Electronic Single Molecule Measurements with the Scanning Tunneling Microscope

by

Jong One Im

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

Approved April 2016 by the
Graduate Supervisory Committee:

Stuart Lindsay, Chair
Peiming Zhang
Robert Ros
Ralph Chamberlin

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Richard Feynman said “There’s plenty of room at the bottom”. This inspired the techniques to improve the single molecule measurements. Since the first single molecule study was in 1961, it has been developed in various field and evolved into powerful tools to understand chemical and biological property of molecules. This thesis demonstrates electronic single molecule measurement with Scanning Tunneling Microscopy (STM) and two of applications of STM; Break Junction (BJ) and Recognition Tunneling (RT). First, the two series of carotenoid molecules with four different substituents were investigated to show how substituents relate to the conductance and molecular structure. The measured conductance by STM-BJ shows that Nitrogen induces molecular twist of phenyl distal substituents and conductivity increasing rather than Carbon. Also, the conductivity is adjustable by replacing the sort of residues at phenyl substituents. Next, amino acids and peptides were identified through STM-RT. The distribution of the intuitive features (such as amplitude or width) are mostly overlapped and gives only a little bit higher separation probability than random separation. By generating some features in frequency and cepstrum domain, the classification accuracy was dramatically increased. Because of large data size and many features, supporting vector machine (machine learning algorithm for big data) was used to identify the analyte from a data pool of all analytes RT data. The STM-RT opens a possibility of molecular sequencing in single molecule level. Similarly, carbohydrates were studied by STM-RT. Carbohydrates are difficult to read the sequence, due to their huge number of possible isomeric configurations. This study shows that STM-RT can identify not only isomers of mono-saccharides and disaccharides, but also various mono-saccharides from a data pool of eleven analytes. In addition, the binding affinity between recognition molecule and analyte was investigated by comparing with surface plasmon resonance. In present, the RT technique is applying to chip type sequencing device onto solid-state nanopore to read out glycosaminoglycans which is ubiquitous to all mammalian cells and controls biological activities.

ACKNOWLEDGMENTS

There are many people who have helped me for the works in this thesis and I would like to give my gratitude to all of them. First, I would like to thank my advisor Professor Stuart Lindsay for his endless and priceless support and guidance. Dr. Lindsay has given to me a precious chance to join this research group and explore the cutting edge techniques. Also I want to express thanks to my committee members Professors Ralph Chamberlin, Peiming Zhang, and Robert Ros.

I would like to express my big gratitude to Dr. Brian Ashcroft for his computational works to analyze the complicate RT data. He has made a robust powerful analysis program using Support Vector Machine and helped me to understand Machine Learning algorithms.

Besides, I give special thanks to Dr. Peiming Zhang for his support in Chemistry. He suggested the promising research to identify Glycosaminoglycan and guided me to design proper experimental procedures.

I want to thank my friends and colleagues from Dr. Lindsay's lab including Dr. Brett Gyarfas, Dr. Shuai Chang, Dr. Pei Pang, Dr. Hao Liu, Dr. Yanan Zhao, Dr. Weisi Song, and Suman Sen for their knowledges and useful discussions about STM-RT experiments. I have to thank to Sovan Biswas and Suddipta Biswas for their helps in organic synthesis and chemistry experiments. I also appreciate to Margaret Black and Michael Dodson for their technical supports in lab.

My gratitude should be given to other friends and colleagues out of Dr. Lindsay's lab including Dr. Inho Kim, Dr. Sangho Koo, Dr. Christopher Bruot, and Liming Xiang for their dedications.

Also, I sincerely thanks to my wife, JuYoung Kim, for her love and efforts to manage sweet home, and my daughter, SeoHyun Im, for her lovely cute smile, and my parents and brother for their supports.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION TO SINGLE MOLECULE MEASREMENTS

1.1. Methods of Single Molecule Measurements

Single molecule research has been evolved into powerful tools to understand chemical and biological property of molecules, due to their unique abilities not only high sensitivity and the insight of molecular features. The conventional ensemble methods provide averaged properties of large numbers of molecules. In contrast with the conventional methods, single molecule methods allow to directly measure properties of individual molecule through measuring molecular force or molecular functional responses to mechanical manipulation etc. High sensitivity of single molecule measurements have direct benefits. Some molecules are easily aggregated in ensemble concentration. Single molecule methods can be done at low concentration, and it allows to study the property of the monomeric species at equilibrium. [1][2]

Richard Feynman said “There’s plenty of room at the bottom”. [3] This inspired the techniques to improve the single molecule detection limit. The first single molecule study was in 1961, “Measurement of activity of single molecules of beta-D-galactosidase”, which observed single beta-D-galactosidase in microdroplets and on fluorogenic substrate. [4][5]

Single molecule measurement techniques have been developed in various fields from physics to biology. As described in figure 1, single molecule techniques can be grouped into two regions by aims of the methods. One is for developing and improving methods and the other is for addressing scientific questions. [2] For instance, near-field approaches such as scanning tunneling microscopy (STM) and atomic force microscopy (AFM), and optical microscopy of single quantum systems such as ion traps and atom traps, and confocal microscopy with fluorescence, and optical tweezers etc. [4] Some of these pioneer researches was selected for Nobel Prize. Though some techniques require extreme conditions, such as low temperature (~1K) or ultra-high vacuum (~10^{-9} Torr), others can be performed in liquid at room temperature. Some methods is able to directly measure the molecule, while others need to label targets for measurable interaction, which may be larger than the target molecules. Also diverse molecules have been studied from single atoms to complicate living cells. [1][6][7][8]

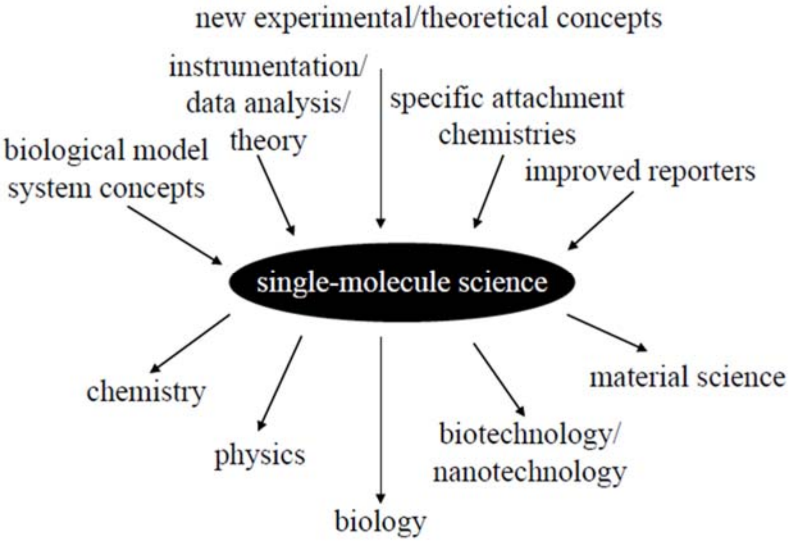

new experimental/theoretical concepts
instrumentation/ data analysis/ theory
specific attachment chemistries
biological model system concepts
improved reporters
single-molecule science
chemistry
material science
physics
biotechnology/ nanotechnology
biology

Figure 1. Components and scope of single molecule science.
Figure taken from ref [2].

Electron is a good probe at short distance scales for single molecule measurements. Its tunneling behavior allows us measurements in a few nanometers with angstrom sensitivity. Charge transport is an important mechanism in chemical and biological processes. Understanding charge transport provides chemical information in single molecule level [9] and it can be used for molecular electronics and sensor applications which are based on electrical detection of molecular binding events [10]. [1][11] The thesis is based on electronic single molecule measures through STM applications. The details are discussed in the following chapters.

1.1.1. Electron Based Measurements

Electron is a very attractive probe for direct and label free detections. STM and TEM are the most popular techniques for electron based single molecule measurements. Some of nano-gap techniques including STM is useful to measure molecular conductance, vibrational energy levels, electronic polarizability and spin states [12][13]. TEM uses high energy electron beams through the sample. Crystallographic structure can be analyzed by diffraction patterns. [1]

The quantum tunneling allows that electrons can be transported across a nanometer-gap such as molecule or insulating layer between electrodes. Electrons on the negatively charged electrode are more likely to be moved to the other electrode positively charged. Then the gap size is critical to tunneling current. The larger gap makes harder to transport electrons, so current decreases exponentially as gap increases. If the gap is filled by a conductive medium instead of non-conducting insulator such as vacuum or air, charges can be transferred easily then the tunneling current increases. Some detail discussions about the tunneling current will be in chapter 1.2. [14]

The sensitivity of tunneling current to the gap size and the medium in the gap means that solution environments can give a range of opportunities to study chemical reactions and to sequence the composition of biopolymers. Liquid environments makes easier to prepare biomolecules such as DNA and proteins. Also, analyte molecules in solution are easily saturated on the electrode surface and it increases the stability of the molecular junction. In contrast to fluorescence spectroscopy, the tunneling based methods are label-free due to the medium sensitivity since molecules have their own unique electronic structure. [14]

1.1.1.1. Scanning Tunneling Microscopy (STM)

As discussed before, STM is based on electron tunneling phenomena at small gap between two electrodes while a bias applied. Because of small gap size (a few nano-meter) and small tunneling current (sub nano-ampere), STM requires precise tip position controller and amplifier as shown in Figure 2. Thus the surface image can be acquired with precise resolution in angstrom unit. The tip position is controlled by piezo-electric material with feedback servo loop to maintain constant current or gap distance. The tunneling current exponentially decreases as the gap size increases. Also the electronic structure of the molecular junction is critical to the tunneling current. It makes that STM can quantify molecule energy level. [1][11]

STM is a most widely used tool for single molecule detection and manipulation, because of the resolution and sensitivity. STM can image an individual molecule which is absorbed on the substrate and is able to manipulate single molecule even an atom on the substrate [15]. By modifying the tip with a molecule, one can offer spectroscopy measurement [16]. [11]

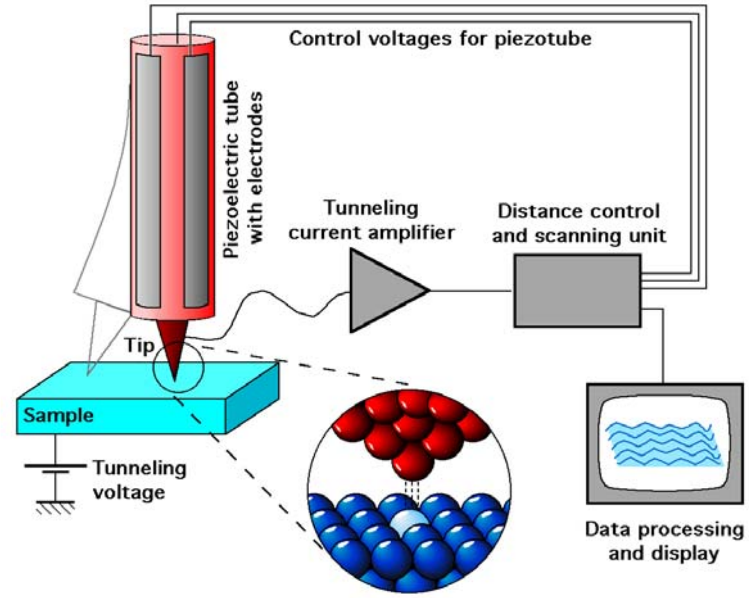

Control voltages for piezotube
Piezoelectric tube with electrodes
Tunneling current amplifier
Distance control and scanning unit
Tip
Sample
Tunneling voltage
Data processing and display

Figure 2. Schematics of Scanning Tunneling Microscopy (STM).

STM is based in tunneling phenomena between metal tip and metal sample. At a few nanometer gap, a few pico-ampere of tunneling current is induced. Amplifier is required to read very small tunneling current, and the small gap is maintained by feedback controller and piezotube. Figure taken from ref [17].

1.1.1.2. Fixed Electrodes

STM is a fascinate tool for single molecule measurements by the facts of the sensitivity, easy to control the gap size, and low operating cost. However, it has a noise issue which comes from the feedback servo controller. It reduces the signal-to-noise ratio of STM. A simple way to over the limit is to fabricate facing electrodes separated by nanogap in molecular scale corresponding to a few nanometers. The fixed electrodes which are fabricated on a solid substrate provide mechanically stable molecular junctions. There are two types of devices, planar and vertical nanogap devices, as shown in Figure 3.

However, fabricating nanogap is beyond the present nanofabrication techniques limit. Researchers have been developed to fabricate such a molecular scale nanogap junctions, or to use nanoparticles interacting with analyte and electrodes as described in Figure 4(b) and 4(c). Not only low fabrication yield due to difficulty, but also large device to device variation should be improved. [11][18]

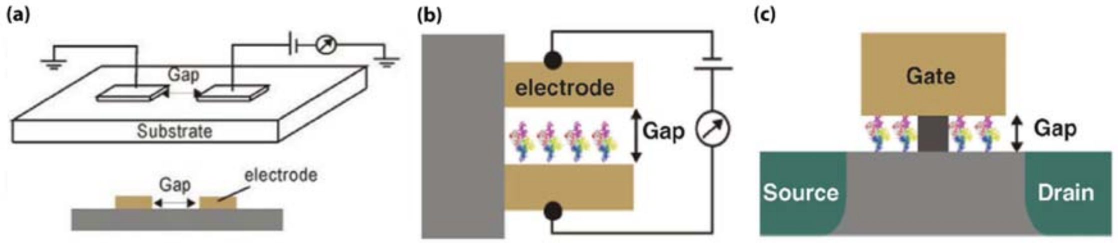


Figure 3. Schematics of Planar and Vertical Nanogap Devices.

(a) Planar nanogap device with two electrodes (b) Vertical nanogap device with two electrodes (c) Vertical nanogap device with three electrodes; Source, Gate, and Drain. Figure taken from ref [18].

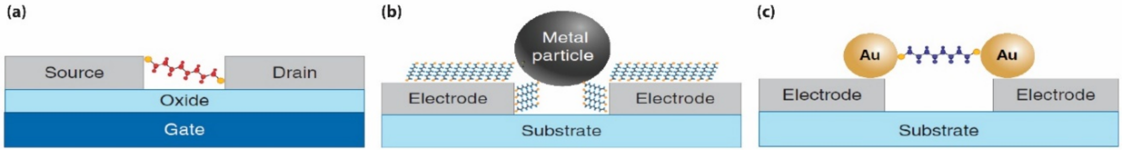


Figure 4. Schematics of Different Molecular Junction Mechanisms.

(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) Dimer structure, consisting of two Au particles bridged with a molecule, assembled across two electrodes. Figure taken from ref [11].

1.1.1.3. Mechanically Controllable Break Junction (MCBJ)

As discussed in the previous section, fixed nanogap fabrication has a limit to make molecular scale separation between electrodes. MCBJ is a novel way to generate nanogap electrodes by mechanically breaking metal wire which is controlled by bending substrate as described in Figure 5. [19] The rod displacement is converted to the gap distance between electrodes with the attenuation factor r. MCBJ gives high stability holding single molecule at room temperatures [20].

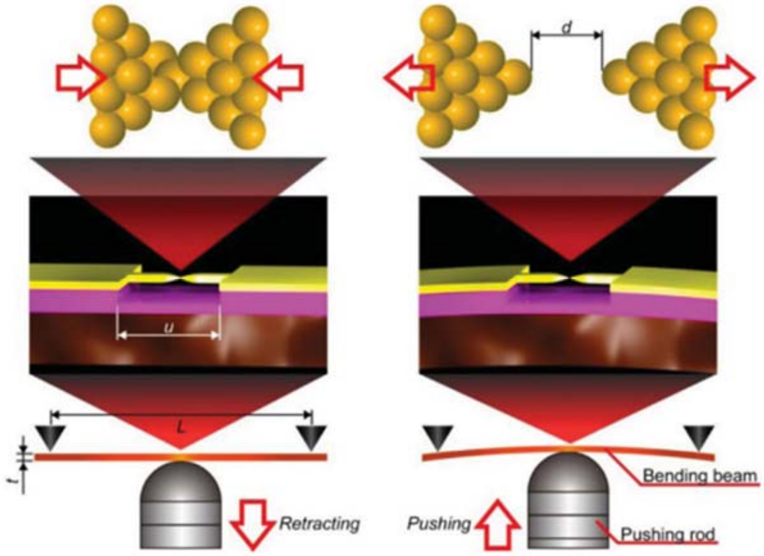


Figure 5. Schematics of MCBJ Fabrication.

A free standing metal junction is fabricated by e-beam lithography and metal deposition on the substrate. The substrate and pushing rod are in a three-point bending configuration, the vertical movement of the pushing rod is controlled by PZT and stepping motor. The gap size d can be tuned by the vertical displacement of the pushing rod D by d=rD, where r is the attenuation factor determined by the configuration as r=3ut/L. Figure taken from ref [19].

1.1.1.4. Conducting Atomic Force Microscopy (CAFM)

AFM was invented to overcome the limit of STM which cannot perform the non-conducting samples. Though AFM is working on the force interaction between cantilever tip and sample, by modifying the probe and substrate with conducting material, it provides conductance spectroscopy of the molecular junction simultaneously. The break-down of individual molecular junction generates step change in the conductance curves with abrupt force decrease shown in Figure 6(b). Cui et. al. showed that the molecular junction conductance is quantized and it is reproducible with CAFM. [21] The average breakdown force of the stepwise conductance decrease is ~1.5 nN [22] which is the same with the Au-Au bond breaking. It implies that the breakdown occurs at the Au-Au bond. [11]

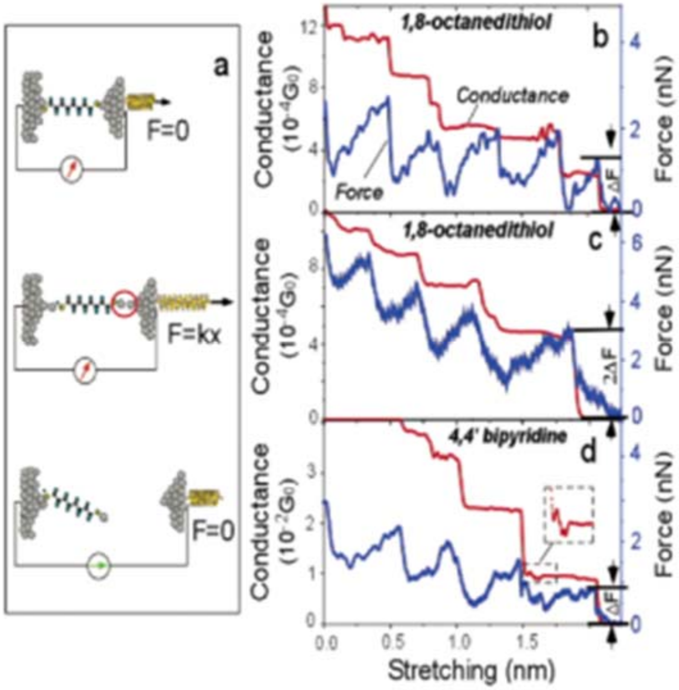


Figure 6. Schematics of Conducting Atomic Force Microscopy (CAFM).

(a) Schematic illustration of a molecule covalently bonded to two Au electrodes under mechanical stretching, during which both the conductance and the force are measured. (b-d) Simultaneously recorded conductance and force curves of C8 (b-c) and BPY junctions (d) during stretching. (c) shows that two molecules can break simultaneously at the last stage, resulting in twice as much change in the conductance and the force. The inset in (d) shows that the force fluctuations are correlated with conductance fluctuations. [11], Figure taken from ref [23].

1.1.1.5. Nanopore

The idea of nanopore method has been originated from patch clamp experiments in 1970s which is possible to measure the currents of single ion channel molecules. The molecular property is measured by passing the analytes through a nanometer scale pore which gives millisecond scale switching signals in sub-nanoampere order of ionic currents with kilohertz frequency. [24]

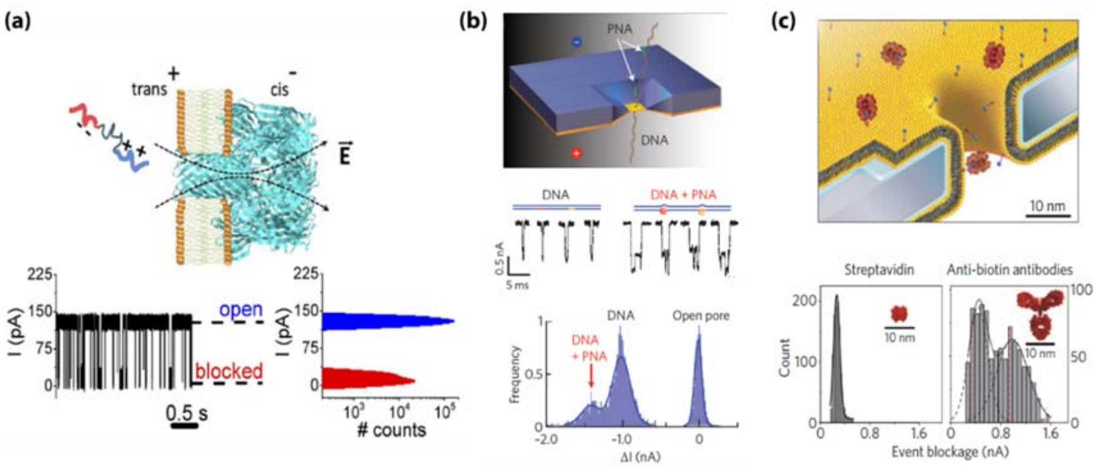


Figure 7. Schematics of Nanopore Mechanism.

(a) biological pore (b) solid state pore (c) hybrid pore. Figures taken from ref *[25] and [26]*.

The nanopore fabrication techniques can be grouped into biological, solid state, and hybrid methods. The protein pores are formed in a membrane such as lipid bilayer. The solid state pores are fabricated on synthetic materials substrate such as silicon nitride or graphene. Also, the protein pores can be formed in electrically resistant membrane bilayer (Figure 7(c)). The Oxford Nanopore sequencing kit which is the first commercially available pore based sequencing device, is fabricated by hybrid methods. The array of alpha-hemolysin protein pores is set in an electrically resistant polymer membrane. [1][25][26]

1.2. Charge Transport in Molecular Junctions

In nano-scale study, the characteristic length scale of the molecular junction is important. According to the Ohm's law, it states that the current through a material is proportional to the voltage and the constant proportionality is the resistance defined by

$$R = \rho \frac{L}{A} \qquad equation\ (2)$$

where ρ is the resistivity of the material, L is the length of the material, and A is the cross-section area.

Then the macroscopic conductance is relative with the current and applied voltage. Then the conductance, the inverse of the resistance, is

$$G = \sigma \frac{A}{L} \qquad equation\ (3)$$

where σ is the conductivity of the material which is the inverse of the resistivity.

In the diffusive regime when junction size is larger than the electron mean free path (*l*), the electrons behave like a random walk between elastic collisions with impurities. However, in the ballistic regime that junction is smaller than *l*, the momentum of electrons becomes constant. In addition, the phase coherence length (L_{φ}) must be considered at small scale, which is the distance that electron's phase information can be preserved. The materials in this scale have coherent charge transport. [27][28]

The following sections will introduce the basic models for the quantum tunneling and Landauer formula.

1.2.1. Quantum Tunneling

The quantum tunneling means that electrons are able to transport across a nano-meter scale gap, insulating layer such as molecule between electrodes. By applying bias voltage onto the source, some of electrons in source have higher energy than drain side, then they can be transported towards the drain electrode, even their energy is smaller than molecular junction barrier. The tunneling exponentially depends on the gap size, the applied bias voltage, and the electronic structure of the bridge junction. The figure 8 shows the schematics of molecular junction circuit (Figure 8(a)) and the relative energy diagrams for electrons energy level (Figure 8(b)) and wave propagation (Figure 8(c)) in a square potential barrier. The source and drain electrodes can be regarded as electron reservoirs bridged by molecule. Then the electrons act like free plane wave propagation through electrodes. When the waves are scattered into the barrier, there is a probability that the electrons will be transmitted through barrier even with the less energy than the barrier. Considered that electrons incident from left to barrier, as described in Figure 8c, the electrons have a probability to be transmitted through the barrier and reflected from the barrier. From the scattering theory, the transmission probability is given by

$$T = \left|\frac{transmitted}{incident}\right|^2 \qquad equation\ (4)$$

For the STM, there were many efforts to predict the tunneling current. In 1961, Bardeen explained the tunneling phenomena as the net of many independent scattering. [29] His theory is based on some assumptions. Two of them is from Oppenheimer’s perturbation theory:

(O-1) tunneling is weak enough that the first-order approximation.

(O-2) tip and sample states are nearly orthogonal.

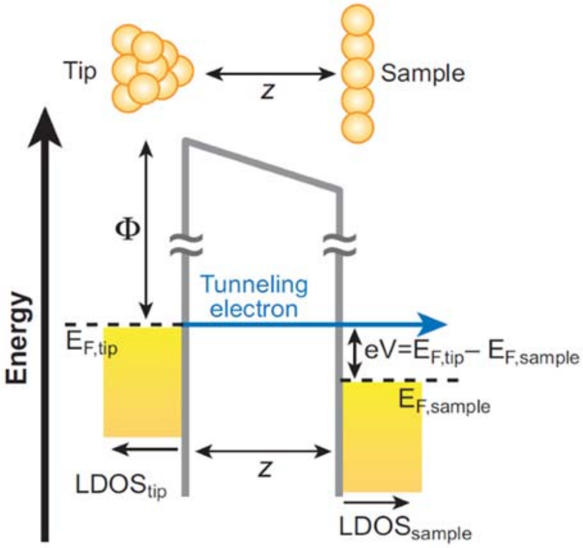


Figure 8. Schematics of Tunneling Current.

An energy level diagram for a one-dimensional electron-tunneling junction. The Fermi energy levels (E_F) of the tip and sample are offset by the applied bias voltage (V) times the electron charge (e). The resultant current is exponentially dependent on the distance between the sample and the tip (z). LDOS, local density of states; Φ, work function of the metal. Figure taken from ref [12].

In addition, he introduced several further assumptions:

(B-1) the electron-electron interaction can be ignored.

(B-2) occupation probabilities for the tip and sample are independent of each other, and do not change, despite the tunneling

(B-3) the tip and sample are each in electrochemical equilibrium.

Thus, the tunneling current can be written as

$$I = \frac{2\pi e}{\hbar}\rho_{tip}\sum_{ts}|M_{ts}|^2 \qquad equation\ (5)$$

where $M_{ts} = \frac{\hbar}{2m} \int dS \, (\psi_t{}^* \nabla \psi_s - \psi_s \nabla \psi_t{}^*)$. Though his model gives valuable intuition for STM imaging, it requires wave functions of the tip and sample states to predict tunneling current.

In 1983, Tersoff and Hamann applied Bardeen's formula to the STM [30] by modelling the electronic wavefunctions of the tip by radially symmetric and Chen improved their model [31] by regarding the tip wavefunction outside of the tip region as a linear combination of the generalized wavefunction and its partial derivatives. For low bias, the tunneling current can be written by

$$I = \frac{eh^3}{m^2} (\mu_t - \mu_s) \cdot \rho_t \cdot \rho_s(0) \qquad equation\ (6)$$

where $\rho_s(0)$ is the local density of sample states per unit volume. Thus the Tersoff-Hamann model predicts that tunneling current is proportional to the local density of sample states.

1.2.2. Landauer Formula

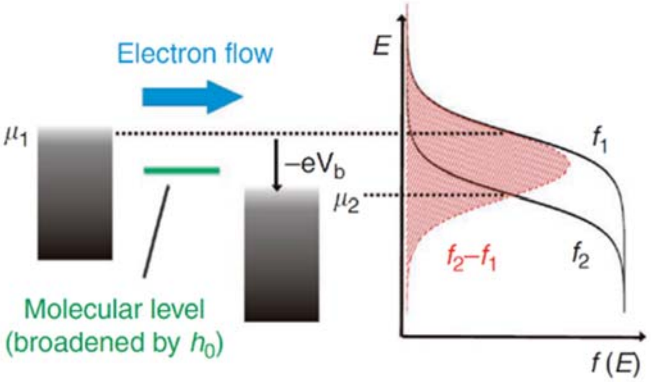


Figure 9. Schematics of Molecular Junction's Energy Level and Corresponding Fermi Distribution.

Two electrodes (labeled by 1 and 2) which separated by potential barrier $-eV_b$ have Fermi energy levels μ_1 and μ_2. The molecular energy level is broadened by h_0 and the Fermi functions of the electrons on two electrodes are shown. The difference of electrodes are shown in red. Figure taken from ref [14].

Landauer theory well explains the charge transportation in quantum system. [32] Considered the molecular junction which shown in Figure 9, the current is the electron flow between electrodes 1 and 2 induced by the Fermi distribution difference. Then the flux is proportional to a transmission function $T(E)$ and the function characterizes the molecular junction property. The net current can be calculated by

$$I_{net} = \frac{q}{h}\int_{-\infty}^{+\infty} dE\ T(E) \cdot \{f_1(E) - f_2(E)\} \qquad equation\ (7)$$

where q is the charge, $f_1 = [1 + \exp(E - \mu_1)/k_B T]^{-1}$, and $f_2 = [1 + \exp(E - \mu_2)/k_B T]^{-1}$ with $\mu_2 = \mu_1 + qV_b$.

The transmission function $T(E)$ can be calculated using the Green's function formalism as

$$T(E) = trace[\Gamma_2 G \Gamma_1 G^+] \qquad equation\ (8)$$

where $\Gamma_{1,2}$ is the broadening matrix of the electrodes, G and G^{+} are the retarded and advanced Green's function of the junction. Taking the assumption that two electrodes are coupled with a single channel of length d and energy h_0 [33], the broadening matrices, the retarded, and the advanced Green's functions are given by

$$\Gamma_{1,2} = \frac{\hbar \cdot v}{d} \qquad \textit{equation}\ (9)$$

$$G = \frac{d}{-i\hbar v - U_0}\ \textit{and}\ G^{+} = \frac{d}{i\hbar v - U_0} \qquad \textit{equation}\ (10)$$

where v is the velocity of electrons and U_0 is an additional potential acting on the channel. Thus, the transmission function is given as

$$T(E) = \frac{\hbar^2 v^2}{\hbar^2 v^2 + {U_0}^2} \qquad \textit{equation}\ (11)$$

For large U_0, $T(E)$ becomes too small to transport charges by the junction and the current is zero. In the case of $U_0 = 0$ which there is no potential barrier by the channel, the transmission function becomes 1. Replaced the Fermi functions by step functions, the distribution difference within the molecular junction is 1. Thus, the total current is

$$I_{net} = \frac{q^2}{h} V_b \qquad \textit{equation}\ (12)$$

Considering the contribution of the two electrons of opposite spin in each level, the quantum conductance is given by

$$G_0 = \frac{dI_{net}}{dV_b} = \frac{2q^2}{h} = 77.46\mu S \qquad \textit{equation}\ (13)$$

Therefore, the conductance of molecular junction is given in units of G_0 and it can characterizes the molecular property.

1.3. Applications of STM for Biological Single Molecule Measurements

1.3.1. Break Junction

Conductance is the most frequently measured physical property of single molecule and Scanning Probe Microscopy (SPM) is one of the most straightforward technique. Break junction method is the most widely used technique to study electron transport of molecular junction. Similar with AFM based break junction method which was described in chapter 1.1.1.4, STM also can be used for break junction method. Xu & Tao reported an STM break junction method [34]. As described in Figure 10, a metal substrate electrode is modified with a self-assembled monolayer and a metal tip is withdrawn. The modified molecules have two end groups which can make a bond with tip and substrate. Thiol (SH) is the mostly used anchor group because of strong Au-S bonds forming stable molecular junction with electrodes. The tip position is controlled by servo feedback and repeatedly moving into and out of the substrate. The tip motion speed is determined by the sweep rate of the piezo-voltage.

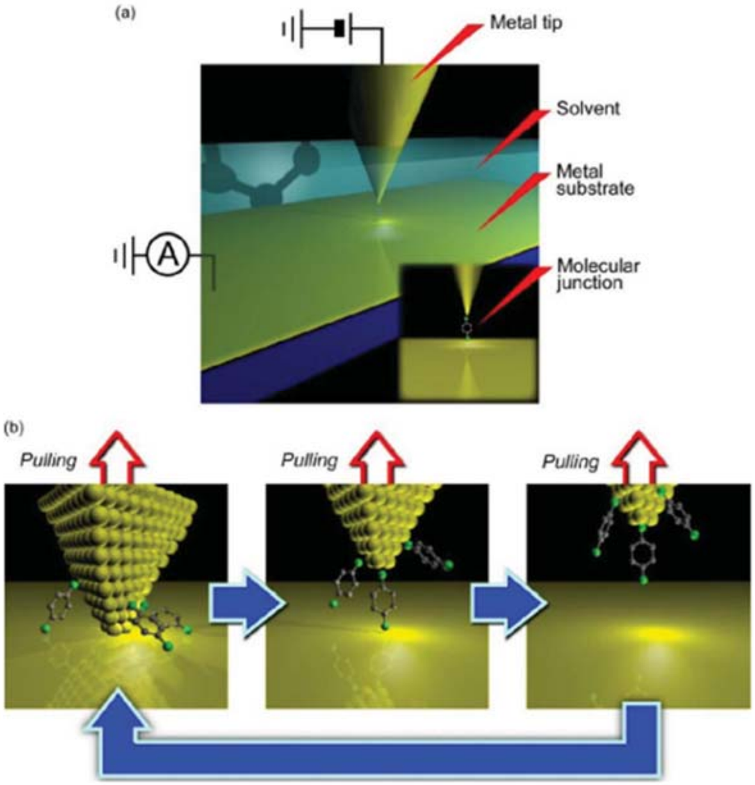


Figure 10. Schematics of Break Junction Method.

*(**a**) The system consists of a metal substrate covered with self-assembled monolayer of a target molecule and a metal tip; (**b**) A tip is indented into a substrate in a solvent. Subsequently, the fused contact is stretched by pulling out the tip and the conductance trace is measured (left). In case of Au junctions, the conductance drops in a stepwise fashion and show a long plateau at 1 G0 signifying formation of Au single atom chains. After breaking the Au contacts, current flows through several molecules bridging the tip and the substrate. Further retracting the tip, the metal-molecule bonds rupture and the number of the current-carrying molecules decreases one by one showing conductance staircases (middle) and finally to zero (right). Thousands of single-molecule junctions can be formed within a relatively short time by repeating the series of processes, which allows deduction of junction-to-junction variations of the single-molecule conductance. Figure taken from ref [19].*

The STM break junction method has some unique advantages rather than MCBJ or AFM-BJ. This method electively measures the conductance of the molecules bound to the both electrodes and focuses on the separation process. Because only the bound molecules can be stretched and broken as the tip is withdrawn. This technique is performed in an organic solvent, so the sample molecules can be easily introduced. In addition, STM can take an image of the substrate surface before break junction measurement. Then it makes us to place the tip on an atomically flat surface.

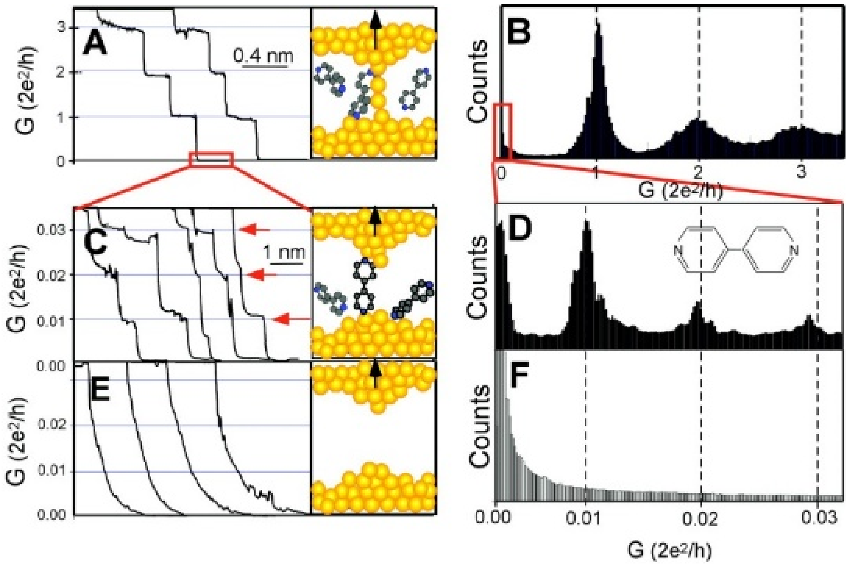


Figure 11. Break Junction Measurements.

