short peptides Abltide (EAIYAAPFAKKK) (Figure 6.1) and the tyrosine phosphorylated peptide Abltide, which are related to chronic myelogenous leukemia [56, 57]. I will illustrate our results about peptide phosphorylation detection.

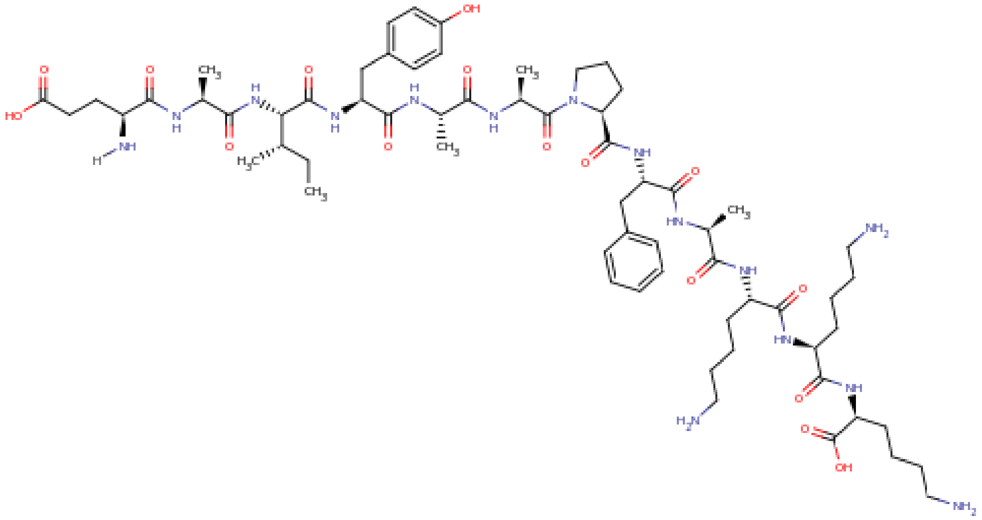


Figure 6.1 Abltide (EAIYAAPFAKKK), consisting of 12 amino acid bases, Glu-Ala-Ile-Tyr-Ala-Ala-Pro-Phe-Ala-Lys-Lys-Lys. The tyrosine phosphorylated peptide happed at the Tyr position.

Before trying to differentiate peptides, we first run test on tyrosine and phosphorylated tyrosine to see if phosphor-group generate distinguished signals.

The structures of Tyrosine and Phosphorylated-tyrosine are shown in Figure 6.2.

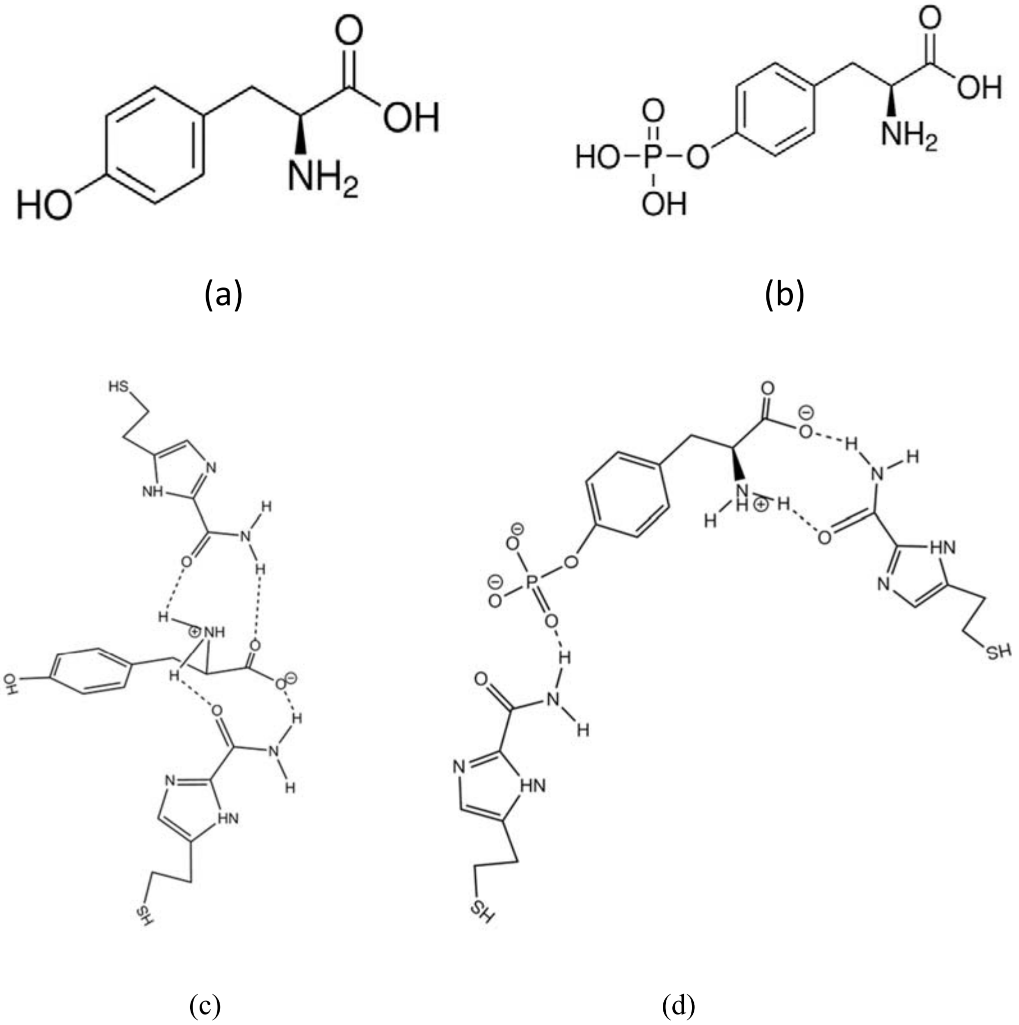


Figure 6.2 Structure of Tyrosine (a) and (b) Phospho-tyrosine (c) proposal bonding with two ICA readers for tyrosine (d) proposal bonding with two ICA readers for phospho-tyrosine

Each molecule was dissolved in 1 mM PB to get 100 μM solution. All the probes fabrication and modification and substrate modification process were the same as the process described in chapter 3. We still used ICA as the "Reader" molecule. One change we made was that the Setpoint of STM was set to 6 pA because we also described that we could not get signals from Tyr at 4 pA as other amino acids can.

Figure 6.3 shows typical raw signals collected by RT method.

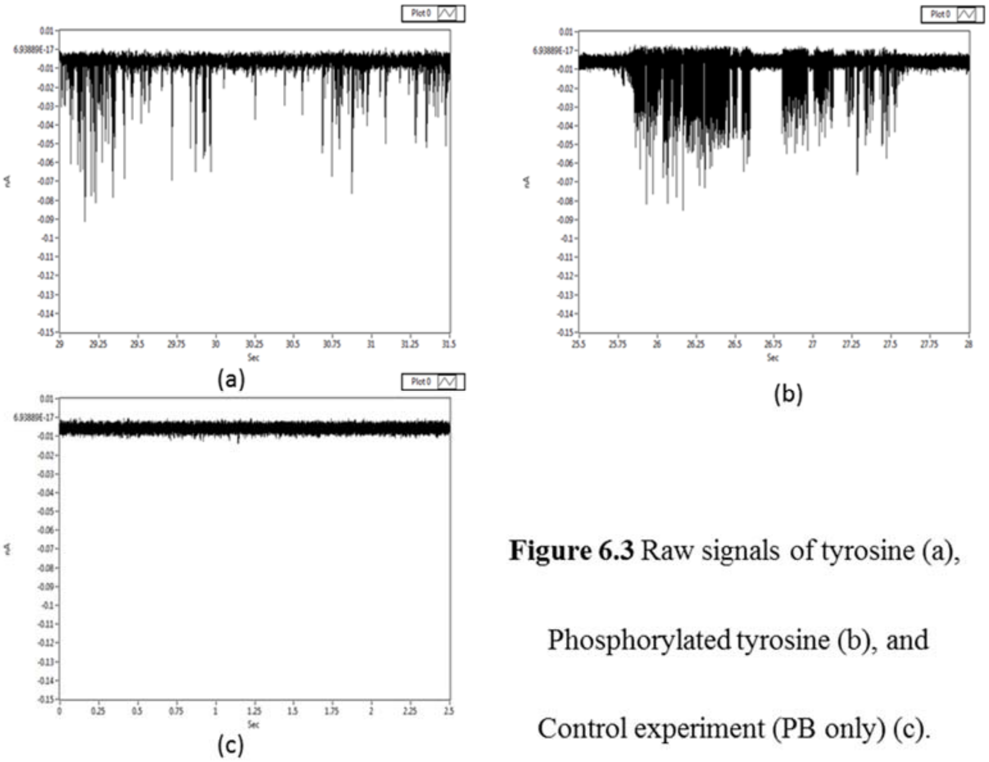


Figure 6.3 Raw signals of tyrosine (a), Phosphorylated tyrosine (b), and Control experiment (PB only) (c).

There were some difference in raw signals visually such as frequency of signals was less for tyrosine compared with phosphorylated tyrosine. After running FFT on the signals, we plotted the Peak_maxAmplitude and Peak_averageAmplitude of tyrosine (Red dot) and phosphorylated-tyrosine (blue dot) spikes. Clearly, two clusters were shown in phosphorylated-tyrosine, one of which almost overlapped with tyrosine data, while the other cluster was unique for phosphorylated-tyrosine. We are still trying to build a model to explain this phenomenon. Currently, DFT calculation was used to build connection models for both measurements Figure 6.2 (c,d). We assume that for Tyrosine, there might be one connection figuration to form H-bonds with two ICA molecules, while for phosphorylated tyrosine, besides the similar connection as tyrosine, another configure would be one ICA molecule forms H-bond with phosphorus in the phosphor-group and the other ICA molecule forms H-bond with amide group. The theoretical calculating is under performing right now. We also employed SVM to run analysis on these two molecules signal data (Figure 6.5).

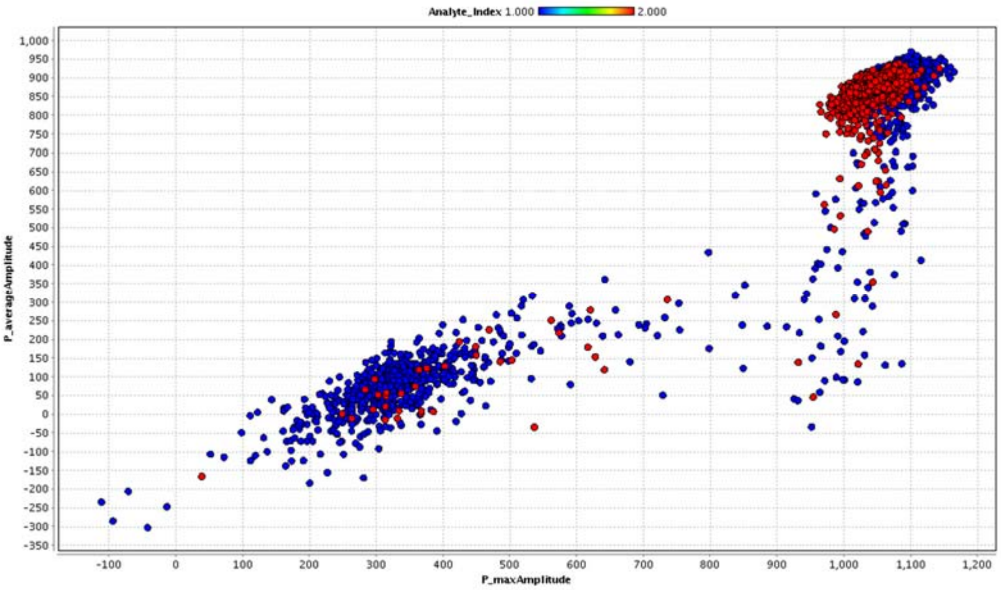


Figure 6.4 Two-dimension plot of Peak_maxAmplitude and Peak_averageAmplitude of tyrosine and phosphorylated-tyrosine spikes. Explanation of parameters can be found at Chapter 3.

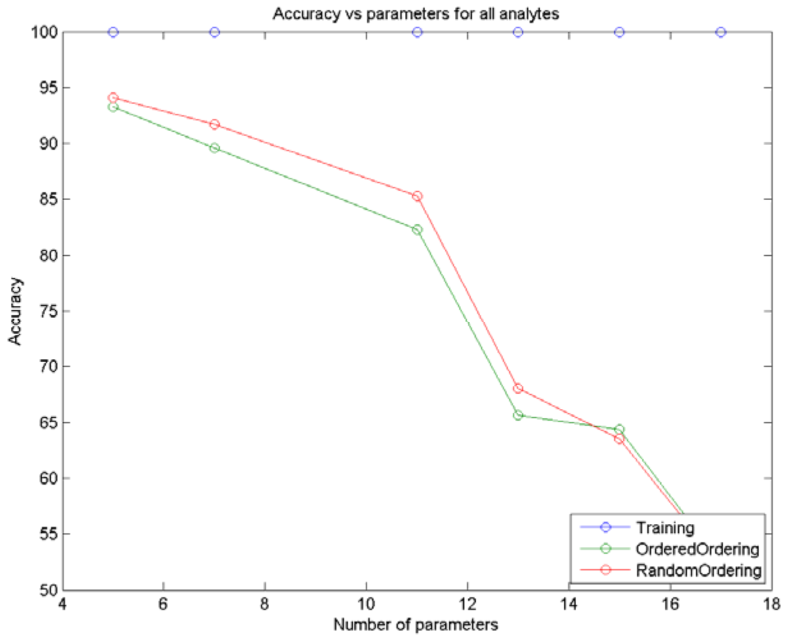


Figure 6.5 SVM analysis on tyrosine and phosphorylated-tyrosine.

Figure 6.5 showed that even using 4 parameters, the accuracy of separation of tyrosine and phosphorylated-tyrosine could reach up to 95%. More analysis are still undergoing. The surprisingly good result for these two molecules enlightened us that if phosphate-group generates so unique spikes, proteolysis would be unnecessary for the two peptides. Considering peptides were much longer than single amino acid, we decreased the setpoint of STM to 2 pA to get a larger gap. Raw signals for these two peptides are shown in Figure 6.6. By using around 30 parameters, the accuracy could reach up to 95%.

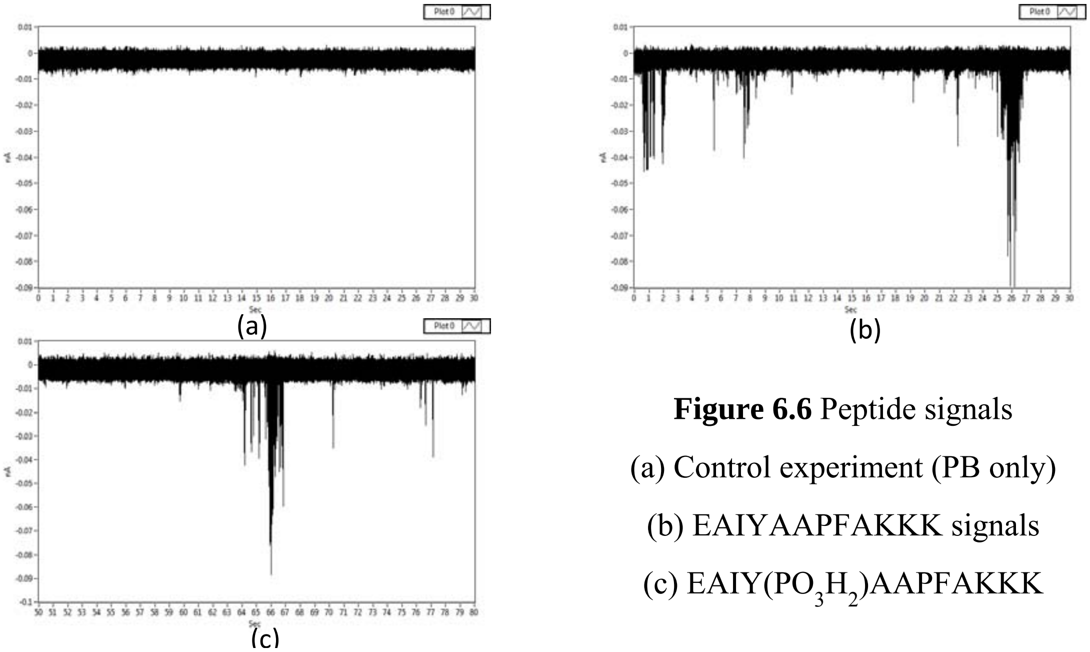


Figure 6.6 Peptide signals

(a) Control experiment (PB only)

(b) EAIYAAPFAKKK signals

(c) EAIY(PO_3H_2)AAPFAKKK

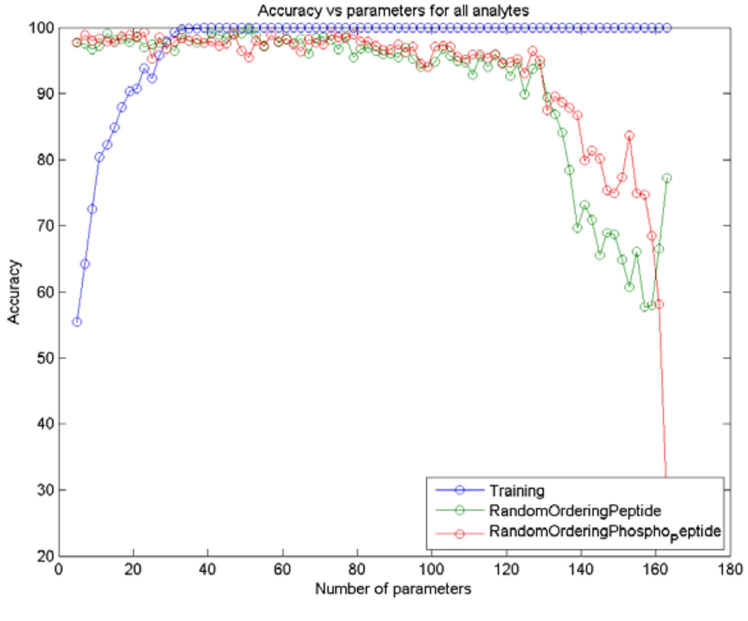
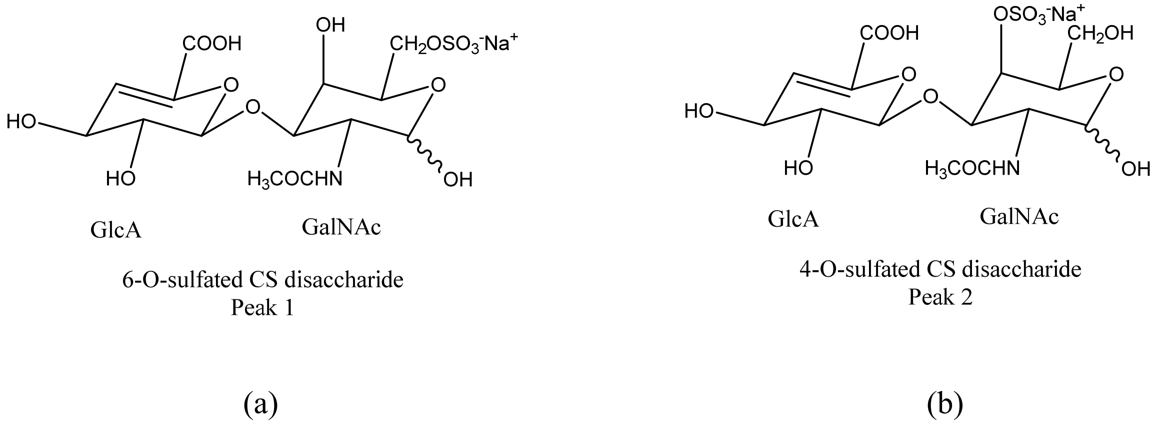


Figure 6.7 Accuracy vs parameters for peptide (EAIYAAPFAKKK) and phosphorylated peptide (EAIY(PO_3H_2)AAPFAKKK). Analyzed by SVM.

6.2 Disaccharide Identification

We have also collaborated with colleagues in biochemistry department and working on the application of differentiation of two glycosaminoglycans disaccharides (Figure 6.8).



6-O-sulfated CS disaccharide
Peak 1

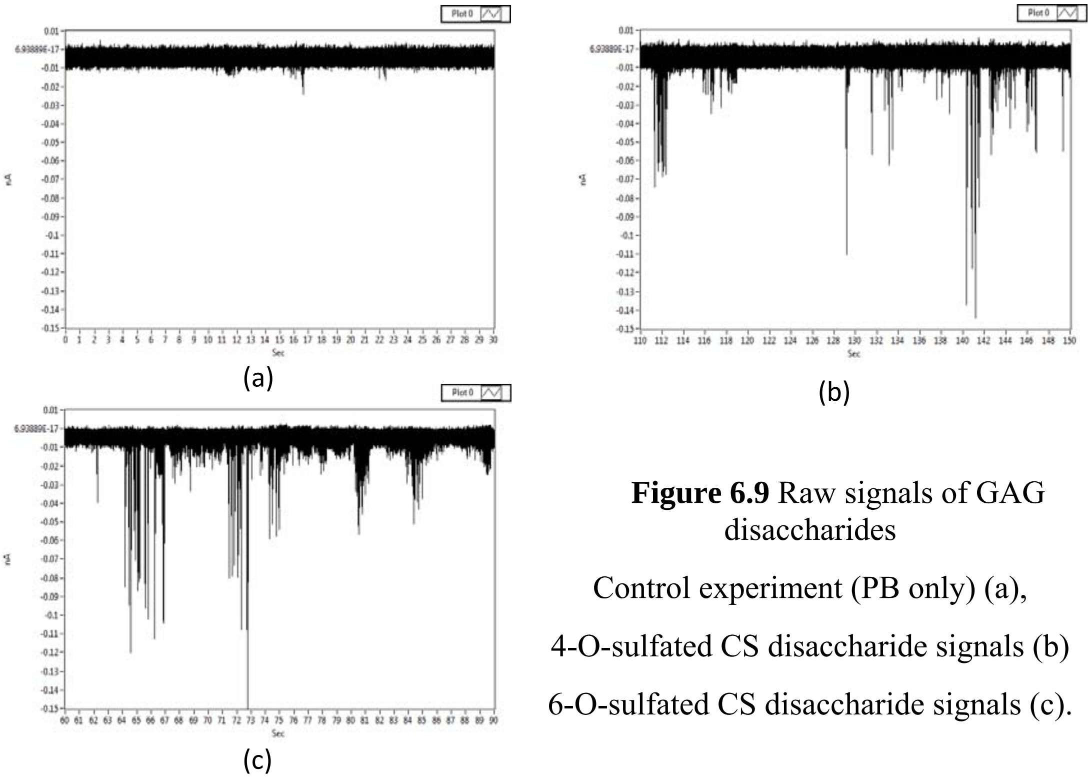
4-O-sulfated CS disaccharide
Peak 2

(a)

(b)

Figure 6.8 Glycosaminoglycans disaccharides. (a) 6-O-sulfated CS disaccharide, (b) 4-O-sulfated CS disaccharide

Standard procedure was followed. Raw signals are shown in Figure 5.9 and SVM analysis of the features of the signal are in Figure 5.10. An accuracy over 90% separation could be achieved.



(a)

(b)

(c)

Figure 6.9 Raw signals of GAG disaccharides

Control experiment (PB only) (a),

4-O-sulfated CS disaccharide signals (b)

6-O-sulfated CS disaccharide signals (c).

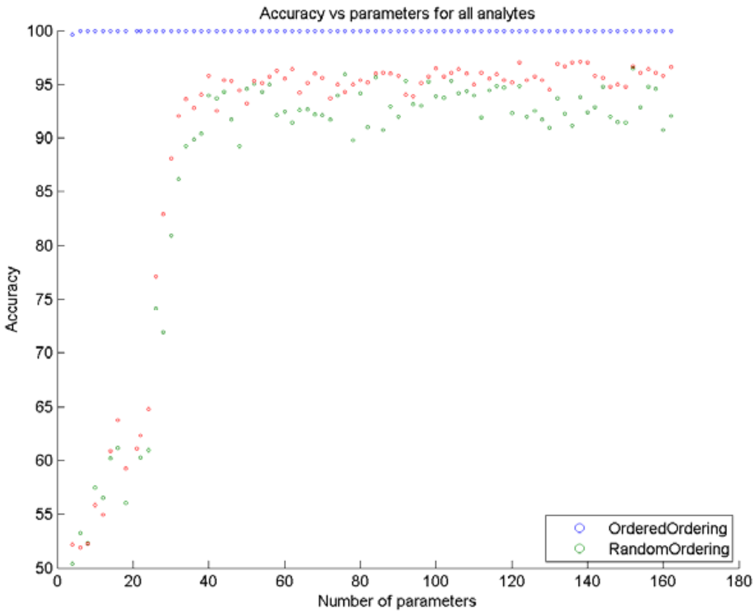


Figure 6.10 Accuracy vs parameters for GAG disaccharides. Analyzed by SVM.

6.3 Ibuprofen Enantiomers Identification

Inspired by the mixture of L/D-ASN measurement, we decided to run some test on ibuprofen enantiomers. There are two types of enantiomers for ibuprofen (Figure 5.11), S-(+)-ibuprofen and R-(-)-ibuprofen. S-ibuprofen is the only found to be the active form both in vitro and in vivo. While ibuprofen products selling in market are actually consist of half amount S-(+)-ibuprofen and half amount R-(-)-ibuprofen. We suggested that if there were a tool to do real time analysis ratio of the mixture during the synthesis, industry might be able to improve product quality significantly.

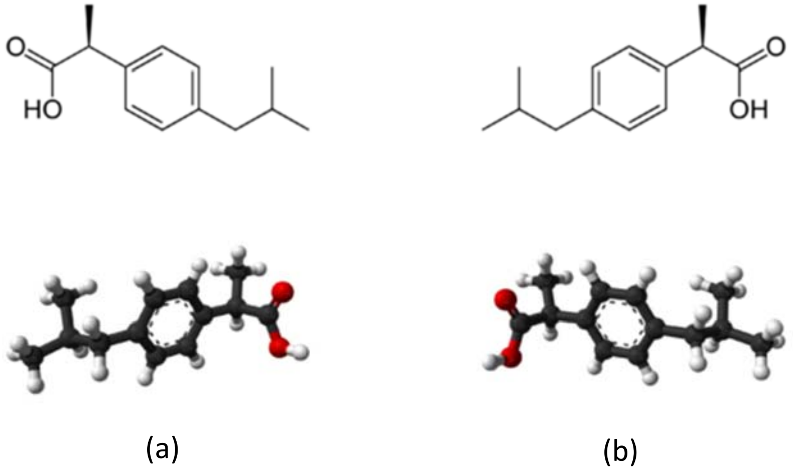
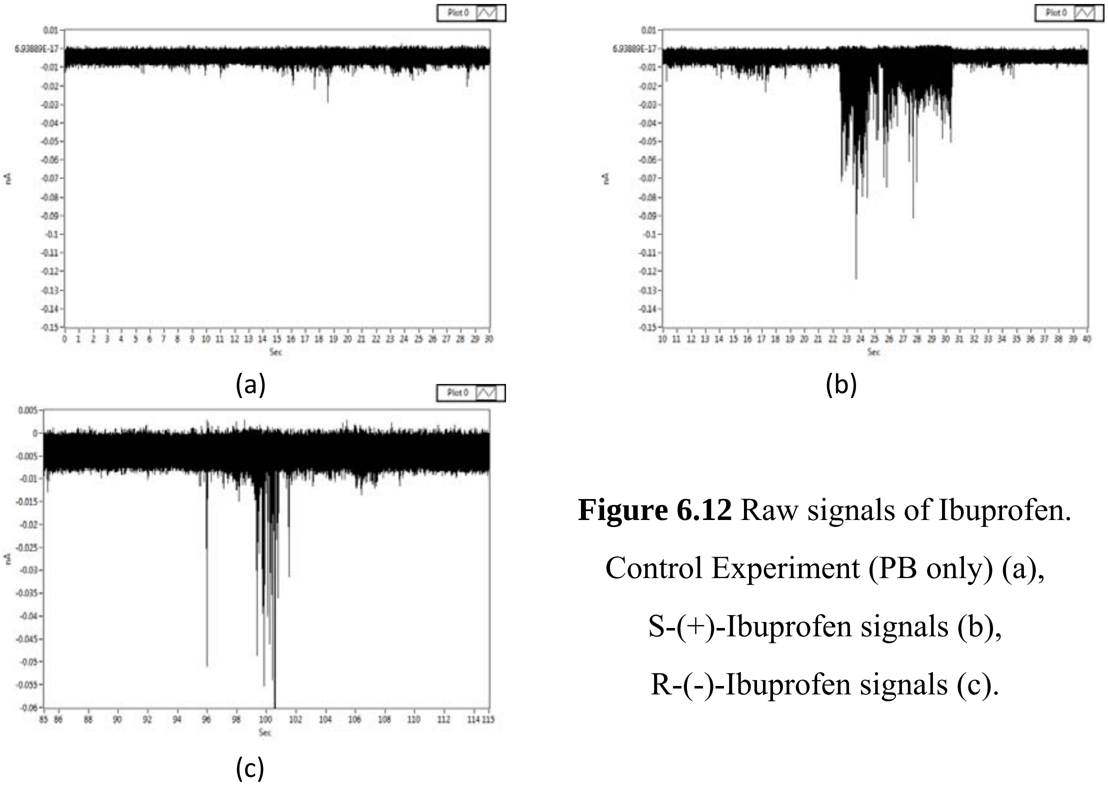


Figure 6.11 Ibuprofen enantiomers. (a) R-Ibuprofen. (b) S-Ibuprofen



(a)

(b)

(c)

Figure 6.12 Raw signals of Ibuprofen. Control Experiment (PB only) (a), S-(+)-Ibuprofen signals (b), R-(-)-Ibuprofen signals (c).

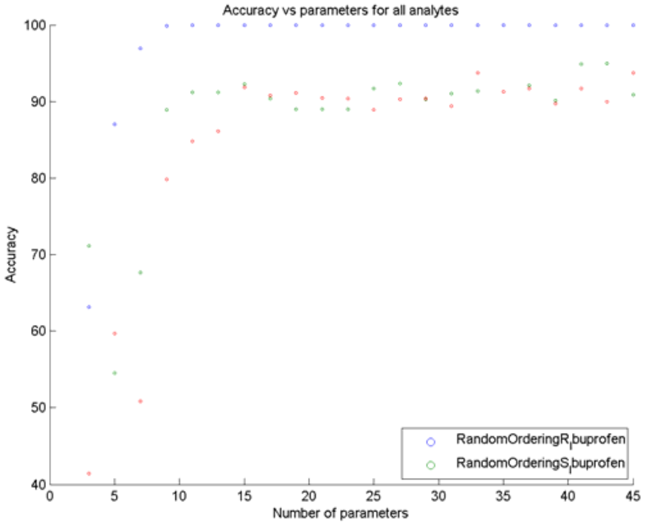


Figure 6.13. Accuracy vs parameters for Ibuprofen. Analyzed by SVM.