Electronic break junction can be used to quantify the conductance of single molecules. As a gold contact is slowly broken, quantized decreases in conductance are observed, first corresponding to changes in the geometry of the gold contact (A,B), then a set of smaller quantized peaks (C,D) resulting from one, two, or three conductive molecules (here dipyridine) bridging the junction. At slightly larger distances, the junction ceases to exhibit conductance (E,F). Figure taken from ref [34].

In Figure 11, it shows quantized conductance measurements via STM-BJ. The gold metal junction (11(a)) provides some set of peaks corresponding to multiples of $G_0 = 2e^2/h$ (11(b)) through a chain of single Au atoms. Molecular junctions also gives similar pattern of peaks (11(c)) and fractions of G_0, corresponding to the molecular conductance. At little larger distance which is longer than molecular junction length, no quantized conductance was observed (11(e),(f)).

In the next chapter, some carotenoids with different functional groups are characterized through the STM-BJ experiments.

1.3.2. Recognition Tunneling

As discussed in the previous chapters, molecular junctions can be characterized at nano-gap device through charge transportation, because of the strong distance dependence and electronic sensitivity. The tunneling is also one of the most power method of sequencing biological analytes. This method does not require any labelling which results in the high cost of competing sequencing methods such as fluorescent or radioactive tags. Additionally, tunneling sequencing method is able to sequence not only DNA, but also peptides or carbohydrates etc.

In Figure 12, it shows the schematics of probing process by tunneling. Freely diffused analytes bridge the two electrodes and enhance the tunneling current. Once the molecule is unbound from the gap, the tunneling current goes back to the original current level, baseline. Through statistical analysis of many tunneling current data, the analytes can be identified by a characteristic conductance. The mean and width of the conductance distribution depend on the molecule junction and not only on the electrical properties and electrodes geometry. However, the dependence between the junction and tunneling current is not always trivial. The distribution of each analytes could be overlapped, so the mean of conductance is not distinguishable to characterize analytes. Or the junction can generate current signals including valuable information about molecular process in the gap such as hydrogen bonding patterns, Recognition Tunneling (RT). Modifying the electrodes makes the current distribution narrowing down showing in Figure 13. [35]

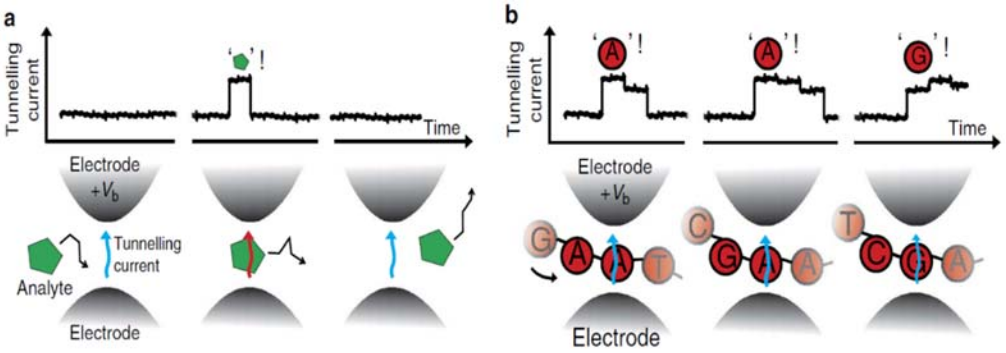


Figure 12. Probing Dynamic Processes at the Solid/Liquid Interface by Tunneling.

*(**a**) Sensing of a freely diffusing analyte that first diffuses into the tunneling junction, then bridges the gap and finally detaches and diffuses out of the junction. (**b**) Tunneling readout of DNA base composition. In this scenario, the DNA strand moves along the tunneling junction in a controlled conformation and at well-defined speed. Each base interacts with the tunneling junction and is identified by a characteristic tunneling conductance. V_b is the bias voltage applied to the two tunneling electrodes. Figure taken from ref [14].*

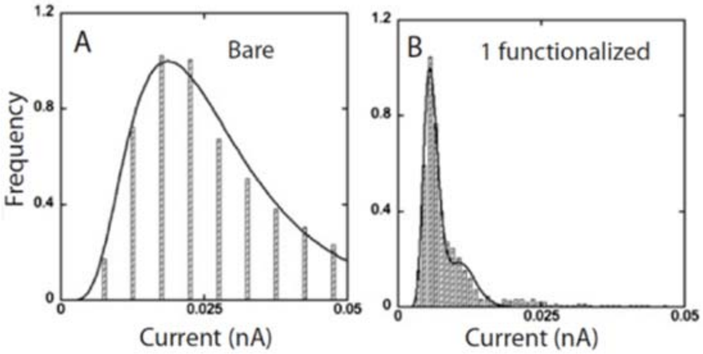


Figure 13. Current Amplitude Distribution Measured with (A) Bare Electrodes and (B) Functionalized Electrodes.

(A) Shows the current distribution measured for a pair of bare gold electrodes for dG. GBL was increased to 20 pS to obtain these reads. (B) Shows the current distribution for dG with just one electrode functionalized. Figure taken from ref [36].

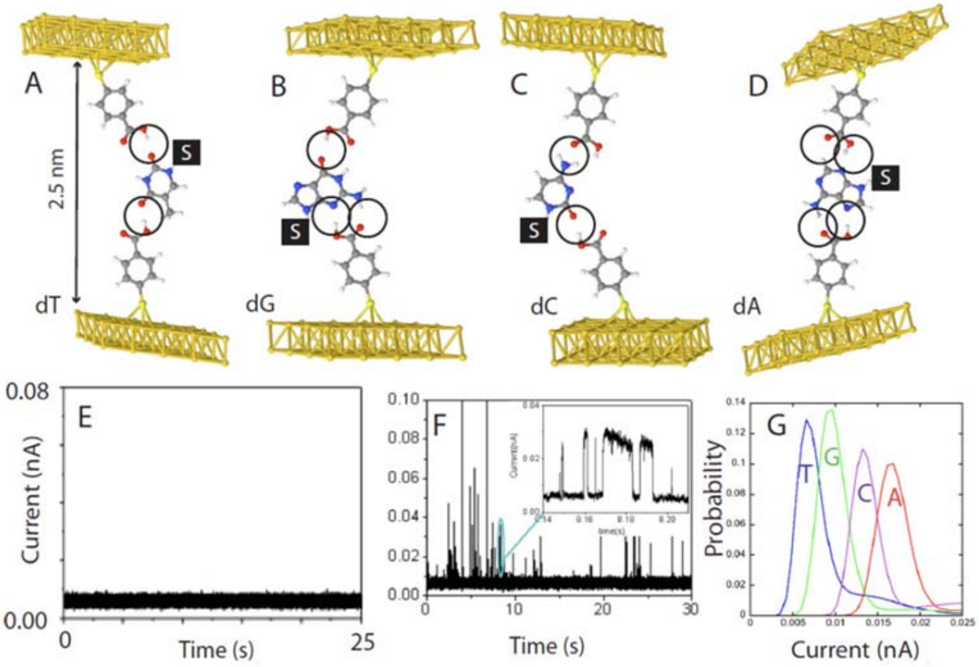


Figure 14. DNA Base Recognition through Recognition Tunneling.

(A)–(D) Show energy-minimized structures for the four nucleosides bound in a 2.5 nm gap with 4-mercapto benzoic acid as the reading reagent. The 'S' stands for the deoxyribose sugar (not shown) and the order (dT, dG, dC, dA) corresponds to the predicted order of increasing tunnel conductance using DFT. (E) Shows the background tunnel current in organic solvent (trichlorobenzene) with the gap set to GBL = 12 pS (6 pA at 0.5 V bias). At this gap there is no indication of interactions between the two benzoic acid readers. (F) Shows an example of the current spikes that are observed when a solution of dG is injected into the tunnel junction. The inset shows details of some of the spikes on a millisecond-timescale. Many of them show the telegraph noise switching characteristic of single molecule binding (the slight slope in the 'on' level reflects the action of the servo used to control the tunnel gap). (G) Measured distributions of current for the four bases. The order agrees with the density functional prediction, but the measured currents are larger than predicted. The overlap between reads limits the probability of a correct assignment on a single read to about 60%. Figure taken from ref [36].

STM-RT has been used to identify DNA bases. [35] The both electrodes are functionalized with reading reagent which has hydrogen bond donors and acceptors, and they can capture a molecule within the gap as described in Figure 14(A~D). The 'S' stands for the deoxyribose sugar (not shown) and the order (dT, dG, dC, dA) follows the increasing order which calculated by using DFT. The black circle represents hydrogen bonds between reading reagents and nucleotides. The measured current distribution shows a limited separation about 60% in Figure 14(G). For the better separation accuracy, further features rather than the current amplitude can be extracted at other domains, such as frequency or quefrency domain. The details of the feature extraction will be discussed in chapter 4.

This study shows that STM-RT can be used for peptide and carbohydrate sequencing which are followed in the chapter 3 and 4.

CHAPTER 2

CONDUCTANCE MEASUREMENTS BY STM-BASED BREAK JUNCTIONS

2.1. Introduction

Research for the electrical properties of single molecules is important not only from fundamental point of view but also for nanoscale electrical applications. The researches aim to make chemically stable molecular configurations, so that to be possible to systematically control the conductance output. There are many molecular parameters what should be controlled, such as length, conjugation, substitutions, conformation, alignment of the highest occupied/lowest unoccupied (HOMO/LUMO) molecular frontier orbitals to the electrode Fermi energy level, and anchoring chemical groups to the electrodes. [37][38][39][40][41][42]

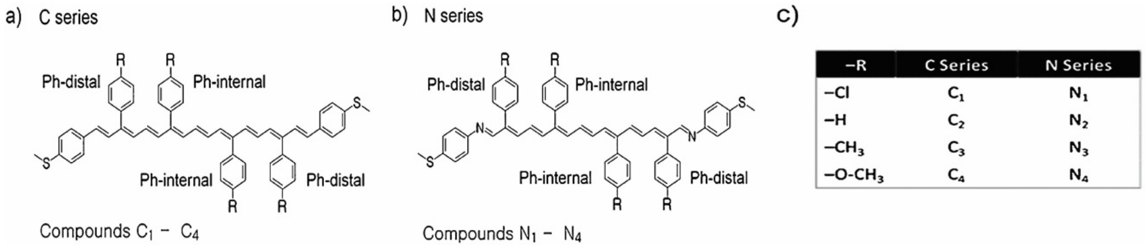
This study reported new series of carotenoid molecular wires whole conductivity is fine-tuned by inserting multiple phenyl substituents and their conductance measurements via STM based break junction method. Furthermore, it presented the molecular modeling based on DFT-optimized structures and distribution of the molecular orbitals to describe the experimental results.

2.2 Experimental Method

2.2.1. Carotenoid Samples

The carotenoid samples were prepared by Dr. Sangho Koo's lab in Myoung Ji University, Korea. The detail synthesis recipes and NMR spectra of the samples can be obtained from the supporting information of the paper. [43]

Two kinds of series were studied in this study, C-series and N-series. All the polyene chain of the C-series is composed of carbon atoms. The N-series replaces the two distal carbons of the chain by nitrogen atoms, as shown in Figure 15. The two series are identical for the position of the phenyl substituents, but the angle of the distal phenyl rings relative to the plane of the conjugated backbone are different as shown in Figure 15. The internal phenyl substituents of the both groups are fixed, but the distal phenyls of the N-series are twisted by 11.4°.



−R	C Series	N Series
−Cl	C_1	N_1
−H	C_2	N_2
−CH_3	C_3	N_3
−O-CH_3	C_4	N_4

Figure 15. Molecular Structures and Substituent Components.

a) C-series and b) N-series with distal and internal phenyls labelled as "Ph-distal" and "Ph-internal".
c) Definition of the R groups.

A key feature of the carotenoid wires in this study is the cementing of phenyl substituents on a pure carotenoid backbone bearing a $-SCH_3$ group at each end which can bind to gold electrodes. The $-SCH_3$ as anchoring groups allow for the formation of robust single molecule junctions and narrow conductance distributions. [37][38][44]

The phenyl substituents have two important functions; chemical stability and electron donating property. The phenyl substituents ensure chemical stability of the highly conjugated system since their conductance is much higher than saturated counterparts. Also the substituents vary the electron donating character from less electron-donating (Ph-Cl) to more electron-donating (Ph-OCH_3) groups. [45][46] The phenyl substituents locations on the polyene chain were confirmed through NMR measurements, as shown in Figure 17. The other NMR data can be downloaded from the publication. [43]

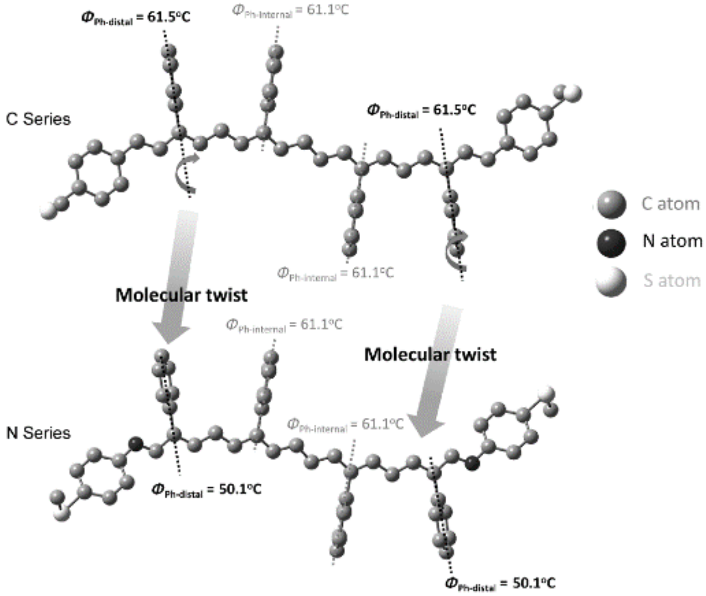

$\Phi_{Ph\text{-}distal}$ = 61.5°C
$\Phi_{Ph\text{-}Internal}$ = 61.1°C
C Series
$\Phi_{Ph\text{-}distal}$ = 61.5°C
C atom
N atom
S atom
$\Phi_{Ph\text{-}internal}$ = 61.1°C
Molecular twist
$\Phi_{Ph\text{-}internal}$ = 61.1°C
Molecular twist
N Series
$\Phi_{Ph\text{-}internal}$ = 61.1°C
$\Phi_{Ph\text{-}distal}$ = 50.1°C
$\Phi_{Ph\text{-}distal}$ = 50.1°C

Figure 16. B3LYP/6-31G(d) Optimized Structures of Ph-H Substituted C-series and N-series.

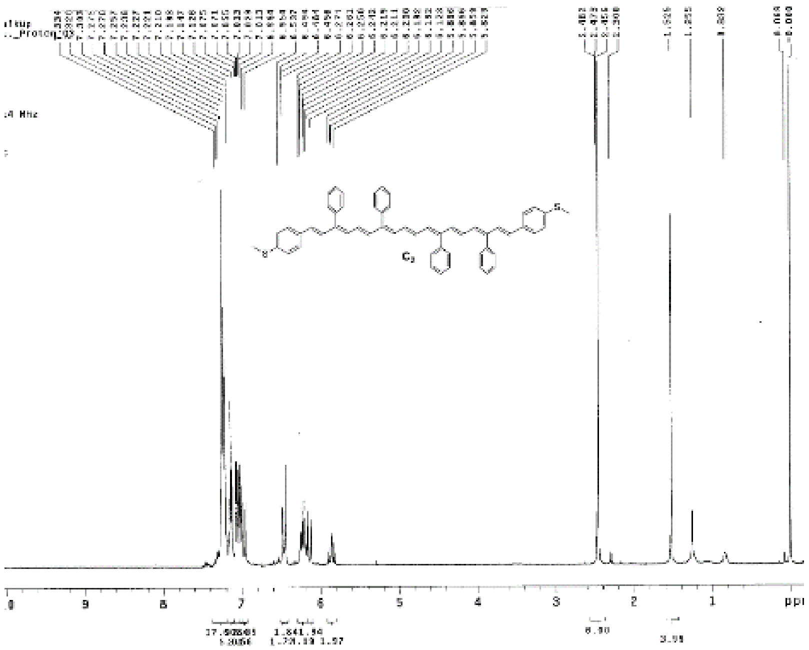


Figure 17. NMR Spectra of C_2 Compounds.

2.2.2. Sample Preparation and STM-based Break Junction

Au(111) monocrystalline substrate (10 mm X 1 mm) was purchased from MaTeck (Germany) which is 99.9999% purity and orientation accuracy <0.1 degrees. The substrate was electropolished to remove residual contaminants and annealed with a H_2 flame. The cleaned substrate was assembled with the STM cell and 80 μL of pure mesitylen (ACROS Organic, Thermo Fisher Scientific USA) was injected into cell. Control experiment was taken first in mesitylen. Next, 6 drops of a few μM of the carotenoids was injected to STM cell for break junction experiments. All glassware and Teflon STM cells were cleaned with piranha solution (H_2O_2:$H_2SO_4 = 1:1$), and rinsed with 18 MΩ cm Milli-Q water (Millipore), and dried in N_2 gas flow.

All the break junction experiments was conducted by using PicoSPM microscope (Agilent) controlled by a Picoscan 2500 (Agilent). Data acquisition was performed by using NI-DAQmx>BNC-2110 National Instruments (LabVIEW data acquisition system). The collected data was analyzed by LabVIEW code which was developed in Dr. Nongjian Tao's research group in ASU. Detail experimental procedure is following. First, the STM tip was approached to clean Au (111) substrate in tunneling distance. Then the feedback system was turned off and the LabVIEW data acquisition code was started to drive the tip into and out of contact with the substrate in 1~2 V/s which results in current decay with junction break indications. 3000 ~ 4000 of current decay curves were collected in each runs. The molecular conductance was determined by $\mathrm{G} = I_{step}/U_{bias}$, where I is the current of step in decay curve and U is the applied voltage bias. The LabVIEW analysis code automatically selects good decay curves which maintain plateaus of break junction and the selected decay curves are accumulated to semi-logarithmic histogram of conductance. By fitting the accumulated histogram to Gaussian distribution function, it gives an average single molecule conductance. The plateaus in current decay curve represents the absence of molecular bridge (or break junction). The same selection criteria was used to compile the conductance histograms. Typically, 8 ~ 15 % of decay curves provide clear plateaus like step. The compiled conductance histograms gave obvious peaks above the background histogram.

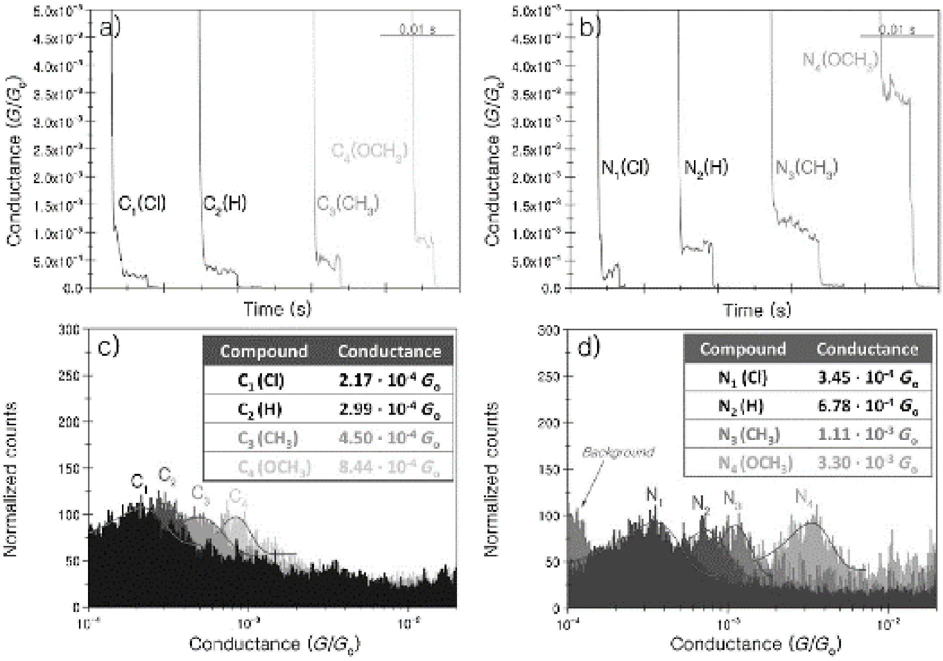
2.2.3. Computational Methods

DFT calculations were conducted by co-worker in University of Barcelona. The geometry optimizations were performed using Density Functional methods (DFT) within the B3LYP/6-31G(d) approximation. The calculated frontier orbitals were performed using higher basis set cc-pVTZ.

2.3 Results and Discussion

I measured eight carotenoids and all measured current traces show well-defined plateaus (Figure 18(a) and (b)) which represents a single molecule bridge formation during the process. The measured single molecule conductance increases as the electron donating capability of the phenyl substituents increases as shown Figure 18(c) and (d). As the HOMO level is closer to the Fermi energy level, the orbital dominates the electron transport in the molecular junction. [37][47][48][49][50]

Also the conductance of the carotenoids with substituents in N-series is higher than the C-series with the corresponding same substituents. Comparing C- and N- series, Ph-Cl substituent gives an indistinguishable conductivity in the both series as shown in Figure 18. It means that the inserted nitrogen atom is not included in the electron transport of the carotenoid molecular junction. However, the conductance of the other phenyl substituents shows increasing trend and larger in the N- series carotenoids (Figure 18). This can be explained by rotation of the distal phenyl groups as shown in Figure 15. The insertion of the nitrogen atom allows more interaction between the distal phenyl groups and the polyene backbone because of their twisted phenyls with lower dihedral angle of the distal functionalized phenyl to the polyene plane (Table 1). It results in the larger electron donating ability of the Ph-distal substituents and HOMO level shift toward Fermi level for the N- series which makes them higher electron donating character.



Compound	Conductance
C_1 (Cl)	$2.17 \cdot 10^{-4}\ G_0$
C_2 (H)	$2.99 \cdot 10^{-4}\ G_0$
C_3 (CH_3)	$4.50 \cdot 10^{-4}\ G_0$
C_4 (OCH_3)	$8.44 \cdot 10^{-4}\ G_0$

Compound	Conductance
N_1 (Cl)	$3.45 \cdot 10^{-4}\ G_0$
N_2 (H)	$6.78 \cdot 10^{-4}\ G_0$
N_3 (CH_3)	$1.11 \cdot 10^{-3}\ G_0$
N_4 (OCH_3)	$3.30 \cdot 10^{-3}\ G_0$

Figure 18. Representative Individual Current Traces and Conductance Histograms.

Semi-log conductance histograms for c) C1–C4 and d) N1–N4 compounds. The conductance values are extracted from Gaussian fits of the peaks. Insets in figures c and d summarize the average conductance magnitude for each compound. The applied bias was set to 50 mV.

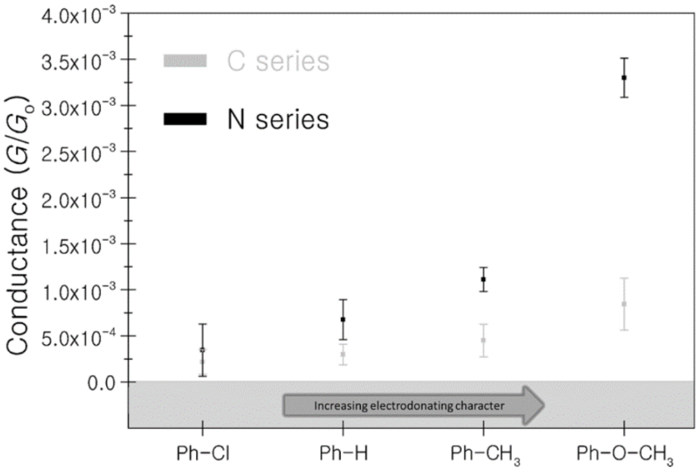
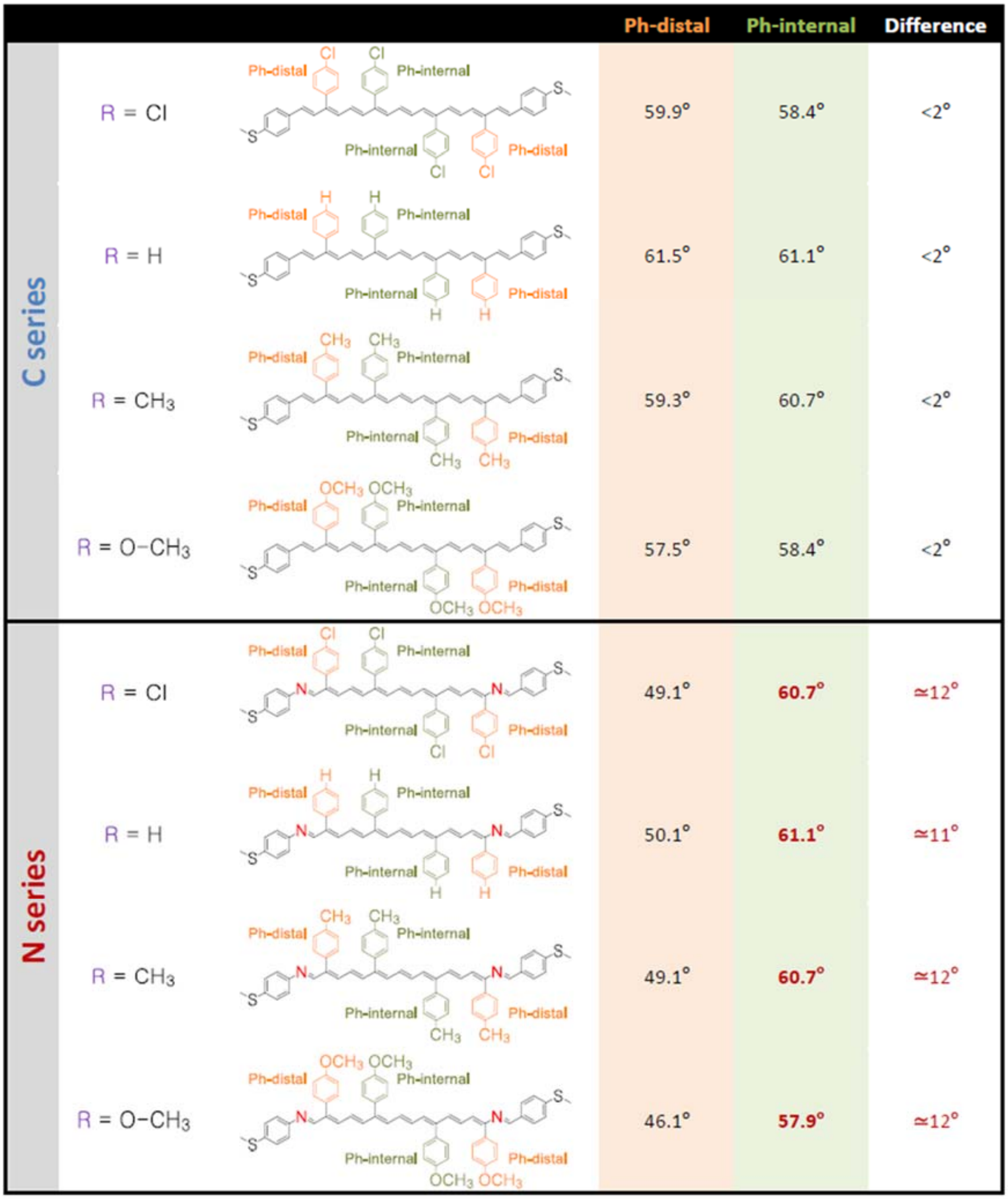


Figure 19. Evolution of the Conductance Values versus the Electron-donating Character of the Phenyl Substituents.

The error bars represent the standard deviation of the conductance values and are derived from the full width at half maximum (FWHM) values of the conductance peaks in the histograms of Figure 18 c,d.

The reported experimental results are supported by the optimized orbital geometry and the calculated energy of the HOMO/LUMO levels of the molecular systems. First, the contribution of the twisted phenyls (distal phenyls) and the non-twisted phenyls (internal phenyls) were studied by comparing pure N-series polyene backbone and Ph-OCH$_3$ substituents. As shown in Table 2, inserting two twisted distal phenyls increases the HOMO energy by 0.29 eV than phenyl free carotenoid and two internal phenyls increases 0.10 eV. The HOMO energy increasing implies the higher electron donating character of the distal phenyl substituents.

Furthermore, the orbital energies of C- and N- series with substituents are calculated (Table 3). The HOMO energy increment from Ph-Cl to Ph-OCH$_3$ of N-series is 0.48 eV and C-series is 0.37 eV. The calculation confirms the conductance evolution of phenyl substituents and C-/N- series as shown in Figure 19.

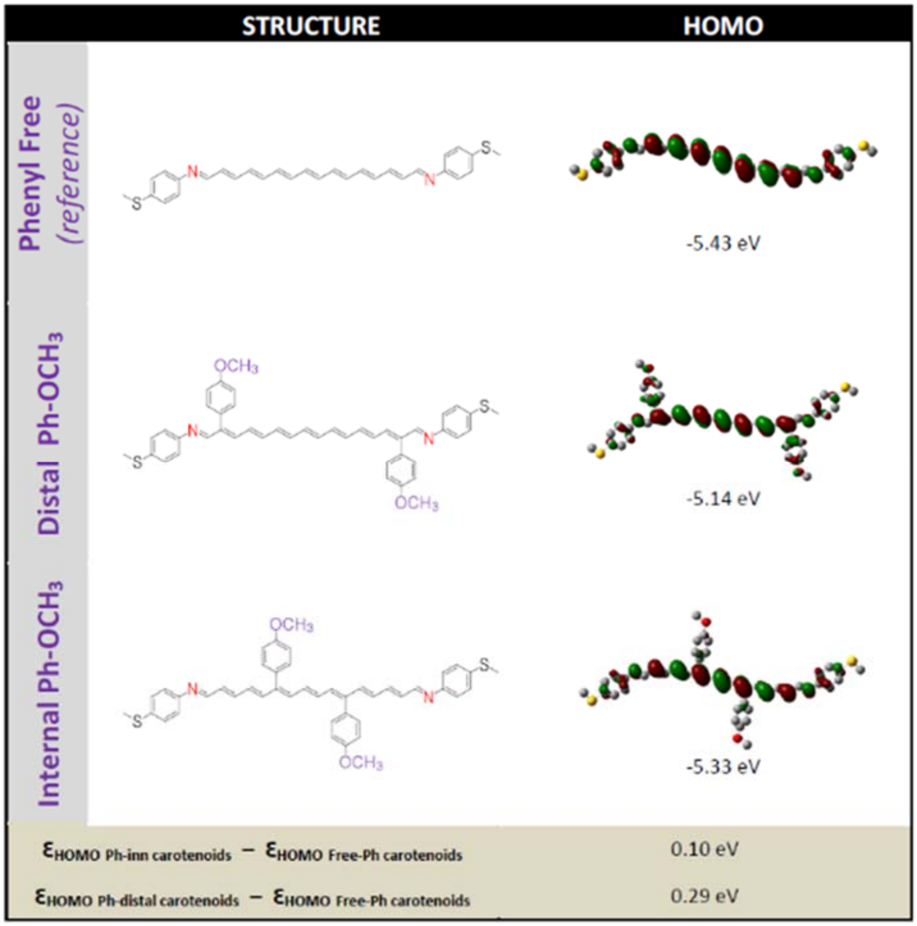


		Ph-distal	Ph-internal	Difference
C series	R = Cl	59.9°	58.4°	<2°
	R = H	61.5°	61.1°	<2°
	R = CH_3	59.3°	60.7°	<2°
	R = O−CH_3	57.5°	58.4°	<2°
N series	R = Cl	49.1°	**60.7°**	≃12°
	R = H	50.1°	**61.1°**	≃11°
	R = CH_3	49.1°	**60.7°**	≃12°
	R = O−CH_3	46.1°	**57.9°**	≃12°

Table 1. Dihedral Angles of Phenyl Substituents of the both C- and N- series.

The dihedral angles for both Ph-distal and Ph-internal groups together with their differences are listed. Angles were obtained from the optimized geometry performed with B3LYP/6-31G(d) level of DFT.

In Table 1, the dihedral angle and substituents dependence of conductance are summarized. All the phenyl substituents in the C-series has about 60° of dihedral angle and this prevents electronic communication between the carotenoid backbone and the substituents. On the other hands, the molecular orbitals of phenyl substituents in the N-series have 50° of dihedral angle. The N- series shows a more effective hybridization and higher conductance than C- series. The HOMO energy increment of the C- series is 0.37 eV and N- series is 0.48 eV from Ph-Cl to Ph-OCH3. It implies that the molecular orbital energy of the N- series carotenoids is closer to the Fermi energy and results in the higher conductance of the molecular junction.


S
N
N
S
OCH3
OCH3

	STRUCTURE	HOMO
Phenyl Free *(reference)*		-5.43 eV
Distal Ph-OCH3		-5.14 eV
Internal Ph-OCH3		-5.33 eV
$\varepsilon_{HOMO\ Ph\text{-}inn\ carotenoids} - \varepsilon_{HOMO\ Free\text{-}Ph\ carotenoids}$		0.10 eV
$\varepsilon_{HOMO\ Ph\text{-}distal\ carotenoids} - \varepsilon_{HOMO\ Free\text{-}Ph\ carotenoids}$		0.29 eV

Table 2. Optimized Geometry and HOMO Energy Calculation of N-series for the Ph-OCH3.

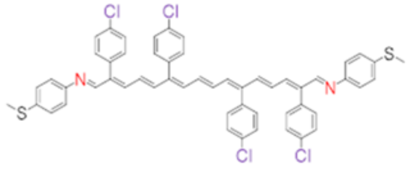
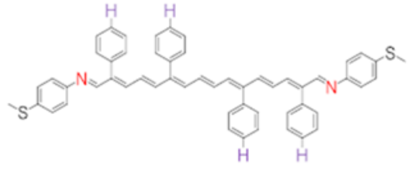
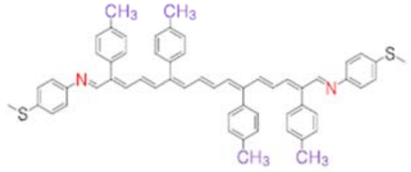
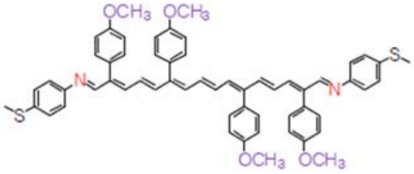
		HOMO-1	HOMO	LUMO
R = Cl		-6.13 eV	-5.44 eV	-3.44 eV
R = H		-5.91 eV	-5.17 eV	-3.17 eV
R = CH_3		-5.82 eV	-5.07 eV	-3.07 eV
R = O-CH_3		-5.70 eV	-4.96 eV	-3.03 eV
	N-series HOMO level Energy increment from Ph-Cl to Ph-OCH_3		0.48 eV	
	C-series HOMO level Energy increment from Ph-Cl to Ph-OCH_3		0.37 eV	

Table 3. HOMO and LUMO Orbital Energies of N- series with Various Substituents at B3LYP/cc-pVTZ of DFT.