Signal data were collected in the same way as before (Figure 6.12) and analysed by SVM (Figure 6.13). The accuracy of separation of these two enantiomers could goes up to 90% using around 15 parameters.

There are many other data mining tools available to verify our SVM results. In the last chapter, I will use Ibuprofen data and a new data mining tool called Decision tree to differentiate these two molecules while I discuss SVM algorithms.

Chapter 7

Ongoing Project-stacked Junction

STM is a robust tool for running RT method. We are able to control the gap size precisely and vary it if needed as shown before when a narrower gap needed for Tyrosine (Tyr) and Tryptophan (Trp) or a wider gap for long peptides identification. But the cost for adjustable gap is that stabilizing the system is rather time consuming. When running RT measurements, we waited 2 hours for its initial stabilization and another 1 hour after analyte added. Another serious drawback is that the system is not ideal designed for flow-through analysis. The flow current could vibrate the probe and since the system is extremely sensitively to the gap size, not only lots of noise spikes would show up but also features of spikes of analyte would change and generate unrecognizable signals.

To solve these problems, we have been developing a new device named stacked junction. A device has a fixed gap size. The size of the gap is around 2.4 nm to perform DNA bases or amino acids analysis. If a larger gap needed such as for peptide measurement, we can increase the thickness of insulating layer that separate two electrodes easily. Besides, the device is also capable of running flow-through experiments without the vibration problem STM has. In general, we are capable of developing specialized devices according to requirements.

A sketch of fabrication process is shown in Figure 7.1.

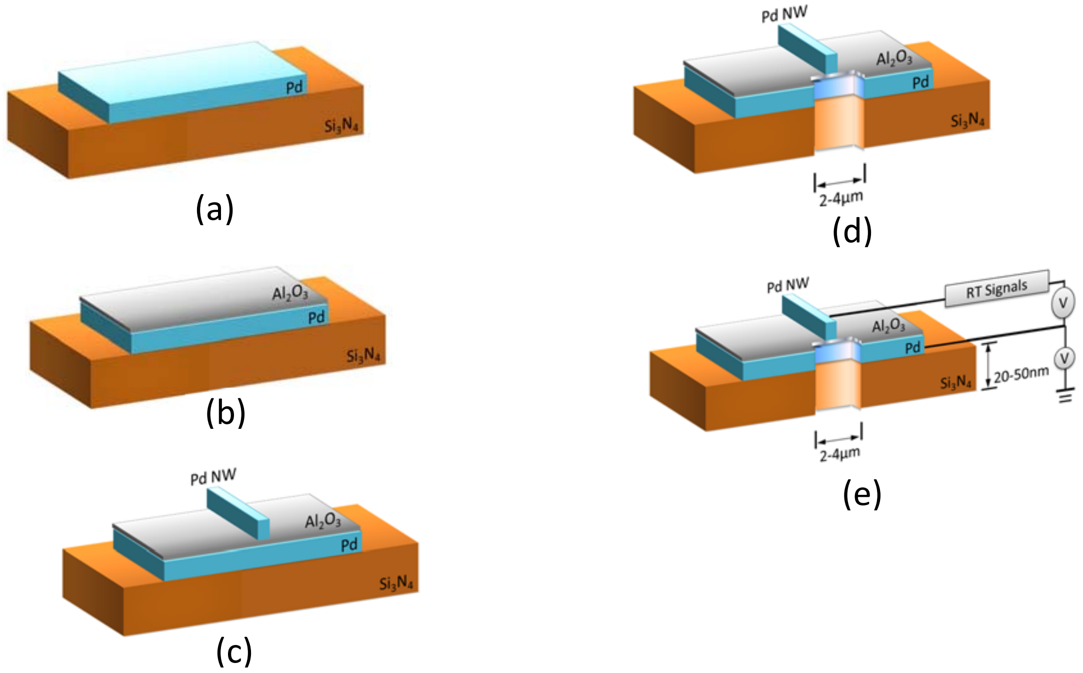


Figure 7.1 Fabrication of tunneling stacked junction

The processes in Figure 7.1 are:

(a) Pattern the bottom electrode by EBL and deposit Ti/Pd (1nm/10nm) on silicon nitride membrane by e-beam evaporation machine, followed by lift-off process.

(b) Deposit Al_2O_3 layer (2.0 to 2.5nm thick) on the whole chip by plasma-enhanced atomic layer deposition (ALD) machine.

(c) Pattern the top nanowire electrode by EBL and metal Ti/Pd (1nm/10nm) is deposited by e-beam evaporation machine, followed by lift-off process.

(d) Trench which is through the stacked junction is made by FIB, He ion beam, or RIE. Two electrodes are exposed on the trench wall.

(e) Test whether the device is shorted or not. If not, analytes are added to collected signals.

A cross-section view is shown in Figure 7.2 and optical image is shown in Figure 7.3.

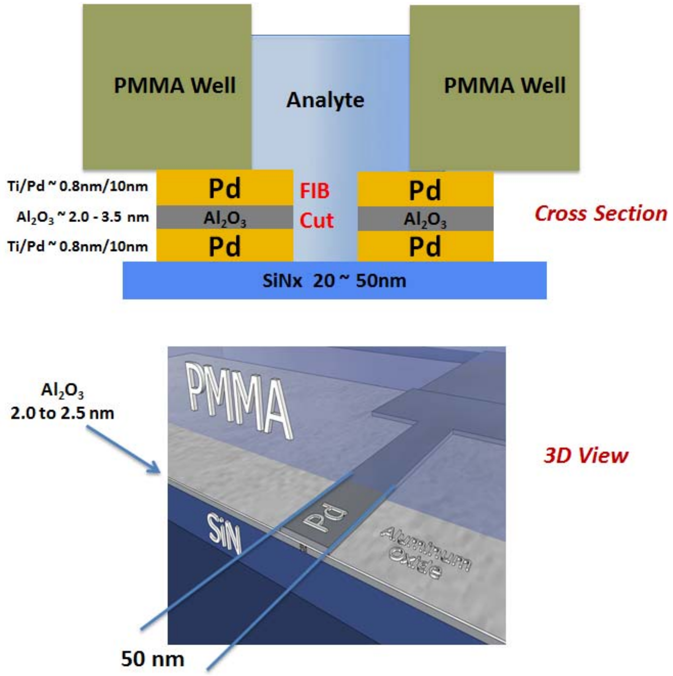


Figure 7.2 Cross-section view of stacked junction

Figure 7.3 is an optical image of the device from the backside view. The junction area is in the sample well. Two vertical electrodes are connected to the back Pd wire and the horizontal electrodes are connected to the top Pd wire. The microfluidic channel are built to run flow through experiments. Tunneling signals of a complete cycle of passing four nucleotides through the junction is shown in Figure 7.4 (a). The current distributions of each base is shown in Figure 7.4 (b). Even the current has already shown pretty good separation, we found that the plotting of wavelet components of the signals gave four isolated regions in 2-D plot (Figure 7.5).

We are also developing a stacked junction device combined with a nanopore so that we could drive single strand DNA or peptide into the nanopore and measure tunneling single sequentially. In the foreseeable future and it should not be far from now, we should be able to start getting oligomer signals and eventually a complete single strand DNA and peptide signals.

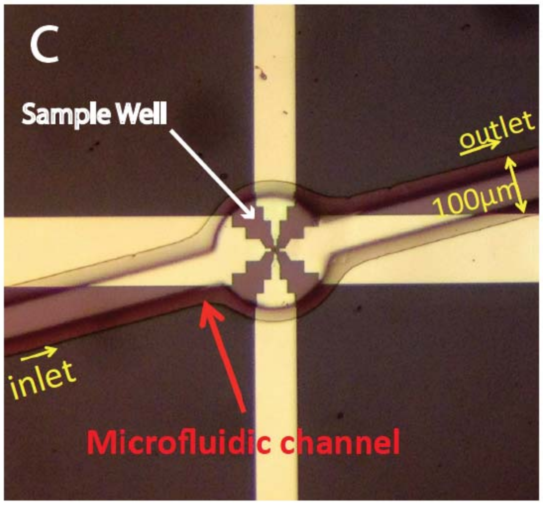
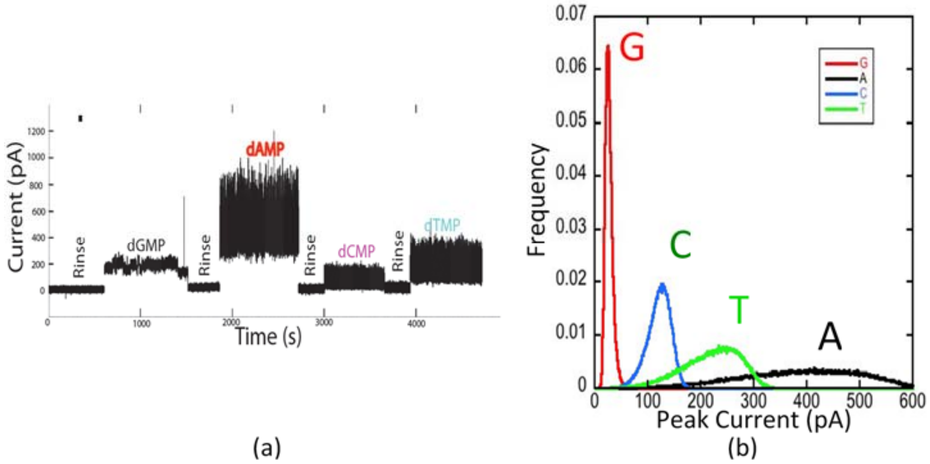


Figure 7.3 Optical image of stacked junction



(a)

(b)

Figure 7.4 Raw data for flow-through cycle of DNA bases (a) and peak current distribution (b).

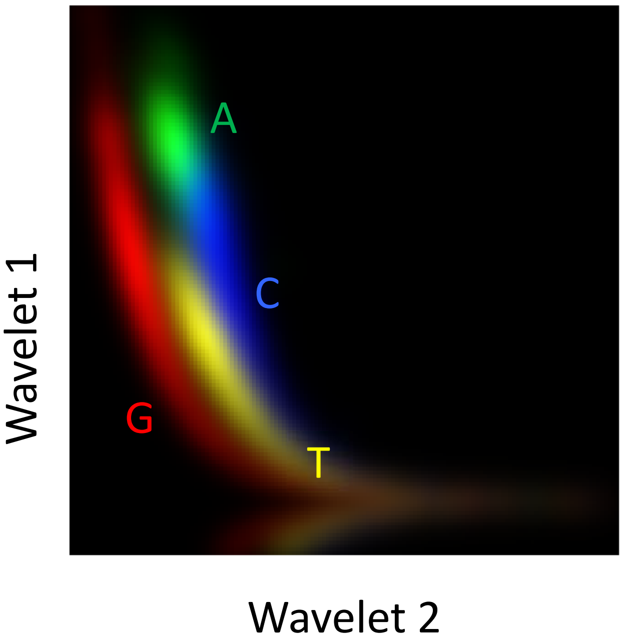


Figure 7.5 SVM analysis for 4 DNA bases

Chapter 8

Physical Model for Recognition Tunneling

The basic assumption of the recognition tunneling is that the signature of the molecular recognition arises from the specificity of the H-bonds, created between the universal molecular readers and the targeted molecule. The useful information appears in form of a characteristic telegraph noise, originating from the thermal fluctuations of both molecule and the surrounding water environment. Using the Oxford DNA model with embedded with Brownian dynamics, we construct a universal reader (ICA) and run the molecular bonding dynamics. We show that fluctuations of only hydrogen bond length, applied in a simple exponential model of electron tunneling can produce the telegraph noise, which can be decoded by support vector machine, resulting in the DNA bases recognition, mimicking those obtained in the single base STM experiment.

This work is done by Dr. Predrag Krstic, Dr. Brian Ashcroft and Dr. Stuart Lindsay and a paper is under preparation. The author of this thesis has obtained the privilege to quote their work in this thesis.

8.1 INTRODUCTION

The basic idea of the so called physics-based third generation methods for DNA sequencing is electrical detection of the DNA sequences, either by reading the electron tunneling current measured across the pore transversally to the translocation single DNA polymer varies with a base passing through the pore between the electrodes.

A typical tunneling current through a DNA base, read between metal substrate and STM tip, functionalized by the imidazole readers (ICA), is shown in Figure 8.1. The apparent telegraphic noise- like signal of the order of tens of pA is observed at the scale of seconds, with the peak features resolved at ms scale (inset in Figure 8.1). Interestingly, our simulation of this experiment, included

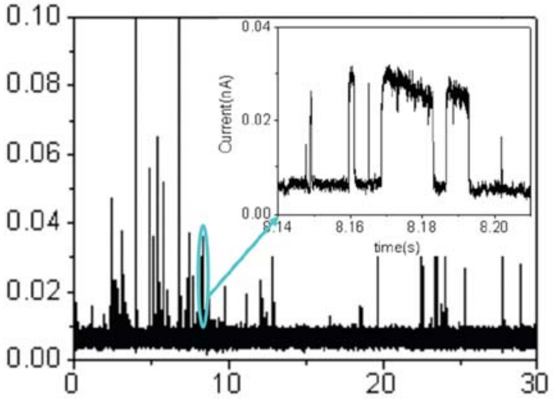


Figure 8.1 A typical recognition tunneling current through a DNA base

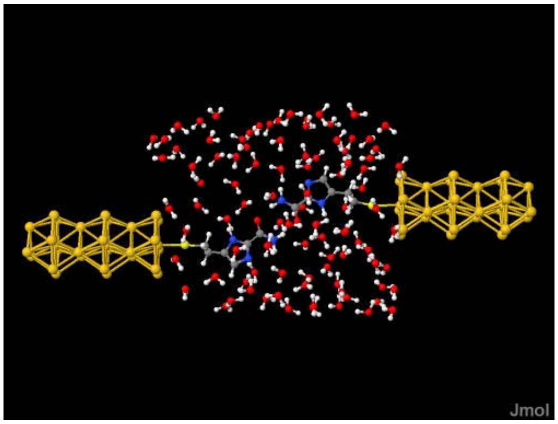


Figure 8.2. Simulation of the recognition tunneling dynamics using gold electrodes, imidazole readers, and T-nucleotide (not shown in the figure) embedded in a bath of 90 water molecules (red-oxygen, white-hydrogen, blue-nitrogen, grey-carbon, yellow-gold atoms)

a simplified geometry at Figure 8.2, where gold wire-electrodes, imidazole (attached with the thiol to the electrodes), thiamin- nucleotide (not shown in the figure) are embedded in the bath of 90 water molecules. We let the system evolve dynamically, using the classical molecular dynamics with the time step of 0.2 fs, using Berendsen thermostat at 25 °C, recording all coordinates each 10 fs. The gold atoms and sulfur atoms are frozen. We apply non-equilibrium Green-function (NEGF) calculation of the tunneling current through the system at the recorded states, thus obtaining time resolved tunneling signal, shown in Figure 8.3. The applied bias voltage is 0.1 V.

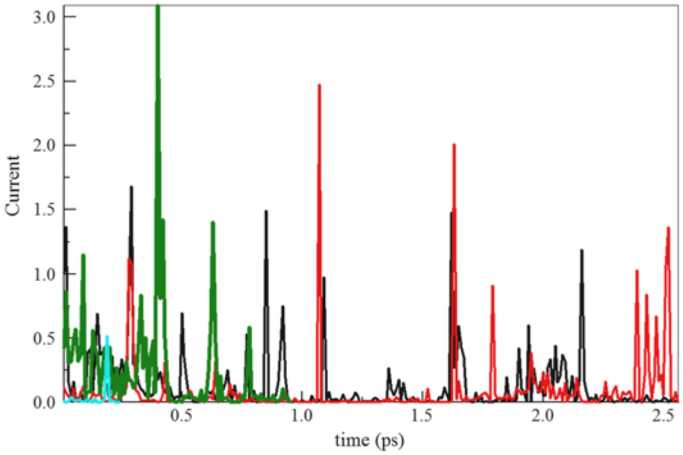


Figure 8.3 Combined classical molecular dynamic and quantum-mechanical calculation of the tunneling using NEGF method. Black lines: readers, no water; Red lines: T nucleotide, H bonded to the readers, no water; Green lines: readers in water; Light-blue lines: T-nucleotide, H-bonded to the readers, in water.

The current signal at figure 8.3 (in pA) shows a number of interesting features. Even in absence of fluctuations caused by the Brownian motion of the water molecules, the signal shows the typical telegraphic-like fluctuations (resembling to the measured one in Figure 8.1) though with lower frequency than in presence of water (green line), caused by the thermal fluctuations of the bonding molecules. Although the classical MD calculations are fast, the computational bottle-necks are the QM NEGF calculations, which becomes formidable in presence of water. This is the reason that the calculations have been performed in the interval ranging 2.5 ps (without water) and below 1ps range in presence of water. The presence of the water changes the signal through readers, increasing the frequency, and even intensity of the peaks. The tunneling signal through the T-nucleotide and readers, in presence of water, is suppressed in the calculated range in comparison to the one through only readers. The important observation is here that the fluctuations of the calculated signal is caused by the thermal fluctuations, and that the tunneling signature is in all cases at figure 8.3 is distinct, and significantly influenced by the presence of water.

Scale of data measured and simulation are all in pA. So the simulation provides a good explanation of the physical mechanism of recognition tunneling measurement.

8.2 Model of DNA

To simulate long time and length scale processes involving DNA it is necessary to use a coarse-grained description, capable of the self-assembly process, bond-making and bond-breaking. OxDNA, a coarse-grained DNA model developed by the University of Oxford, is particularly suited to this task. Coarse-grained modelling of DNA has now reached an exciting stage. Models are now available that allow DNA nanosystems to be systematically and accurately probed, thus opening up the field of DNA nanotechnology to molecular simulations. Even complete DNA origami with over ten thousand nucleotides can be structurally characterized. Probing the self-assembly mechanisms of such large nanostructures is still a real challenge for simulations because of the long time scales involved, and assumes significant improvements in the computational speed, in particular with inclusion of the GPU threading.

OxDNA model and its interaction potentials have been described in detail by Ouldridge in a number of publications, while the code implementing OxDNA model is available for public download. The model represents DNA as a string of nucleotides, where each nucleotide (sugar, phosphate and base group) is a rigid body with interaction sites for backbone, stacking and hydrogen-bonding interactions. The potential energy of the system can be decomposed as

$$V_{oxDNA}=\sum_{<ij>}(V_{b.b.}+V_{stack}+V_{exc}^{'})+\sum_{i,j\notin<ij>}(V_{HB}+V_{cr.st.}+V_{exc}+V_{cx.st.}) \quad \text{(Eq. 8.1)}$$

where the first sum is taken over all nucleotides that are nearest neighbors on the same strand and the second sum comprises all remaining pairs. The interactions between nucleotides are schematically shown in Figure 8.4.

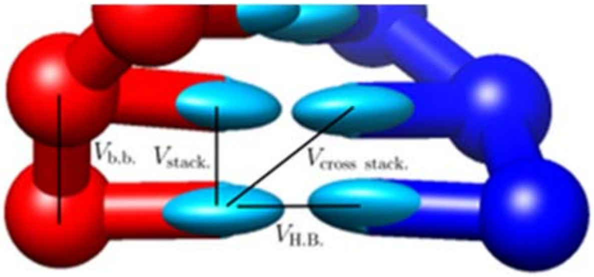


Figure 8.4 Interactions in the oxDNA model

The parameter values used in these models have been chosen to reproduce physically reasonable behavior for a range of systems. Since the main focus of the present work is the hydrogen bonding, we pay a more detailed attention to how it is modelled. The model reproduces the base pairing through the V_{HB} term in Eq. (1), which incorporates a radial term dependent on the separation of hydrogen bonding sites, R. This interaction is modulated by terms that encourage the co-linear alignment of the hydrogen bonding sites (quantified by three angles, θ_1, θ_2, θ_3). Further modulation comes from the terms that encourage the planes of the bases to be antiparallel (θ_4). Small modulation also comes from the terms that penalize the hydrogen bonding between the bases that are not opposite to each other (θ_5 and θ_6)

$$V_{HB} = f_1(R)F(\overline{\theta}) \qquad \text{(Eq. 8.2)}$$

where $\overline{\theta}$ is the set of the angles $\theta_i, i = 1-6$. We present in Figure 5 the radial dependence of the $V_{HB}(R)$ for randomly chosen sets $\overline{\theta}$ for both Watson-Crick pairs A-T and G-C, to show the sensitivity of the radial potential and its bottom on the angular variation of the base positions. The range of coupling extends to more than 6 Angstroms, while the absolute minimum of the coupling is close to 3.4 Angstroms and is about 0.32 eV for G-C and about 0.23 eV for A-T hydrogen bonding (not achieved in the figure). We note that these values include two-H bonds of A-T and three ones for G-C.

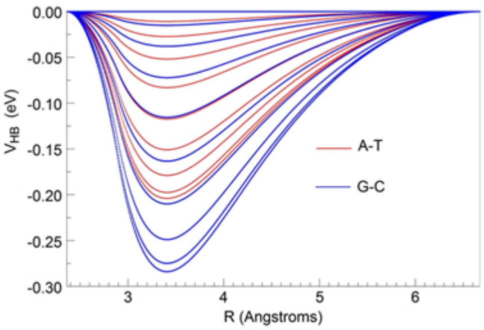


Figure 8.5 The hydrogen bonding potential energy V_{HB} as function of the distance between the interacting bases, for random sets of the orientation angles.

Figure 8.6 shows the 3D plot of potential $V_{HB}(R)$ for uniform variation $\delta\theta$ of the angles from the minimum position.

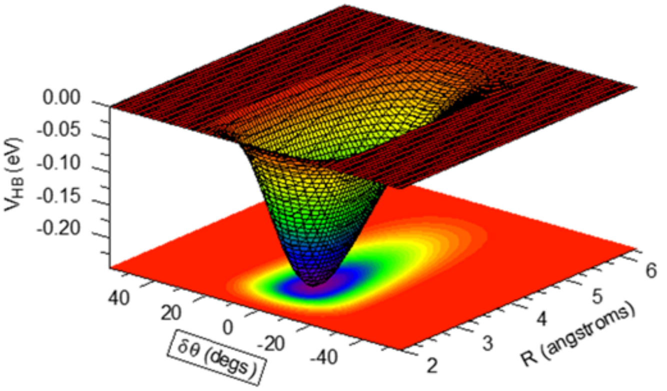


Figure 8.6 3D representation of the V_{HB}

Finally, the sensitivity of the $V_{HB}(R)$ to the external field, applied in the direction of a perfectly aligned bonding (with $\delta\theta=0$ in Figure 8.6) is shown in Figure 8.7. A typical electrophoretic field for a DNA translocation though a nanopore is between 10^6 and 10^7 V/m, and can somewhat influence the creation and breaking the H-bonds. However, the tunneling bias of 0.5V across 2.5 nm, leads to the fields of the order of 10^8 V/m, which obviously has a strong effect to the bond life-times, decreasing the activation barrier by about 30%.

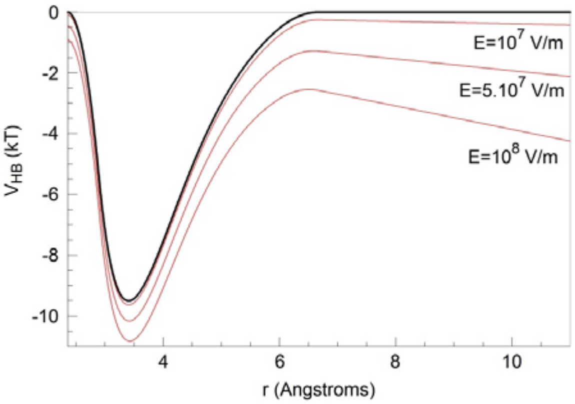


Figure 8.7 Sensitivity of V_{HB}(R) on the electric field, in the direction of the perfectly aligned bond.

Model of the reader-base interaction

We develop a simple model of a universal molecular reader interacting with all DNA bases, we develop a relative simple deviation of the oxDNA model, implementing a new “base” Z in the model which bonds to all four DNA base. The base-independent angular modulation of the hydrogen bonding is left as in oxDNA model. Following the experimental observation, Z-T and Z-A bond strengths (minima in the curves in figure 8.4, for ideally antiparallel angles) are given the smallest values of 0.23 eV and 0.26 eV respectively, while Z-C and Z-T are assigned somewhat larger values, of 0.292 eV and 0.32 eV, respectively. The adopted H-bond strengths are between maximum (G-C) and minimum (A-T) values of the Watson-Crick pairs, which are kept as in the original oxDNA model. The coupling base-pair dependent stacking strength for the Z-A, Z-T, Z-C, Z-G and Z-Z stacking is here taken 0.424 eV, which is average value of all original stacking strength between various base pairs (which are kept here intact). These parameters are expected to give a widened distribution of the tunneling conductance for Z-A and Z-T bonds, in comparison to those of Z-C and Z-G bonds.

8.3 Description of the Dynamics

Langevin Dynamics is here the approach of choice in the oxDNA model for generating diffusive motion of coarse-grained models with implicit solvent. The principle is that the solvent exerts both random forces and dissipative drag on the solute, and that the two are related by a fluctuation-dissipation relation to ensure that the steady-state distribution is given by Boltzmann equation. Newton's equations, with the addition of these solvent-mediated forces, can then be integrated to give dynamical trajectories. In this model, each nucleotide is a 3D rigid body so that the state space of *N* model nucleotides has 6*N* spatial dimensions and 6*N* momenta. Pairs of nucleotides interact through a number of pairwise effective interactions, described in the previous Section. oxDNA uses quaternion-based algorithm of David Chack et al. as an efficient methodology for simulating rigid body. In this algorithm it is necessary to specify a friction tensor relating the drag forces experienced by a particle to its momenta. For simplicity, it is assumed that each nucleotide interacts with the solvent in a spherically symmetric manner, meaning that the task is reduced to identifying linear and rotational damping coefficients. Although slow this approach reproduces diffusive motion.

The thermostat implemented in the simulation code is a simple thermostat that emulates Brownian dynamics. The system is evolved integrating Newton's equations of motion ('NVE' ensemble) for a given (small) number of steps. Then the velocity and momentum of each particle are refreshed, with a given fixed probability. The new velocities and momenta are chosen according to the Maxwell distribution of the temperature at which the simulation is run. This approximates a Brownian dynamics on time scales much longer than the refresh interval, taken here each 103 steps (one step used was ~15 fs). The presence of random collisions causes the velocities of particles to decorrelate ("lose memory" of their initial values from some previous point in time) much faster than the NVE dynamics. As a result, true molecular kinetics are not preserved by the Andersen thermostat. For example, the computed diffusion constants for particles would give erroneous values. However, to implement in the model diffusion limited dynamics it is alternatively used an additional input coefficient, diff-coeff, the monomer diffusion coefficient, and probability is adapted to that value. The algorithm works as follows: the system is evolved for a number of steps equal to newtonian_steps according to Newton's equations of motion. Then for each particle a random number is extracted; if it is larger than the value for pt (either set explicitly or derived from diff_coeff) the particle is left untouched. If the random number extracted is lower than the value of pt, each of the components of the velocity and angular momentum of the particle are refreshed according to the Maxwell distribution dictated by the value of the temperature. Whether velocities get updated or not is independent of the current temperature. Because, the system evolves at constant energy between the collisions, this method generates a succession of microcanonical simulations, interrupted by small energy jumps corresponding to each collision (refresh interval), leading to the canonical assemble at the time scale much longer of the refresh interval (here ~ps). The system becomes ergodic at the same time scale.

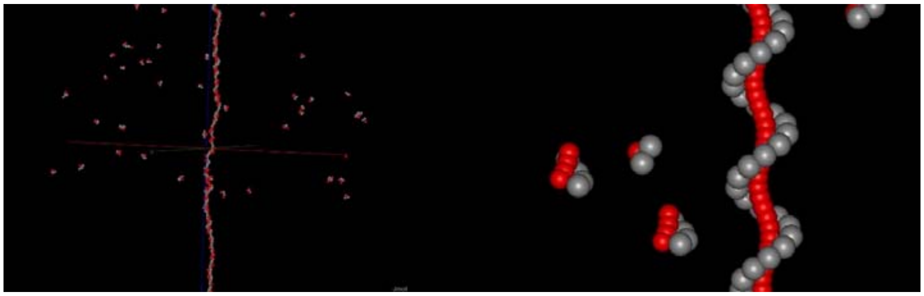
The applied thermostat is a variant of Andersen thermostat and requires further discussion. A stochastic element to the temperature is included by having random collisions of molecules with an imaginary heat bath at the desired temperature (in refresh intervals). In the Andersen scheme, one does not perform a collision with each molecular dynamics time step, but rather it is customary to adopt a collision frequency H or collision time. In that sense, the refresh interval chosen above is a chosen collision frequency. This should be chosen so as not to be too short to the time scales of molecular motions. Assuming that the collision frequency in water is 0.1 ps, the chosen interval in our calculation of about 1 ps is accelerating the dynamics calculation, which is an advantage.

The above described dynamics produces trajectories of the system, damping them in a prescribed step. Our typical calculation have ran 10^9 to 10^{10} time steps, i.e. tens to hundreds of μs, dumped each 200 steps, i.e in steps of 3 ps. Besides the trajectories the regular output of the dynamics are the components of energy, including the hydrogen bonding energy. However to obtain the tunneling current associated to the particular hydrogen bonding, we used the estimate, obtained by the first principle calculations. People extracted the tunneling current that corresponds to the hydrogen bonding and found the exponents β that well fit the calculated results in a simple exponential tunneling formula

$$G = G_0 \exp(-\beta R) \qquad \text{(Eq. 8.3)}$$

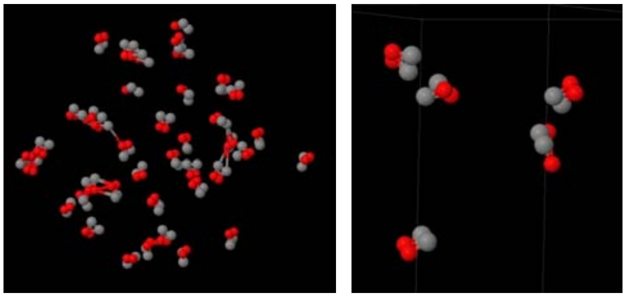
where R is the hydrogen bonding length, and G_0 is quantum of conductance (77μS). They obtained $\beta_{G-C} = 3.3A^{-1}$ and $\beta_{A-T} = 2.5A^{-1}$. We will use these two relations to identify tunneling conductance through established hydrogen bonding. Still, the oxDNA dynamics doesn't provide the hydrogen bonding length. We develop a code to extract form the trajectories the evolution of the hydrogen bonding length in time.