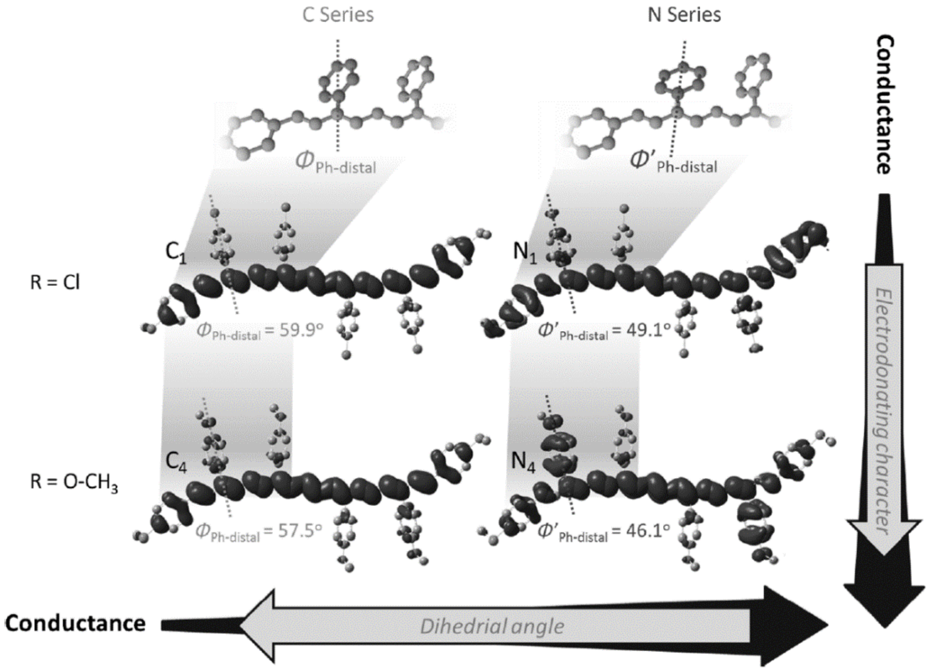

C Series
N Series
Conductance
$\Phi_{\text{Ph-distal}}$
$\Phi'_{\text{Ph-distal}}$
C1
N1
R = Cl
$\Phi_{\text{Ph-distal}}$ = 59.9°
$\Phi'_{\text{Ph-distal}}$ = 49.1°
Electrodonating character
C4
N4
R = O-CH3
$\Phi_{\text{Ph-distal}}$ = 57.5°
$\Phi'_{\text{Ph-distal}}$ = 46.1°
Conductance
Dihedrial angle

Figure 20. Distribution of the Orbital versus the Electron-donating Character of the Phenyl Substituent for both C- and N- series.

Arrows indicate the evolution of the conductance as a function of the phenyl dihedral angle (x-axis) and phenyl electro-donating character (y-axis).

In conclusion, a series of carotenoid wires were synthesized to show the conductance fine-tuning ability by replacing two of carbon atoms in polyene backbone to nitrogen and inserting various phenyl substituents with controlled conformation. The distal phenyl substituents of the N- series rotate more toward the backbone rather than one of the C- series. It enhances their electron donating role and increase the conductivity of the wires. In addition, the substituent replacement amplifies the conductance of the molecular wires. By adjusting the composition and geometry of the phenyl substituents, the conductivity of the wires were increased over an order of magnitude, from $2.1 * 10^{-4} G_0$ conductance to $3.5 * 10^{-3} G_0$, where the conductance quantum $G_0 = 77.4\ mS$ for a 3 nm length wire. This study offers a general method to fine-tune the conductivity of a molecular wire in a wide conductance range by adjusting conformation of the side group substituents.

CHAPTER 3

SINGLE MOLECULE SPECTROSCOPY OF AMINO ACIDS AND PEPTIDES BY RECOGNITION TUNNELING

3.1. Introduction

The proteome is a promising powerful tool indicating the health status rather than the genome. [51] However, the proteome has a limit to be developed, because of the lack of protein amplification technique. [52] Thus, there may be many protein variants which are not discovered yet. The difficulty can be overcome by a single molecule techniques which are able to identify biomarkers and real-time diagnostic. The STM-RT has been developing as an electronic single molecule sequencing technique for DNA recognition. This study shows that STM-RT is able to identify individual amino acids and peptides, and may open the protein sequencing. [53]

3.1.1. Recognition Tunneling

The Recognition Tunneling was discussed in chapter 1.3.2. In brief, two electrodes are modified with recognition molecules and separated by ~2 nm gap (Figure 21(a)). The recognition molecules capture target molecules by weaker non-covalent contacts. The trapped molecule makes a stochastic current signals (pA-nA) from thermal vibrations of the molecule (Figure 21(c),(d)). The current traces characterize the bonding of the molecular junction (Figure 21(e),(f)). The complicate tunneling current data is decoded with high accuracy via supporting vector machine which is a machine-learning algorithm for big data analysis. [54][55]

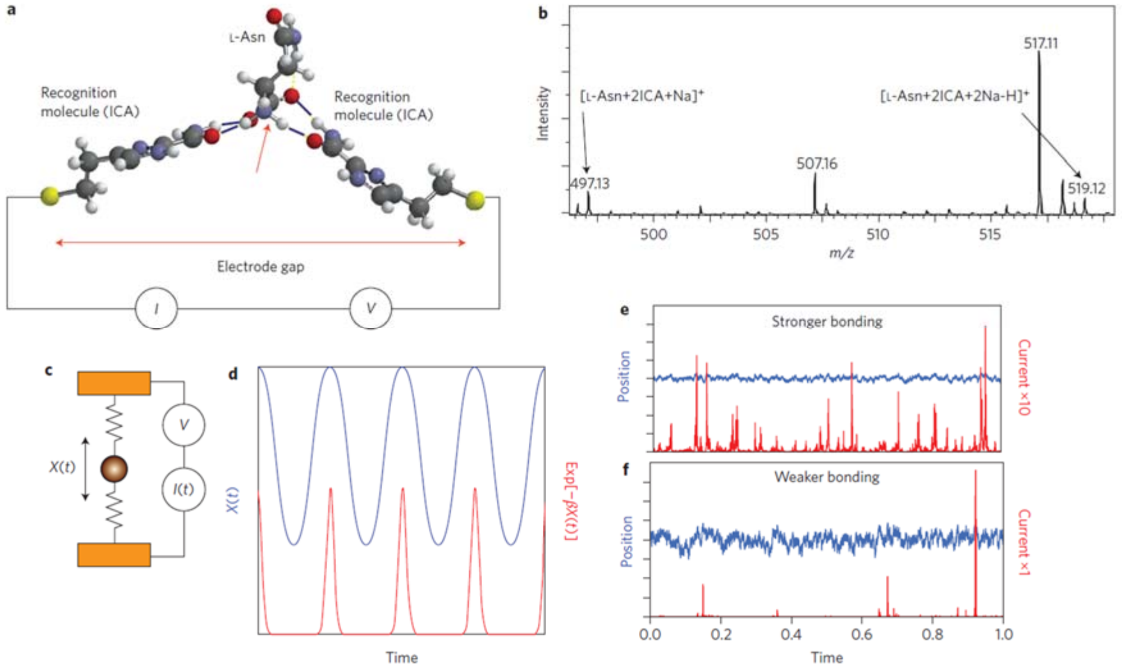


Figure 21. STM Recognition Tunneling (STM-RT).

(a) *Recognition molecules (1H-imidazole-2-carboxamide, 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 interactions (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. (**b)** ESIMS shows that stoichiometric adducts form between reader molecules, 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 extent 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). The blue traces show the random thermal fluctuations in position of the analyte. The simulations are carried out as described by Huang and colleagues [56].*

3.1.2. Amino Acids and Peptide Sequencing

Protein or peptide sequencing has many difficulties than DNA sequencing. First, there are many fundamental blocks, 21 amino acids (Figure 22) for protein but 4 nucleotides for DNA. It makes more complicate to identify the uniqueness of each basis. Second, no technique is developed to amplify low concentration of proteins. In contrast, for DNA sequencing, low concentrations of DNA can be easily amplified through the polymerase chain reaction (PCR). Thus it is hard to detect small amount proteome and there may be many undiscovered rare proteins which are below the detection limits of present methods. [57] Once being able to read protein sequencing, it will make us to understand cellular processes and design drugs for specific metabolic pathways. [51][52]

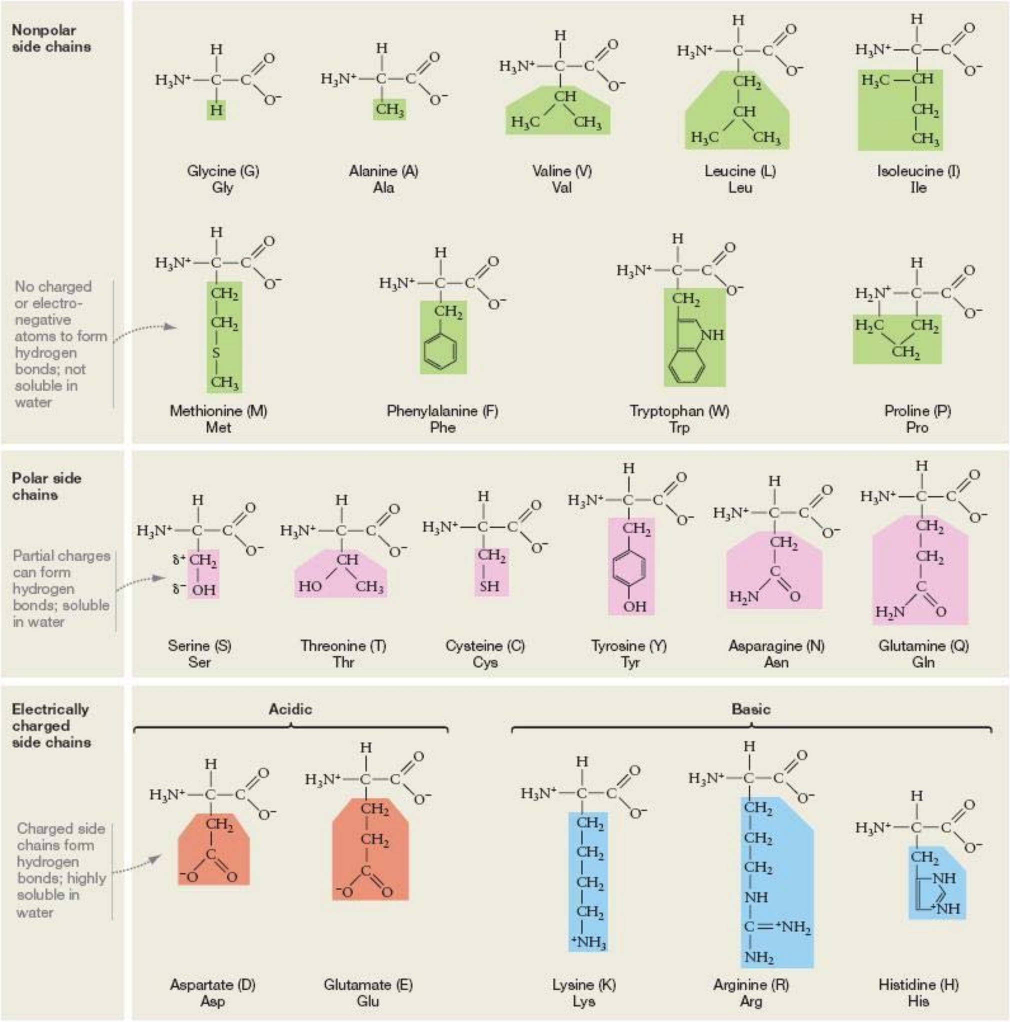


Figure 22. The 21 Amino Acids Found in Eukaryotes Grouped according to Their Functional Side Chains.

3.2 Experimental Method

3.2.1. Preparation of Analytical Solutions

Amino acids compounds were purchased from Sigma Aldrich which purity is higher than 98%. The buffer solution was made to be 1 mM phosphate buffer at pH 7.4 using purified water from a Milli-Q system with ~18 MΩ *cm* and total organic carbon contamination below 5 ppb. Peptides were purchased from CPC Scientific and dissolved in the same buffer solution.

3.2.2. Preparation of Probes and Substrates

STM probes were etched [58] from 0.25 mm Pd wire (purchased from California Fine Wires) and insulated with polyethylene to avoid ionic current leakage. The metal tip apex opens only a few tens nanometers. Substrates were prepared by depositing palladium on silicon wafer with the following recipe. First, a 10 nm Ti adhesion layer was deposited on 750 μm silicon wafer using electron-beam evaporator (Lesker PVD 75). Next, 100 nm Pd is fabricated on the Ti film. Tunneling current is very sensitive to the distance, then leakage of probes make errors in the set point current. So probes were tested to ensure that ionic current leakage was less than 1 pA in buffer solution at -500 mV bias. For electrodes functionalization, electrodes were immersed in ICA solution (~0.5 mM) in ethanol [59]. After ~20 hours, the probe and substrate were rinsed with ethanol, and gently dried with nitrogen gas. All the probes and substrates are freshly prepared, right before the experiments.

3.2.3. STM-RT

Two different PicoSPMs (Agilent Technologies) with custom LabView interfaces were used for data acquisition. Tunnel current was collected at 50 kHz sampling rate. The STM cells were cleaned in piranha, and rinsed with Milli-Q water and ethanol, and dried in N_2 gas flow. A probe was approached to a substrate with 4 pA tunneling set-point and -

0.5 V bias applied on substrate under integral and proportional gains of 1.0. The surface of substrate was scanned to ensure that the probe provides good image and the grain structure of Pd is clearly visible. After 2 hours stabilization of the microscope, the integral and proportional gains were reduced to 0.1 and tunneling current was recorded. The control (1.0 mM phosphate buffer at pH 7.4) was run before an amino-acid solution was measured. Freshly prepared probes and substrates were used for each run, usually recording four runs for each analyte.

3.2.4. ESIMS

Sample solutions were prepared to be the same molar ratio of the capturing mechanism in the STM electrodes gap in 1:1 and 2:1 of ICA to amino acid. The compounds were dissolved in water to be 100 mM. The prepared amino acids solutions were injected into a Bruker MicrOTOF-Q electrospray ionization quadrupole time-of-flight (ESI-Q-TOF) mass spectrometer (MS) and tandem MS was used to confirm the mass peaks of the corresponding ratio mixtures.

3.2.5. SVM Analysis

The kernel-mode SVM [55] was used in this work and available from https://github.com/vjethava/svm-theta. The first step of data analysis is feature extraction from the raw current signal. There are two main groups, peak and cluster. Peak represents thermal fluctuation of a captured single molecule maintaining within the gap. Cluster is composed by some of close neighbor peaks which would be from the same molecule that anchors at one electrode and the other end binding and unbinding. Once a molecule is unbound from any electrode, current signal goes back to the baseline current. The baseline of raw signal, 4 pA, was shifted to zero and all the spikes above 15 pA was determined as peaks and characterized using the features listed in Table 4. The shape of each spike was characterized in frequency domain by a fast Fourier transformation (FFT). FFT amplitudes were averaged at three sections that were equally spaced (0–2.7 kHz, 8.4–11.1 kHz and 22.3–25 kHz), and theses averages were used to define more features, such as the ratio of the highest to lowest FFT bins (High Low Ratio in Table 4). The Fourier transformed trace was down-sampled into nine bins that equally spaced from 0 kHz to 25 kHz.

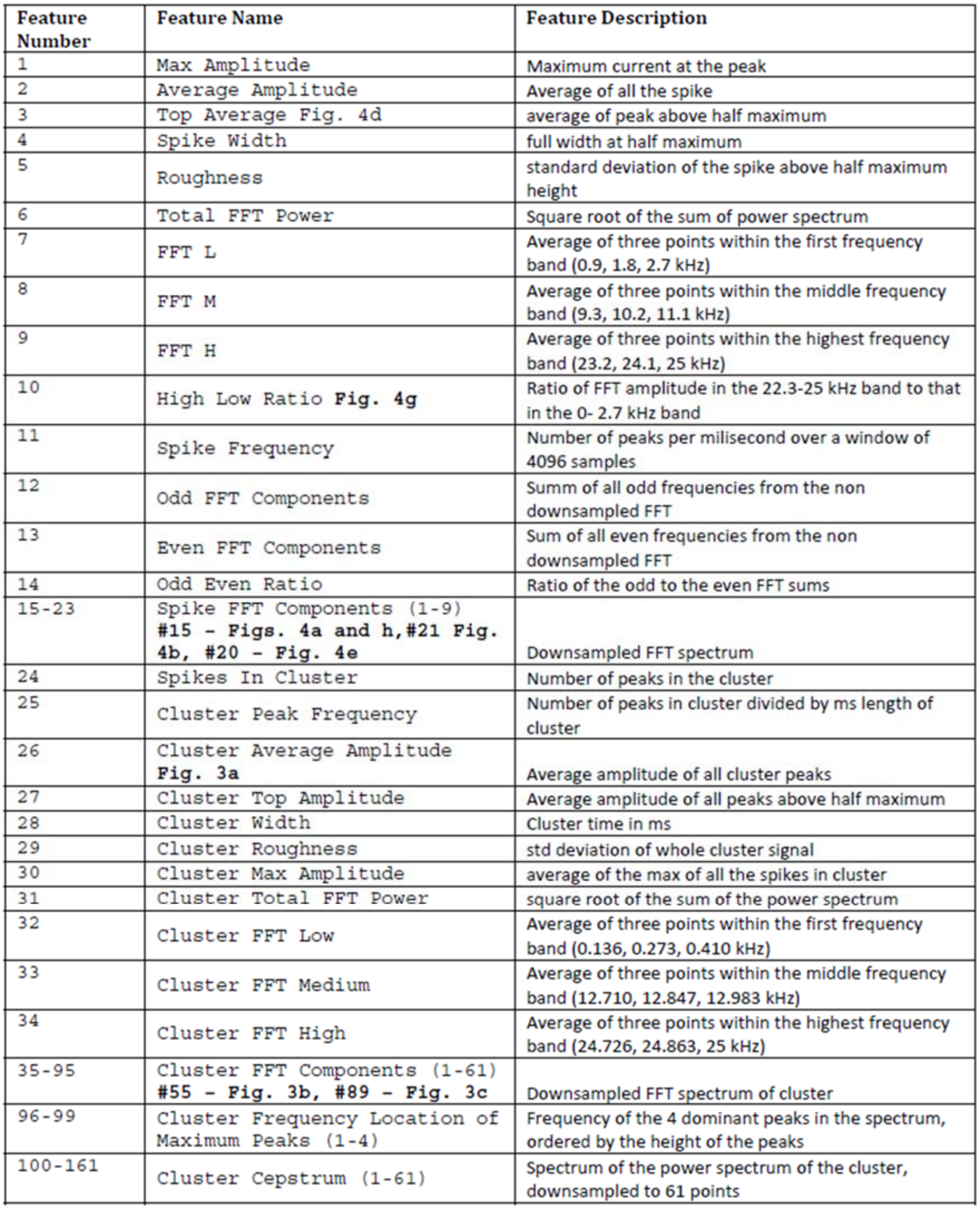
Clusters were automatically identified by applying a Gaussian window to the detected peaks as described in Figure 23 (ref. [54]). While determining the peaks, the spikes smaller than 15 pA threshold was ignored as random noise signals. However, once a cluster was identified, all the spikes in the cluster range were used to define cluster features. The Fourier spectrum of the whole cluster is deconvolution of instrumental response by spectral division. Since clusters contain more data points than spikes, the cluster FFT spectrum was down-sampled into bigger number of bins (61 bins) than the one of peak (9 bins). Each bin in cluster down-sampling corresponds to 25 kHz/61 or 410 Hz in width. The Noll method [60] was used to calculate the cepstrum amplitudes from the FFT spectrum and the resulting spectrum was down-sampled again into 61 bins.

To avoid that some bigger (or huge) numerical value features bias others, all the extracted features were rescaled as follows. The distribution of each individual feature was calculated for one amino acid which arbitrarily selected (in this study, arginine for amino acids and glycine for peptide analysis). The scaling factor was determined by shifting the mean of the distribution to be zero and the standard deviation to be 1.0. All the features of all other analytes were adjusted by using the same scaling factor.

Feature selection follows three steps. First, the features which are highly correlated with other features were removed. To calculate the normalized correlation between different feature pairs (feature x and y), $\sigma_{xy} = \langle(x - \bar{x})(y - \bar{y})\rangle$ was used. σ_{xx} is 1, because the features were normalized. All the data was used to make a correlation matrix (Figure 36) and all feature combinations of $\sigma_{xy} \geq 0.7$ were removed.

Second, feature variations was compared which named 'in-group' and 'out-group' fluctuation. The 'in-group' fluctuation is a variation between repeated experiments of the same analyte. The 'out-group' fluctuation is a comparison with all possible pairs of other analytes. The ratios of 'out-group' fluctuation to 'in-group' fluctuation were ranked and the lowest 15 features were removed. Finally, the usefulness of the survived features was tested by evaluating the identification accuracy with randomly selected feature sets. A tree search was used in the process.

The details of the SVM written in Matlab can be obtained from https://svmsignalanalysis.codeplex.com/.



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 Fig. 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 **Fig. 4g**	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 milisecond over a window of 4096 samples
12	Odd FFT Components	Summ 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) **#15 - Figs. 4a and h,#21 Fig. 4b, #20 - Fig. 4e**	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 **Fig. 3a**	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) **#55 - Fig. 3b, #89 - Fig. 3c**	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

Table 4. 161 Starting Features Used in the Signal Analysis.

Details of their calculation are given in ref [54].

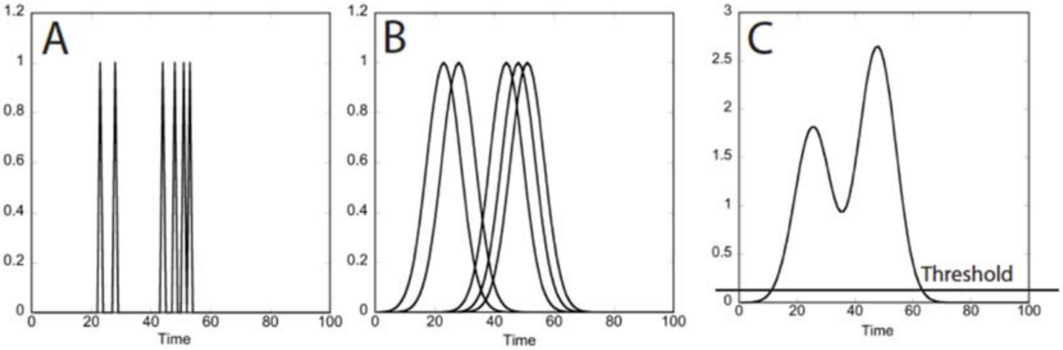


Figure 23. Automatic Cluster Identification.

Automatic cluster identification was carried out by placing Gaussians of unit height (A) and full width of 4,096 data points (1 data point = 20 µs) at the location of each spike (B), summing them (C), and assigning a cluster to regions where this sum exceeds a threshold (0.05 in this study). Figure taken and adjusted from ref [54].

3.3 Results and Discussion

3.3.1. Identifying Amino Acids

In this study, three application experiments were conducted. First, a data pool of seven different analytes was examined to show how well single amino acid can be identified which is a first step for sequencing technique. Second, it was shown that a modified amino acid, sarcosine (N-methylglycine or mGly, a promising cancer marker [61]) can be identified from glycine (Gly). Third, enantiomers (L- and D- asparagine) and isobaric amino acids (leucine and isoleucine) were distinguished through STM-RT.

The example RT current traces are shown through Figure 24 ~ Figure 27. The tyrosine and tryptophan provide RT signal under different conditions rather than other molecules as described in Figure 27. All other amino acids give RT signal under 4 pA set point at 0.5 V bias which corresponds to ~2 nm gap. However tyrosine is at 6 pA and tryptophan is as 10 pA.

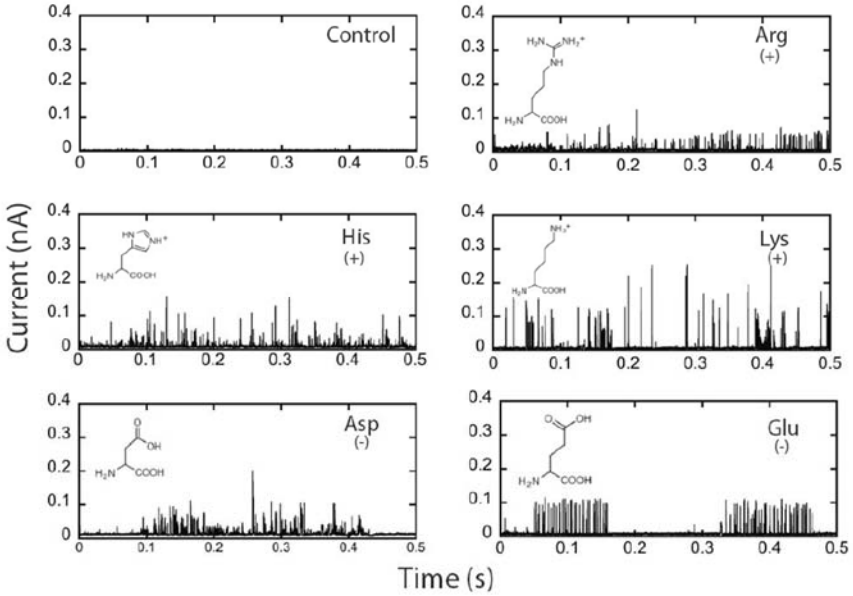


Figure 24. RT Current Traces of the Charged Amino Acids.

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.

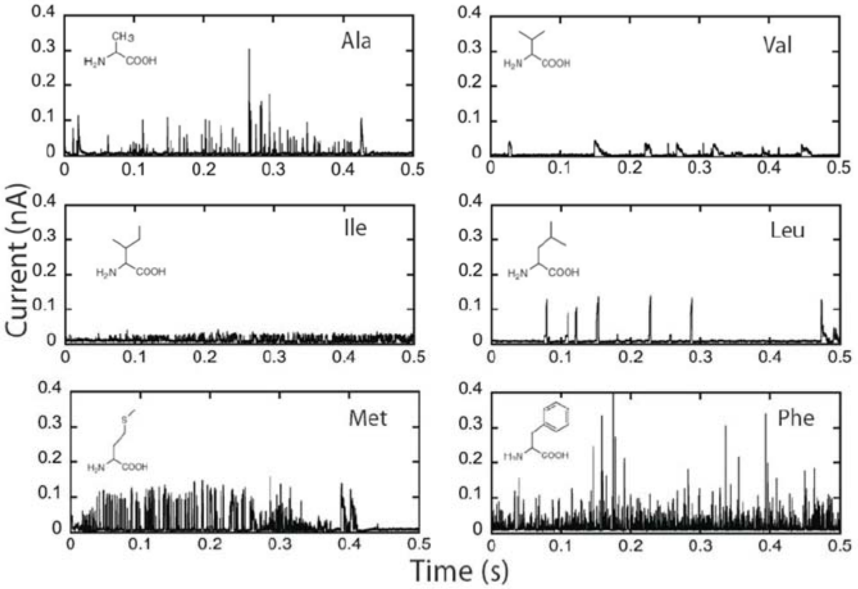


Figure 25. RT Current Traces of the Hydrophobic Amino Acids.

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. (excluding tyrosine and tryptophan)

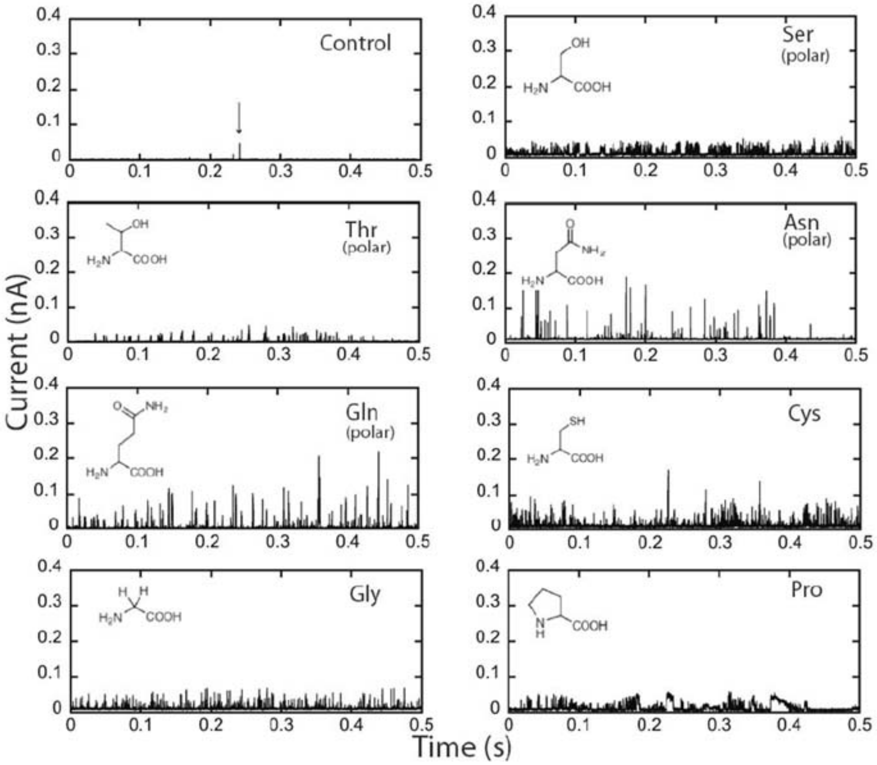


Figure 26. RT Current Traces of the Remaining Amino Acids.

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. A trace for buffer alone is shown as the control in the upper left. The arrow points to a "water" spike.

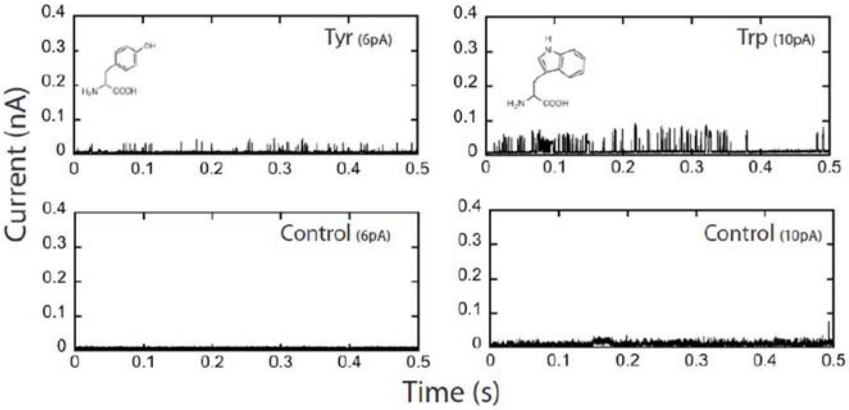


Figure 27. RT Current Traces of the Tyrosine and Tryptophan.

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 for 100µM solutions in 1 mM phosphate buffer.

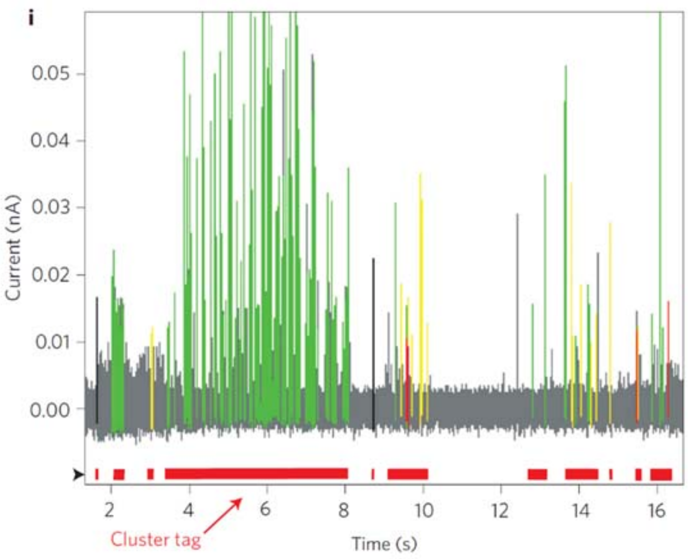


Figure 28. Signal Trace of Arg with Color Code.

The color code corresponds to the peak assignment made by a machine learning algorithm (green: correct call, red: wrong call, black: water peak, yellow: common to all amino acids). The red bars at the bottom mark signal clusters generated by a particular single-molecule binding event.

The RT signal includes important information and enable to extract unique features to identify the molecule. A computer algorithm was made to define clustered data automatically as depicted in Figure 23. Clusters is single molecule binding events by the following reasons. First, cluster width (or duration) is the order of 0.2 sec (Figure 29) which is comparable to the hydrogen bond lifetime in a nanogap. [62] Second, signals within clusters are more strongly correlated rather than signals from other clusters. Finally, each cluster of mixture sample is assigned into only one molecule (Figure 31).

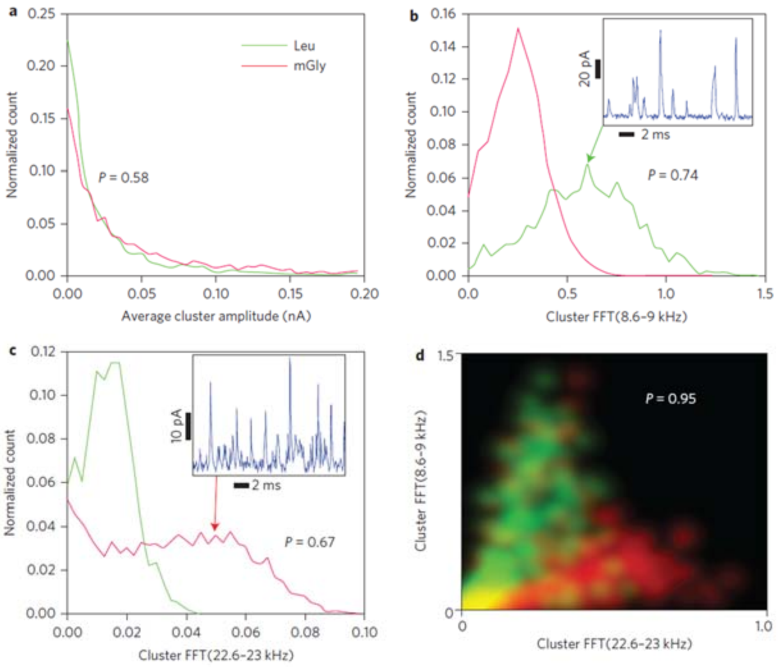


Figure 29. Feature Distribution of mGly and Leu of STM-RT.

***a**, Peak amplitudes are exponentially distributed so provide little discrimination. Identification accuracy between Leu and mGly is 0.58 only slightly better than random (0.5). **b**,**c**, Particular Fourier components of the clusters show more separation, producing 74% (**b**) and 67%(**c**) accuracies. The way in which these Fourier components reflect peak shapes in a cluster is illustrated by the signal traces inset in **b** and **c**, each trace having the feature value indicated. The high amplitude of high-frequency components of the mGly signals (inset in **c**) is evident in the sharper spikes. Accuracy improves when multiple features are used together. **d**, Two-dimensional plot of probability density. The color scale shows mGly data as red and Leu as green. Calling all the spikes with pairs of feature values that fall in the green regions as Leu and all the spikes with pairs of features that fall in the red regions as mGly produces a correct call 95% of the time. Only the yellow regions yield ambiguous calls.*

Figure 29 shows features distribution of mGly and Leu STM-RT data. Though the distribution of the average amplitude of cluster is almost overlapped (Figure 29a), some features which are associated with signal shapes are less overlap (Figure 29(b),(c)). However, the accuracy of single molecule true positive calling is limited to ~70% where the random calls probability is 50%). When the two features are considered together, it provides much higher accuracy than single feature identification. The two dimensional scatter map gives ~95% calling accuracy (Figure 29(d)). Only a small fraction is overlapped near origin, colored by yellow in which red represents mGly and green is for Leu. It follows the Cover's theorem that pattern recognition accuracy increases with higher number of features. [63] In this study, SVM, a machine learning algorithm, was used to identify the analytes from a pool of multiple data sets. SVM trains on a subset of the data to find a hyper-plan separating the analytes in the subset and tests the determined hyper-plan on the rest of the data. [54][64] With a large number of features, SVM is able to provide high accuracy.