We perform final computations in three reservoir configurations, presented schematically in Figs. 8.8-8.10. In the configuration 1 (Figure 8.8), we set a short ssDNA segment of a length of few tens of bases, applying at its ends stretching force of about 100 pN, whose goal is to prevent the folding of the DNA segment. In the computing box (7-12 nm) we also set some number of dimers containing either pair of Z-nucleotides (universal readers), or a pair of Z-nucleotide and one of the A,G,C,T nucleotides. The system was let running the dynamics (at 300K) for long time (step 15 fs, 10^9 to 10^{10} steps). The state of the system is damped each 200 steps, including positions and velocities of all particles, energies etc. The red-spheres in Figure 8.8(b) represent the sugar-phosphors back bone, while the grey spheres are the bases. We note that the DNA segment is composited of either single or of mixed bases. Configuration 2, in Figure 8.9(a) contains only mix of five of dimers of nucleotides A, T, C, and/or G and 25 dimers of the readers ZZ, in a box of 12 nm. Finally, configuration 3 in Figure 8.9(b), contains 4 of the basis dimers and one ZZ dimer, in a box of 4 nm lateral dimension.



(a) (b)

Figure 8.8 Configuration of the DNA segment and reader dimers in Configuration 1.

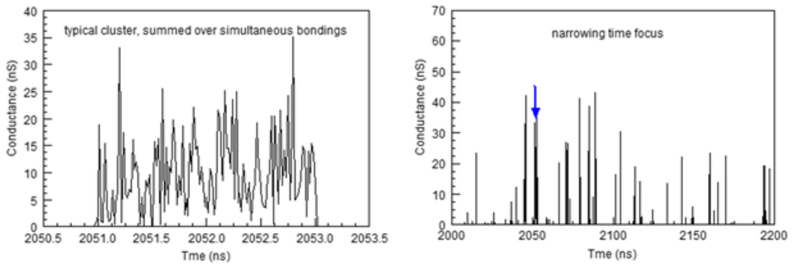


(a) (b)

Figure 8.9 Configurations of the nucleotide and reader dimers

8.4 Recognition of the Unknown Analytes

Unlike to the experiment, our model simulation has zero background currents. This simplifies the analysis, in particular the recognition by simulating tunneling form Eq. (8.3), upon recognizing the hydrogen bonding done by the probe dimers with the unknown DNA analyte. With this in mind, our analysis of the trajectories eliminates any bonds created among the dimers. Also, the SVM analysis is strongly dependent on the analysis of the properties of the peaks, as discussed in Section 4. We define a peak as a structure (Figure 8.10a), between two zero conductance values at the time scale.



(a) (b)

Figure 8.10 A peak structure (a), contained in the conductance time spectrum.

As discussed in Section 3, multiple bonds are frequent, and these may introduce a significant noise in the SVM recognition analyses. The noise is particularly damaging when homogeneous DNA segments are used as the training analytes, to recognize content in the heterogeneous DNA segment. Conditions of bonding are in these two cases quite different, and multiple bonds in the training may contain a strong noise to the recognition procedure. Thus, having in model a possibility of independent resolution of the nature of the hydrogen bonding even in each point, we extract the single bonds at the background of a mixture of multiple bonds. Thus, when ZZ dimer bonds to the analyte, the peak content belonging to each of the Z readers is followed in time, thus devising two peaks from one. Similar procedure is applied in case of triple, quadruple bonds, etc.

Once the training "multiples" are decoupled, we perform the same procedure for the unknown analyte, in this case the unknown analyte is 70-base heterogeneous ssDNA segment, GCCGTTCGCACGGCGCGAAGGAGCGGCTGCCAGTTCCAAGTGCGG ACGCGGCTGCCGCAACGGAGCTCGT. We separate peaks which belong to the various bases, obtaining thus four separate recognition "signals", following in each peak signal form each of the readers in the dimer, each of which can be analyzed, as well as can be all together. The result of such separation is shown in Figure 8.12, each of the bases is color coded.

After analyzing the training signals (Figure 8.11) and the unknown analyte, Figure 8.12, the results obtained as functions of the used number of parameters (Section 4) are shown in Figure 8.13. While the training accuracy has reached 100%, the recognition probability of the various bases in DNA varies between 85% (for A and T) and 100% (for G). Since G has a larger number and higher quality (more points) peaks, it looks like A and T recognition can be much improved by increasing the number of cases (i.e. extending the time). However, number of points in the A and T peaks cannot be improved, and it is not certain that increasing number of time points will improve the SVM probabilities. In fact, from Figure 8.14 could be seen that A and G bases are well separable, but A and T are not. This may be coming from the nature of the oxDNA model: Only strengths (single parameters) are used to distinguish them in building the H-bonding, which obviously is not enough for a superior recognition. Our simulation experiments show that A+T and G+C are in all cases recognized with very high accuracy.

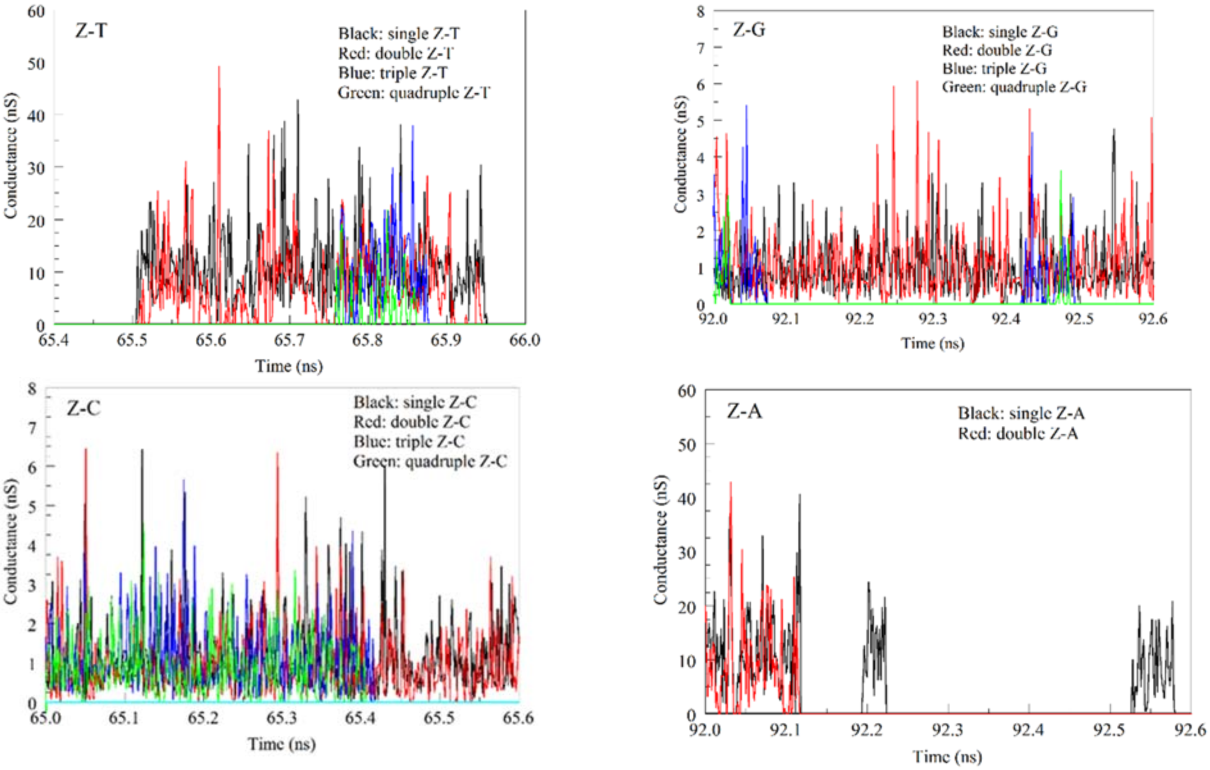


Figure 8.11 Decoupling of the multiple bonds for the training analytes

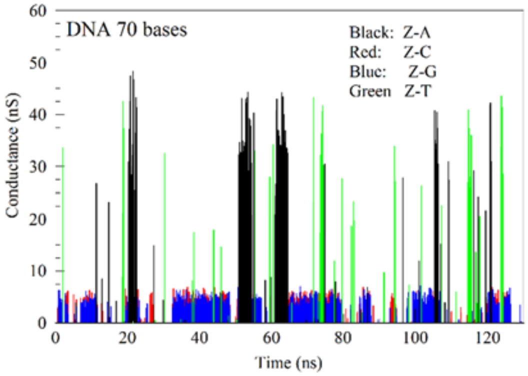


Figure 8.12 Conductance over time probed by the ZZ dimers at the 70-ple heterogeneous ssDNA segment.

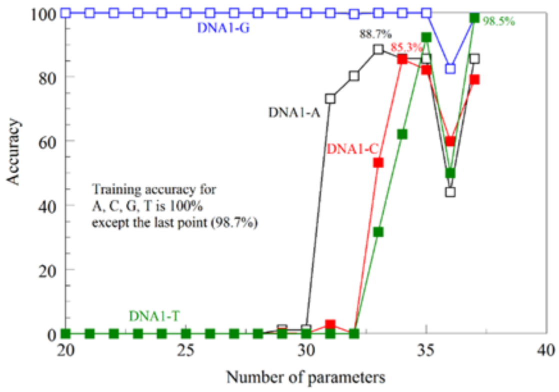


Figure 8.13 SVM recognition of the 70-uple DNA

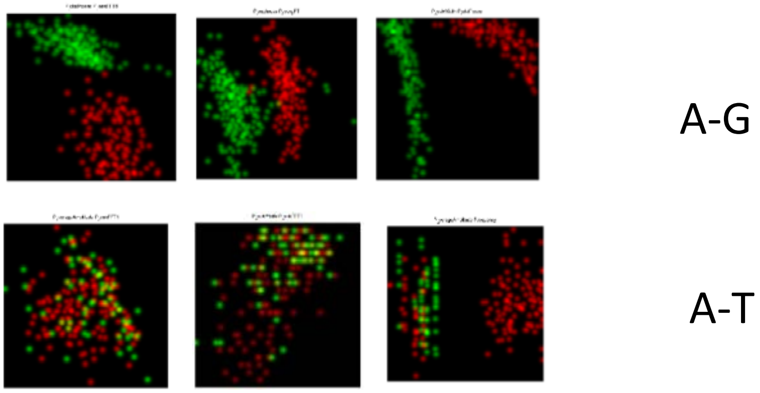


Figure 8.14 2D projections of the SVM results

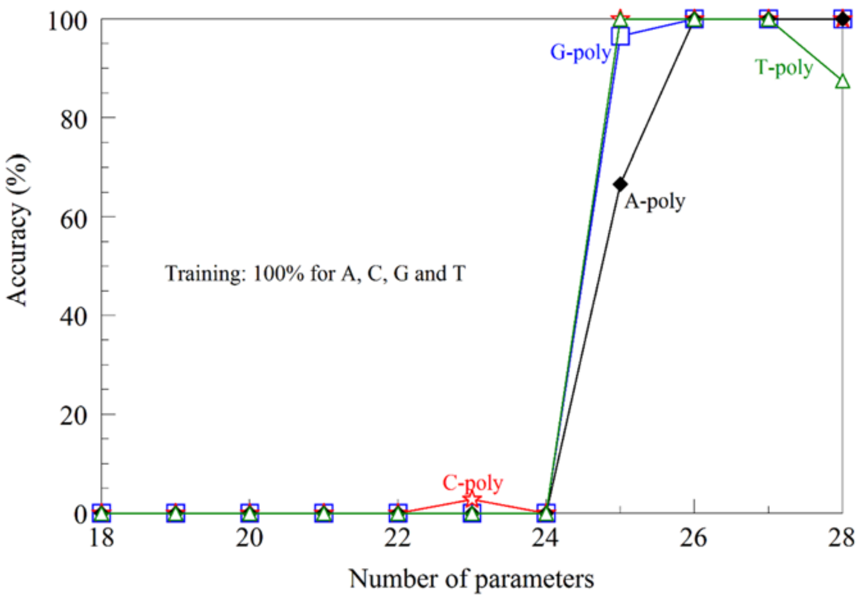


Figure 8.15 Recognition accuracy of homogeneous ssDNA segments

However, the homogeneous ssDNA segments of length 30, were recognized by SVM, using the same training as in Figure 8.11, show accuracy of almost 100% for all four bases.

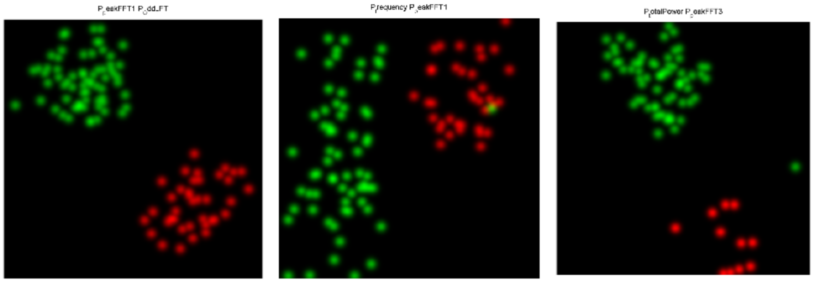


Figure 8.16 2D projections of the SVM parameter space, for the PolyA-PolyC separability

We note that the "swarm" approaches, described in Figure 8.9, are also providing and excellent recognition accuracy of all four bases, without a need to do the multiple-bonds decoupling procedure. The case described at Figure 8.9(a) is shown in Figure 8.17, with recognition ranging from 92% to 94%, with excellent 2D separability. Similar results were obtained for case in Figure 8.9(b).

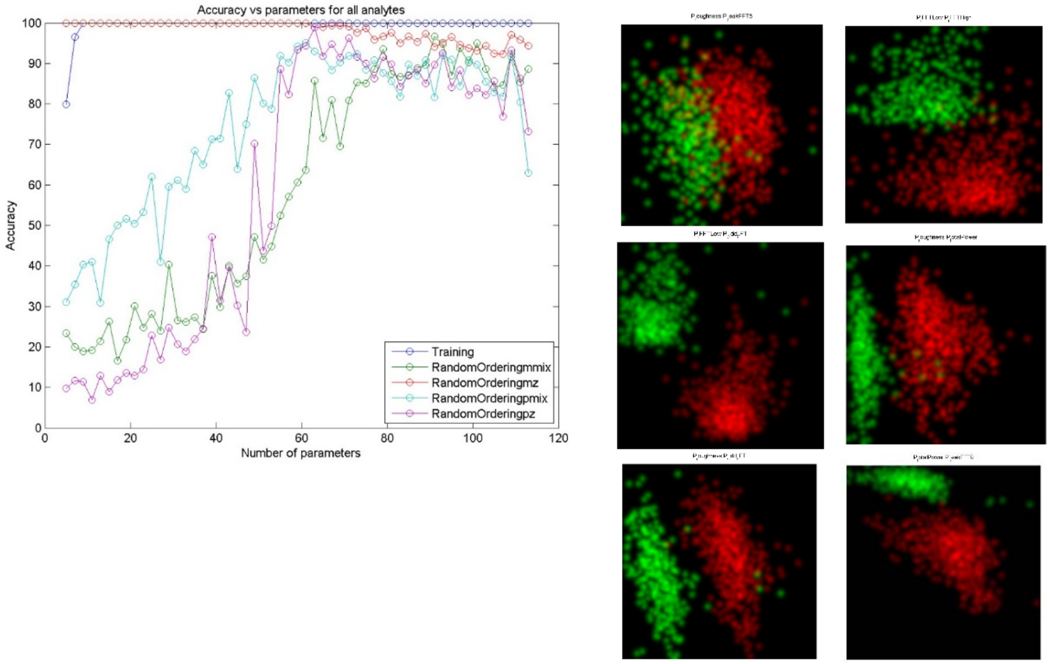


Figure 8.17 Recognition all 4 bases in the “swarm” Model of Figure 8.9(a)

Chapter 9

Introduction of Data Mining Methods-Support Vector Machine and Decision Tree

1-D and 2-D plot are the most popular plots used when dealing with differentiation, while 3-D plot is rarely used in most literature related to physical science. For the parameters plotted, people tended to focus on the data related to explicit physical properties as conductivity, etc. Statistics is useful, while when talking about sequencing ONE single strand DNA or peptide and detect the corresponding nucleotides or amino acids, we have to do the measurement hundreds or thousands of time to get a distribution for each base and use 1-D/2-D distribution information to recognize what kind of base it is. This process would sacrifice most of advantages that nanotechnology have, such as fast and low dose. In our previous analyze methods, we approach same results using a different method. Instead of running the experiments multiple times and focusing on one or two parameters, we study all the features the signals for one single measurement, not only including the traditional current amplitude, duration time of the peaks, etc. but also some features that we are still not clearly understand how they are related to the physical or chemical properties. These features were constant through all the experiments including using different STM and different probes, different substrate and different person conducting the experiment. We suggested these features must be related to some intrinsic properties of the molecule. We used these features in our data analysis process without elaborating its physical/chemical meaning. While hoping that sometime later we could connect the dots to bridge the theory and experiments results. Another endeavor we did was to use data mining methods in our analysis. Even though data mining methods are blooming in the fields as Computer Science, Bioinformatics, Engineering, and Economics, etc. Researchers in solid science seem to ignoring these powerful tools. However, we decided to embrace it not because it was easy, but because it was promising. I will introduce two data mining methods in this chapter, one of which is called Decision Tree and the other one is Support Vector Machine (SVM). SVM performed better in our analysis that was the reason we use SVM as our major analytical method in the previous chapters.

9.1 Decision Tree

A decision tree is shown in Figure 9.1. The purpose of this specific decision tree is to make a decision if you should go outside play tennis or not. In a decision tree, each node specifies a test for some attribute of the instance (as humidity, wind). Each branch corresponds to an attribute value and each leaf node assigns a classification (as Yes/No).

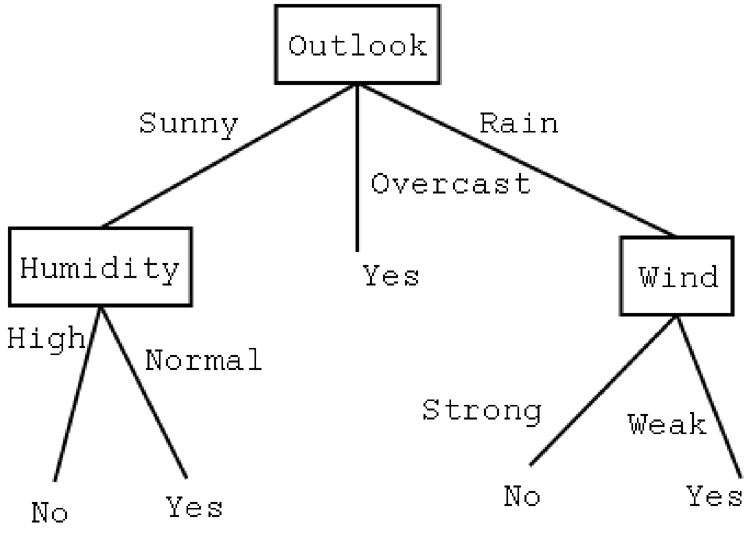
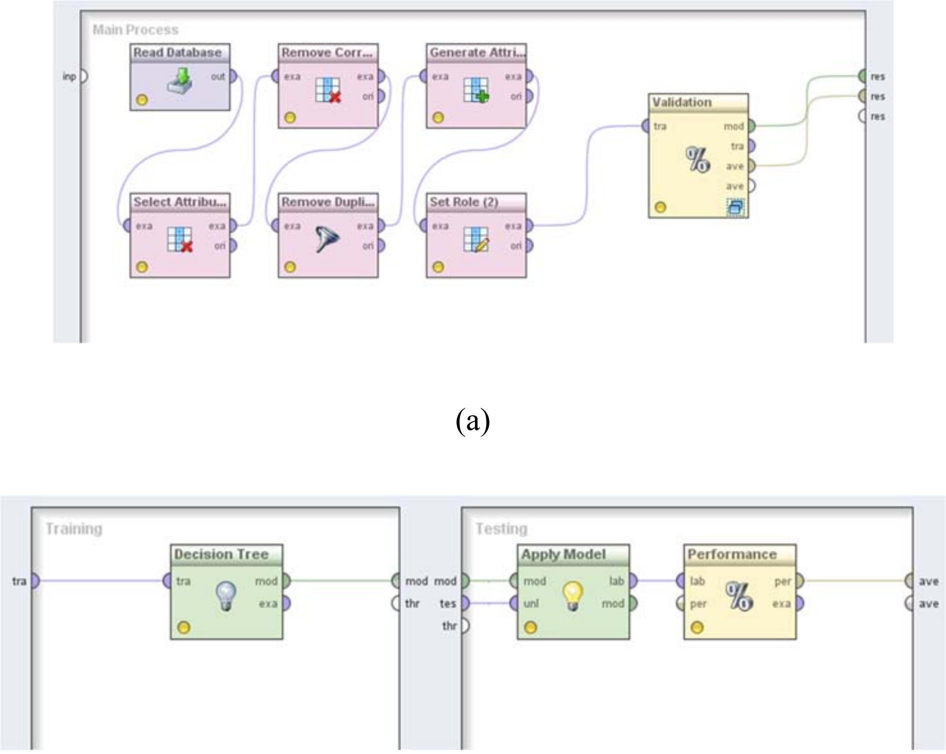


Figure 9.1 A simple decision tree

To use decision tree to perform classification, a small data pool with classes each of them that they belong to are randomly selected from the original data. This small portion of original data are then used to training to build the Decision Tree first. Lots of criterion are available to build the tree [63]. Then, the rest of data are tested by passing through the tree and a class comparison of what the tree predicts and what it should be are done to give the performance of the tree.

Luckily, there is no need to build everything from ground up because a commercial available data mining software called RapidMiner, which allows us to perform Decision Tree analysis by building operator blocks. I will explain how it works using Ibuprofen data.



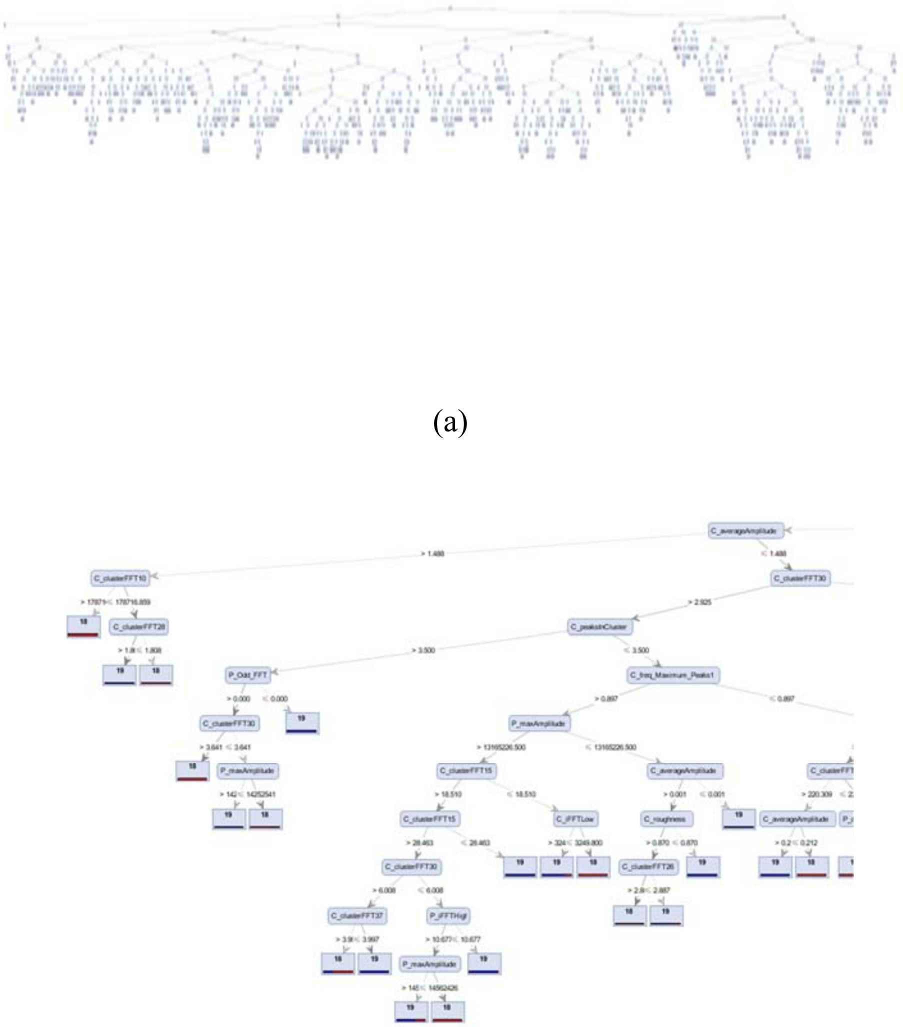
(a)

(b)

Figure 9.2 Decision Tree analysis process (RapidMiner)

Figure 9.2 Shows the Decision Tree analysis process I built to analysis Ibuprofen data (Figure 8.2 (a)). The starting operator “Read Data” retrieved all the 161 features of the spikes (Chapter 3) from a database. Then the data was processed as removing the highly correlative attributes and duplicates. At last the data were processed by an operator call “Validation”. The “Validation” operator was consist two parts – Training and Testing (Figure 9.2 (b)). The “Training” block was an operator of decision tree. Here a small portion of data was selected and a decision tree was build based on the data. I used a criterion named “gini_index” to build the tree. After the tree was built, the rest data was passed to the testing block. The “Apply Model” block would predict what kind of class

the data belongs to and a comparison of what should belongs to were done in the "Performance" block.

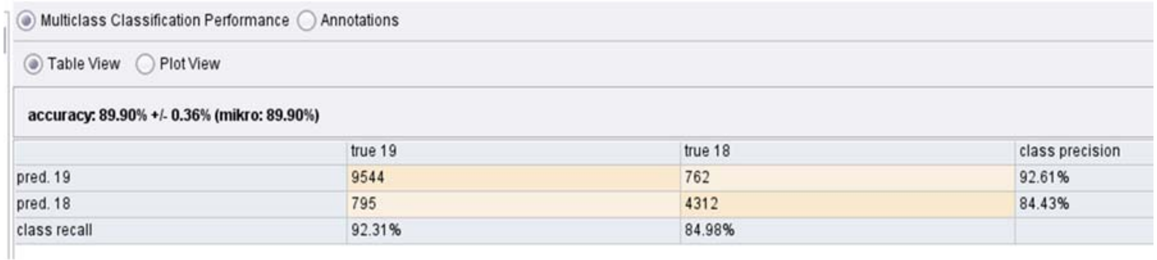


(a)

(b)

Figure 9.3 The complete Decision tree for Ibuprofen analysis (a), partial of the tree shows the branch detail (b).

Complete decision tree and a partial of tree showing the detail of branch are given in Figure 8.3. The performance of the decision are given in Figure 8.4.



◉ Multiclass Classification Performance ○ Annotations

◉ Table View ○ Plot View

accuracy: 89.90% +/- 0.36% (mikro: 89.90%)

	true 19	true 18	class precision
pred. 19	9544	762	92.61%
pred. 18	795	4312	84.43%
class recall	92.31%	84.98%	

Figure 9.4 Performance of Decision tree

Precision is defined as (true positive)/(true positive + false positive). Recall is defined as (true positive)/(true positive + false negative)).

The accuracy of differentiating S-Ibuprofen and R-Ibuprofen is 89.9%.

9.2 Support Vector Machine

A linear support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin) [64]. Take a Figure 9.5 for example. There are two classes – solid black dot and empty circle. Hyperplane H_1 does not separate the two classes. H_2 does, but only with a small margin. H_3 is ideal hyperplane that we are looking for, which gives a maximum margin.

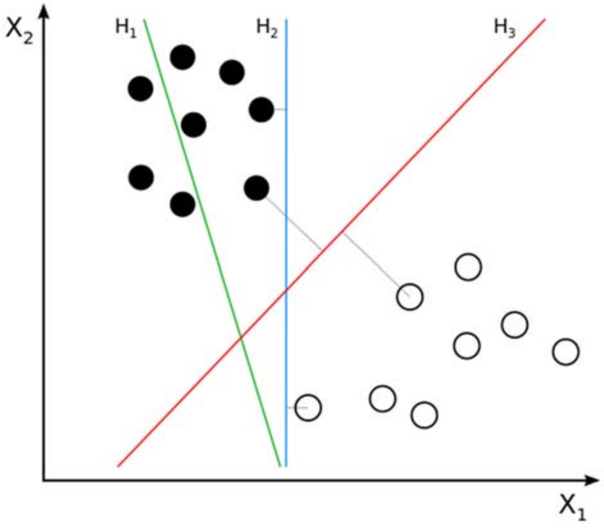


Figure 9.5 Hyperplane to separate different classes

Suppose each of the training data $\mathcal{D}$ has p dimension. We need to build a binary classifier (a p-1 dimension hyperplane) to separate two classes y_i .

$$\mathcal{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

The value of y_i indicates which class the data belongs to. Any hyperplane

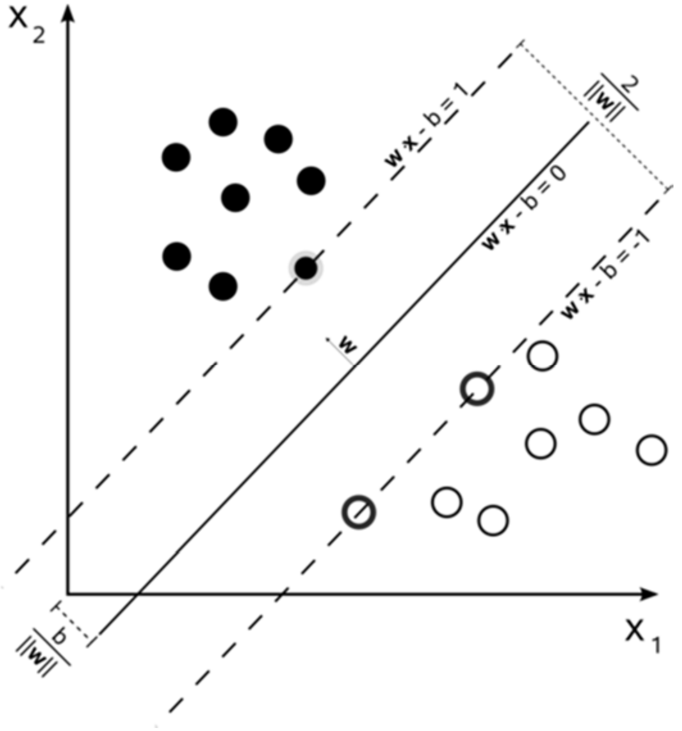


Figure 9.6 Maximum-margin hyperplane and margins for an SVM trained with samples from two classes.

could be written in the form of

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$$

w is the normal vector to the hyperplane. If the data is linearly separable, we have already found such a hyperplane. Then we could also find two other hyperplanes lay on each side of this hyperplane to separate these two classes, shown in Figure 8.6. The distance between these two hyperplanes is 2/||w||. So, our purpose is to minimize ||w|| (thus maximize 2/||w||), subject to $y_i(\boldsymbol{w} \cdot x_i - b \geq 1$, which prevent data point falling into the margin.