In the case of chemically similar pairs of analytes, L-Asn and D-Asn in Figure 30, the Cover's theorem is more dramatic. The six of single feature distribution are overlapped and give a low identification accuracy in Figure 30(a),(b),(d),(e),(g),(h). Using two features increases the separation accuracy, higher than 80% in all three cases shown in Figure 30(c),(f),(i). If there were more analytes to be identified, it becomes more complicate and harder than two analytes identification. The SVM is a powerful tool to analyze complicate and big data. Once numerical features are extracted from tunneling current data, each single spike is plotted in N dimensional space, where N is the total number of features. A randomly selected subset of the data is used to make a hyperplane of N-1 dimension (support vectors) classifying the known data. The SVM is originally developed for a binary classifier, but multiclass SVMs can be constructed. There are a number of methods to construct multiclass SVMs from binary SVMs and is still researching. The SVM used in this study is based on 'One against the Rest approach'.[65] Assuming *M* classes to be classified, the *M* of binary classifiers is created where each classifier is trained to identify one class from the rest *M-1* classes. In other words, the best partition to separate signals from each analyte from a pool of data set. The support vectors are determined by finding margin from the hyperplane. Other classifiers (support vectors) are constructed in the same manner. Once trained with a subset of the data, the determined support vectors are tested on the rest of data set.

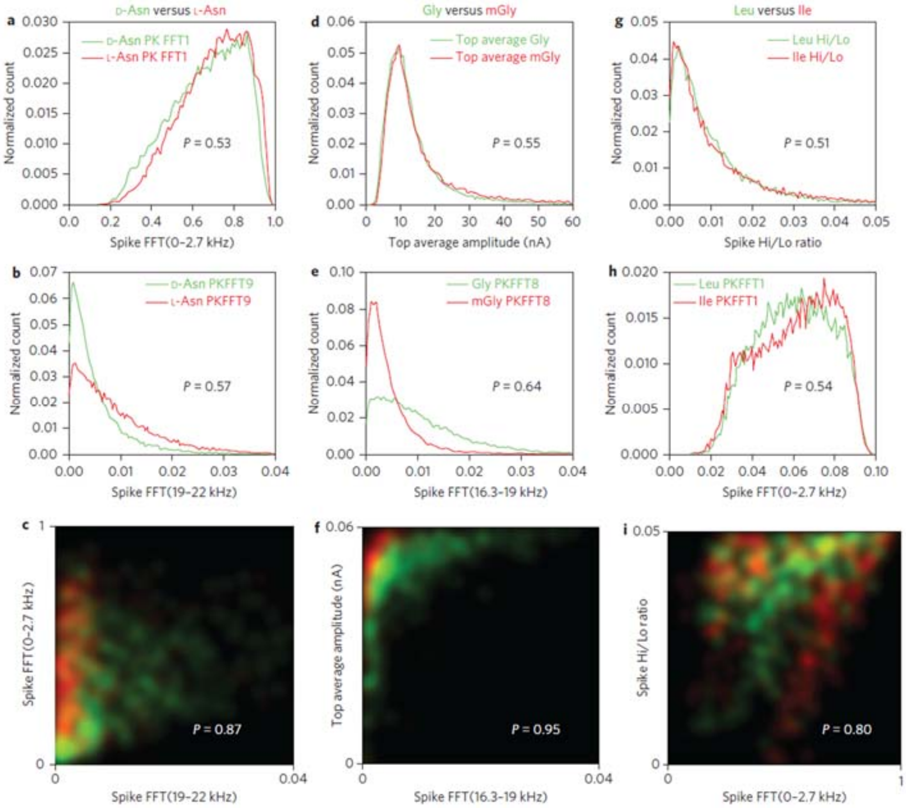


Figure 30. Feature Distributions of Chemically Similar Pairs of Analytes, D-Asn and L-Asn.

Closely related pairs of analytes can be significantly separated (>80%) using just two signal features together. All data are for pure solutions of one analyte. ***a–i****, Chiral enantiomers D-Asn and L-Asn (a–c), Gly and mGly (****d–f****) and the isobaric isomers Leu and Ile (****g–i****) are quite well separated in two dimensional probability density maps (****c,f,i****), even when the distributions of any one signal feature are almost completely overlapped in one dimension (****a,b,d,e,g,h****; see Methods and Supplementary Table 4 for a description of these features). The two-dimensional maps plot probability densities for the analyte pairs (color coded as listed at the top) as a function of both features, which, by themselves, produce separations only a little above random (0.51 to 0.64). Probabilities of making a correct call based on the probability densities are marked on* ***c, f*** *and* ***i****, and calculated as described in the caption for Figure 29.*

3.3.2. Analyzing Mixtures of Analytes

The STM-RT and SVM analysis were tested to identify an amino acid from mixture. The mixture sample was prepared with L- and D-Asn in various ratios of 1:1, 2:1 and 3:1. The pure L- and D-Asn RT signals were used to constructed support vectors and the support vectors was used to assign the spikes from mixture signal. Figure 31(a) shows a trace of raw RT signal which is color-labeled with respect to the assignment (L-Asn is yellow, D-Asn is purple, Common spike is black). The identical clusters are marked by the red bars at the bottom. It clearly shows that each cluster is composed by one type of analyte. By applying the common noise filter, only pure cluster remain. It implies that the clusters represent single molecule binding events. Using assignment of SVM makes us to count molecules in mixtures. Figure 31(c) shows the trend between the measured L/D ratio and the actual ratio by counting peak assignments (red points) and cluster assignments (blue points). Counting by peaks overcounts the L-Asn as described in Figure 31(c), the slope is 2.7. Counting by clusters undercounts the L-Asn which the slope is 0.2. It might be explained by binding strength between analyte and ICA and local reduction of L-Asn on the electrode surface.

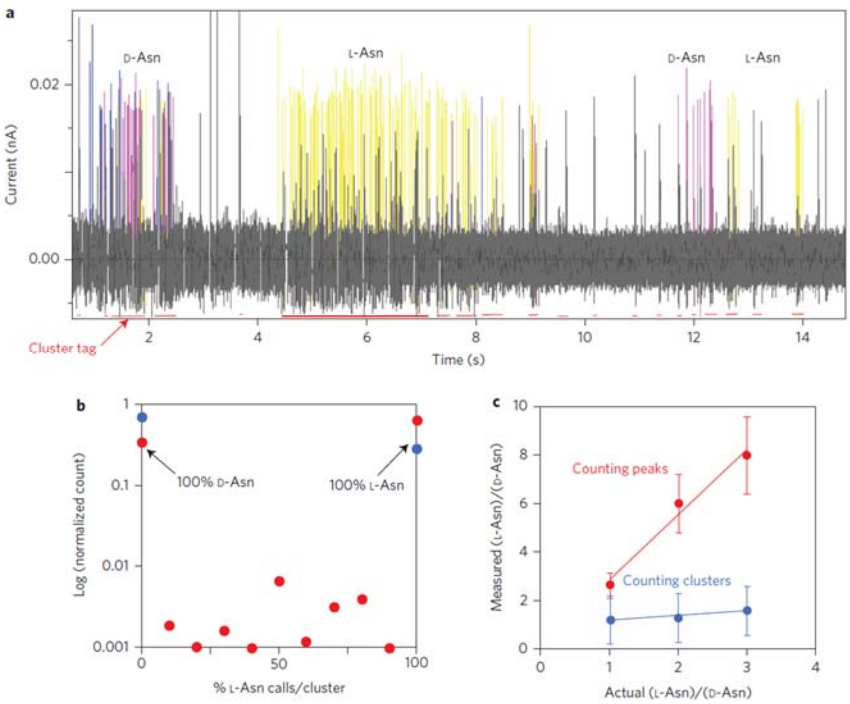


Figure 31. Mixture Analysis to Recognize Pure Amino Acids.

*A mixture produces alternating cluster signals as different molecules diffuse into and out of the gap. **a**, Signal trace obtained with a 1:1 mixture of L- and D-asparagine. The SVM assignments are coded purple (D-Asn) and yellow (L-Asn) (black spikes are unassigned). **b**, Each cluster (red tags in **a**) contains only one type of signal, as shown statistically. The red points are for 556 raw data clusters and the blue points are for 400 clusters that remain after filtering for common signals. After filtering (blue points), no mixed clusters survive, with all of the clusters being 100% L- or D-Asn signals. **c**, Quantification of the L/D ratio using SVM trained on pure samples. The measured ratio increases with actual ratio in the samples, but the calibration depends on whether the number of signal spikes (red) or clusters (blue) is used, probably reflecting differential binding. Error bars are from repeated runs and repeated samplings.*

3.3.3. RT Signals from Peptides

In the previous sections, they show that the possibility of STM-RT to recognize pure single amino acids and components from mixture in various ratio. Furthermore, this technique was tested to identify some short peptides; GGGG and GGLL. In the case of amino acids measurements, the hydrogen bonding (recognition mechanism) sites are the zwitterionic center as depicted in Figure 31(a). Peptides may be more spatially separated between N and C termini, so it may be hard to generate RT signals. However, peptides produced obvious RT signals and the examples of peptide RT signals are shown in Figure 32.

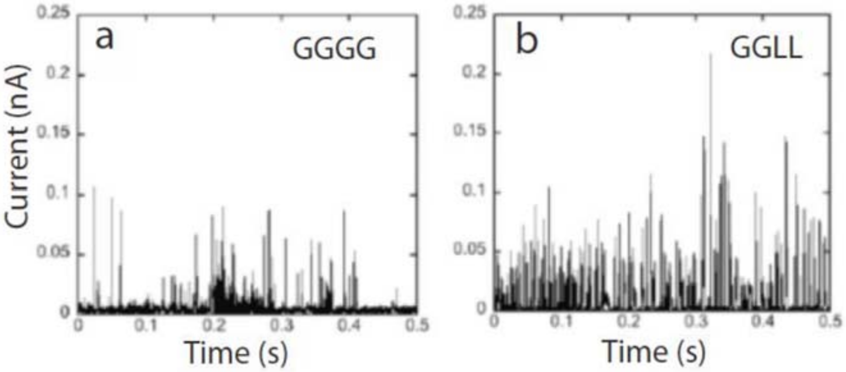
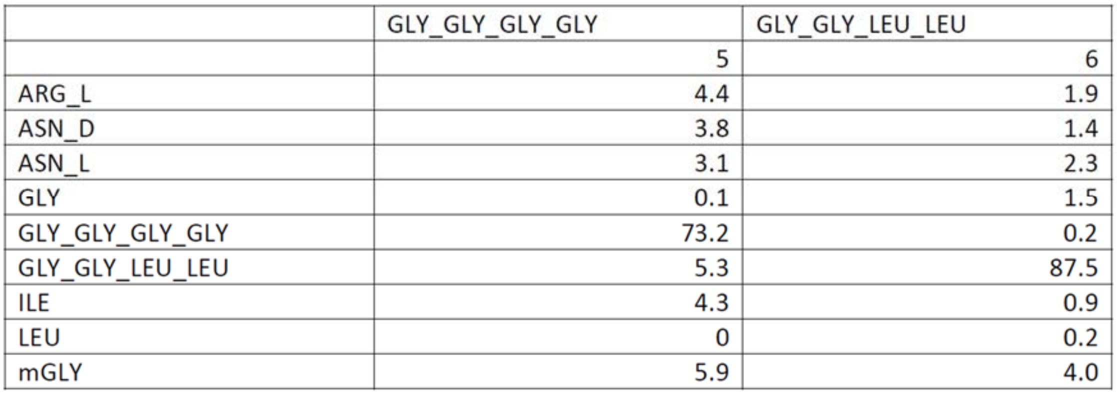


Figure 32. Examples of Peptide RT Signal.
(a) GGGG and (b) GGLL.

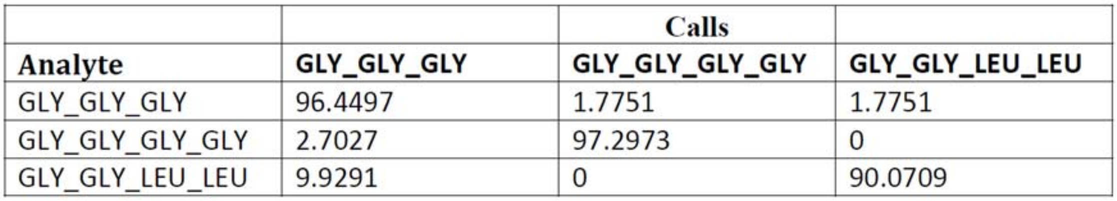
The support vectors from pure amino acids was applied onto the peptide data, but it could not identify their constituent amino acids (Table 5). It means that the binding mechanism of the amino acids in peptides are different from the pure amino acids in solution. So it has been tried to identify a peptide from the pool of three peptides data (GGG, GGGG and GGLL). Through SVM analysis, one produced >90% accuracy with 65% of common peak rejection (Table 6). Thus, though the RT mechanism is different with pure amino acids, multiple peptides can be recognized from others, even the difference is only one residue among four constituents. It suggests that single molecule sequencing of protein is possible through RT.



	GLY_GLY_GLY_GLY	GLY_GLY_LEU_LEU
	5	6
ARG_L	4.4	1.9
ASN_D	3.8	1.4
ASN_L	3.1	2.3
GLY	0.1	1.5
GLY_GLY_GLY_GLY	73.2	0.2
GLY_GLY_LEU_LEU	5.3	87.5
ILE	4.3	0.9
LEU	0	0.2
mGLY	5.9	4.0

Table 5. Peptide Recognition Calling from SVM of Pure Amino Acids and Peptides.

Distribution of calls among the peptides and amino acids, showing percentages of the signal spikes from each peptide called as one of the seven amino acids, the correct peptide, or the wrong peptide. The vast majority of calls are correct (73 and 87%) showing how each peptide it distinct form the other and distinct from the amino acids.



		Calls	
Analyte	**GLY_GLY_GLY**	**GLY_GLY_GLY_GLY**	**GLY_GLY_LEU_LEU**
GLY_GLY_GLY	96.4497	1.7751	1.7751
GLY_GLY_GLY_GLY	2.7027	97.2973	0
GLY_GLY_LEU_LEU	9.9291	0	90.0709

Table 6. Peptide Recognition Calling from SVM of Three Peptides Pool.

Separation of signals from three peptides. Samples are listed in the left hand column with the distribution of calls among the three possible calls listed in the three right columns. This accuracy was achieved with 65% of the signal spikes rejected as "common".

3.3.4. Bonding in the RT Junctions

Even though ICA was designed to bind with DNA bases, it has been demonstrates ICA also can capture amino acids through hydrogen bonding. It is confirmed with the density functional theory (Figure 21(a)) and ESIMS measurements (Table 7, Table 8 and Figure 33, Figure 34).

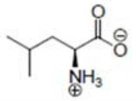
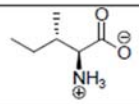
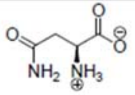
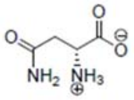
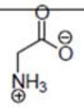
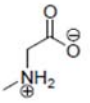
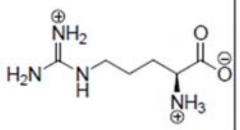
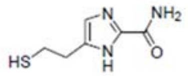
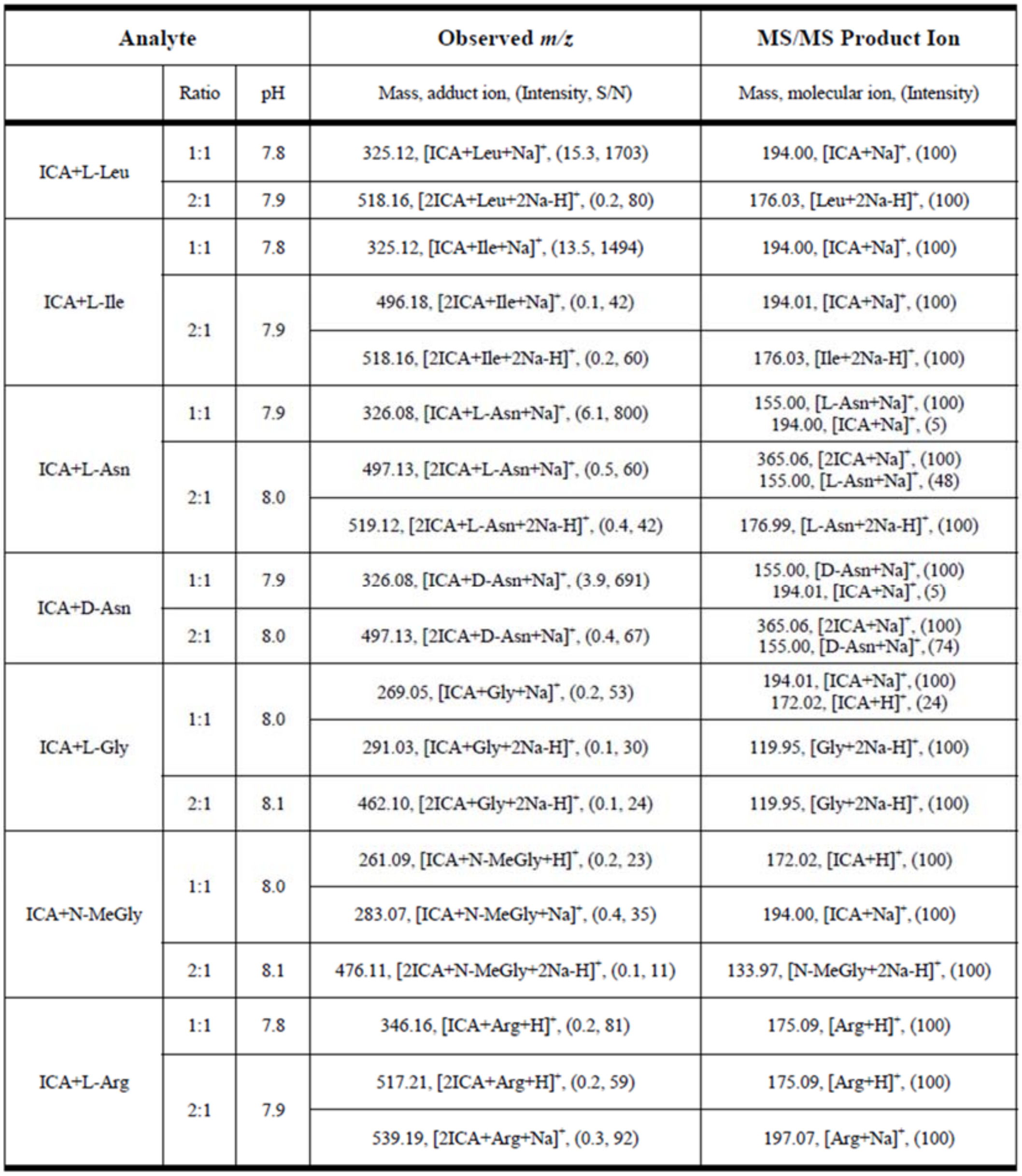
Analyte	Calculated Monoisotopic Mass	Solution pH	Molecular form	[1] Observed *m/z*
L-Leu	131.0946	8.1		154.04, $[M+Na]^+$, (82) 176.03, $[M+2Na-H]^+$, (85) 285.16, $[2M+Na]^+$, (100)
L-Ile	131.0946	8.0		154.04, $[M+Na]^+$, (65) 176.03, $[M+2Na-H]^+$, (100) 285.16, $[2M+Na]^+$, (50)
L-Asn	132.0535	8.1		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, $[ICA'+Na]^+$

Table 7. Structure Information and MS Data of Individual Amino Acids and ICA.

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.



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)

Table 8. Characteristic ESIMS of ICA-Amino Acids 1:1 & 2:1 Mixtures and Their MS/MS Products.

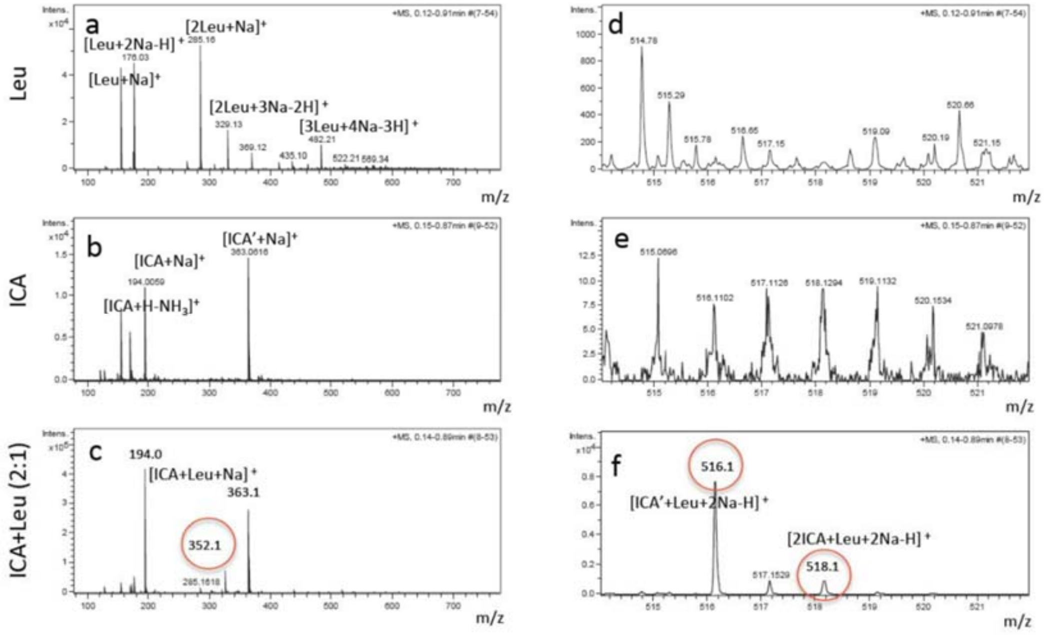


Figure 33. Examples of ES-MS Spectra of Pure Compounds and Complexes.

(a) Leucine, (b) ICA, (c) ICA+ Leucine at 2:1 ratio. (d), (e), (f) show spectra at higher resolution.

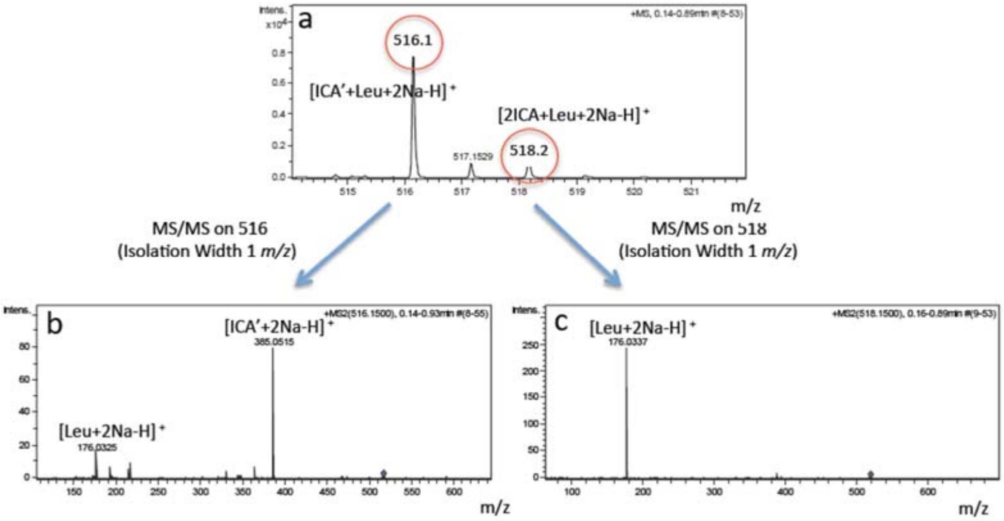


Figure 34. Examples of MS-MS Spectra.

Two peaks are found in 2:1 mixtures of ICA with Leucine, circled in (a). MS-MS shows that the peak at 516 Daltons 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 nonoxidized ICA molecules with one Leucine.

3.3.5. Reproducibility of the SVM Analysis

The previous sections demonstrate that STM-RT is able to recognize amino acids and peptides through hydrogen bonding between ICA and target analytes and SVM analysis. The bonding mechanism was confirmed through mass spectroscopy measurements that ICA can form a complex with amino acids in 2:1 ratio. Then, it can reach the question about analysis method, such as 'How reproducible are the tunneling data?' and 'How transferrable is the SVM training?'.

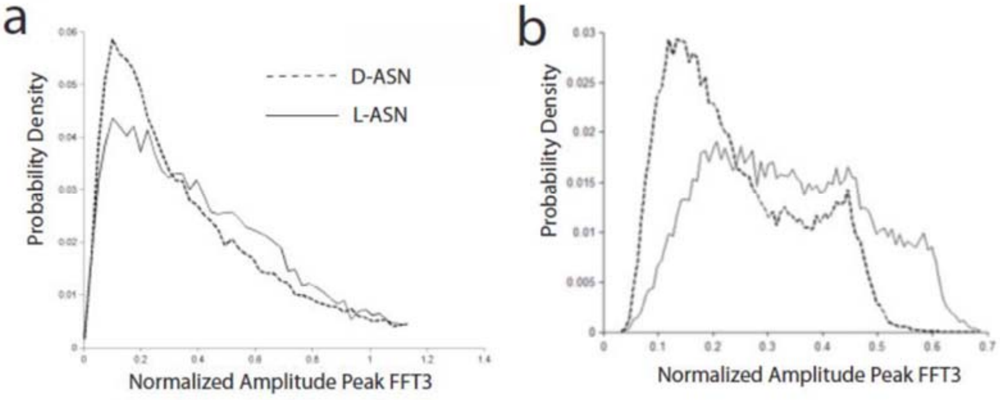


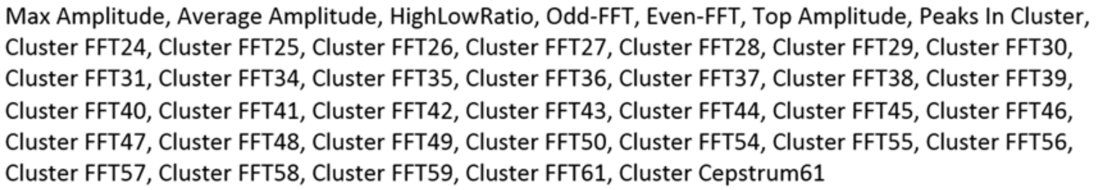
Figure 35. Correction for Instrumental Frequency Response.

Showing the amplitude distribution for FFT3 (5.6 – 8.3 kHz) for L- and D-ASN before (a) and after (b) correction of the Fourier amplitudes by division of the signals by the Fourier amplitudes of the background signal. Large differences between the analytes at low amplitudes were masked by the instrumental response in (a).

In order to check the reproducibility of RT and SVM, we analyzed multiple data sets for each analyte by selecting features and setting SVM running parameters which give robust results. There is 161 starting features as described in Table 4. There are two types of features; for individual spikes (or clusters) and for shapes of spikes (or clusters). The Fourier and cepstrum components are used for shape features.[60] Here, Fourier and cepstrum components were corrected for the instrumental frequency response as shown in Figure 35. The seven pure amino acids RT data produced total 30,000 spikes (3,000 clusters) and corresponding 161 features. Figure 36 shows the correlation among all the features which gives 40 features are highly correlated as shown in Table 9. After removing highly correlated features, it reduced the total features to 121. In addition, the variance from run to run of the same analyte was considered which does not vary from one analyte to others. The features in Table 10 is the bottom 15 features ranked of out-of-group fluctuations to in-group fluctuations from the seven pure amino acids data pool. By removing the bad features of in-/out-group ranking, it reduces the sensitivity to experimental artifacts.

Next, noise spikes were removed. The noise spikes were determined by common peak assignments of all analytes by SVM. The noise filtering stiffness was adjusted by the soft margin of SVM running parameters which determined the broadening of the hyper plane boundaries. The higher soft margin improves the classification accuracy with removing more signals as shown in Figure 37. After filtering process, the SVM trains on a subset (~10%) of the data and test the supporting vectors on the rest of data.

The training data set was randomly selected and all the described data process was repeated to make sure that the outcomes fluctuations were small.



Max Amplitude, Average Amplitude, HighLowRatio, Odd-FFT, Even-FFT, Top Amplitude, Peaks In Cluster, Cluster FFT24, Cluster FFT25, Cluster FFT26, Cluster FFT27, Cluster FFT28, Cluster FFT29, Cluster FFT30, Cluster FFT31, Cluster FFT34, Cluster FFT35, Cluster FFT36, Cluster FFT37, Cluster FFT38, Cluster FFT39, Cluster FFT40, Cluster FFT41, Cluster FFT42, Cluster FFT43, Cluster FFT44, Cluster FFT45, Cluster FFT46, Cluster FFT47, Cluster FFT48, Cluster FFT49, Cluster FFT50, Cluster FFT54, Cluster FFT55, Cluster FFT56, Cluster FFT57, Cluster FFT58, Cluster FFT59, Cluster FFT61, Cluster Cepstrum61

Table 9. Features which are Highly Correlated with σ≥0.7.

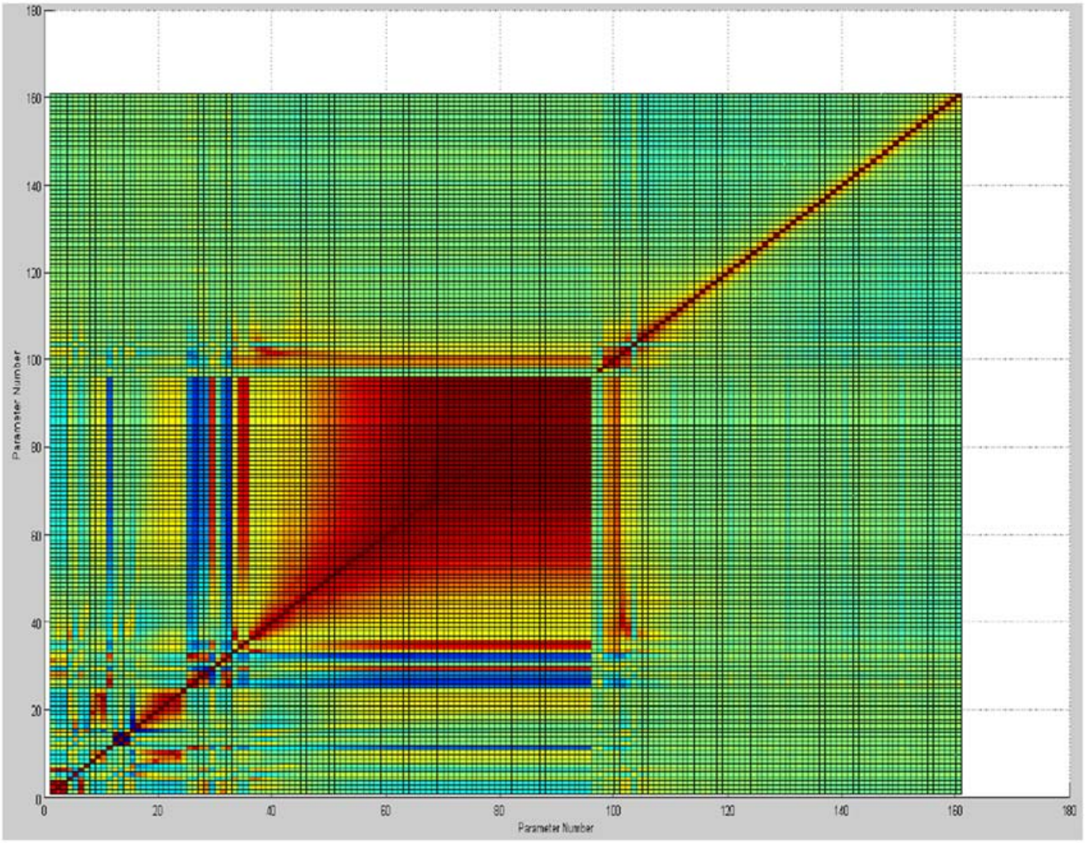
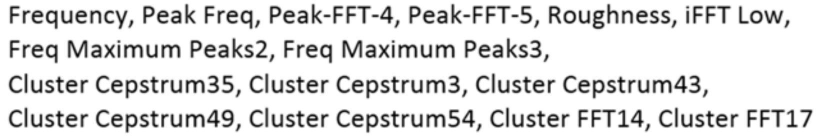


Figure 36. Correlation Map for All 161 Features.

Each axis lists the feature number as labeled in Table 4. Blue = -1, red = +1. The large red region in the middle reflects a high degree of correlation among the higher frequency cluster FFT components.



Frequency, Peak Freq, Peak-FFT-4, Peak-FFT-5, Roughness, iFFT Low,
Freq Maximum Peaks2, Freq Maximum Peaks3,
Cluster Cepstrum35, Cluster Cepstrum3, Cluster Cepstrum43,
Cluster Cepstrum49, Cluster Cepstrum54, Cluster FFT14, Cluster FFT17

Table 10. Features Removed by Ranking of Out-of-group Fluctuations to In-group Fluctuations.

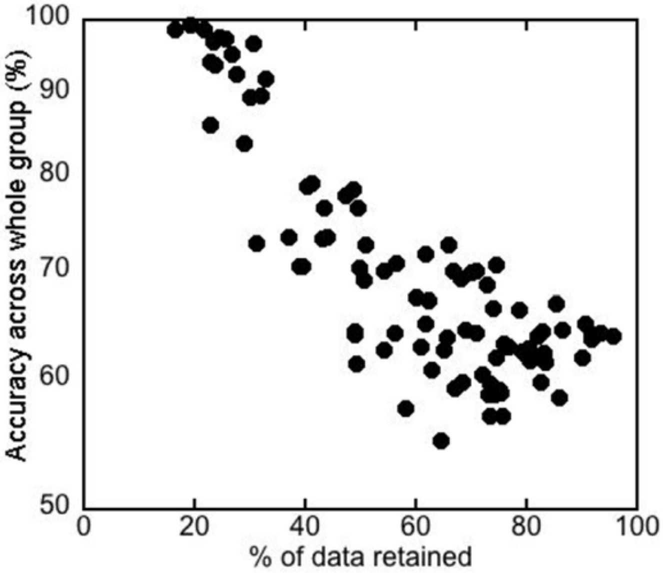


Figure 37. Scatter Plot of Data Rejection by Sort Margin Value and Accuracy.

Scatter plot of the average accuracy for calling all seven analytes from a single spike as a function of the percent of data rejected as common by broadening the soft margins of the SVM rejection filter. Repeated points are for different feature combinations. There is a "sweet point" at about 30% data retention. Further filtering of common signals does little to improve accuracy beyond this point.

3.4 Conclusion

STM has been fascinated for single molecule measurements and can be a powerful tool for sequencing device. Even the recognition molecule (ICA) was designed to capture DNA bases, it also enable to bond amino acids and generates RT signals. RT is complicate signals that gives a possibility of a new molecular spectroscopy at single molecule level. This study shows that STM-RT is able to identify pure amino acids, distinguish the constituent amino acids from mixture and count the ratio, and recognize not only enantiomers and isobaric isomers also peptides. Though RT signals are complicate and includes many information about the captured molecule between ICA, SVM enables to analyze RT signals and gives high accuracy of the classification of the analytes.

One has studied that nanopore device is able to carry out continuous strand sequencing.[66] Integrated RT junctions into nanopores, it may prove the limit of the present single molecule techniques.

CHAPTER 4

ELECTRONIC SINGLE MOLECULE IDENTIFICATION OF CARBOHYDRATE ISOMERS BY RECOGNITION TUNNELING

4.1. Introduction

Though glycans play an important role in most biological processes, it is difficult to read the sequence of glycans. This is because of the fact that there are huge number of possible isomeric configurations for a short oligosaccharide. Recently, ion-mobility spectrometry-mass spectrometry showed possibility of stereoisomers identification.[67] STM-RT is also promising technique for glycan analysis. The present work shows how to identify stereoisomeric carbohydrates and individual carbohydrate from 11 different molecules data pool.[68]

4.1.1. Carbohydrates

A carbohydrate broadly means a biological molecule which is consisted of carbon, hydrogen and oxygen atoms, usually in a constant ratio of 2:1 for hydrogen and oxygen. The term of saccharide is commonly used in biochemistry. It can be categorized into four chemical groups; monosaccharides, disaccharides, oligosaccharides and polysaccharides sorted by numbers of carbohydrates. The term of glycan means polysaccharide linked glycosidically to proteins. By the glycosylation linker, it can be grouped into N-linked glycans and O-linked glycans.

Glycosylation is one of the most important post-translational modifications of cell proteins. Glycan modified proteins play a central role as mediators in a wide range of biological and physiological processes, such as protein folding, cell adhesion, cell communication, gene expression, pathogen recognition and cellular immunity. [69][70][71][72][73][74][75] Figure 38 shows some functions of glycans. Glycan functions depends on the structure of the oligosaccharides which are covalently attached to proteins through two motifs. The N-linked glycans are attached to the amide group of an asparagine and the O-linked glycans are attached to the hydroxyl group on serine or threonine. Structural isomerism (epimers, anomers, regioisomers and branched sequences) makes complicate glycan structures and hard to analyze glycan.[76] For instance, hexa-saccharide can have more than 10^{12} structures. Mass-Spectrometry (MS), Nuclear Magnetic Resonance (NMR), High Performance Liquid Chromatography (HPLC) and High-Performance Anion-Exchange chromatography with Pulsed Amperometric Detection (HPAE-PAD) are commonly used for glycan analysis. [77][78][79][80][81][82]

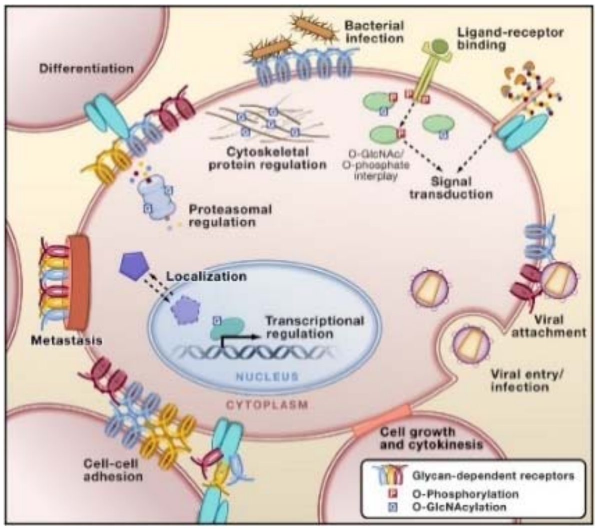


Figure 38. Glycans Permeate Cellular Biology.

Complex glycans at the cell surface are targets of microbes and viruses, regulate cell adhesion and development, influence metastasis of cancer cells, and regulate myriad receptor: ligand interactions. Glycans within the secretory pathway regulate protein quality control, turnover, and trafficking of molecules to organelles. Nucleocytoplasmic O-linked N-acetylglucosamine (O-GlcNAc) has extensive crosstalk with phosphorylation to regulate signaling, cytoskeletal functions, and gene expression in response to nutrients and stress. Figure taken from ref [69].

4.1.2. Current Technique to Analyze Carbohydrates

One of the most popular methods for glycan structural analysis is Nuclear Magnetic Resonance (NMR). NMR is non-destructive and measures the magnetic distortion of glycan. The combination of one-dimensional proton and carbon NMR spectra and two-dimensional homonuclear and heteronuclear NMR methods provides the ratio of the components for monosaccharides and their anomeric bonds.[83] Nano-NMR analysis has been shown high resolution spectrum which enables distinguishing linkage site of samples and mixtures of N- / O-linked glycans.[84] Though NMR is fascinate to study glycan-protein interactions because of its fast exchange[85], it requires a quite large amounts of sample (~milligrams) and long data acquisition time (~hours or days). And it cannot distinguish small amounts of coexisting isomers.[86]

Mass-Spectrometry (MS) is a power technique to investigate glycans due to its high resolution and mass accuracy which provides glycans profiling and structural information. In contrast to the HPLC method, it requires a large amount of glycan samples for a single MS spectrum. Matrix-assisted laser desorption/ionization time-of-filight (MALDI-TOF) MS is the most popular technique. The MALDI-TOF MS provides mass weight data of the sample, which can assign monosaccharide structures in a pure oligosaccharide. However, it is unable to identify coexisting isomers without additional chemical steps, since they share a molecular weight.[87][88] The problem has recently been addressed with ion-mobility spectrometry-mass spectrometry (IM-MS) by measuring collision cross-section of isomer samples.[67] However, IM-MS cannot identify epimers which have almost identical collision cross-sections. This study shows that recognition tunneling with STM is able to identify carbohydrates with label-free and at single molecule level.

4.1.3. Recognition Tunneling

As discussed in chapter3, recognition tunneling is capable to analyze and sequence biological samples. Briefly, capturing a molecule tethered to two electrodes in a few nanometers generates characteristic electron tunneling current spikes. The STM-RT has been used to identify individual nucleobases, amino acids, and peptides.[35][36][53][89][90] In the present study, the capture molecule was 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide which has multiple hydrogen bond donors and acceptors for recognition and alkyl chain terminated with a thiol group to bind with electrodes. [59]

4.1.4. Machine Learning Algorithm for Data Analysis

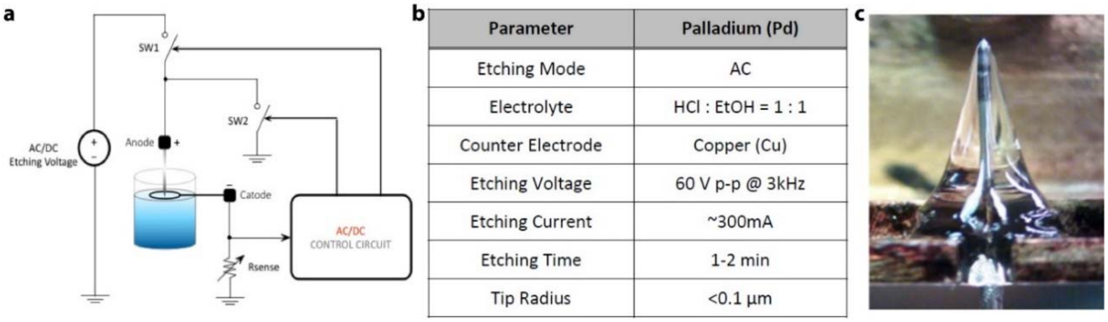
Supporting Vector Machine (SVM) was used for STM-RT data analysis in this study. The details of the machine learning and SVM will be discussed in the following chapter 4.2.7.

4.2. Experimental Methods

4.2.1. Preparation of Probes and Substrates

4.2.1.1 Preparation of Electrodes; STM Probe and Substrate

The electrode preparation followed the procedure that developed in my lab. [91] The STM probes were made from 0.25 mm Pd wire (California Fine Wires) by AC electrochemical etching in mixed solutions of HCl and ethanol (1:1) as shown in Figure 39. The etched tip was insulated with high density polyethylene (HDPE) leaving a few tens of nanometers tip apex. The substrate was prepared by using electron-beam evaporator (Lesker PVD 75). First, a 10 nm titanium was deposited on a 750 µm silicon wafer as an adhesion layer, and a 100 nm thick palladium was deposited.



Parameter	Palladium (Pd)
Etching Mode	AC
Electrolyte	HCl : EtOH = 1 : 1
Counter Electrode	Copper (Cu)
Etching Voltage	60 V p-p @ 3kHz
Etching Current	~300mA
Etching Time	1-2 min
Tip Radius	<0.1 μm

Figure 39. STM Probe Etching Circuit and Conditions.

(a) Schematic circuit diagram. (b) Etching conditions for Pd etching. Figure taken from ref [92]. (c) Image of coating a tip with HDPE.

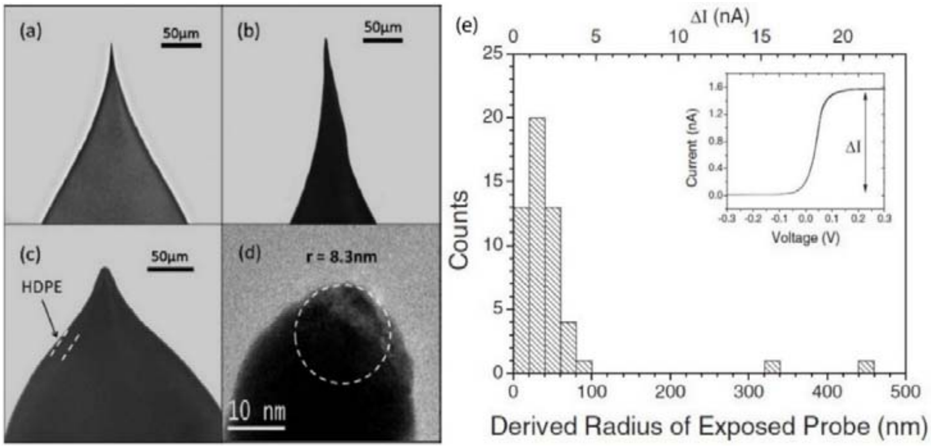


Figure 40. Optical Images of Probes and Saturation Current.

Optical images under 250× magnification of (a) a good etched gold STM probe which is smooth, straight, and sharp such that it reaches an apex of radius less than 1 μm, (b) a poor etched gold STM probe which is not smooth or sharp, such that the apex of this tip is visibly rounded, and (c) a good coated STM probe with a smooth and continuous coating which comes to a point at the apex, and has no visible protrusions there. An TEM image (d) of typical good STM probe with radius of curvature equal to 8.3 nm in this case. (e) The saturation current (ΔI) of coated STM probes was measured by cyclic voltammetry in 100 mM $K_3Fe(CN)_6$ in 1 M KCl (see inset). Figure taken from ref [91].

4.2.1.2 Functionalization and Characterization of Palladium Electrodes

The insulated probes were gently cleaned by ethanol (200 proof), dried with a nitrogen flow, immersed in an ethanolic solution of ICA (0.5 mM, degassed by argon) for 20 hours at room temperature, and then gently rinsed with ethanol and dried with nitrogen.

All the STM probes and substrate were freshly prepared before each experiment. Palladium substrates were functionalized with ICA in the same way as the STM probes were prepared and characterized with various physical and chemical tools.

The modified substrate was characterized with ellipsometry. The Gaertner L 123b Ellipsometer (Gaerner Scientific Corporation) was used for measuring the thickness of ICA monolayer. The palladium substrate was hydrogen flame annealed immediately prior to baseline measurements, and was modified with ICA in the same manner which explained in the previous section. A refractive index of the organic thin films was assumed 1.50. [93] The measured ICA monolayer was 9.10±0.41 [Å] which was collected on five spots of two samples. The ICA molecule was estimated by ChemDraw 3D to be ~8.3 Å long.

Static water contact angle were measured using Easydrop Drop Shape Analysis System (KRÜSS GmbH, Hamburg). The palladium substrate was annealed with hydrogen and modified for SAM formation on the substrate. 1μL of water was dropped on the surface. 5-6 measurements were taken on different locations of the each functionalized and bare palladium substrates. The contact angle for the bare palladium substrate was 8.3 ± 2.0^o, for the ICA monolayer 33.1 ± 5.1^o

The FTIR spectra was obtained by using a Nicolet 6700 FT-IR (Thermo Electron Corporation) with a surface grazing angle device (Smart SAGA, Thermo Electron Corporation) at 4 cm^{-1} resolution with 256 scans for ICA monolayer and with an attenuated total reflection accessory (Smart Orbit, Thermo Electron Corporation) for ICA powder under 4 cm^{-1} resolution with 128 scans in the 6000-400 cm^{-1} (shown 3500-700 cm^{-1} in Figure 41). The spectrum of the ICA powder sample gives two of broad bands;

3400-2800 cm^{-1} and 1700-700 cm^{-1}. The higher band (3400-2800 cm^{-1}) implies the intermolecular hydrogen bonding interactions. In contrast, the ICA monolayer shows very sharp peaks in the same region due to the removal of the intermolecular hydrogen bonds. Both spectra show the vibrations of the amide function in the region of 1700-1600 cm^{-1}.

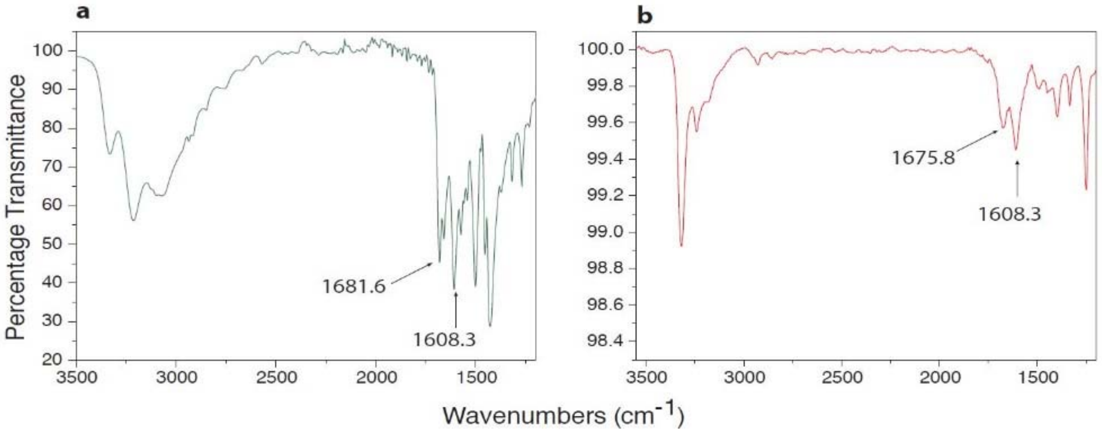
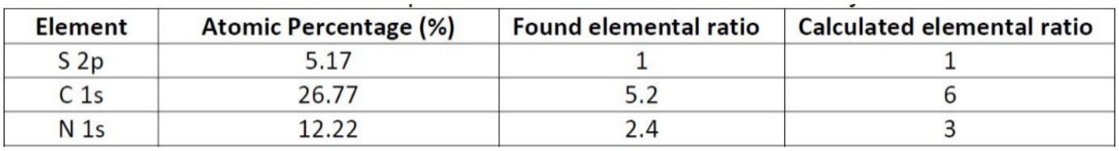


Figure 41. FTIR Spectrum of ICA.
(a) ICA powder and (b) ICA monolayer.

X-ray photoelectron spectra were obtained using a VG ESCALAB 220i-XL photoelectron spectrometer and Al-Kα radiation (15 keV) at 6×10^{10} mbar base pressure. C(1s), Pd(3d), N(1s) and S(2p) core level high resolution spectra were recorded at a pass energy of 20 eV and wide scan spectra were obtained at pass energy of 150 eV. CasaXPS software was used for data analysis. C(1s), N(1s), and S(2p) core peaks were fitted and the ICA element ratio was calculated through area integral of peaks. Table 11 shows the found elemental ratio, which is close to the calculated ratio.



Element	Atomic Percentage (%)	Found elemental ratio	Calculated elemental ratio
S 2p	5.17	1	1
C 1s	26.77	5.2	6
N 1s	12.22	2.4	3

Table 11. Element Compositions of the Imidazole Monolayer from XPS

4.2.2. Chemicals and Reagents

All the monosaccharides, maltose and cellobiose were purchased from Sigma-Aldrich (99% purity). The two disaccharides compounds, 4-O-sulfated-chondroitin sulfate disaccharide and 6-O-sulfated chondroitin sulfate disaccharide, were synthesized by Dr. Xu Wang's lab. All the samples are dissolved in pH 7.4, 1 mM sodium phosphate buffer solution. Water was purified by a Milli-Q system for ~18MΩ-cm and less than 5 ppb of total organic carbon contamination. All the sample solution's concentration is 100 µM and was prepared freshly right before the measurements.

4.2.3. STM Experimental Details

PicoSPM (Agilent Technologies) was used with customized LabView interface for data acquisition. The tip was tested to ensure current leakage is less than 1 pA in PB solution at 500 mV bias. Current set point is 4 pA which corresponds to ~2.5 nm gap size between two electrodes [94] and the probe approached to substrate under 1.0 integral and proportional gain servo control. The surface was scanned to ensure that the probe is not over-coated by high density poly-ethylene (HDPE), so electrodes are good condition for RT measurement. After the clear grain structure of Pd substrate was obtained, probe was withdrawn 1 um and the bias was turned off to avoid possible damages on ICA layer during

2 hour instrument stabilization. Probe was re-engaged, and the integral and proportional gain were set to 0.1. The control (1mM PBS at pH 7.4) was collected before every sugar experiment and we usually recorded four runs for each analyte. Different batches of probes and substrates were used for each run. Analyte tunneling data was collected 5-10 min for control (buffer solution) and 30-40 min for analyte solution. The gain values were determined by noise spectrum (Power Spectral Density) under various gain values. The Figure 42 shows that servo control distorts the signal. With 0.1 for the integral and proportional gain, spectrum under 30 Hz (corresponding 33 ms) is suppressed. The gain is long enough not to distort all spikes but some of long spikes. Tunneling current was Fourier transformed and plotted as a spectral density calculated by

$$PSD = \frac{2}{N \cdot \Delta t} \frac{Re^2 + Im^2}{f} \qquad equation\ (14)$$

where $N = 50{,}000$ and $\Delta t = 20\ \mu s$.

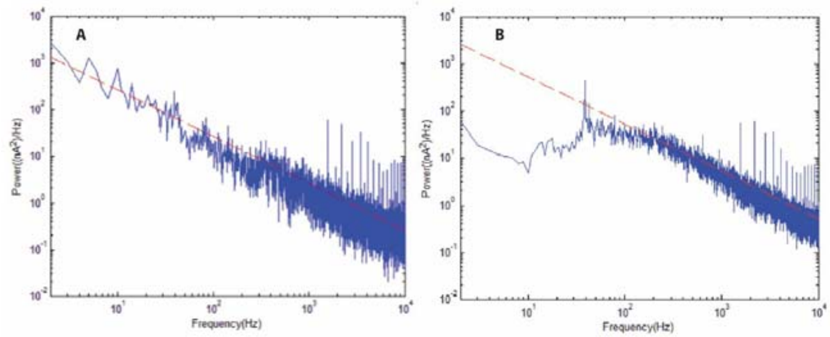


Figure 42. Noise Spectrum of STM.

(a) without servo control. (b) with servo control. Blue lines are the noise spectrum, and red lines are fits to 1/f spectrum.

4.2.4. ESI-MS

ICA (200 μM) and carbohydrate (100 μM each) solutions were respectively prepared in water and sparged with argon. Each sample solution was injected into a Bruker maXis 4G electrospray ionization quadrupole time-of-flight (ESI-Q-TOF) mass spectrometer at a 3 μL/min infusion rate via syringe pump. Tandem (MS/MS) mass spectrometry was used to observe product ion peaks from molecular complex ion peaks to confirm the composition of the molecular complex. The ESI source was equipped with a microflow nebulizer needle operated in a positive ion mode. The spray needle was held at ground and the inlet capillary set to -4500 V. The end plate offset was set to -500 V. The nebulizer gas and dry gas (N_2) were set to 1.2 Bar and 1.5 L/min, respectively, and the dry gas was heated to 220°C. In TOF-only mode the quadrupole ion energy was set to 4 eV and the collision energy was set to 1 eV. Collision gas (Ar) was set to a flow rate of 20%. In most cases MS/MS experiments were conducted with a precursor ion isolation width of 3 *m/z* units. However, if other ions were present in this range precursor ion isolation width was set to 1 *m/z* unit. Collision energy was set to 10-20 eV, which was sufficient to fragment non-covalent complexes. Each spectrum was recorded over a time period of 0.5 to 1 min. Typically a spectrum acquired for one minute is an accumulation of 60 separate recorded mass spectra averaged across 1 min time period. Signal to noise ratio greater than three (S/N>3) was used to define the limit of detection. Due to the lack of an acid modifier in the infused solutions, most carbohydrates and molecular complexes were observed as single or multiply sodium ions $[M+nNa-(n-1)H]^+$ rather than as protonated molecular form $[M + H]^+$. Average mass accuracy was within 0.025 Da.

4.2.5. Feature Extraction

Once the tunneling current signals were collected, some features should be defined to represent the signals and to analyze the data (to classify the analyte molecules). For example, Fourier transformation is the most popular feature for electric signal analysis. This study used three different domains; time, frequency, and cepstrum. The primary features are defined in time domain, such as peak amplitude and peak width etc. (Figure 43(a)) The secondary features are defined in transformed domain such as frequency domain for FFT components and cepstrum(quefrency) domain for cepstrum components.

The primary features are in time domain. The baseline of raw tunneling current 4 pA was shifted to zero and all the current spikes above 15 pA was characterized as described in Supplement Table 12. The clusters were identified by applying Gaussians window (4096 data points and unit height) to each peaks. The Gaussian traces were summed and a cluster was assigned when the sum exceeds 0.1.[54] Though the peak features were characterized with the spikes above 15 pA, the cluster features include all the spike data within the assigned region. An example of a determined cluster is shown in Figure 43(a) and some example features are labeled on the figure. The cluster was furrier transformed with 25 kHz window which is the Nyquist frequency of amplifier and the whole frequency range is down-sampled to small windows. [Nature Nanotech paper] The feature names peakFFT or clusterFFT was used the same window size (red in Figure 43(b)) for sampling, but peakFFT_Whole and clusterFFT_Whole used various different window size (green in Figure 43(b)) that the lower frequency has smaller sampling window size. (In the figure, it shows only six windows for the simplicity.) The third domain is cepstrum which is the inverse Fourier transform of the logarithm of the Fourier transform signal. The spectrum in quefrency domain is also down sampled into the even size of windows as shown in Figure 43(c).

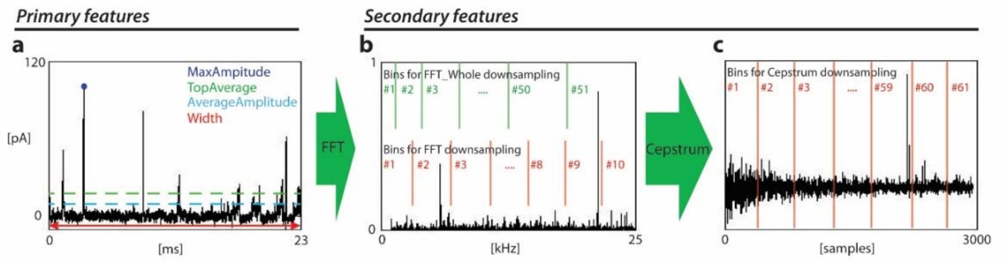
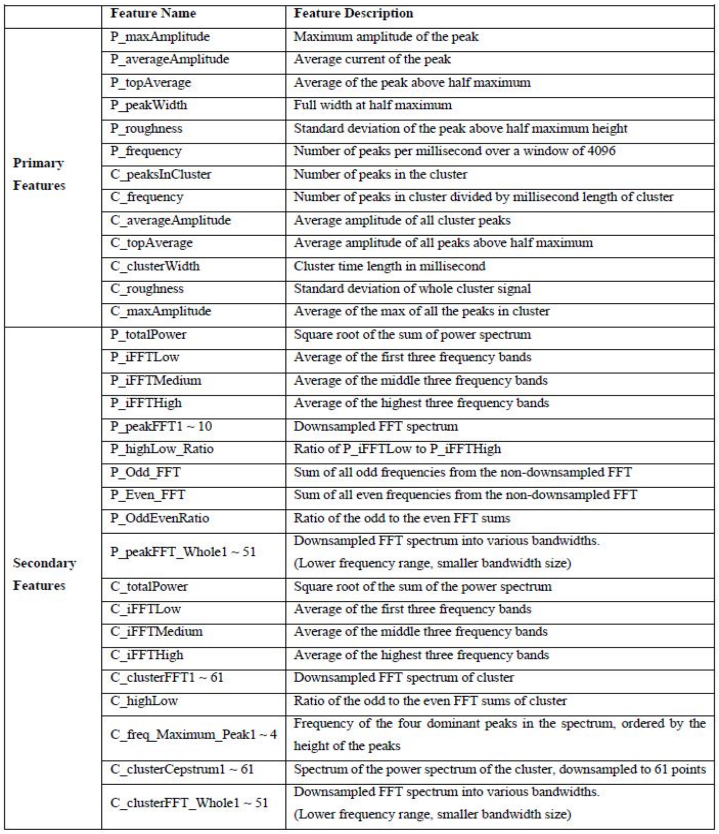


Figure 43. Feature Extraction at Three Domains.

(a) primary features in time domain, secondary features in (b) frequency domain and (c) cepstrum domain.



	Feature Name	Feature Description
Primary Features	P_maxAmplitude	Maximum amplitude of the peak
	P_averageAmplitude	Average current of the peak
	P_topAverage	Average of the peak above half maximum
	P_peakWidth	Full width at half maximum
	P_roughness	Standard deviation of the peak above half maximum height
	P_frequency	Number of peaks per millisecond over a window of 4096
	C_peaksInCluster	Number of peaks in the cluster
	C_frequency	Number of peaks in cluster divided by millisecond length of cluster
	C_averageAmplitude	Average amplitude of all cluster peaks
	C_topAverage	Average amplitude of all peaks above half maximum
	C_clusterWidth	Cluster time length in millisecond
	C_roughness	Standard deviation of whole cluster signal
	C_maxAmplitude	Average of the max of all the peaks in cluster
Secondary Features	P_totalPower	Square root of the sum of power spectrum
	P_iFFTLow	Average of the first three frequency bands
	P_iFFTMedium	Average of the middle three frequency bands
	P_iFFTHigh	Average of the highest three frequency bands
	P_peakFFT1 ~ 10	Downsampled FFT spectrum
	P_highLow_Ratio	Ratio of P_iFFTLow to P_iFFTHigh
	P_Odd_FFT	Sum of all odd frequencies from the non-downsampled FFT
	P_Even_FFT	Sum of all even frequencies from the non-downsampled FFT
	P_OddEvenRatio	Ratio of the odd to the even FFT sums
	P_peakFFT_Whole1 ~ 51	Downsampled FFT spectrum into various bandwidths. (Lower frequency range, smaller bandwidth size)
	C_totalPower	Square root of the sum of the power spectrum
	C_iFFTLow	Average of the first three frequency bands
	C_iFFTMedium	Average of the middle three frequency bands
	C_iFFTHigh	Average of the highest three frequency bands
	C_clusterFFT1 ~ 61	Downsampled FFT spectrum of cluster
	C_highLow	Ratio of the odd to the even FFT sums of cluster
	C_freq_Maximum_Peak1 ~ 4	Frequency of the four dominant peaks in the spectrum, ordered by the height of the peaks
	C_clusterCepstrum1 ~ 61	Spectrum of the power spectrum of the cluster, downsampled to 61 points
	C_clusterFFT_Whole1 ~ 51	Downsampled FFT spectrum into various bandwidths. (Lower frequency range, smaller bandwidth size)

Table 12. 264 Starting Features Used in the Signal Analysis.

Details of their calculation are given in Chang et al. The first letter of each feature name means peak (for P) or cluster (for C). The primary features are defined in time domain, the secondary features are defined in frequency domain or others, after applying Fourier transformation to the raw time domain trace. Table taken from ref [54].

Once all the features are determined, they were normalized and scaled to be avoid that large numeric features dominate those in small range. The mean of each feature was shifted to be zero and scaled to make standard deviation to 1.

4.2.6. Data Analysis (Machine Learning)

STM-RT provides a huge data. I collected four data sets for each analyte and each data set is consisted of 5~10 min current trace of control (buffer solution) and 40~60 min trace of the analyte. From the feature extraction, each peak has 264 starting features. Then, for example, the data pool of methyl α-D-glucopyranoside and methyl β-D-glucopyranoside has ~32,000 data points (total number of features data points is 8.5 millions.) As the number of analytes increases, the data analysis becomes more complicate.

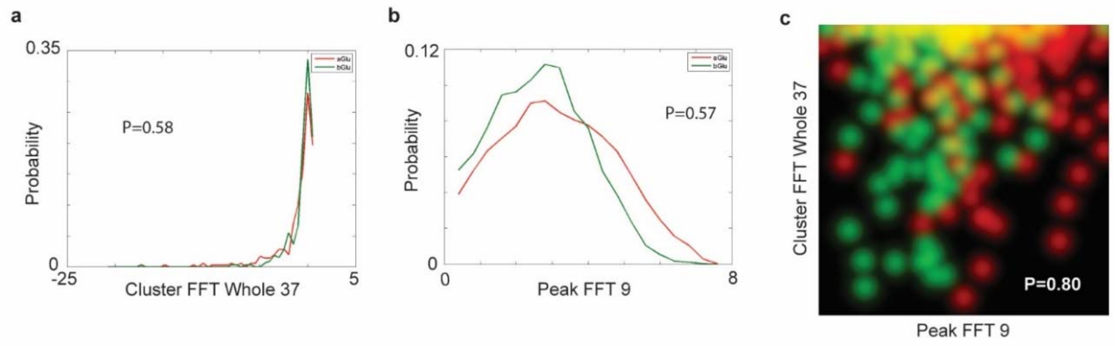
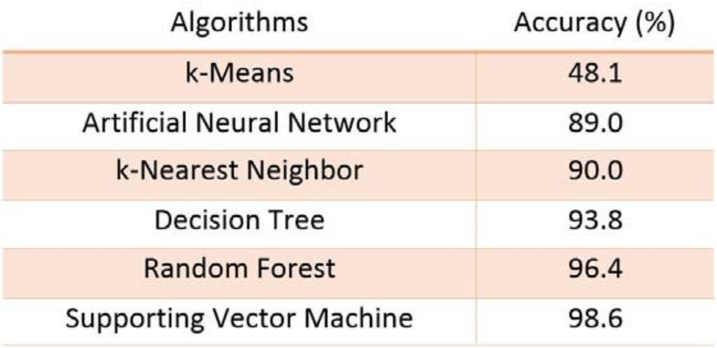


Figure 44. Distribution of Features.

(a) cluster FFT Whole 37 and (b) peak FFT 9 and (c) scatter plot of the two features. Distributions of signal features are broad and overlapped (red = α-^{M}Glu, green = β-^{M}Glu) as shown here for one frequency band in the Fourier transform of signal clusters (cluster FFT whole 37 –a) and for a band in the Fourier transform of individual peaks (Peak FFT 9 –b). Data can only be assigned to one analyte or the other with a probability only marginally above random, P=0.5 (see Methods for details of the signal analysis). However, when the same two distributions are plotted together in a 2D histogram (c) where the brightness of each point represents the frequency with which a particular pair of values occur, the accuracy with which data can be assigned increases to 80%. This accuracy can be improved to ~ 99% using additional signal features. Colors in (c) are mixed so overlapped points are yellow.

The data analysis becomes more difficult and complicate in the case of isomeric molecule recognition, such as methyl α-D-glucopyranoside and methyl β-D-glucopyranoside identification. Because isomers have same molecular components and share many molecular properties (molecular weight and charge etc.). As shown in Figure 44, classification with single features is almost random separation (0.5), 0.58 from Cluster FFT Whole 37 and 0.57 from Peak FFT 9. However, by plotting the two features at the same time, the classification accuracy reached to 0.80. It can be explained by Cover's theorem.[63] Using more features can give higher separation.

For the complicate STM-RT data analysis, there are many well-known computer based analysis technique for big data, Machine Learning. A variety of machine learning algorithms has been developed from the field of computer science. Table 13 shows some accuracy results from six kinds of the algorithms. All the calculations was conducted with the packages in Matlab R2015a. The machine learning can be categorized by two, supervised learning and unsupervised learning.[96] Supervised learning is to find an inferring function from labeled training data, which means all the data is known and determine the best partition or clustering boundary. Unsupervised learning is to discover hidden structure in unlabeled data. In Table 13, k-Means is unsupervised learning algorithm, and all others are supervised one. SVM gives the best classification accuracy.



Algorithms	Accuracy (%)
k-Means	48.1
Artificial Neural Network	89.0
k-Nearest Neighbor	90.0
Decision Tree	93.8
Random Forest	96.4
Supporting Vector Machine	98.6

Table 13. Machine Learning Algorithm Comparison with methyl α-D-glucopyranoside and methyl β-D-glucopyranoside Data Pool.

The k-Means is unsupervised clustering method to separate the data into k clusters. Each cluster is represented by its centroid and defined as the center of the points in the cluster. Each data point is assigned to the cluster whose center is nearest. The calculation is based on the equation (1), minimizing intra-cluster variance or the sum of squares of distances between data and the corresponding cluster centroid.[97]

$$\sum_{i=1}^{N}\left(argmin\left\|x_i - c_j\right\|^2\right) \qquad equation\ (15)$$

Though it is simple and fast, it has some limits. First, it does not yield the same result with different run. This is because of the initial centroid point dependence. Figure 45 shows how the results can be different by the initial centroid points. Second, it requires choosing appropriate number of clusters. Finally, k-Means does not guaranteed to find the optimal configuration.

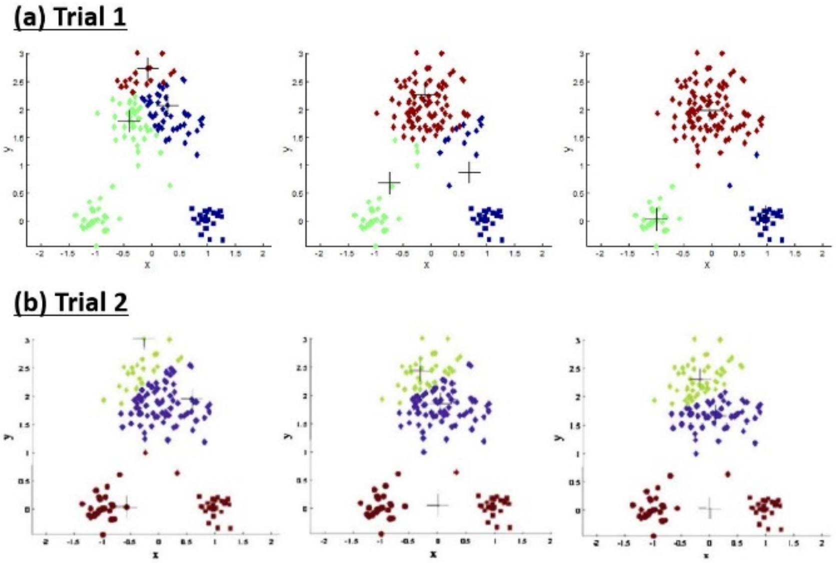


Figure 45. Dependence of Starting Centroid Points in k-Means Clustering.

(a) The initial centroid points are close and within the biggest data point group. The lower two centroids move far away from the top one, and the lower two data groups (green and blue) are assigned into two different clusters. (b) Two of initial centroid points are within the biggest group and the other is away from two centroids and close to the lower data group (brown). Finally, the k-Means assigns the lower two data groups into the same cluster (brown). This shows how the initial centroids selection is critical in k-Means. Figure taken from ref [97].

Artificial Neural Network (ANN) is a mimic of human brain system, network of neurons. Figure 46 shows schematics of neural network. ANN has multiple hidden layers which are networks of transfer functions. The multiple weighted inputs are evaluated by their success at discriminating the classes in training. While the network is training, the weights are adjusted by the separation error between inputs and predetermined classes. Convergence proceeds until the reduction error reaches to threshold.[98] Though ANN is one of the most popular algorithm and gives good performance, it is hard to visualize the network model.[99]

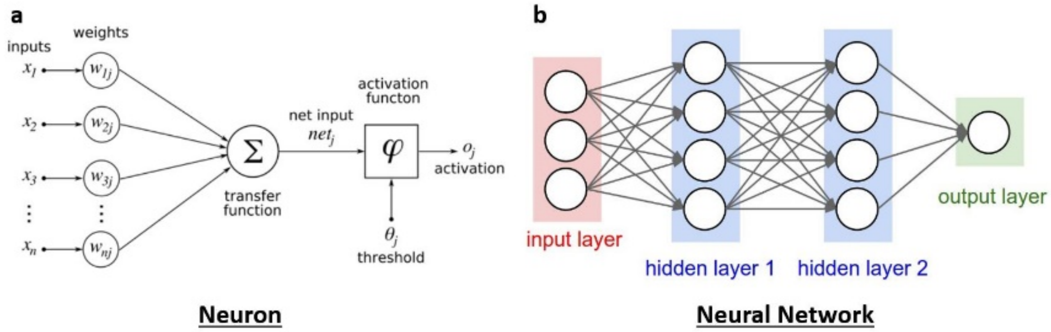


Figure 46. Schematic Diagrams of Neural Network.