This problem could be described in primal form, which transform the problem into a quadratic programming optimization problem:

$$\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2$$

Subject to

$$y_i(\boldsymbol{w} \cdot x_i - b) \geq 1$$

By introducing Lagrange multiplier α, then problem now can be interpreted as:

$$\arg\min_{\mathbf{w},b} \max_{\alpha\geq 0} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1] \right\}$$

The "stationary" karush-Kuhn-Tucker condition implies that the solution can be expressed as a linear combination of the training vectors

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

The corresponding $\mathbf{x_i}$ satisfy $y_i(\boldsymbol{w} \cdot x_i - b) = 1$ are support vectors. And also from the support vectors, we have

$$w \cdot x_i - b = \frac{1}{y_i} = y_i \iff b = w \cdot x_i - y_i$$

Which could be used to calculate offset b.

If majority of the data are separable, a modified maximum margin idea are proposed by Corinna Cortes and Vladimir N. Vapnik [65]. A "Soft Margin" method will choose a hyperplane that splits the examples as cleanly as possible by introducing non-negative slack variables ξ_i, which measures the degree of misclassification of the data $\boldsymbol{x_i}$.

$$y_i(\boldsymbol{w} \cdot x_i - b) \geq 1 - \xi_i$$

And the final optimization problem becomes:

$$\arg\min_{\mathbf{w},\xi,b} \max_{\alpha,\beta} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \right\}$$

with $\alpha_i, \beta_i > 0$

If the data is not linear separable, a kernel trick is applied to transform the original data into a higher-dimensional space, presumable making the separation easier in that space (Figure 9.7). The details could be found in Aizerman and Boser's work [66, 67].

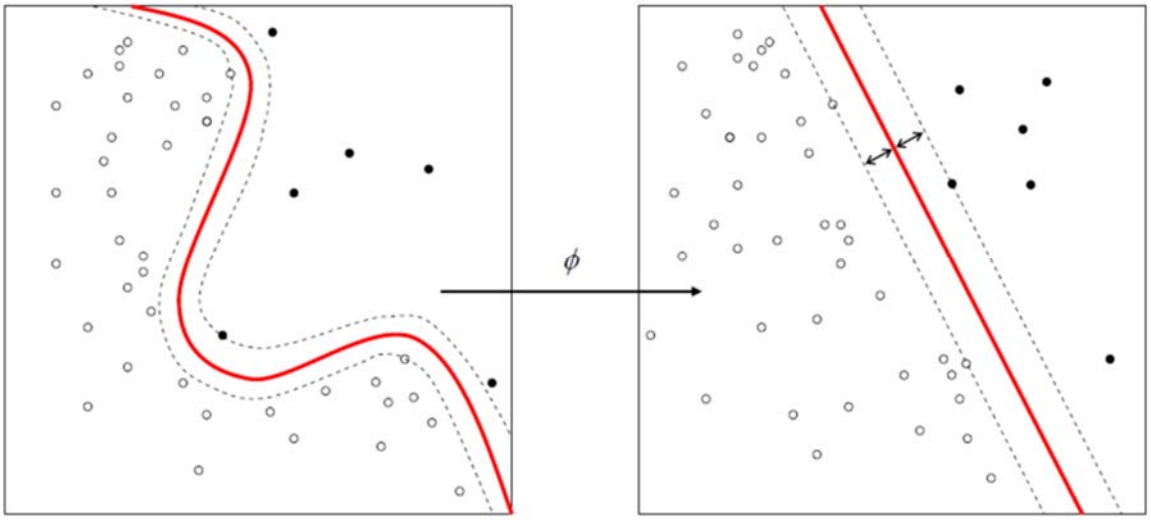


Figure 9.7 Kernel Machine

For multiclass SVM, there are many methods available, the one that we used is as following:

Building binary classifiers which distinguish between (i) one of the labels and the rest (one-versus-all) or (ii) between every pair of classes (one-versus-one). Classification of new instances for the one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class (it is important that the output functions be calibrated to produce comparable scores). For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification

SUMMARY AND CONCLUSION

In this thesis, I summarized the major work that I have done during my PhD studies, including using break-junction methods and recognition tunneling method to perform molecule conductance measurements and molecule identification.

In the break-junction methods part, I used carotenoids with different polar aromatic substituents (OMe, Me, H and Br) conjugated polyene chain as analytes to study the combine effects of these of these polar aromatic substituents. We found that the electron-releasing anisyl substituent increase the nearby π-electron density along the conjugated polyene chain, whereas the electron-withdrawing bromophenyl group would decrease it. When combine these two groups, the flow of π-electrons along the carotenoid wire was controlled by the more electron-withdrawing group, acting like a "bottle-neck effect".

In the recognition methods part, I illustrated that by modifying the probe and substrate with reagent molecules, tunneling signals were enriched by information of physical and chemical properties of analytes bonding with reagent molecules. A data mining tool named Support Vector Machine was employed to do classification and prediction based on the features of the signal. The accuracy of separation of amino acids was over 95%. Thus, this enlightened us to develop commercializable device to do peptide sequencing.

More analytes were analyzed by the similar way as disaccharide, ibuprofen, and phosphorylated-peptide identification. The preliminary results seemed very promising.

The application RT methods of DNA sequencing was also briefly introduced. Even though SVM was not used to do the analysis, the conductance of each DNA (except dTMP had no signal) base had already given us a pretty good separation.

With so many advantages, the RT method carried out by STM had its own drawbacks. The major one was that STM needed a very long time to stabilize and during the experiment, the gap formed by STM might fluctuated during the measurement, which introduced some noise signals. The second drawback was that analytes were captured in the gap majorly by thermo fluctuation. When using very low concentration solutions, we have to run very long experiments to collect enough data. Although active methods as “clock scanning” was used [1], it only applied to substrate with relatively large flat areas such as gold. To overcome these difficulties, we sacrificed the adjustable gap that STM can form and designed a stacked-junction. The signals collected for stacked-junction showed even more promising result.

In the later chapters, I briefly introduced the physical models we build to simulate RT methods and the process applying data mining methods to differentiate molecules.

In the near future, we will improve the fabrication techniques of stacked-junctions to achieve massive production and collect more data for different analytes to have a robust conclusion. Another aspect we will focus on is the theory of the RT methods. We will build more sophisticated models to explain the characters of raw signals and how they are related to the parameters of SVM used to differentiate analytes, thus finally bridge the gap between experiment and theory.

REFERENCES

[1] Xu, B.; Tao, N., Measurement of single-molecule resistance by repeated formation of molecular junctions. ***Science*** 2003, 1221-1223.

[2] B.J. van Wees, H. van Houten, etc, Quantized conductance of point contacts in a two-dimensional electron gas, ***Phys. Rev. Lett.*** 60. 848. 1988

[3] He, J. Ph.D thesis. Arizona State University, Tempe, 2005.

[4] Fang Chen, N.J. Tao, Measurement of Single-Molecule conductance, ***Annu. Rev. Phys. Chem.*** 2007, 58:535-64

[5] Park J. Pasupathy AN, Goldsmith JL, Change C, Yasih Y, et al. 2002. Coulomb blockade and the Kondo effect in single-atom transistors, ***Nature*** 417:722-25

[6] Liang WJ, Shores M, Bockrath M, Long JR, Park H. 2002. Kondo resonance in a single-molecule transistor. ***Nature*** 417:725-29

[7] Li CZ, Bogozi A, Huang W, Tao NJ. 1999. Fabrication of stable metallic nanowires with quantized conductance. ***Nanotechnology*** 10:221–23

[8] Morpurgo AF, Marcus CM, Robinson DB. 1999. Controlled fabrication of metallic electrodes with atomic separation. ***Appl. Phys. Lett.*** 14:2084–86

[9] Kervennic YV, Thijssen JM, Vanmaekelbergh D, Dabirian R, Jenneskens LW, et al. 2006. Charge transport in three-terminal molecular junctions incorporating sulfur-end-functionalized tercyclohexylidene spacers. ***Angew. Chem. Int. Ed. Engl.*** 45:2540–42

[10] Li XL, He HX, Xu BQ, Xiao X, Tsui R, et al., 2004. Electron transport properties of electrochemically and mechanically formed molecular junctions. ***Surf. Sci.*** 573:1–10

[11] Lee J-O, Lientschnig G, Wiertz F, Struijk M, Janssen RAJ, et al., 2003. Absence of strong gate effects in electrical measurements on phenylene-based conjugated molecules. ***Nano Lett.*** 3:113–17

[12] Luber SM, Strobel S, Tranitz HP, Wegscheider W, Schuh D, Tornow M. 2005. Nanometre spaced electrodes on a cleaved AlGaAs surface. ***Nanotechnology*** 16:1182–85

[13] Kubatkin S, Danilov A, Hjort M, Cornil J, Bredas J-L, et al., 2003. Singleelectron transistor of a single organic molecule with access to several redox states. ***Nature*** 425:698–701

[14] Zhitenev NB, Meng H, Bao Z. 2002. Conductance of small molecular junctions. ***Phys. Rev. Lett.*** 92:186805

[15] Blum AS, Kushmerick JG, Long DP, Patterson CH, Yang JC, et al., 2005. Molecularly inherent voltage-controlled conductance switching. ***Nat. Mater.*** 4:167–72

[16] Amlani I, Rawlett AM, Nagahara LA, Tsui RK. 2002. An approach to transport measurements of electronic molecules. Appl. ***Phys. Lett.*** 80:2761–63

[17] Guisinger NP, Greene ME, Basu R, Baluch AS, Hersam MC. 2004. Room temperature negative differential resistance through individual organic molecules on silicon surfaces. ***Nano Lett.*** 4:55–59

[18] Piva PG, DiLabio GA, Pitters JL, Zikovsky J, Rezeq M, et al., 2005. Field regulation of single-molecule conductivity by a charged surface atom. ***Nature*** 435:658–61

[19] Reichert J, Ochs R, Beckmann D, Weber HB, Mayor M, L¨ ohneysen HV. 2002. Driving current through single organic molecules. ***Phys. Rev. Lett.*** 88:176804

[20] Weber HB, Reichert J, Weigend F, Ochs R, Beckmann D, et al., 2002. Electronic transport through single conjugated molecules. ***Chem. Phys.*** 281:113–25

[21] Xu BQ, Tao NJ. 2003. Measurement of single molecule conductance by repeated formation of molecular junctions. ***Science*** 301:1221–23

[22] Xu BQ, Xiao XY, Tao NJ. 2003. Measurement of single molecule electromechanical properties. ***J. Am. Chem. Soc.*** 125:16164–65

[23] Moresco F. 2004. Manipulation of large molecules by low-temperature STM: model systems for molecular electronics. ***Phys. Rep.*** 399:175–25

[24] Stipe BC, Rezaei MA, Ho W. 1998. Inducing and viewing the rotation motion of a single molecule. ***Science*** 279:1907–9

[25] Mazur U, Hipps KW. 1999. Orbital-mediated tunneling, inelastic electron tunneling, and electrochemical potentials for metal phthalocyanine thin films. ***J. Phys. Chem. B.*** 103:9721–27

[26] Datta S, Tian WD, Hong SH, Reifenberger R, Henderson JI, Kubiak CP. 1997. Current-voltage characteristics of self-assembled monolayers by scanning tunneling microscopy. ***Phys. Rev. Lett.*** 79:2530–33

[27] Bumm LA., Arnold JJ, Cygan MT, Dunbar TD, Burgin TP, et al., 1996. Are single molecular wires conducting? ***Science*** 271:1705–7

[28] Zhao, Y., Lindsay, S., Jeon, S., Kim, H.-J., Su, L., Lim, B. and Koo, S. (2013), Combined Effect of Polar Substituents on the Electronic Flows in the Carotenoid Molecular Wires. ***Chem. Eur. J.***, 19: 10832–10835.

[29] G. Leatherman, E. N. Durantini, D. Gust, T. A. Moore, A. L. Moore, S. Stone, Z. Zhou, P. Rez, Y. Z. Liu, S. M. Lindsay, ***J. Phys. Chem. B*** 1999, 103, 4006 – 4010

[30] J. He, F. Chen, J. Li, O. F. Sankey, Y. Terazono, C. Herrero, D. Gust, T. A. Moore, A. L. Moore, S. M. Lindsay, ***J. Am. Chem. Soc.*** 2005, 127, 1384 – 1385

[31] Nichols, R.; Haiss, W.; Higgins, S.; Leary, E.; Martin, S.; Bethell, D., The experimental determination of the conductance of single molecules. ***Physical Chemistry Chemical Physics*** 2010, 2801-2815

[32] Elgar G, Vavouri T (July 2008). "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes". ***Trends Genet.*** 24 (7): 344–52

[33] Coon, Joshua J. (April 13, 2009). "Collisions or Electrons? Protein Sequence Analysis in the 21st Century". ***Anal. Chem.*** 81 (9): 3208–3215

[34] Edman P, Begg G (March 1967). "A protein sequenator". ***Eur J Biochem.*** 1 (1): 80–91

[35] Edman, P.; Högfeldt, Erik; Sillén, Lars Gunnar; Kinell, Per-Olof (1950). "Method for determination of the amino acid sequence in peptides". ***Acta Chem. Scand.*** 4: 283–293.

[36] Hao, L. Ph.D thesis, Arizona State University, Tempe, 2013

[37] Love, J. C.; Estroff, L. A.; Kriebel, J. K.; Nuzzo, R. G.; Whitesides, G. M., SelfAssembled Monolayers of Thiolates on Metals as a Form of Nanotechnology. **Chemical Reviews** 2005, 105 (4), 1103-1170.

[38] FTIR User Guide. http://mmrc.caltech.edu/FTIR/Nicolet/Nicolet%20Software/Nicolet%202/4700_6700_User.pdf

[39] Shuai Chang, Jin He, Peiming Zhang, Brett Gyarfas, and Stuart Lindsay (2011) Gap Distance and Interactions in a Molecular Tunnel Junction. ***Journal of the American Chemical Society*** 2011, 133 (36), pp 14267–14269

[40] Kwon, S. et al., Characterization of cyclodextrin complexes of camostat mesylate by ESI mass spectrometry and NMR spectroscopy. ***Journal of Molecular Structure*** 938, 192-197 (2009).

[41] Brivio, M., Oosterbroek, R. E., Verboom, W., van den Berg, A. & Reinhoudt, D. N. Simple chip-based interfaces for on-line monitoring of supramolecular interactions by nano-ESI MS. ***Lab Chip*** 5, 1111-1122 (2005).

[42] Cooks, R. G., Zhang, D., Koch, K. J., Gozzo, F. C. & Eberlin, M. N. Chiroselective Self-Directed Octamerization of Serine: Implications for Homochirogenesis. ***Anal. Chem.*** 73, 3646-3655 (2001).

[43] Koch, K. J. et al., Chiral Transmission between Amino Acids: Chirally Selective Amino Acid Substitution in the Serine Octamer as a Possible Step in Homochirogenesis. ***Angew. Chem. Int. Ed.*** 41, 1721-1724 (2002).

[44] Qiu, B., Liu, J., Qin, Z., Wang, G. & Luo, H. Quintets of uracil and thymine: a novel structure of nucleobase self-assembly studied by electrospray ionization mass spectrometry. ***Chem. Commun.*** (Camb) 20, 2863-2865 (2009).

[45] Sherman, C. L., Brodbelt, J. S., Marchand, A. P. & Poola, B. Electrospray ionization mass spectrometric detection of self-assembly of a crown ether complex directed by π-stacking interactions. ***Journal of the American Society for Mass Spectrometry*** 16, 1162-1171 (2005).

[46] Huang, S.et al., Identifying single bases in a DNA oligomer with electron tunneling. ***Nature Nanotech.*** 5, 868–873 (2010).

[47] Chang, S. et al., Chemical Recognition and Binding Kinetics in a Functionalized Tunnel Junction. ***Nanotechnology*** 23, 235101-235115 (2012).

[48] Sreekumar, A.et al., Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. ***Nature*** 457, 910–914 (2009).

[49] Uhlen, M. & Ponten, F. Antibody-based proteomics for human tissue profiling. ***Mol. Cell. Proteom.*** 4, 384–393 (2005)

[50] Friddle, R. W., Noy, A. & De Yoreoa, J. J. Interpreting the widespread nonlinear force spectra of intermolecular bonds. ***Proc. Natl. Acad. Sci.*** USA 109, 13573–13578 (2012).

[51] Fuhrmann, A.et al., Long lifetime of hydrogen-bonded DNA basepairs by force spectroscopy. ***Biophys. J.*** 102, 2381–2390 (2012).

[52] Ku¨hnle, A., Linderoth, T. R., Hammer, B. & Besenbacher, F. Chiral recognitionin dimerization of adsorbed cysteine observed by scanning tunneling microscopy. ***Nature*** 415,891–893 (2002).

[53] Analysis of phosphorylated proteins and peptides by mass spectrometry, Derek T McLachlin, Brian T Chait, ***Current Opinion in Chemical Biology*** 2001, 5:591–602

[54] Marks F (Ed): Protein Phosphorylation. New York: VCH; 1996.

[55] Hunter T: Signaling — 2000 and beyond. ***Cell*** 2000, 100:113-127.

[56] Interleukin-3 protects Bcr-Abl-transformed hematopoietic progenitor cells from apoptosis induced by Bcr-Abl tyrosine kinase inhibitors, J Wu, etc, ***Leukemia*** (2002) 16, 1589-1595.

[57] An anticancer C-Kit kinase inhibitor is reengineered to make it more active and less cardiotoxic, Ariel Fernandez, etc. ***J. Clin. Invest.*** 2007, 117(12):4044–4054

[58] Shuo, H. Ph.D thesis, Arizona State University, 2011

[59] Shuai, C. Ph.D thesis, Arizona State University, 2012

[60] Shuai C, Shuo H, He J, Liang F, Zhang PM, Li SQ, Chen X, Sankey O, Lindsay SM (2010) Electronic Signatures of all Four DNA Nucleosides in a Tunneling Gap. ***Nano letters*** 10(3), 1070-1075

[61] Lee, M.; Sankey, O., Insights into electron tunneling across hydrogen-bonded base-pairs in complete molecular circuits for single-stranded DNA sequencing. ***Journal of Physics-Condensed Matter*** 2009.

[62] Lee, M.; Sankey, O., Theory of tunneling across hydrogen-bonded base pairs for DNA recognition and sequencing. ***Physical Review E*** 2009.

[63] Tom M. Mitchell. Machine Learning. McGraw Hill. 1997. Chapter 3

[64] C.J.C Burges, A tutorial on support vector machines for pattern recognition, ***Data mining and knowledge discovery***, 1998, 2(2): 121~167

[65] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". ***Machine Learning*** 20 (3): 273

[66] Aizerman, Mark A.; Braverman, Emmanuel M.; and Rozonoer, Lev I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". ***Automation and Remote Control*** 25: 821–837

[67] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. p. 144.

[68] Michael Tuchband, Jin He, Shuo Huang, and Stuart Lindsay (2012) Insulated gold scanning tunneling microscopy probes for recognition tunneling in an aqueous environm. ***Review of Scientific Instruments*** 2012, 83, 015102

[69] Lundstrom, M., Moore's law forever? ***Science*** 2003, 210-211

[70] http://en.wikipedia.org/wiki/Transistor_count

[71] http://en.wikipedia.org/wiki/Semiconductor_device_fabrication

[72] Dadosh T, Gordin Y, Krahne R, Khivrich I, Mahalu D, et al. 2005. Measurement of the conductance of single conjugated molecules. ***Nature*** 436:677–80

[73] Yanan Z. et al., Single-molecule spectroscopy of amino acids and peptides by recognition tunneling, ***Nature Nanotech.*** 2014, 10.1038

[74] Michael Z. Ludwig. 2002. Functional evolution of noncoding DNA. ***Current Opinion in Genetics & Development***, Vol. 12, Issue 6, Pg634-639

[75] Christine P Bird, Barbara E Stranger, Emmanouil T Dermitzakis, Functional variation and evolution of non-coding DNA, ***Current Opinion in Genetics & Development***, Volume 16, Issue 6, December 2006, Pages 559-564

[76] Chen, W., X. Yin, J.Q. Mu, and Y. Yin, Subfemtomole level protein sequencing by Edman degradation carried out in a microfluidic chip. ***Chem. Commun.***, 2007: p. 2488-2490

[77] Tanaka, H.; Kawai, T., Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. ***Nature Nanotechnology*** 2009, 518-522

[78] Love, J.; Estroff, L.; Kriebel, J.; Nuzzo, R.; Whitesides, G., Selfassembled monolayers of thiolates on metals as a form of nanotechnology. ***Chemical Reviews*** 2005, 1103-1169

[79] He, J.; Lin, L.; Zhang, P.; Spadola, Q.; Xi, Z.; Fu, Q.; Lindsay, S., Transverse tunneling through DNA hydrogen bonded to an electrode. ***Nano Letters*** 2008, 2530-2534

[80] Ramachandran, G.; Hopson, T.; Rawlett, A.; Nagahara, L.; Primak, A.; Lindsay, S., A bond-fluctuation mechanism for stochastic switching in wired molecules. ***Science*** 2003, 1413-1416

[81] Xiao, X.; Xu, B.; Tao, N., Conductance titration of single-peptide molecules. ***Journal of the American Chemical Society*** 2004, 5370-5371

[82] Venkataraman, L.; Klare, J.; Tam, I.; Nuckolls, C.; Hybertsen, M.; Steigerwald, M., Single-molecule circuits with well-defined molecular conductance. ***Nano Letters*** 2006, 458-462

[83] Zwolak, M.; Di Ventra, M., Colloquium: Physical approaches to DNA sequencing and detection. ***Reviews of Modern Physics*** 2008, 141-165

[84] Zwolak, M.; Di Ventra, M., Electronic signature of DNA nucleotides via transverse transport. ***Nano Letters*** 2005, 421-424

[85] He, J.; Lin, L.; Zhang, P.; Lindsay, S., Identification of DNA base pairing via tunnel-current decay. ***Nano Letters*** 2007, 3854-3858

[86] Chang, S.; He, J.; Kibel, A.; Lee, M.; Sankey, O.; Zhang, P.; Lindsay, S., Tunneling readout of hydrogen-bonding-based recognition. ***Nature Nanotechnology*** 2009, 297-301

[87] Huang, S., Recognition Tunneling Measurement of the Conductance of DNA Bases Embedded in Self-Assembled Monolayers. ***The Journal of Physical Chemistry C*** 2010, 114 (48), 20443–20448

[88] Tao, N., Electron transport in molecular junctions. ***Nature Nanotechnology*** 2006, 173-181

[89] Xia, J.; Diez-Perez, I.; Tao, N., Electron transport in single molecules measured by a distance-modulation assisted break junction method. ***Nano Letters*** 2008, 1960-1964

[90] Bayley, H., Sequencing single molecules of DNA. ***Current Opinion in Chemical Biology*** 2006, 628-637

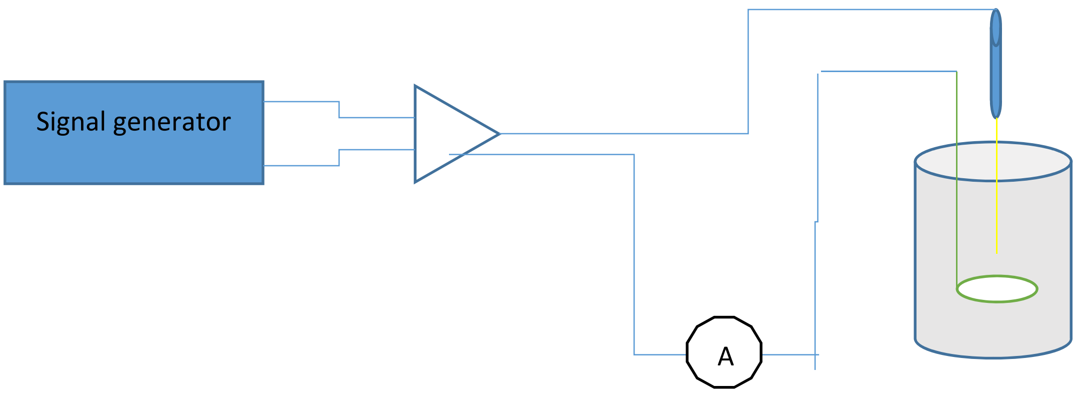
[91] He, J.; Chen, F.; Li, J.; Sankey, O. F.; Terazono, Y.; Herrero, C.; Gust, D.; Moore, T. A.; Moore, A. L.; Lindsay, S. M. (2005) Electronic Decay Constant of Carotenoid Polyenes from Single-Molecule Measurements. ***J. Am. Chem. Soc.*** 127(5), 1384-1385.

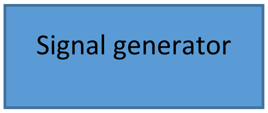
[92] Cui, X.D., et al., Changes in the Electronic Properties of a Molecule when it is wired into a circuit. ***J. Phys. Chem. B*** 106 8609-8614, 2002.

APPENDIX A

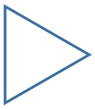
PROTOCOL FOR MARKING STM GOLD TIPS

Setup:

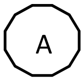




: Signal generator



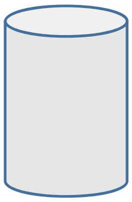
: Amplifier



: Ammeter

: Tip holder

: Platinum ring



: Container filled with HCL and ethanol mixture

The system is consisted of a signal generator, an amplifier, an ammeter, a tip holder, a platinum ring and a gold wire that needs to be shaped (diameter = 0.25mm, purity = 99.999%). The etching solution is mixture of HCl (37%) and ethanol (1:1 mixture).

The diameter of the platinum ring used in this experiment is 1.6cm.

Procedure of making tips (specified for Lindsay's Lab equipment):

1. Set up the connection:

   For the signal generator, the parameters are as following:

   Waveform: Square wave

   Frequency: 5 KHz

   Amplitude: 1.05V-1.60V

   For the amplifier:

   Set the Gain to 3dB

2. Sharp the wire:

   Dip the Au wire into the mixture rapidly and stop dipping the wire when the current reaches to some value around 350mA~400mA. Wait around 1~2min until the etching process stops.

3. Re-etch the tip:

   Decrease the amplitude of the square wave to some value around 0.35V~0.53V. Dip the tip into the mixture rapidly again and raise the tip as quickly as you can once the current reaches to some value around 50mA~80mA. So, the tip needle of the wire would be re-sharped.

4. Check the shape of the tip:

   Check the tip under microscope, if the tip is not sharp enough, repeat step 2 and step 3.

Supplement material:

- ❖ For step 1:

  The key parameter of making good tips is the amplitude of the square wave from the signal generator and the Gain of the amplifier. If we set the amplitude of the square wave to 1.55V and the Gain is set to 3dB, then the amplitude of the signal after passing the amplifier should be $V_{out} = 1V*10^{3/20} = 1.41V$. However, the measured amplitude is 29.5V, which indicates that the Gain should be 25.6dB. Let's stick to 25.6dB as the real Gain because this setup uses a CD player as the amplifier and that amplifier probably gives the wrong Gain. So, in other setups, once the voltage after amplified is between [20V, 30V], that setup should work. Either higher (fast etching) or lower (slow etching) voltage would result in a blunt tip.

- ❖ For step 2:

  The proper current may vary from device to device (depends on the amplitude of the voltage, the size of the platinum rings, the length of tip dipped in the mixture), while a higher current usually indicates a longer sharp tip.

- ❖ For step 3:

  Try to raise the tip as quickly as you can. Otherwise the sharp needle would be etched and what you will observe under microscope is a blunt tip. The corresponding voltage after the amplifier could be calculated by using this formula:

Gain = 20*lg (V_{out}/V_{in})

While, Gain = 25.6

The voltage range after the amplifier is [6.67V, 10V]

APPENDIX B

STANDARD PROTOCOL FOR RUNNING EXPERIMENT

1. Insert the tip into the scanner and connect to the microscope head.
2. Zero the current by using the screw on the top of the scanner while checking the physical oscilloscope.
3. If PicoView is not reading 0pA then adjust the offset in the PicoView software. At this point the oscilloscope, PicoView, and LabView should all be reading 0pA/0V.
4. Add 120 µl PB solution to the liquid cell.
5. Verify that no more than 1pA of leakage current is showing with an applied bias of -0.5V. If more than 1pA discard tip and try another. Also check for 60Hz noise, if any is visible then discard the tip.
6. Change the bias from -0.5V to 0V, if the current decay lasts longer than 0.1sec discard the tip, it is most likely unexposed.
7. After approach (4pA, -0.5V, I=1, P=1) and withdraw of 20 µm, stabilize the system for 2 hours (-0.1V).
8. If necessary, re-zero the scanner using the screw.
9. Approach surface (4pA, -0.5V, I=1, P=1), wait 5 min.
10. Gather control signal (4pA, -0.5V, I=0.1, P=0.1). Withdraw tip.
11. Add amino acid solution (100 µM) you are going to study to the liquid cell.
12. Wait 1 hour to re-stable the system.
13. Approach the tip (4pA, -0.5V, I=1, P=1) and wait 2~5 minutes begin capture of data.

14. Set the data recording to 300 seconds of data and capture at least 12 data files corresponding to different locations on the substrate for each molecule.

BIOGRAPHICAL SKETCH

Yanan Zhao was born on February, 11th, 1986 in XuChang, China. He received his Bachelor of Science in Physics in University of Science and Technology of China in 2008. In 2008, Yanan Zhao applied the Ph.D program of Physics in Arizona State University and got admitted to pursue his Ph.D degree in Physics. He joined Dr. Stuart Lindsay's group later and studied applying recognition tunneling in the areas of molecule identification especially in DNA and Protein sequencing. During the study in ASU, he maintained good academic records, awarded with Molecular Imaging Corporation fellowship and published two papers in the best journals including Nature Nanotechnology. He also filed a patent about a revolutionary peptide sequencing system.