(a) neuron and (b) neural network system. Figure taken from https://en.wikibooks.org/wiki/Artificial_ Neural_Networks/Activation_Functions.

The decision tree is non-parametric supervised learning which does mapping observations to conclude target values, in other words Divide-and-Conquer algorithm. As shown in Figure 47, input v follows the tree branch and each node has a condition to assign the input into a class (output). Decision tree is simple to understand and can be easily visualized. However, it can create over-complex trees (overfitting) which do not provide good prediction. Also it is very sensitive to even small variations in the data which results in completely different tree. It may not work well for complicate large data with small internal data variation. [96]

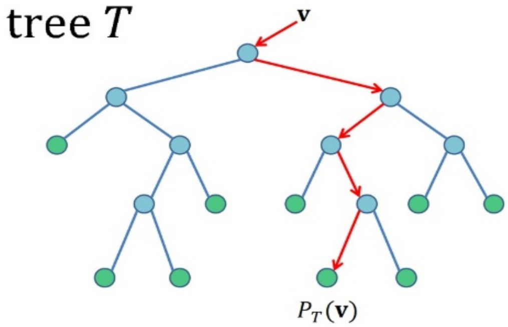


Figure 47. Diagram of Decision Tree.

Random forest is an ensemble classification. It fits multiple decision tree classifiers on various sub-set of the all data. During training, it averages to improve the predictive accuracy and control over-fitting. [96] Random forest is one of the most accurate algorithm and effectively runs with large data. However, it can over-fit even averaging many subsets. Unlike decision tree, it is difficult to visualize because of many subsets and various weighting factors. [97]

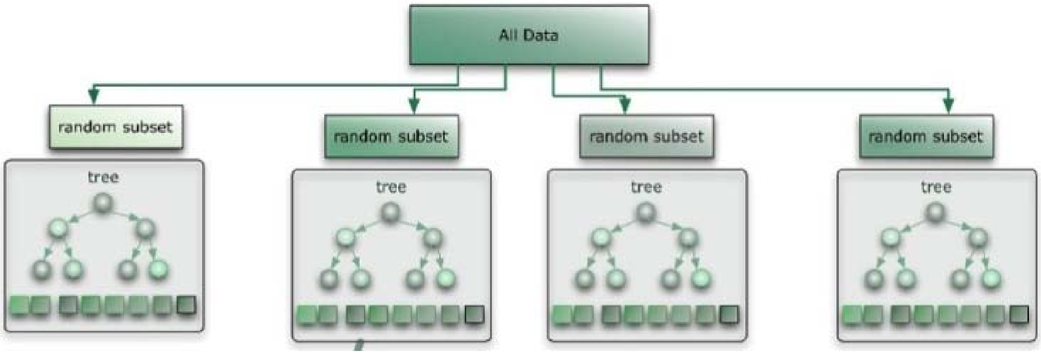


Figure 48. Schematic Diagram of Random Forest. Figure taken from ref [96]

Supporting Vector Machine (SVM) was suggested by Vapnik in 1995.[100] SVM finds a non-linear partitions in high-dimensional space by solving a quadratic optimization problem, equation (2) or (3). The hyperplane is defined by support vectors which defines hyperplane and is a subset of training samples.

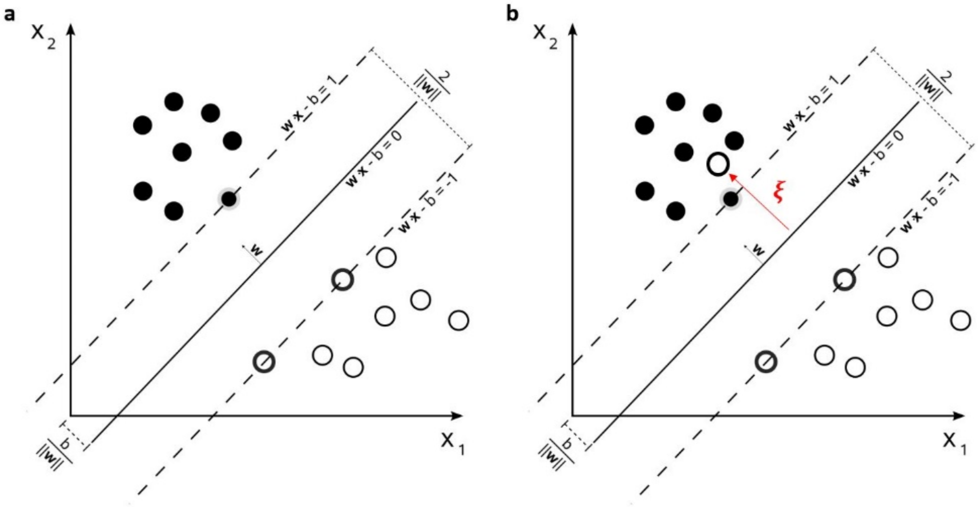

a
X_2
X_1
$\mathbf{w}\mathbf{x}-b=1$
$\mathbf{w}\mathbf{x}-b=0$
$\mathbf{w}\mathbf{x}-b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$
b
X_2
X_1
$\mathbf{w}\mathbf{x}-b=1$
$\mathbf{w}\mathbf{x}-b=0$
$\mathbf{w}\mathbf{x}-b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$
ξ

Figure 49. Schematics of SVM to Determine Maximum Margin Hyperplane.
(a) separable linear case (b) non-separable linear case.

In the case of Figure 49a, the maximum marginal hyperplane is determined by the equation (2)

$$\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 \qquad \textit{equation}\ (16)$$

where $\vec{w}$ is the normal vector to the hyperplane.

For the non-separable linear case of Figure 49(b), SVM finds the margin by introducing a cost parameter C and slack variable ξ. If $0 < \xi \leq 1$, it represents that the data point is between the marginal boundary and the correct side of hyperplane. If $\xi > 1$, the data point is misclassified. The parameter C controls the importance of minimizing $\vec{w}$, equivalent to the maximizing the margin. In other words, creating wide range of safety margin around the partition makes us to maximize the margin. If C is close to 0, there is no cost for the margin constraint. If C is large or close to infinite, the running should pay lots of data points which don't satisfy the constraint. The cost function can be minimized by selecting small number of support vectors.[98][101] Figure 50 shows the relationship between C and width of support vectors.

$$\arg\min_{\mathbf{w},\xi,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \qquad equation\ (17)$$

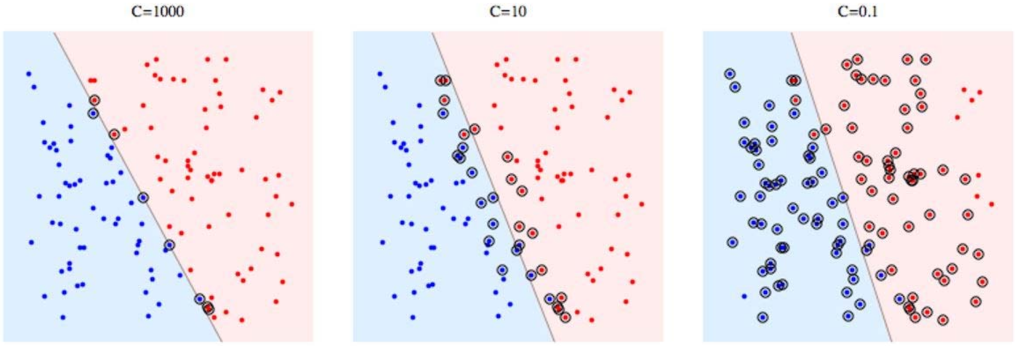
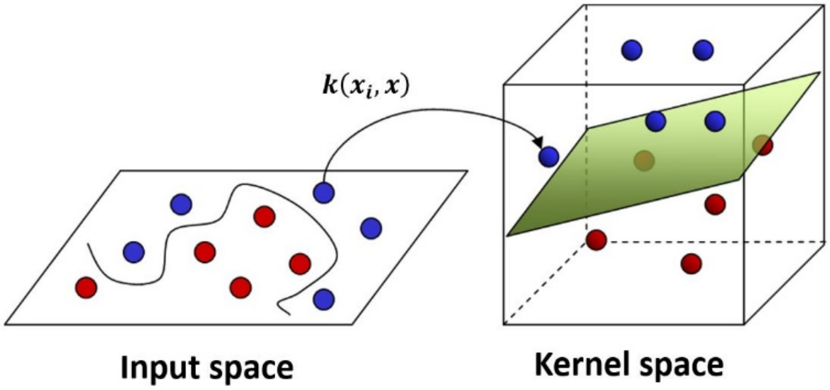


Figure 50. Relations between Cost Parameter and Support Vectors.

Figure taken from ref [102].

Also it is necessary to consider the case of non-linearly separable data, because data is always linear separable. It can be treated by introducing kernel function which is a transformation of input data. Kernel functions make SVM to enable classify non-linear support vectors using a linear hyperplane as described in Figure 51.[96] There are the most commonly used kernel functions in Figure 52. This study used RBF for data analysis.

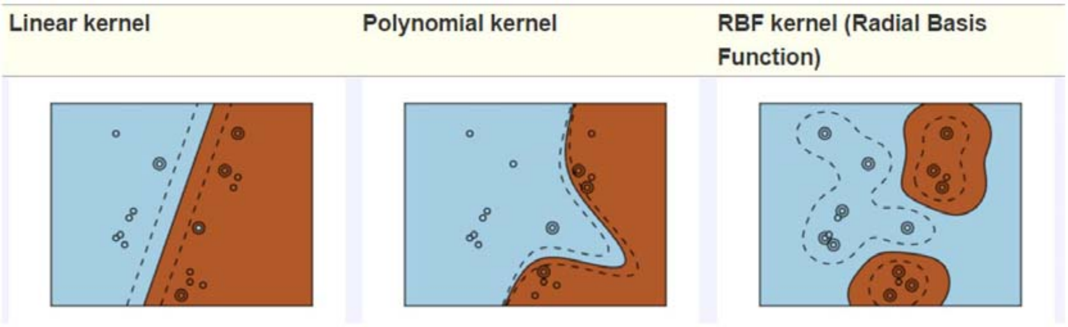


$$\mathrm{w}(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) - b$$

α_i; *weighting parameter*
$k(x_i, x)$; *kernel function*
x_i; *support vectors*

Figure 51. Schematics of Kernel Function Method.

Figure taken from ref [96].



Polynomial Function; $k(x, x') = (\gamma\langle x, x'\rangle + r)^d$

Radial Basis Function(RBF); $k(x, x') = \exp(-\gamma\|x - x'\|^2)$

Figure 52. Examples of Kernel Functions

Figure taken from ref [96].

4.2.7. Details of Data Analysis

A class of the water signals was determined through SVM from control data and all the peaks of analyte signals within the water class is removed. For the SVM analysis, randomly selected 10% data is used to construct support vectors (hyper plane to separate analyte data points) calling as training process, and the rest 90% of data is tested with the previously determined supported vectors calling as testing process.

Once features are extracted, there are 264 starting features. Some features are strongly correlated with other features, so they were removed through the normalized correlation calculation between feature pairs. The features which coefficient is bigger than 0.7 were rejected. The feature variation of the repeated experiments and different analytes are calculated by comparing the single feature histogram with the accumulated histogram. The difference between the repeated runs histogram and the accumulated histogram of the given analyte is assigned as 'in-group' fluctuation (variation of the repeats). The difference of the normalized histogram between the possible analyte pairs is 'out-group' fluctuation (variation of the analytes).[54] The features were ranked by the ratio between the in-group fluctuation and the out-group fluctuation, and the low ranked features were dropped. The survived features were evaluated the classification accuracy and optimized to get the maximum true positive accuracy.

4.2.8. STM-RT Experiment for Binding Affinity

The α-^{M}Glu was dissolved in sodium phosphate buffer (pH 7.4, 1 mM) to make a 1 mM stock solution, which was diluted to various concentrations from 500 μM to 100 pM. For each measurement, an analyte solution (200 μL) was injected into the liquid cell using a syringe attached to a micro filter. After the measurement, the liquid cell and electrodes were rinsed with the phosphate buffer solution (3 mL) through the fluidic channels to obtain a clean control signal. A pair of electrodes was able to carry out three measurements with different concentrations from lower to higher concentration. The measurement at each concentration was repeated at least 2 times. The isotherm absorption data were analyzed in software OriginPro 2016 using the Levenberg-Marquardt algorithm for fitting to a Langmuir equation: $f(x) =a*(b\cdot x)/(1+b\cdot x)$.

4.2.9. SPR Experiment for Binding Affinity

A gold chip was immersed into an absolute ethanol solution of ICA (100 μM) for 24 h, followed by rinsing with absolute ethanol and drying with a nitrogen flow, and used immediately. The instrument Bi 2000 from Biosensing Instrument was used for SPR measurements. An ICA modified gold chip was mounted on the instrument and calibrated with 1% ethanol in a PBS buffer, pH 7.4. A solution of α-^{M}Glu (500 μM) was flowed onto the chip at a rate of 50 μl/min over a period of 1.5 min. Association (k_{on}) and dissociation rate constants (k_{off}) were determined using built-in Biosensing Instrument SPR data analysis software version 2.4.6.

4.3. Results and Discussion

4.3.1. Isomeric Carbohydrates Identification

First, two of anomeric isomers (methyl α-D-glucopyranoside (α-^{M}Glu) and methyl β-D-glucopyranoside (β-^{M}Glu)) were tested to identify the molecules through STM-RT. The only difference between these two molecules is the relative orientation of methoxy group which is colored in red as shown in Figure 53(b),(c). The theoretically calculated capturing configuration and corresponding hydrogen bonding energy shows that isomers can be distinguishable. α-^{M}Glu forms two hydrogen bonds with each ICA, but β-^{M}Glu makes a single hydrogen bond. It results in that α-^{M}Glu makes a more stable complex than β-^{M}Glu. The complexes between ICA and the anomers in a solution was confirmed by ESI-MS measurements, as shown in Table 14 and Table 15 (it includes all the data for other monosaccharides and disaccharides which were used in this study).

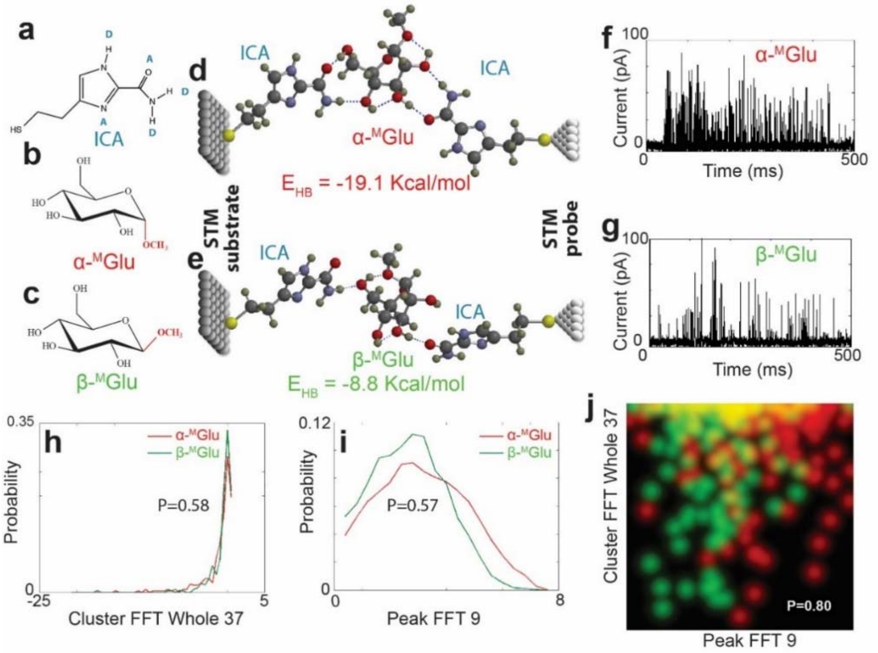
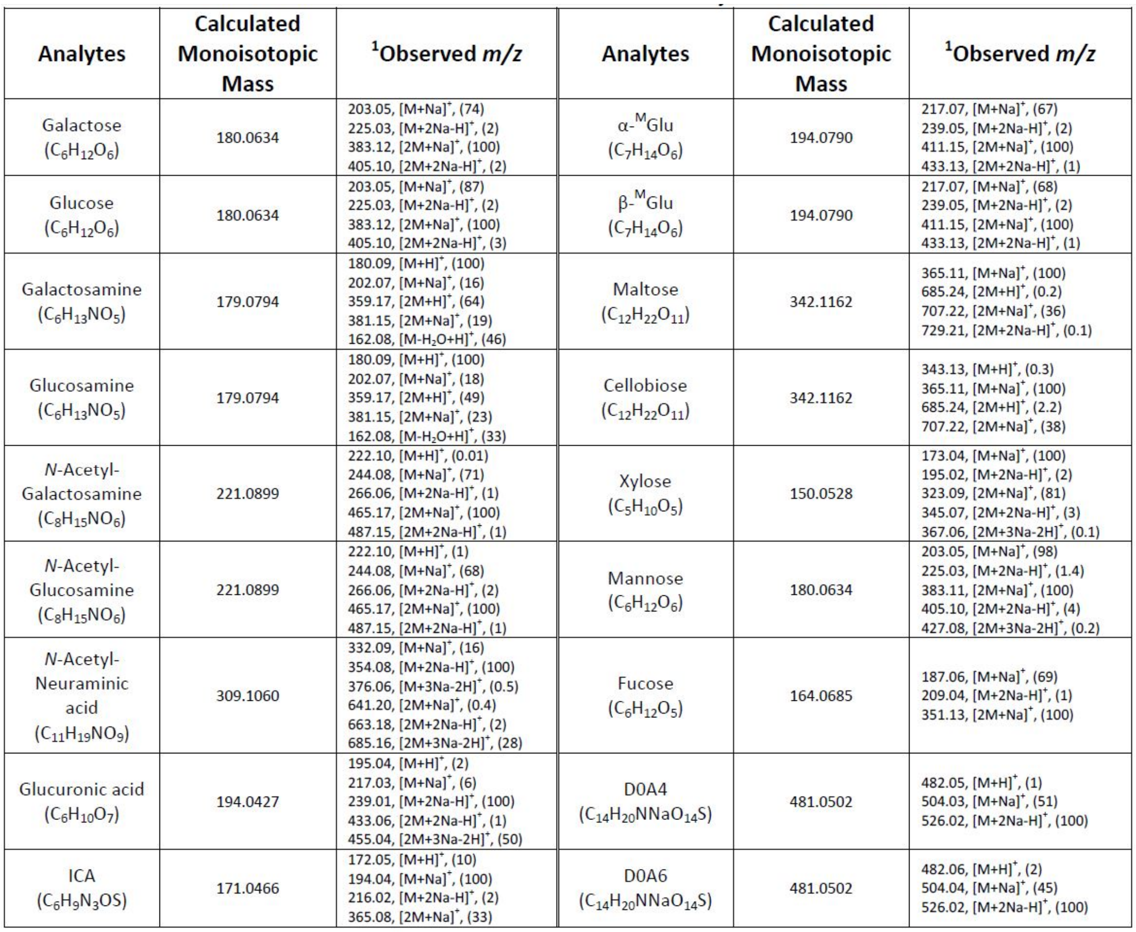


Figure 53. Structures of Molecules (a, b, c). Schematics of the Capturing Configuration (d, e). Example Current Signal Trace (f, g). Distribution of Features (h-j).

(a) The recognition molecule ICA contains a thiol linkage to bond to metal electrodes, as well as a number of hydrogen bonding donors and acceptors through which a large range of analytes can be can captured by a diversity of spatial arrangements. (b) Structure of α-^{M}Glu and (c) Structure of β-^{M}Glu, both of which can form hydrogen-bonded triplets with ICA molecules spanning a tunnel gap of 2.2 nm, as shown by simulations in (d) and (e). Evidence of these complexes is provided by the current-spikes that appear only after an analyte solution is added to pure buffer solution in a tunnel gap (f and g). Distributions of signal features are broad and overlapped (red = α-^{M}Glu, green = β-^{M}Glu) as shown here for one frequency band in the Fourier transform of signal clusters (cluster FFT whole 37 –h) and for a band in the Fourier transform of individual peaks (Peak FFT 9 –i). Data can only be assigned to one analyte or the other with a probability only marginally above random, P=0.5 (see Methods for details of the signal analysis). However, when the same two distributions are plotted together in a 2D histogram (j) where the brightness of each point represents the frequency with which a particular pair of values occur, the accuracy with which data can be assigned increases to 80%. This accuracy can be improved to ~ 99% using additional signal features. Colors in (j) are mixed so overlapped points are yellow.

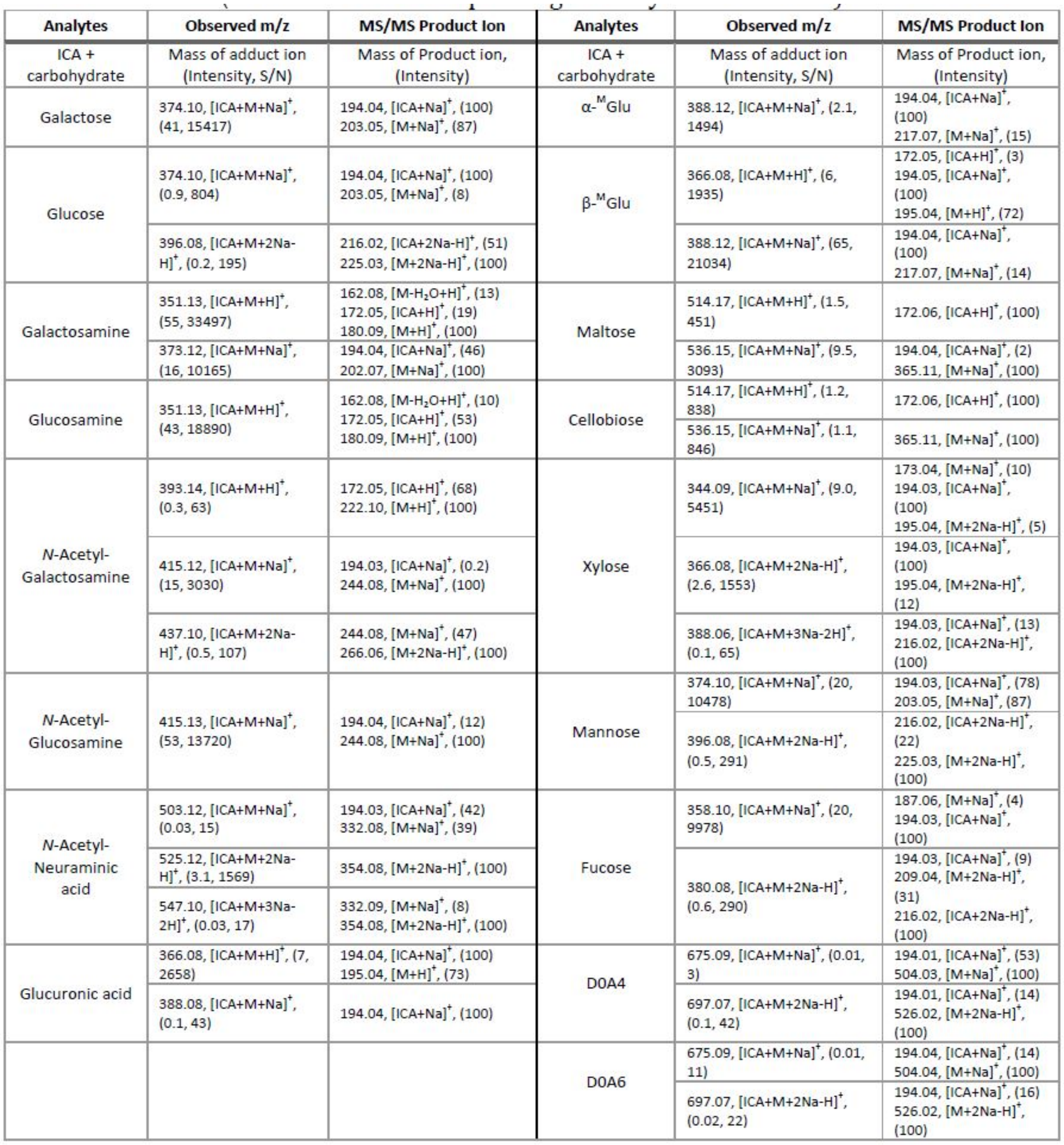


Analytes	Calculated Monoisotopic Mass	^{1}Observed *m/z*	Analytes	Calculated Monoisotopic Mass	^{1}Observed *m/z*
Galactose ($C_6H_{12}O_6$)	180.0634	203.05, [M+Na]$^+$, (74) 225.03, [M+2Na-H]$^+$, (2) 383.12, [2M+Na]$^+$, (100) 405.10, [2M+2Na-H]$^+$, (2)	α-^{M}Glu ($C_7H_{14}O_6$)	194.0790	217.07, [M+Na]$^+$, (67) 239.05, [M+2Na-H]$^+$, (2) 411.15, [2M+Na]$^+$, (100) 433.13, [2M+2Na-H]$^+$, (1)
Glucose ($C_6H_{12}O_6$)	180.0634	203.05, [M+Na]$^+$, (87) 225.03, [M+2Na-H]$^+$, (2) 383.12, [2M+Na]$^+$, (100) 405.10, [2M+2Na-H]$^+$, (3)	β-^{M}Glu ($C_7H_{14}O_6$)	194.0790	217.07, [M+Na]$^+$, (68) 239.05, [M+2Na-H]$^+$, (2) 411.15, [2M+Na]$^+$, (100) 433.13, [2M+2Na-H]$^+$, (1)
Galactosamine ($C_6H_{13}NO_5$)	179.0794	180.09, [M+H]$^+$, (100) 202.07, [M+Na]$^+$, (16) 359.17, [2M+H]$^+$, (64) 381.15, [2M+Na]$^+$, (19) 162.08, [M-H_2O+H]$^+$, (46)	Maltose ($C_{12}H_{22}O_{11}$)	342.1162	365.11, [M+Na]$^+$, (100) 685.24, [2M+H]$^+$, (0.2) 707.22, [2M+Na]$^+$, (36) 729.21, [2M+2Na-H]$^+$, (0.1)
Glucosamine ($C_6H_{13}NO_5$)	179.0794	180.09, [M+H]$^+$, (100) 202.07, [M+Na]$^+$, (18) 359.17, [2M+H]$^+$, (49) 381.15, [2M+Na]$^+$, (23) 162.08, [M-H_2O+H]$^+$, (33)	Cellobiose ($C_{12}H_{22}O_{11}$)	342.1162	343.13, [M+H]$^+$, (0.3) 365.11, [M+Na]$^+$, (100) 685.24, [2M+H]$^+$, (2.2) 707.22, [2M+Na]$^+$, (38)
N-Acetyl-Galactosamine ($C_8H_{15}NO_6$)	221.0899	222.10, [M+H]$^+$, (0.01) 244.08, [M+Na]$^+$, (71) 266.06, [M+2Na-H]$^+$, (1) 465.17, [2M+Na]$^+$, (100) 487.15, [2M+2Na-H]$^+$, (1)	Xylose ($C_5H_{10}O_5$)	150.0528	173.04, [M+Na]$^+$, (100) 195.02, [M+2Na-H]$^+$, (2) 323.09, [2M+Na]$^+$, (81) 345.07, [2M+2Na-H]$^+$, (3) 367.06, [2M+3Na-2H]$^+$, (0.1)
N-Acetyl-Glucosamine ($C_8H_{15}NO_6$)	221.0899	222.10, [M+H]$^+$, (1) 244.08, [M+Na]$^+$, (68) 266.06, [M+2Na-H]$^+$, (2) 465.17, [2M+Na]$^+$, (100) 487.15, [2M+2Na-H]$^+$, (1)	Mannose ($C_6H_{12}O_6$)	180.0634	203.05, [M+Na]$^+$, (98) 225.03, [M+2Na-H]$^+$, (1.4) 383.11, [2M+Na]$^+$, (100) 405.10, [2M+2Na-H]$^+$, (4) 427.08, [2M+3Na-2H]$^+$, (0.2)
N-Acetyl-Neuraminic acid ($C_{11}H_{19}NO_9$)	309.1060	332.09, [M+Na]$^+$, (16) 354.08, [M+2Na-H]$^+$, (100) 376.06, [M+3Na-2H]$^+$, (0.5) 641.20, [2M+Na]$^+$, (0.4) 663.18, [2M+2Na-H]$^+$, (2) 685.16, [2M+3Na-2H]$^+$, (28)	Fucose ($C_6H_{12}O_5$)	164.0685	187.06, [M+Na]$^+$, (69) 209.04, [M+2Na-H]$^+$, (1) 351.13, [2M+Na]$^+$, (100)
Glucuronic acid ($C_6H_{10}O_7$)	194.0427	195.04, [M+H]$^+$, (2) 217.03, [M+Na]$^+$, (6) 239.01, [M+2Na-H]$^+$, (100) 433.06, [2M+2Na-H]$^+$, (1) 455.04, [2M+3Na-2H]$^+$, (50)	D0A4 ($C_{14}H_{20}NNaO_{14}S$)	481.0502	482.05, [M+H]$^+$, (1) 504.03, [M+Na]$^+$, (51) 526.02, [M+2Na-H]$^+$, (100)
ICA ($C_6H_9N_3OS$)	171.0466	172.05, [M+H]$^+$, (10) 194.04, [M+Na]$^+$, (100) 216.02, [M+2Na-H]$^+$, (2) 365.08, [2M+Na]$^+$, (33)	D0A6 ($C_{14}H_{20}NNaO_{14}S$)	481.0502	482.06, [M+H]$^+$, (2) 504.04, [M+Na]$^+$, (45) 526.02, [M+2Na-H]$^+$, (100)

Table 14. Characteristic MS Peaks of 1:1 ICA-Carbohydrate Complexes and Their MS/MS Products.

M denotes the corresponding carbohydrate molecule.

The Relative Intensity (I%) and Signal to Noise Ratio (S/N) values are given in parentheses respectively next to each complex ion in observed m/z column. I% values are reported in parentheses next to each complex ion in MS/MS product ion column. The most intense peak is considered as 100.



Analytes	Observed m/z	MS/MS Product Ion
ICA + carbohydrate	Mass of adduct ion (Intensity, S/N)	Mass of Product ion, (Intensity)
Galactose	374.10, [ICA+M+Na]$^+$, (41, 15417)	194.04, [ICA+Na]$^+$, (100) 203.05, [M+Na]$^+$, (87)
Glucose	374.10, [ICA+M+Na]$^+$, (0.9, 804)	194.04, [ICA+Na]$^+$, (100) 203.05, [M+Na]$^+$, (8)
	396.08, [ICA+M+2Na-H]$^+$, (0.2, 195)	216.02, [ICA+2Na-H]$^+$, (51) 225.03, [M+2Na-H]$^+$, (100)
Galactosamine	351.13, [ICA+M+H]$^+$, (55, 33497)	162.08, [M-H$_2$O+H]$^+$, (13) 172.05, [ICA+H]$^+$, (19) 180.09, [M+H]$^+$, (100)
	373.12, [ICA+M+Na]$^+$, (16, 10165)	194.04, [ICA+Na]$^+$, (46) 202.07, [M+Na]$^+$, (100)
Glucosamine	351.13, [ICA+M+H]$^+$, (43, 18890)	162.08, [M-H$_2$O+H]$^+$, (10) 172.05, [ICA+H]$^+$, (53) 180.09, [M+H]$^+$, (100)
N-Acetyl-Galactosamine	393.14, [ICA+M+H]$^+$, (0.3, 63)	172.05, [ICA+H]$^+$, (68) 222.10, [M+H]$^+$, (100)
	415.12, [ICA+M+Na]$^+$, (15, 3030)	194.03, [ICA+Na]$^+$, (0.2) 244.08, [M+Na]$^+$, (100)
	437.10, [ICA+M+2Na-H]$^+$, (0.5, 107)	244.08, [M+Na]$^+$, (47) 266.06, [M+2Na-H]$^+$, (100)
N-Acetyl-Glucosamine	415.13, [ICA+M+Na]$^+$, (53, 13720)	194.04, [ICA+Na]$^+$, (12) 244.08, [M+Na]$^+$, (100)
N-Acetyl-Neuraminic acid	503.12, [ICA+M+Na]$^+$, (0.03, 15)	194.03, [ICA+Na]$^+$, (42) 332.08, [M+Na]$^+$, (39)
	525.12, [ICA+M+2Na-H]$^+$, (3.1, 1569)	354.08, [M+2Na-H]$^+$, (100)
	547.10, [ICA+M+3Na-2H]$^+$, (0.03, 17)	332.09, [M+Na]$^+$, (8) 354.08, [M+2Na-H]$^+$, (100)
Glucuronic acid	366.08, [ICA+M+H]$^+$, (7, 2658)	194.04, [ICA+Na]$^+$, (100) 195.04, [M+H]$^+$, (73)
	388.08, [ICA+M+Na]$^+$, (0.1, 43)	194.04, [ICA+Na]$^+$, (100)

Analytes	Observed m/z	MS/MS Product Ion
ICA + carbohydrate	Mass of adduct ion (Intensity, S/N)	Mass of Product ion, (Intensity)
α-^{M}Glu	388.12, [ICA+M+Na]$^+$, (2.1, 1494)	194.04, [ICA+Na]$^+$, (100) 217.07, [M+Na]$^+$, (15)
β-^{M}Glu	366.08, [ICA+M+H]$^+$, (6, 1935)	172.05, [ICA+H]$^+$, (3) 194.05, [ICA+Na]$^+$, (100) 195.04, [M+H]$^+$, (72)
	388.12, [ICA+M+Na]$^+$, (65, 21034)	194.04, [ICA+Na]$^+$, (100) 217.07, [M+Na]$^+$, (14)
Maltose	514.17, [ICA+M+H]$^+$, (1.5, 451)	172.06, [ICA+H]$^+$, (100)
	536.15, [ICA+M+Na]$^+$, (9.5, 3093)	194.04, [ICA+Na]$^+$, (2) 365.11, [M+Na]$^+$, (100)
Cellobiose	514.17, [ICA+M+H]$^+$, (1.2, 838)	172.06, [ICA+H]$^+$, (100)
	536.15, [ICA+M+Na]$^+$, (1.1, 846)	365.11, [M+Na]$^+$, (100)
Xylose	344.09, [ICA+M+Na]$^+$, (9.0, 5451)	173.04, [M+Na]$^+$, (10) 194.03, [ICA+Na]$^+$, (100) 195.04, [M+2Na-H]$^+$, (5)
	366.08, [ICA+M+2Na-H]$^+$, (2.6, 1553)	194.03, [ICA+Na]$^+$, (100) 195.04, [M+2Na-H]$^+$, (12)
	388.06, [ICA+M+3Na-2H]$^+$, (0.1, 65)	194.03, [ICA+Na]$^+$, (13) 216.02, [ICA+2Na-H]$^+$, (100)
Mannose	374.10, [ICA+M+Na]$^+$, (20, 10478)	194.03, [ICA+Na]$^+$, (78) 203.05, [M+Na]$^+$, (87)
	396.08, [ICA+M+2Na-H]$^+$, (0.5, 291)	216.02, [ICA+2Na-H]$^+$, (22) 225.03, [M+2Na-H]$^+$, (100)
Fucose	358.10, [ICA+M+Na]$^+$, (20, 9978)	187.06, [M+Na]$^+$, (4) 194.03, [ICA+Na]$^+$, (100)
	380.08, [ICA+M+2Na-H]$^+$, (0.6, 290)	194.03, [ICA+Na]$^+$, (9) 209.04, [M+2Na-H]$^+$, (31) 216.02, [ICA+2Na-H]$^+$, (100)
D0A4	675.09, [ICA+M+Na]$^+$, (0.01, 3)	194.01, [ICA+Na]$^+$, (53) 504.03, [M+Na]$^+$, (100)
	697.07, [ICA+M+2Na-H]$^+$, (0.1, 42)	194.01, [ICA+Na]$^+$, (14) 526.02, [M+2Na-H]$^+$, (100)
D0A6	675.09, [ICA+M+Na]$^+$, (0.01, 11)	194.04, [ICA+Na]$^+$, (14) 504.04, [M+Na]$^+$, (100)
	697.07, [ICA+M+2Na-H]$^+$, (0.02, 22)	194.04, [ICA+Na]$^+$, (16) 526.02, [M+2Na-H]$^+$, (100)

Table 15. Characteristic MS Peaks of 2:1 ICA-Carbohydrate Complexes and Their MS/MS Products.

M denotes the corresponding carbohydrate molecule.

The Relative Intensity (I%) and Signal to Noise Ratio (S/N) values are given in parentheses respectively next to each complex ion in observed m/z column. I% values are reported in parentheses next to each complex ion in MS/MS product ion column. The most intense peak is considered as 100.

The tunneling current signals are stochastic driven by thermal fluctuations and includes a lot of information about the capturing configuration and captured analytes.

Individual signal features are broadly distributed and give mostly overlapped distributions. Figure 53(h) and (i) show two of individual feature distribution and they give only a little bit higher separation probability, 0.58 with Cluster Whole FFT 37 and 0.57 with Peak FFT 9, when random calling probability is 0.5. However, scatter plot of those two features together provides much higher separation 0.80 (Figure 53(j)). It can be explained by Cover's theorem as discussed in chapter 3. On the plot, the red points represent α-^{M}Glu and the green points are β-^{M}Glu. The overlapped area is appeared as yellow. The plot are complicate with many islands of data points, but the two analytes are well separated with ~80%. Like Cover's theorem, by considering more features, the separation accuracy can be getting higher. Here, Supporting Vector Machine (SVM) was used for data analysis to identify each analytes. The α-^{M}Glu and β-^{M}Glu identification accuracy is shown in Table 16.

In addition to the glucose isomers, five more anomeric molecule pairs were measured as shown in Table 16. First, three of monosaccharide anomeric pairs were used. STM-RT is able to distinguish glucose, glucosamine are N-acetylglucosamine from their C-4 epimers. Though there can be an equilibrium ratio of anomeric isomers in solution, STM-RT identifies the epimers with high accuracy. This is because the training data set also contains the same equilibrium anomer mixture. STM-RT also was tested with disaccharide anomeric pairs that are the pair of maltose (α-D-glucopyranosyl-(1→4)-D-glucopyranose) and cellobiose (β-D-glucopyranosyl-(1→4)-D-glucopyranose), and the pair of 4-O-sulfated-chondroitin sulfate disaccharides (D0A4) and 6-O-sulfated chondroitin sulfate disaccharides (D0A6) which are repeating disaccharide unit of glycosaminoglycans (GAGs). The feature sets that provides the accuracies on Table 16 are shown in Table 17.

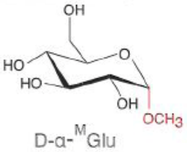
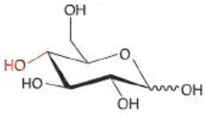
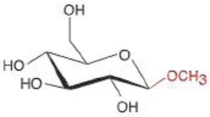
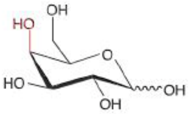
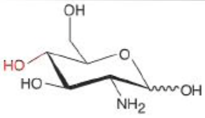
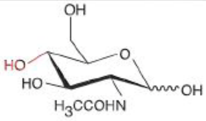
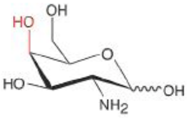
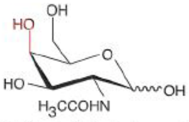
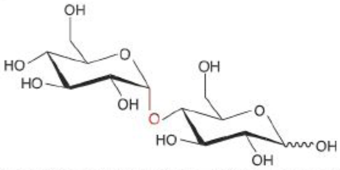
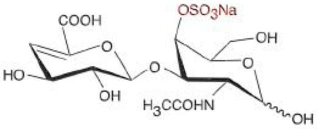
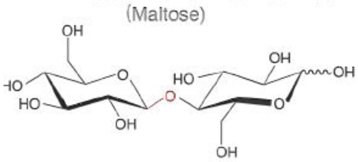
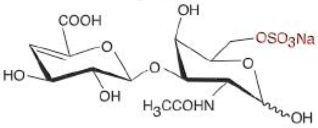
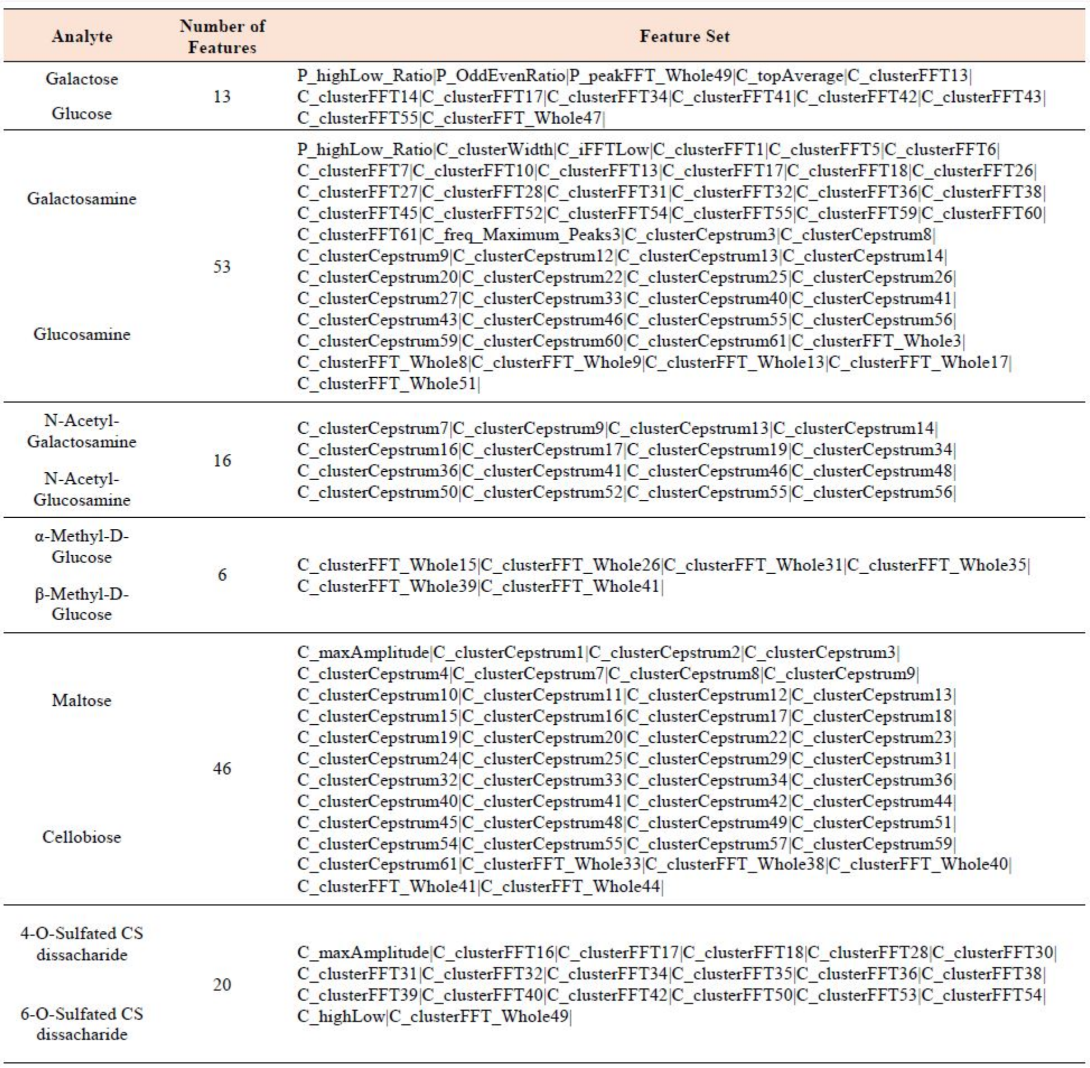
Analyte	Accuracy (%)	Analyte	Accuracy (%)
 D-α-^{M}Glu	99.1	 D-Glucose	99.7
 D-β-^{M}Glu	99.4	 D-Galactose	99.5
 D-Glucosamine	98.9	 D-N-Acetyl-Glucosamine	99.3
 D-Galactosamine	98.7	 D-N-Acetyl-Galactosamine	99.5
 α-D-Glucopyranosyl-(1→4)-D-glucopyranose (Maltose)	99.5	 β-Δ^{4,5}-D-UA(1→3)-D-GalNAc-4-O-Sulfate (D0A4)*	98.8
 β-D-glucopyranosyl-(1→4)-D-glucopyranose (Cellobiose)	99.8	 β-Δ^{4,5}-D-UA(1→3)-D-GalNAc-6-O-Sulfate (D0A6)*	99.6

Table 16. Accuracy of Isomeric Molecular Pairs by SVM Analysis.



<table>
<thead>
<tr><th>Analyte</th><th>Number of Features</th><th>Feature Set</th></tr>
</thead>
<tbody>
<tr><td>Galactose
Glucose</td><td>13</td><td>P_highLow_Ratio|P_OddEvenRatio|P_peakFFT_Whole49|C_topAverage|C_clusterFFT13|C_clusterFFT14|C_clusterFFT17|C_clusterFFT34|C_clusterFFT41|C_clusterFFT42|C_clusterFFT43|C_clusterFFT55|C_clusterFFT_Whole47|</td></tr>
<tr><td>Galactosamine
Glucosamine</td><td>53</td><td>P_highLow_Ratio|C_clusterWidth|C_iFFTLow|C_clusterFFT1|C_clusterFFT5|C_clusterFFT6|C_clusterFFT7|C_clusterFFT10|C_clusterFFT13|C_clusterFFT17|C_clusterFFT18|C_clusterFFT26|C_clusterFFT27|C_clusterFFT28|C_clusterFFT31|C_clusterFFT32|C_clusterFFT36|C_clusterFFT38|C_clusterFFT45|C_clusterFFT52|C_clusterFFT54|C_clusterFFT55|C_clusterFFT59|C_clusterFFT60|C_clusterFFT61|C_freq_Maximum_Peaks3|C_clusterCepstrum3|C_clusterCepstrum8|C_clusterCepstrum9|C_clusterCepstrum12|C_clusterCepstrum13|C_clusterCepstrum14|C_clusterCepstrum20|C_clusterCepstrum22|C_clusterCepstrum25|C_clusterCepstrum26|C_clusterCepstrum27|C_clusterCepstrum33|C_clusterCepstrum40|C_clusterCepstrum41|C_clusterCepstrum43|C_clusterCepstrum46|C_clusterCepstrum55|C_clusterCepstrum56|C_clusterCepstrum59|C_clusterCepstrum60|C_clusterCepstrum61|C_clusterFFT_Whole3|C_clusterFFT_Whole8|C_clusterFFT_Whole9|C_clusterFFT_Whole13|C_clusterFFT_Whole17|C_clusterFFT_Whole51|</td></tr>
<tr><td>N-Acetyl-Galactosamine
N-Acetyl-Glucosamine</td><td>16</td><td>C_clusterCepstrum7|C_clusterCepstrum9|C_clusterCepstrum13|C_clusterCepstrum14|C_clusterCepstrum16|C_clusterCepstrum17|C_clusterCepstrum19|C_clusterCepstrum34|C_clusterCepstrum36|C_clusterCepstrum41|C_clusterCepstrum46|C_clusterCepstrum48|C_clusterCepstrum50|C_clusterCepstrum52|C_clusterCepstrum55|C_clusterCepstrum56|</td></tr>
<tr><td>α-Methyl-D-Glucose
β-Methyl-D-Glucose</td><td>6</td><td>C_clusterFFT_Whole15|C_clusterFFT_Whole26|C_clusterFFT_Whole31|C_clusterFFT_Whole35|C_clusterFFT_Whole39|C_clusterFFT_Whole41|</td></tr>
<tr><td>Maltose
Cellobiose</td><td>46</td><td>C_maxAmplitude|C_clusterCepstrum1|C_clusterCepstrum2|C_clusterCepstrum3|C_clusterCepstrum4|C_clusterCepstrum7|C_clusterCepstrum8|C_clusterCepstrum9|C_clusterCepstrum10|C_clusterCepstrum11|C_clusterCepstrum12|C_clusterCepstrum13|C_clusterCepstrum15|C_clusterCepstrum16|C_clusterCepstrum17|C_clusterCepstrum18|C_clusterCepstrum19|C_clusterCepstrum20|C_clusterCepstrum22|C_clusterCepstrum23|C_clusterCepstrum24|C_clusterCepstrum25|C_clusterCepstrum29|C_clusterCepstrum31|C_clusterCepstrum32|C_clusterCepstrum33|C_clusterCepstrum34|C_clusterCepstrum36|C_clusterCepstrum40|C_clusterCepstrum41|C_clusterCepstrum42|C_clusterCepstrum44|C_clusterCepstrum45|C_clusterCepstrum48|C_clusterCepstrum49|C_clusterCepstrum51|C_clusterCepstrum54|C_clusterCepstrum55|C_clusterCepstrum57|C_clusterCepstrum59|C_clusterCepstrum61|C_clusterFFT_Whole33|C_clusterFFT_Whole38|C_clusterFFT_Whole40|C_clusterFFT_Whole41|C_clusterFFT_Whole44|</td></tr>
<tr><td>4-O-Sulfated CS dissacharide
6-O-Sulfated CS dissacharide</td><td>20</td><td>C_maxAmplitude|C_clusterFFT16|C_clusterFFT17|C_clusterFFT18|C_clusterFFT28|C_clusterFFT30|C_clusterFFT31|C_clusterFFT32|C_clusterFFT34|C_clusterFFT35|C_clusterFFT36|C_clusterFFT38|C_clusterFFT39|C_clusterFFT40|C_clusterFFT42|C_clusterFFT50|C_clusterFFT53|C_clusterFFT54|C_highLow|C_clusterFFT_Whole49|</td></tr>
</tbody>
</table>

Table 17. List of Feature Sets which Generates the Accuracy on Table 16.

4.3.2. Pure Carbohydrates Identification

STM-RT was also able to identify many different monosaccharides from their pooled data with high identification accuracy as shown in Table 18. Among the samples, D-Glucosamine, D-Galactosamine, and D-Mannose are abundant in the mammalian glycome. [103][104] N-acetyl-neuraminic acid is the predominant sialic acid in mammalian cells. The eleven monosaccharides were well identified in overall 94%

accuracy (random would be 9%). In contrast, IM-MS does not effectively discriminate the molecules between galactose and mannose [105] or glucose and galactose [67].

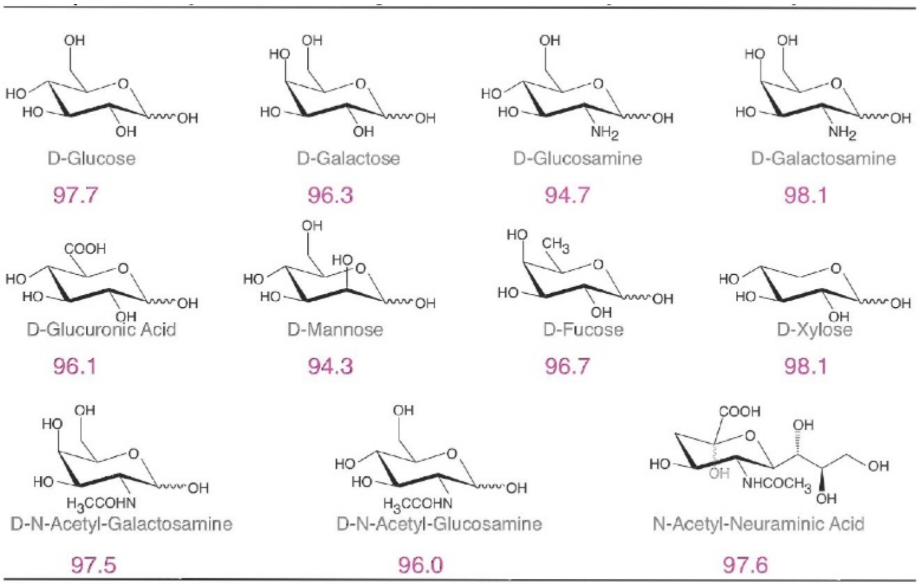


Table 18. Accuracy of Individual Mono-saccharides from Their Pooled Data.

Analyte	Number of Features	Feature Set
11 Sugars	13	C_topAverage|C_totalPower|C_clusterFFT16|C_clusterFFT18|C_clusterCepstrum9| C_clusterCepstrum10|C_clusterCepstrum26|C_clusterCepstrum39| C_clusterCepstrum42|C_clusterCepstrum46|C_clusterCepstrum52| C_clusterCepstrum61|C_clusterFFT_Whole51|

Table 19. List of Feature Set which Generates the Accuracy on Table 18.

4.3.3. Binding Affinity of ICA to α-^{M}Glu

The binding affinity of ICA to α-^{M}Glu in the gap was determined by STM-RT and Surface Plasmon Resonance (SPR). In Figure 54, it shows a trend of the normalized single peak counts at various concentrations under STM-RT. The data was well fitted to a Langmuir isotherm equation with $R^2 = 0.976$. The dissociation constant was obtained from the fitting parameter, $K_d = 0.74 \pm 0.25\ \mu M$. However, the one measured by SPR is

$4\ mM$ as the adsorption of α-^{M}Glu on an ICA monolayer (Figure 55). Compared the two values, STM-RT gives less dissociation constant which means better affinity between ICA and analyte, α-^{M}Glu in here. The enhanced affinity comes from the simultaneous interaction with a pair of ICA at two electrodes. Assuming that entropy changes for ICA at two electrodes are the same and single binding has a constant entropy change, the adsorption free energy would be doubled for binding at two electrodes. Thus K_d of two-site binding can be square of the single site binding, $(4\ mM)^{\ 2}\ or\ 16\ \mu M$, but it is still much larger than the observed $0.74\ \mu M$. There are some more possible enhancement factors. The electric field in the gap enhance the capturing of bound molecules. Also the molecular dipole can increase the free energy upto 20% of thermal energy.

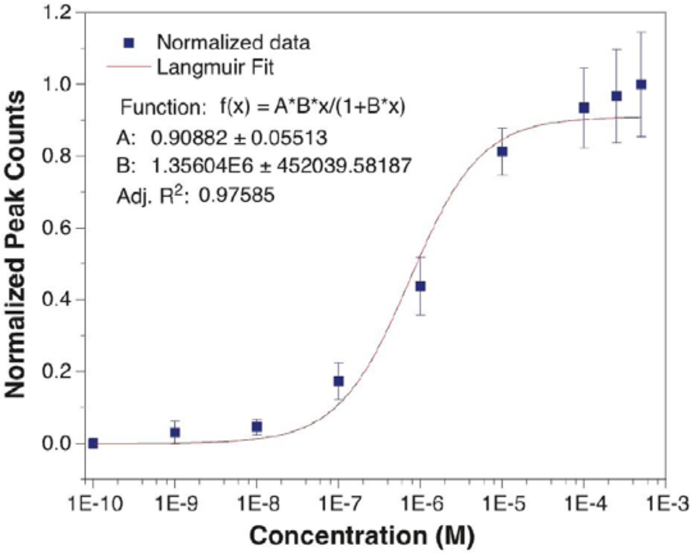
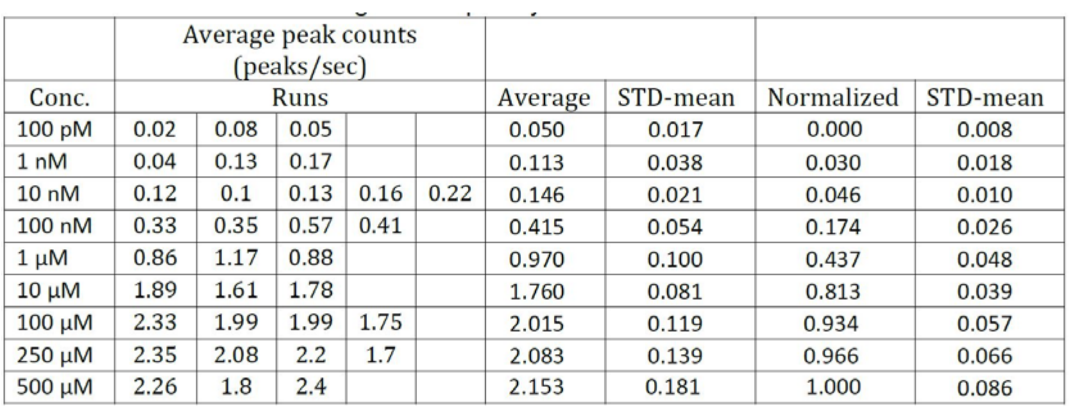


Figure 54. Plot of Normalized Peak Counts and Concentration of α-^{M}Glu.

Plot of normalized RT counting rates vs concentration of α-MGlu for trapping the analyte in an RT gap functionalized with ICA molecules and a fit to a Langmuir isotherm.



	Average peak counts (peaks/sec)								
Conc.	Runs					Average	STD-mean	Normalized	STD-mean
100 pM	0.02	0.08	0.05			0.050	0.017	0.000	0.008
1 nM	0.04	0.13	0.17			0.113	0.038	0.030	0.018
10 nM	0.12	0.1	0.13	0.16	0.22	0.146	0.021	0.046	0.010
100 nM	0.33	0.35	0.57	0.41		0.415	0.054	0.174	0.026
1 μM	0.86	1.17	0.88			0.970	0.100	0.437	0.048
10 μM	1.89	1.61	1.78			1.760	0.081	0.813	0.039
100 μM	2.33	1.99	1.99	1.75		2.015	0.119	0.934	0.057
250 μM	2.35	2.08	2.2	1.7		2.083	0.139	0.966	0.066
500 μM	2.26	1.8	2.4			2.153	0.181	1.000	0.086

Table 20. Peak Counts of α-^{M}Glu at Various Concentrations and Adjusted Counts.

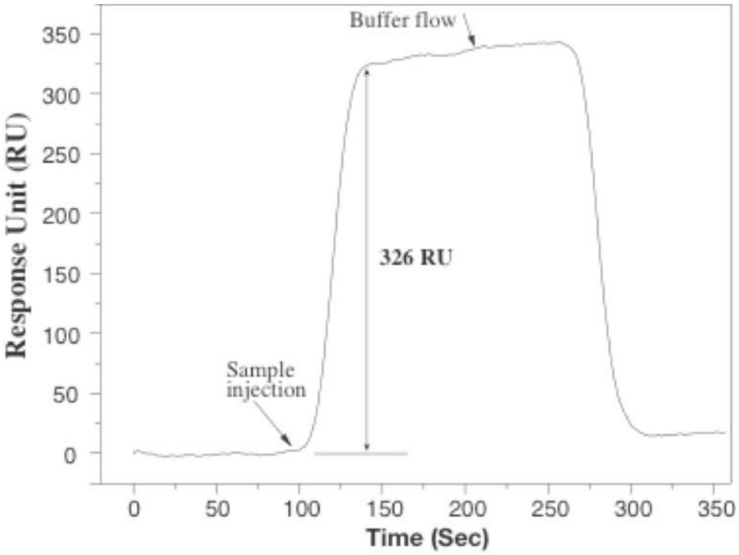


Figure 55. Example Signal of the SPR Sensogram for ICA and α-^{M}Glu.

α-^{M}Glu	k_{on} $[M^{-1}s^{-1}]$	$k_{off}[s^{-1}]$	$k_d{}^2$ $[mM]$
500 μM	23.05 ± 1.34	0.09 ± 0.01	4.00 ± 0.14

Table 21. Kinetic Parameters of α-^{M}Glu on the ICA Monolayer.

1. Each datum listed is an average of two measurements.

2. $k_d = k_{off}/k_{on}$

The number of peak counting at a given concentration is reproducible shown in Table 21. In the range of 10 ~ 100 nM, there is a significant peak count rates increasing. It implies a quantitative ability to quantify the relative amount of a given analyte. This provides an overcome limits of current techniques. For instance, mass spectroscopy is not quantitative and requires additional techniques labelling isotope for quantification.

STM-RT needs 200 μL of sample solutions to fill the liquid cell. However, the recent research in this research group has showed that it can be reduced upto microliter volumes of sample with micron-scale solid-state tunnel junctions. [106]

4.4. Conclusion

In spite of the enormous biological importance of glycans, their sequence analysis remains one of the most challenging areas of chemistry, lagging behind genomic sequencing. This is because of their extensive isomerism, which leads to astronomical numbers of possible configurations for even a short oligosaccharide. This study demonstrates that STM-RT is able to discriminate individual saccharide from a pool of data, even they are isomeric molecules. No current technique could analyze large number of isomers in a long oligosaccharide. [80] However, RT can provide a single molecule recognition in linear oligosaccharides. In the present work, we illustrate the technique with two glucopyranoside anomers and go on to show how six other pairs of “difficult” isomers are readily separated, show how an individual glycan can be identified in data pooled from 11 different molecules, and demonstrate quantitative measurement over a dynamic range of concentrations. It has been proposed that a nanopore with RT electrodes can be used for

DNA sequencing. [107] The present work shows that RT is able to identify many different individual glycans. Thus, if it is combined with a nanopore, we can obtain the compositional sequence of linear oligosaccharides directly without any kinds of molecular labelling. It also opens a path to direct, single-molecule sequencing of linear oligosaccharides (an on-going project).

CHAPTER 5

SINGLE MOLECULE SEQUENCING OF GLYCOSAMINOGLYCANS USING RECOGNITION TUNNELING NANOPORES

5.1. Introduction

Carbohydrates play significant roles in mediating extracellular interactions and their functions are determined by the sequence of the carbohydrates. However, sequencing polysaccharides is more difficult than the one of DNA and protein. This study uses RT junction embedded nanopore to read sequence of linear polysaccharides known as glycosaminoglycan (GAG). GAG's linearity is attractive for the nanopore sequencing. GAGs are linear and sulfated polysaccharides are common to all mammalian cells. They interact with enzymes, signaling proteins and pathogenic microbes which make GAGs important modulators of biological phenomena. However, it requires further improvement to understand the relation between their structure and function. GAGs' structure is difficult to analyze due to their large size and random nature of sulfation modifications. The current ensemble analytical techniques are not able to identify structures. A single molecule sequencing technique provides advantages in polysaccharides analysis.

The goal of this study is to develop a nanopore based recognition tunneling device for single molecule sequencer as shown in Figure 56. Analytes pass an electrodes pore modified with recognition molecules which make non-covalent capturing with the analytes. The RT chip fabrication was published by colleagues in this research group[106] and I am working to incorporate it into nanopore.

The step one of this work is to investigate translocations of GAG through solid state nanopores and to study how translocation time and blockade current are determined by length of GAG molecules and sulfation densities. The step two is to identify GAG fragment of charged disaccharides using the RT nanopore functionalized with ICA.

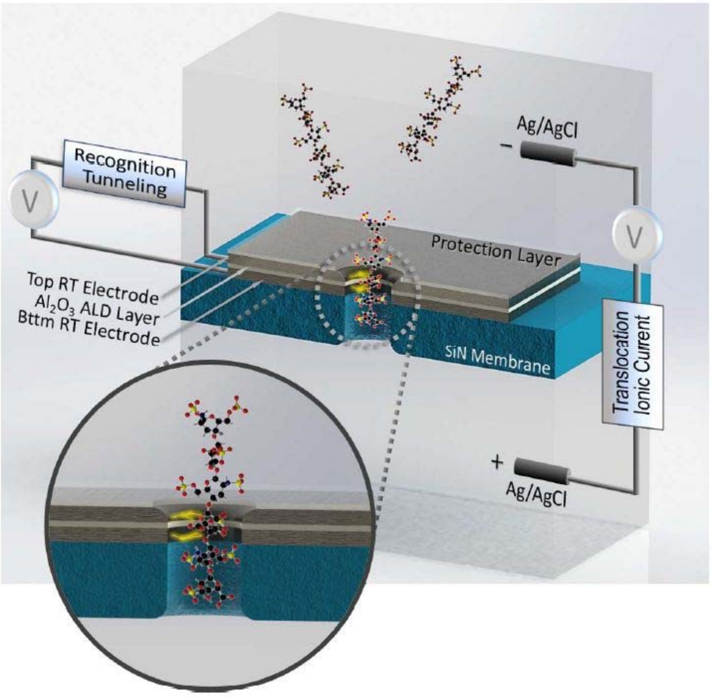


Figure 56. Schematic of Recognition Tunneling Nanopore Device.

GAGs are translocated from one side (Cis) of the chip to the other (Trans). Their recognition tunneling signals can be recorded for determination of their sequence structures.

5.1.1. Glycosaminoglycans

Polysaccharides play significant roles in mediating extracellular interactions in every organism. Glycosaminoglycans (GAGs) is a linear and sulfated polysaccharide ubiquitous to all mammalian cells. GAG molecules control many biological activities, resulting from their extracellular localization and acidic nature. These properties make them to attract signaling molecules through electrostatic forces, and modulate proteins' interactions with cell surface receptors. GAGs have been shown to be vital to tissue repair and regeneration[108]; cancer cells are also known to express over-sulfated GAGs to attract growth factors during metastasis[109][110][111][112]; GAGs have also been identified as receptors for a growing list of pathogenic bacteria, protozoa and viruses[113].

Most GAGs are long polysaccharides composed of repeating disaccharide units of an uronic acid and an amino sugar. Classification of GAGs is based on the type of uronic acid and amino sugar contained in its disaccharide unit. Heparan sulfate (HS) or heparin (HP) contains glucosamine (GlcN) as the amino sugar and either glucuronic acid (GlcA) or iduronic acid (IdoA) as the uronic acid (Figure 57). Chondroitin sulfate (CS), another major type of GAGs, contains N-acetylgalactosamine (GalNAc) as the amino sugar and mostly GlcA as the uronic acid (Figure 57). Despite their repetitive and linear structures, GAGs are among the most complex biomolecules in nature. Their structural complexity is the result of their large size (each chain can contain hundreds of saccharide residues) and semi-random sulfations of monosaccharides. Most GAGs are found anchored to the cell surface through covalent linkages to serine hydroxyl groups on proteins, forming entities known as proteoglycans.

Because of GAGs' multitude of biological activities, interest in utilizing GAG as diagnostic tools and therapeutics in disorders such as cancer, inflammation and even Alzheimer's disease have been high for some time.[114][115] But before these potential applications of GAGs can be realized, the structure-activity relationships between GAG sulfation patterns and its activity need to be fully understood. Indeed many basic questions regarding the relationship between GAG structure and its activity remain unresolved. For instance, N-sulfation of HS/HP is known to have important consequences on the biological activity of GAGs. Although enzymatic degradation analysis of HS/HP indicates N-acetylated and N-sulfated GlcNs often exist in clusters, the sizes and distributions of these clusters in intact GAGs have never been definitively measured because structural analysis of long GAG chains remains challenging.

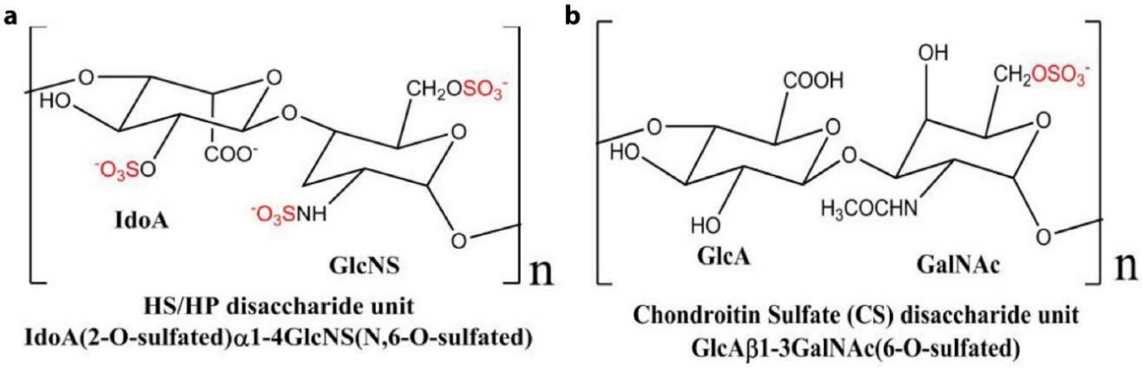


Figure 57. Chemical Structures of Representative HS/HP and CS Disaccharide Units.

5.1.2. Nanopore

A nanopore can be described as a short tube with a diameter of nanometers. It can be a nanofluidic channel for charged molecules transportation. Nanopores have become a single molecule tool for DNA sequencing [116], and sensing and unfolding proteins [66][117][118]. In the introduction, nanopores were categorized into three groups sorted by fabrication methods; biological, solid stat, and hybrid pores. Here, let's assign them into three types: (a) protein nanopore, (b) solid-state nanopores, and (c) nanogap nanopores, as shown in Figure 58.[119] When an analyte is driven through the pore, it will partially block the channel (nanopore) and ionic current will be changed. Then the ionic current can be characterized by blockade amplitude (I_p) and dwell time (t_d). A protein nanopores sequences DNA in single nucleotide resolution[120], but solid-state nanopore has not achieved it yet[121]. On the other hand, a nanogap nanopore can be used to identify single nucleotides by measuring RT of single analyte captured by two electrodes in the pore.

Compared to DNA, there are limited number of studies on polysaccharides through nanopores. Kullman *et al* studied interactions of the maltoporin protein pore with maltodextrins from triose to heptaose by translocation.[122] Bayley and his coworkers studied effects of electroomosis on cyclodextrin binding to α-hemolysin.[123] Teixeira *et al* showed that heparins and dextran sulfates blocked α-hemolysin protein pores in the presence of Ca^{2+} cations.[124] Bacri *et al* first demonstrated that neutral polysaccharids maltose and dextran with molecular weights ranging from 504 to 10,300 g mol^{-1} can be translocated through α-hemolysin protein pores.[125] They demonstrated that oligosaccharides differing by two disaccharide repeats could be distinguished by their dwell times. There is only one report of the translocation of charged polysaccharides through silicon nitride (SiN) nanopores.[126] Translocation of sulfated GAGs has not been reported yet.

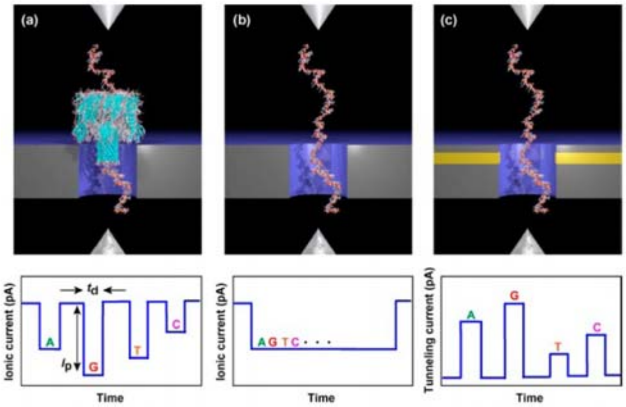


Figure 58. Types of Nanopores.

(a) protein nanopore (b) solid-state nanopore, and (c) nanogap nanopore. Figure taken from ref [119].

5.1.3. Recognition Tunneling (RT)

As discussed in the previous chapters, RT is powerful tool for single molecular sequencing. This can be integrated with nanopores to electronically read individual biomolecular components. Electrons can tunnel through the nanogap when a bias is applied between two electrodes. The tunneling can be enhanced and more sensitive by modifying the electrodes with recognition molecule. 4-(2-Mercaptoethyl)-1*H*-imidazole-2-carboxamide (ICA) was designed as a universal reader to read DNA bases via hydrogen bonds between ICA and analytes.[59] The RT is tunneling current signal generated by capturing analyte between functionalized electrodes. Unexpectedly, the universal reader is also able to recognize amino acids and short peptides (Chapter 3) and to identify carbohydrates (Chapter 4). The complex of ICA and analyte generates a time-dependent current due to thermal fluctuations of the bonds. As the current goes back to the baseline current when a captured molecule leaves the gap, there is no frequent and dense signals in control experiments (without analyte in solution). The signals reflect thermal noise, but they are not random. The signals depend on the structure of analytes and types of bonding in the junction. The collected RT data is analyzed by using Support Vector Machine (SVM), one of machine learning algorithms, to identify unique signatures of analytes. It has been shown that the RT can distinguish among monosachaarides with different charges and between two anomeric isomers of a disaccharide (Chapter 4). These results demonstrate that the RT has the power to identify subtle differences in structures of carbohydrates.

5.2. Experimental Methods

5.2.1. Preparation of GAG Solutions

All the GAG samples; chondroitin sulfate and heparin polysaccharides, are prepared by Dr. Xu Wang's research group. The GAGs is prepared in various length. In this discussion, two molecules (heparin dp10 and dp60) data will be discussed. The heparin dp10 (HEP10) means the five of HP disaccharide unit (corresponding to ten of monosaccharides) which is shown Figure 57(a). The heparin dp30 (HEP60) is a mixture of various length heparins, ranging from dp10 to dp60 (corresponding 20 to 120 monosaccharides). The GAG powder compounds are dissolved into a solution of 400 mM KCl buffered with 1 mM PB (pH 7.0) to be 1 μM. Water was purified by a Milli-Q system for ~18MΩ-cm and less than 5 ppb of total organic carbon contamination.

5.2.2. Fabrication of Layered Junction Device and Nanopores

The fabrication of layered junction device has been developed by colleagues in this research group.[106] In the previously published paper, the pore size is ~20 nm diameter resulting in a bunch of analytes pass through the pore. It is hard to capture single molecule RT signals with large nanopore. The colleagues is working to introduce nanopores to tunneling junction. The details of current layered junction fabrication is following.

Layered junction device is fabricated on <100> polished Si wafers or silicon nitride chip with 50 nm thick windows (purchased from Norcada). Au leads and pads which are used for connection with instrument were fabricated by photolithography or electron beam lithography (EBL) with a JEOL JBX 6000FS/E. The 6 μm width bottom electrodes (1 nm

Ti adhesion layer and 10 nm Pd) were deposited by electron beam evaporator (Lesker PVD75) (Figure 59(a-i)). After 2 nm thick Al_2O_3 layer was deposited by atomic layer deposition (ALD) (Figure 59(a-ii)), 50 ~ 80 nm width of Pd nanowire with 1 nm Ti adhesion layer was fabricated by EBL on 60 nm thick patterned PMMA. The nanowire patterns were exposed at $500\ \mu C/cm^2$. The metal on PMMA layer was removed by soaking dicholoromethane for 15 min, and rinsed with acetone, isopropyl alcohol (IPA), and DI water, and gently dried by nitrogen gas flow.

Making nanopores on the layered junction device is on process by using focused ion beam (FIB), reactive-ion etching (RIE), and TEM.

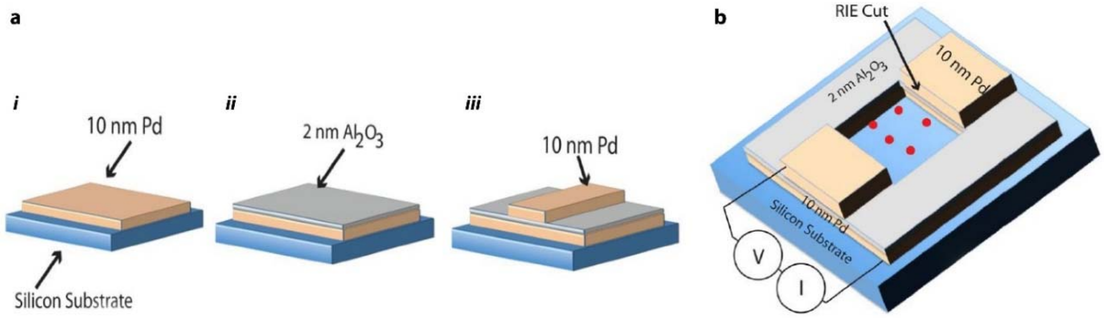


Figure 59. Process of the Fabrication of Layered Tunnel Junctions.

(a) Process for fabrication of layered tunnel junction: i. 10 nm thick Pd electrode is defined on a SiN support; ii. 2 nm thick Al_2O_3 layer is deposited by atomic layer deposition (ALD); iii. A second 10 nm thick Pd nanowire is deposited on top of the dielectric layer. (b) A nanopore cut into the sandwich ('RIE Cut') exposes the junction giving access to analyte molecules (red dots). Figure taken from ref [106].

5.2.3. Preparation of Nanopores on Silicon Membrane

Silicon nitrides chips were purchased from Norcada, Alberta. Silicon frame is $5 \times 5\ mm$ with $200\ \mu m$ thickness, and the silicon nitrides window is $200 \times 200\ \mu m$. Nanopores are drilled using the electron beam in a JEOL 2010FEG transmission electron microscope (TEM) at 200 kV. The size of the pores is controlled by the focused electron beam. The drilled nanopores are imaged right after drilling for the pore size estimation. Prior to every experiment, the chip with a nanopore is immersed in hot piranha (H_2O_2: $H_2SO_4 = 1:3$) for 10 minutes and rinsed with water. The chip is dried with N_2 gas flow, and assembled in a piranha cleaned home-made Polytetrafluoroethylene (PTFE) cell, and sealed with a quick-curing silicone elastomer gasket to reduce capacitance.

5.2.4. Translocation Measurement and Data Analysis

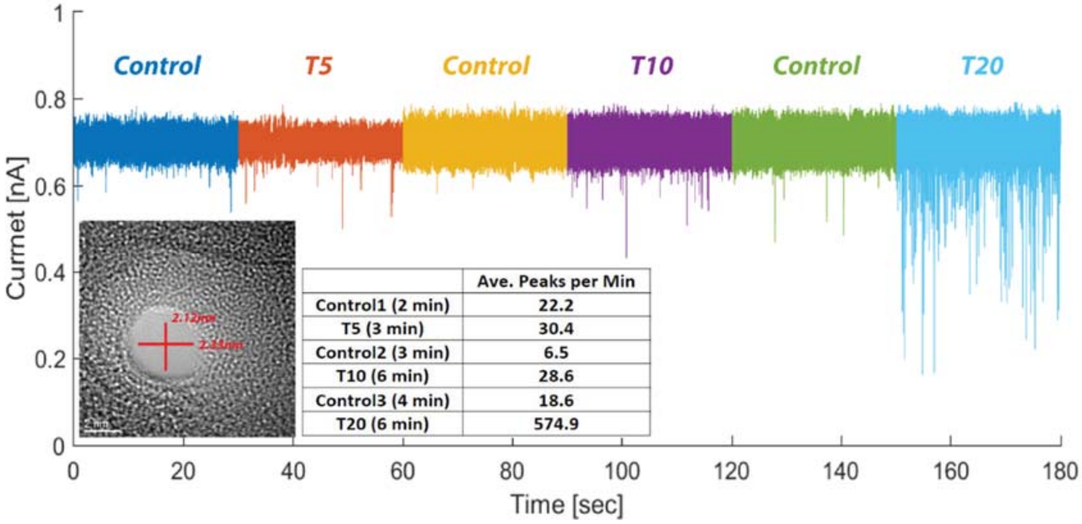
The sample solution is injected into the *cis* chamber and positive bias is applied into *trans* chamber through freshly made Ag/AgCl electrodes. Ionic currents are taken with 100 kHz sampling rate with 10 kHz low pass filter by using patch clamp amplifier Axon Axopatch 200B and digitizer DigiData 1550A (Axon Inc.). AxoScope 10.4 software is used to control those instruments and to record ionic current data. For data analysis, OpenNanopore is used, which is based on Matlab and developed by Laboratory of Nanoscale Biology (LBEN) of Ecole Polytechnique Fédérale de Lausanne (EPFL).

5.3. Results

As the first step of the project, it was studied to show possibility that solid-state nanopore enables to translocate GAG molecules. In the following chapters, it discusses the effect of molecular length and sampling rate to translocation. Furthermore, the nanopore size, voltage bias, and the charges on molecules will be investigated in future.

5.3.1. Length Dependence of Translocation

Colleagues in the research group have published nanopore study with peptide-poly T20 conjugate.[127] It shows that poly-T20 gave frequent translocation signals. It is necessary to determine appropriate length of GAG molecules for nanopore experiments. If the molecule is too short relative to the size and length of nanopore, it does not give obvious translocation signal. First, various length of poly-thymine was measured as control experiments, ranging from 5 to 20 (T5, T10, and T20). As expected from the previous research with peptide conjugate, T20 gives frequent translocation signals, but T5 and T10 are not (Figure 60(a)). Though there are some translocation events, the average number of peaks show that they would be random background noise. Figure 60(c) shows that the average number of peaks of T5 and T10 is close with the one in control measurements. Though the number of T5 and T10 is little bit higher than controls, it is hard to conclude that the events are from the GAG's translocation. Similar statistic results are shown in Figure 61. It also gives some events with short molecule, but it is not as frequent as longer molecules. It implies there would be a critical length which drives frequent translocation events. This limit can be overcame through the RT junction embedded nanopore. It is able to identify the events whether they are random noises or analyte unique RT signals.

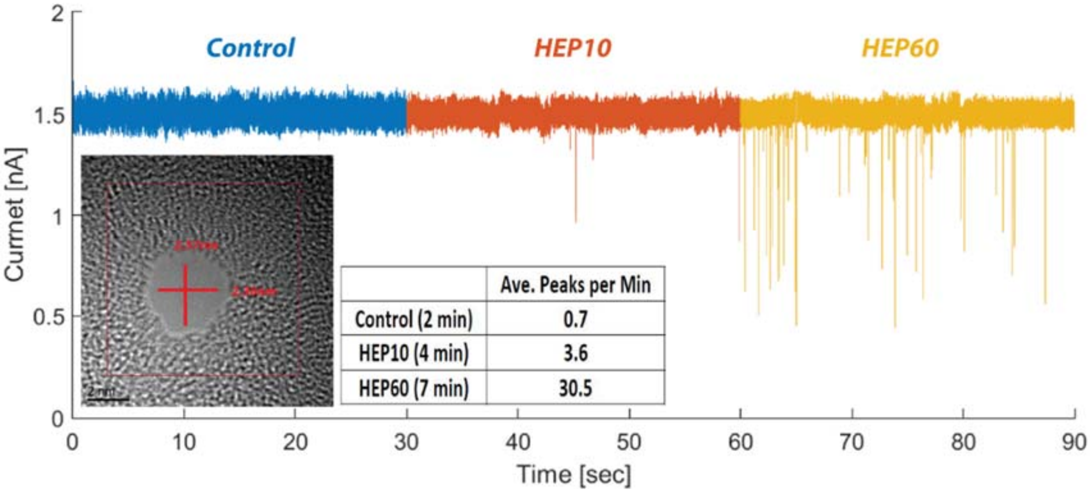


	Ave. Peaks per Min
Control1 (2 min)	22.2
T5 (3 min)	30.4
Control2 (3 min)	6.5
T10 (6 min)	28.6
Control3 (4 min)	18.6
T20 (6 min)	574.9

Figure 60. Analyte Length Effect with DNA Oligomer.

(a) Example of translocation signals of various length oligomers, (b) TEM image of the nanopore right after drilling, (c) Statistics of each measurement.

Example signal is a 30 sec trace from each analyte's entire data. Translocation signals were measured in 100kHz sampling rate with 10 kHz low pass filter. After every oligomer experiments, the cell (trans reservoir) was rinsed by 1 ml of buffer solution and control data was collect to confirm no analyte residues in the cell. The diameter of the nanopore was estimated by Gartan software.



	Ave. Peaks per Min
Control (2 min)	0.7
HEP10 (4 min)	3.6
HEP60 (7 min)	30.5

Figure 61. Analyte Length Effect with GAG.

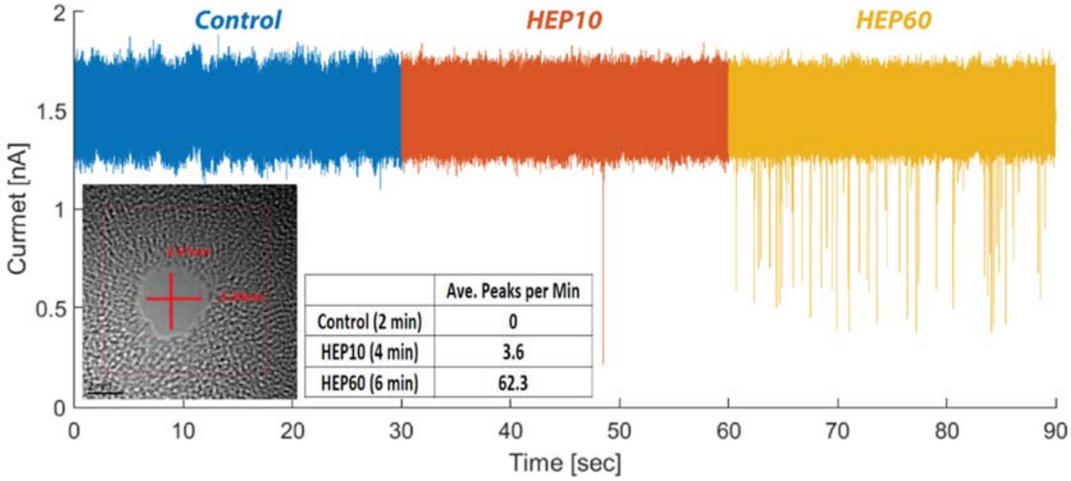
(a) Example of translocation signals of various length GAGs, (b) TEM image of the nanopore right after drilling, (c) Statistics of each measurement.

Example signal is a 30 sec trace from each analyte's entire data. Translocation signals were measured in 100kHz sampling rate with 10 kHz low pass filter. The diameter of the nanopore was estimated by Gartan software.

5.3.2. Sampling Rate Dependence

If an analyte is short, translocation time (dwell time) could be shorter too. It might be the reason why the shorter molecules (T5, T10, and HEP10) did not give frequent signals under 100 kHz sampling rate. The same analytes were measured at higher sampling frequency, 500 kHz (Figure 62). Even with higher sampling frequency, there is no significant translocation event detection with short analyte, HEP10. The average peaks of HEP60 which is a mixture of various length of heparin became doubled, but HEP10 is the same with 100 kHz measurement. It may imply that 500 kHz is still not enough for short molecules, however it can break some of long dwell time translocations of longer molecules that results in the average peaks increasing (from 30 peaks/min at 100 kHz to

62 peaks/min at 500 kHz). It was also observed with poly-thymine oligomers (not shown here).



	Ave. Peaks per Min
Control (2 min)	0
HEP10 (4 min)	3.6
HEP60 (6 min)	62.3

Figure 62. Sampling Rate Effect with GAG.

(a) Example of translocation signals of various length GAGs, (b) TEM image of the nanopore right after drilling, (c) Statistics of each measurement.

Example signal is a 30 sec trace from each analyte's entire data. Translocation signals were measured in 500kHz sampling rate with 50 kHz low pass filter. The higher low pass filter results in wider width of the baseline. The diameter of the nanopore was estimated by Gartan software.

5.4. Conclusion

From several experiments with oligomers and GAGs, it has been shown a possibility of GAG molecule. However, it needs further experiments not only to determine detail conditions such as pore size and length of analyte etc., but also to confirm translocation events. The project is at beginning step. The goal of the first step is demonstrate the possibility of the GAG molecule translocation on solid-state nanopore. The second goal is to identify GAGs through the RT junction embedded nanopore device. Several strategies are being considered to fabricate RT junction nanopore. The method shown in Figure 57 has been studied for a while in this group by colleagues, however current nanopore size is a few tenth nanometer which is too wide to measure in single molecule level. I am also developing a recipe drilling a nanopore on the metal deposited thin silicon membrane by TEM.

REFERENCES

[1] Shelley A. Claridge, Jeffrey J. Schwartz, and Paul S. Weiss (2011). Electrons, Photons, and Force: Quantitative Single-Molecule Measurements from Physics to Biology. *ACS Nano*, **5**, 693-729.

[2] Ashok A. Deniz, Samrat Mukhopadhyay and Edward A. Lemke (2008). Single-molecule biophysics: at the interface of biology, physics and chemistry. *J.R.Soc.Interface*, **5**, 15-45.

[3] Richard P. Feynman (1961). 'There's Plenty of Room at the Bottom', in *Miniaturization*, ed. H.D. Gilbert, *Reinhold Publishing Corporation*, New York, 282–296.

[4] Leandro C. Tabares, Ankur Gupta, Thijs J. Aartsma and Gerard W. Canters (2014). Tracking Electrons in Biological Macromolecules: From Ensemble to Single Molecule, *Molecules*, **19**, 11660-11678.

[5] Boris Rotman (1961). Measurement of activity of single molecules of β-D-galactosidase, *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 1981–1991.

[6] Massimiliano Di Ventra1 and Masateru Taniguchi (2016). Decoding DNA, RNA and peptides with quantum tunnelling, *Nature Nanotechnology*, **11**, 117-126.

[7] W. E. Moerner and Michel Orrit (1999). Illuminating Single Molecules in Condensed Matter, *Science*, **283**, 1670-1676.

[8] Ben N. G. Giepmans, Stephen R. Adams, Mark H. Ellisman, and Roger Y. Tsien (2006). The Fluorescent Toolbox for Assessing Protein Location and Function, *Science*, **312**, 217-224.

[9] Joshua Hihath*, Bingqian Xu*, Peiming Zhang†, and Nongjian Tao (2005). Study of single-nucleotide polymorphisms by means of electrical conductance measurements, *PNAS*, **102**, 16979-16983.

[10] Xiaoyin Xiao, Bingqian Xu, and Nongjian Tao (2004). Changes in the Conductance of Single Peptide Molecules upon Metal-Ion Binding, *Angew. Chem. Int. Ed.*, **43**, 6148-6152.

[11] Fang Chen, Joshua Hihath, Zhifeng Huang, Xiulan Li, and N.J. Tao (2007). Measurement of Single-Molecule Conductance, *Annu. Rev. Phys. Chem.*, **58**, 535-564.

[12] Amanda M. Moore and Paul S.Weiss (2008). Functional and Spectroscopic Measurements with Scanning Tunneling Microscopy, *Annu. Rev. Phys. Chem.*, **1**, 857-882.

[13] Gavin David Scott and Douglas Natelson (2010). Kondo Resonances in Molecular Devices, *ACS Nano*, **4**, 3560-3579.

[14] T. Albrecht (2012). Electrochemical tunnelling sensors and their potential applications, *Nature Communications*, **3**, 829-837.

[15] Francesca Moresco (2004). Manipulation of large molecules by low-temperature STM: model systems for molecular electronics, *Physics Reports*, **399**, 175-225.

[16] Supriyo Datta and Weidong Tian (1997). Current-Voltage Characteristics of Self-Assembled Monolayers by Scanning Tunneling Microscopy, *Physical Review Letters*, **79**, 2530-2533.

[17] https://en.wikipedia.org/wiki/Scanning_tunneling_microscope

[18] Xing Chen, Zheng Guo, Gui-Mei Yang, Jie Li, Min-Qiang Li, Jin-Huai Liu, Xing-Jiu Huang (2010). Electrical nanogap devices for biosensing, *Materials Today*, **13**, 28-41.

[19] Makusu Tsutsui and Masateru Taniguchi (2012). Single Molecule Electronics and Devices, *Sensors*, **12**, 7259-7298.

[20] Nicolas Agrait, Alfredo Levy Yeyati, JanM. van Ruiten beek (2003). Quantum properties of atomic-sized conductors, *Physics Reports*, **377**, 81-279.

[21] X. D. Cui, A. Primak, X. Zarate, J. Tomfohr, O. F. Sankey, A. L. Moore, T. A. Moore, D. Gust, G. Harris, S. M. Lindsay (2001). Reproducible Measurement of Single-Molecule Conductivity, *Science*, **294**, 571-574.

[22] G. Rubio, N. Agrait, and S. Vieira (1996). Atomic-Sized Metallic Contacts: Mechanical Properties and Electronic Transport, *Physical Review Letters*, **76**, 2302-2305.

[23] Bingqian Xu, Xiaoyin Xiao, and Nongjian J. Tao (2003). Measurements of Single-Molecule Electromechanical Properties, *J. Am. Chem. Soc.*, **125**, 16164-16165.

[24] Frederick J. Sigworth and Erwin Neher (1980). Single Na+ channel currents observed in cultured rat muscle cells, *Nature*, **287**, 447-449.

[25] Alina Asandei1, Mauro Chinappi, Jong-kook Lee, Chang Ho Seo, Loredana Mereuta, Yoonkyung Park and Tudor Luchian (2015). Placement of oppositely charged aminoacids at a polypeptide termini determines the voltagecontrolled braking of polymer transport through nanometer-scale pores, *Scientific Reports*, **5**, 10419-10431.

[26] Bala Murali Venkatesan1,2 and Rashid Bashir (2011). Nanopore sensors for nucleic acid analysis, *Nature Nanotechnology*, **6**, 615-624.

[27] R. Landauer (1988). Spatial variation of currents and fields due to localized scatterers in metallic conduction, *IBM J. Res. Develop.*, **32**, 306-316.

[28] M. Buttiker and R. Landauer (1982). Traversal Time for Tunneling, *Physical Review Letters*, **49**, 1739-1742.

[29] Alex D Gottlieb and LisaWesoloski (2006). Bardeen's tunnelling theory as applied to scanning tunnelling microscopy: a technical guide to the traditional interpretation, *Nanotechnology*, **17**, 57-65.

[30] J. Tersoff and D. R. Hamann (1983). Theory and Application for the Scanning Tunneling Microscope, *Physical Review Letters*, **50**, 1998-2001.

[31] C. Julian Chen (1990). Tunneling matrix elements in three-dimensional space: The derivative rule and the sum rule, Physical Review B, 42, 8841-8857.

[32] Eldon G. Emberly and George Kirczenow, Landauer theory (2000). Inelastic scattering, and electron transport in molecular wires, *Physical Review B*, **61**, 5740-5750.

[33] Datta, S. (2006). *Quantum Transport: Atom to Transistor*. 1st rev. edn (Cambridge University Press).

[34] Bingqian Xu and Nongjian J. Tao (2003). Measurement of Single-Molecule Resistance by Repeated Formation of Molecular Junctions, *Science*, **301**, 1221-1223.

[35] Shuai Chang, Shuo Huang, Jin He, Feng Liang, Peiming Zhang, Shengqing Li, Xiang Chen, Otto Sankey, and Stuart Lindsay (2010). Electronic Signatures of all Four DNA Nucleosides in a Tunneling Gap, *Nano Lett*., **10**, 1070-1075.

[36] Stuart Lindsay, JinHe, Otto Sankey, Prokop Hapala, Pavel Jelinek, Peiming Zhang, Shuai Chang and Shuo Huang (2010). Recognition tunneling, *Nanotechnology*, **21**, 262001-262012.

[37] Fang Chen, Xiulan Li, Joshua Hihath, Zhifeng Huang, and Nongjian Tao (2006). Effect of Anchoring Groups on Single-Molecule Conductance: Comparative Study of Thiol-, Amine-, and Carboxylic-Acid-Terminated Molecules, *J. Am. Chem. Soc.*, **128**, 15874-15884.

[38] Xiulan Li, Jin He, Joshua Hihath, Bingqian Xu, Stuart M. Lindsay, and Nongjian Tao (2006). Conductance of Single Alkanedithiols: Conduction Mechanism and Effect of Molecule-Electrode Contacts, *J. Am. Chem. Soc.*, **128**, 2135-2141.

[39] Vincent B. Engelkes, Jeremy M. Beebe, and C. Daniel Frisbie (2004). Length-Dependent Transport in Molecular Junctions Based on SAMs of Alkanethiols and Alkanedithiols: Effect of Metal Work Function and Applied Bias on Tunneling Efficiency and Contact Resistance, *J. Am. Chem. Soc.*, **126**, 14287-14296.

[40] Yu-Mo. Zhanga, Xiaojun Wanga, Weiran Zhanga, Wen Lia, Bing Yanga, Minjie Lia and Sean Xiao-An Zhang (2014). Cross polarization effect of donor–acceptor group on a potential single-molecule transistor, *J. Phys. Org. Chem.*, **27**, 834-840.

[41] Ismael Diez-Perez, Joshua Hihath, Thomas Hines, Zhong-Sheng Wang, Gang Zhou, Klaus Mullen and Nongjian Tao (2011). Controlling single-molecule conductance through lateral coupling of p orbitals, *Nature Nanotechnology*, **6**, 226-231.

[42] Adi Salomon, David Cahen, Stuart Lindsay, John Tomfohr, Vincent B. Engelkes, and C. Daniel Frisbie (2003). Comparison of Electronic Transport Measurements on Organic Molecules, *Adv. Matter.*, **15**, 1881-1890.

[43] Albert C. Aragones, Nadim Darwish, JongOne Im, Boram Lim, Jeongae Choi, Sangho Koo, and Ismael Diez-Perez (2015). Fine-Tuning of Single-Molecule Conductance by Tweaking Both Electronic Structure and Conformation of Side Substituents, *Chem. Eur. J.*, **21**, 7716-7720.

[44] Wenjing Hong, David Zsolt Manrique, Pavel Moreno-García, Murat Gulcur, Artem Mishchenko, Colin J. Lambert, Martin R. Bryce, and Thomas Wandlowski (2012). Single Molecular Conductance of Tolanes: Experimental and Theoretical Study on the Junction Evolution Dependent on the Anchoring Group, *J. Am. Chem. Soc.*, **134**, 2292-2304.

[45] Kai Golibrzuch, Florian Ehlers, Mirko Scholz, Rainer Oswald, Thomas Lenzer, Kawon Oum, Hyungjun Kim and Sangho Koo (2011). Ultrafast excited state dynamics and spectroscopy of 13,13’-diphenyl-β-carotene, *Phys. Chem. Chem. Phys.*, **13**, 6340-6351.

[46] Juwan Maeng, Soo Bong Kim, Nam Joo Lee, Eunho Choi, Se-Young Jung, Inseok Hong, Sung-Hee Bae, Jung Taek Oh, Boram Lim, Joon Woo Kim, Chi Jung Kang, and Sangho Koo (2010). Conductance Control in Stabilized Carotenoid Wires, *Chem. Eur. J.*, **16**, 7395-7399.

[47] N. J. Tao (2006). Electron transport in molecular junctions, *Nature Nanotechnology*, **1**, 173-181.

[48] F. Chen and N. J. Tao (2009). Electron Transport in Single Molecules: From Benzene to Graphene, *Accounts of Chemical Research*, **42**, 429-438.

[49] Latha Venkataraman, Young S. Park, Adam C. Whalley, Colin Nuckolls, Mark S. Hybertsen, and Michael L. Steigerwald (2007). Electronics and Chemistry: Varying Single-Molecule Junction Conductance Using Chemical Substituents, *Nano Lett.*, **7**, 502-506.

[50] Yanan Zhao, Stuart Lindsay, Sunhwa Jeon, Hyung-Jun Kim, Liang Su, Boram Lim, and Sangho Koo (2013). Combined Effect of Polar Substituents on the Electronic Flows in the Carotenoid Molecular Wires, *Chem. Eur. J.*, **19**, 10832-10835.

[51] Mathias Uhlen and Fredrik Ponten (2005). Antibody-based Proteomics for Human Tissue Profiling, *Molecular & Cellular Proteomics*, **4.4**, 384-394.

[52] National Research Council (US) Committee on Intellectual Property Rights in Genomic and Protein Research and Innovation. Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health (National Academies Press, 2006).

[53] Yanan Zhao, Brian Ashcroft, Peiming Zhang, Hao Liu, Suman Sen, Weisi Song,

JongOne Im, Brett Gyarfas, Saikat Manna, Sovan Biswas, Chad Borges and Stuart Lindsay (2014). Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling, *Nature Nanotechnology*, **9**, 466-473.

[54] Shuai Chang, Shuo Huang, Hao Liu, Peiming Zhang, Feng Liang, Rena Akahori, Shengqin Li, Brett Gyarfas, John Shumway, Brian Ashcroft, Jin He and Stuart Lindsay (2012). Chemical recognition and binding kinetics in a functionalized tunnel junction, *Nanotechnology*, **23**, 235101-235114.

[55] Chih-chung Chang and Chih-jen Lin (2011). LIBSVM: A Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, **2**.

[56] Shuo Huang, Jin He, Shuai Chang, Peiming Zhang, Feng Liang, Shengqin Li, Michael Tuchband, Alexander Fuhrmann, Robert Ros and Stuart Lindsay (2010). Identifying single bases in a DNA oligomer with electron tunneling, *Nature Nanotechnology*, **5**, 868-873.

[57] Alexander Ivanovich Archakov, Yurii Dmitrievich Ivanov, Andrey Valerevich Lisitsa, Victor Gavrilovich Zgoda (2007). AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics, *Proteomic*, **7**, 4-9.

[58] Shuai Chang, Suman Sen, Peiming Zhang, Brett Gyarfas, Brian Ashcroft, Steven Lefkowitz, Hongbo Peng and Stuart Lindsay (2012). Palladium electrodes for molecular tunnel junctions, *Nanotechnology*, **23**, 425202-425206.

[59] Feng Liang, Shengqing Li, Stuart Lindsay, and Peiming Zhang (2012). Synthesis, Physicochemical Properties, and Hydrogen Bonding of 4(5)-Substituted 1-*H*-Imidazole-2-carboxamide, A Potential Universal Reader for DNA Sequencing by Recognition Tunneling, *Chemistry*, **18**, 5998-6007.

[60] A. Michael Noll (1964). Short-Time Spectrum and “Cepstrum” Techniques for Vocal-Pitch Detection, *The Journal of the Acoustical Society of America*, **36**, 296-302.

[61] Arun Sreekumar, Laila M. Poisson, Thekkelnaycke M. Rajendiran, Amjad P. Khan, Qi Cao, Jindan Yu, Bharathi Laxman, Rohit Mehra, Robert J. Lonigro, Yong Li, Mukesh K. Nyati, Aarif Ahsan, Shanker Kalyana-Sundaram, Bo Han, Xuhong Cao, Jaeman Byun, Gilbert S.Omenn, Debashis Ghosh, Subramaniam Pennathur, Danny C. Alexander, Alvin Berger, Jeffrey R. Shuster, John T. Wei, Sooryanarayana Varambally, Christopher Beecher & Arul M. Chinnaiyan (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression, Nature Letters, 457, 910-915.

[62] Alexander Fuhrmann, Sebastian Getfert, Qiang Fu, Peter Reimann, Stuart Lindsay, and Robert Ros (2012). Long Lifetime of Hydrogen-Bonded DNA Basepairs by Force Spectroscopy, *Biophysical Journal*, **102**, 2381-2390.

[63] THOMAS M. COVER (2006). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition, *IEEE Transaction on Electric Computers*, **14**, 326-334.

[64] Jethava, Vinay, Anders Martinsson, Chiranjib Bhattacharyya, and Devdatt Dubhashi (2012). "The Lovász ϑ function, SVMs and finding large dense subgraphs." In *Advances in Neural Information Processing Systems*, 1160-1168.

[65] Erin L. Allwein, Robert E. Schapire, and Yoram Singer (2000). Reduing Multiclass to Binary: A Unifying Approach for Margin Classifiers, *Journal of Machine Learning Research*, **1**, 113-141.

[66] Jeff Nivala, Couglas B Marks, and Mark Akeson (2013). Unfolase-mediated protein translocation through an alpha-hemolysin nanopore. *Nature Biotechnology*, **11**, 247-251.

[67] J. Hofmann, H. S. Hahm, P. H. Seeberger, and K. Pagel (2015). Identification of carbohydrate anomers using ion mobility-mass spectrometry. *Nature* **526**, 241-244.

[68] JongOne Im, Sovan Biswas, Hao Liu, Yanan Zhao, Suman Sen, Sudipta Biswas, Brian Ashcroft, Chad Borges, Xu Wang, Stuart Lindsay and Peiming Zhang (2016). Electronic Single Molecule Identification of Carbohydrate Isomers by Recognition Tunneling, arXiv:1601.04221 [physics.chem-ph].

[69] Gerald W. Hart1,* and Ronald J. Copeland (2010). Glycomics Hits the Big Time, *Cell* **143**, 672-676.

[70] A. J Parodi (2000). Protein glucosylation and its role in protein folding. *Annu. Rev. Biochem.* **69**, 69-93.

[71] Yan-Yang Zhao, Motoko Takahashi, Jian-Guo Gu, Eiji Miyoshi, Akio Matsumoto, Shinobu Kitazume and Naoyuki Taniguchi (2008). Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer sci.* **99**, 1304-1310.

[72] Kazuaki Ohtsubo1 and Jamey D. Marth (2006). Glycosylation in cellular mechanisms of health and disease. *Cell* **126**, 855-867.

[73] Floriana Rosati, Antonietta Capone, Cinzia Della Giovampaola, Cecilia Brettoni and Riccardo Focarelli (2000). Sperm-egg interaction at fertilization: glycans as recognition signals. *Int. J. Dev.Biol.* **44**, 609-618.

[74] John B. Lowe1 and Jamey D. Marth (2003). A genetic approach to Mammalian glycan function. *Annu. Rev. Biochem.* **72**, 643-691.

[75] Taro Kawai and Shizuo Akira (2009). The roles of TLRs, RLRs and NLRs in pathogen recognition. *Int. Immunol*. **21**, 317-337.

[76] Bertozzi, C. R. & Rabuka (2009). Essentials of glycobiolgy (Second edition), Cold Spring Harbor Laboratory Press, New York.

[77] J. S. Rohrer, L. Basumallick, and D. Hurum (2013). High-Performance Anion-Exchange Chromatography with Pulsed Amperometric Detection for Carbohydrate Analysis of Glycoproteins, *Biochemistry (Moscow)* **78**, 697-709.

[78] Shuuji Hara, Masatoshi Yamaguchi, Yasuyo Takemori, Kimio Furuhata, Haruo Ogura, and Masuru Nakamura (1989). Determination of Mono-O-acetylated N-Acetylneuraminic Acids in Human and Rat Sera by Fluorometric High-Performance Liquid Chromatography, *Analytical Biochemistry* **179**, 162-166.

[79] David J. Ashline, Anthony J. Lapadula, Yan-Hui Liu, Mei Lin, Mike Grace, Birendra Pramanik, and Vernon N. Reinhold (2007). Carbohydrate Structural Isomers Analyzed by Sequential Mass Spectrometry, *Anal. Chem.* **79**, 3830-3842.

[80] Roger A. Laine (1994). A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4**, 759-767.

[81] Xu Han, Yeting Zheng, Catherine J Munro, Yiwen Ji and Adam B Braunschweig (2015). Carbohydrate nanotechnology: hierarchical assembly using nature's other information carrying biopolymers. *Curr. Opin. Biotechnol.* **34**, 41-47.

[82] Muchena J. Kailemia, L. Renee Ruhaak, Carlito B. Lebrilla, and I. Jonathan Amster (2014). Oligosaccharide Analysis by Mass Spectrometry: A Review of Recent Developments, *Anal. Chem.* **86**, 196-212.

[83] Rahul Raman, S Raguram, Ganesh Venkataraman, James C Paulson & Ram Sasisekharan (2005). Glycomics: an integrated systems approach to structure-function relationships of glycans, *Nature Methods* **2**, 817-824.

[84] Adrianan E. Manzi, Karin Norgard-Sumnicht, Sulabha Argade, Jamey D. Marth, Herman van Halbeek and Ajit Varki (2000). Exploring the glycan repertoire of genetically modified mice by isolation and profiling of the major glycan classes and nano-NMR analysis of glycan mixtures, *Glycobiology* **10**, 669-689.

[85] Heide Kogelberg, Dolores Solis and Jesus Jimenez-Barbero (2003). New structural insights into carbohydrate–protein interactions from NMR spectroscopy, *Current Opinion in Structural Biology* **13**, 646-653.

[86] Jens O. Duus, Charlotte H. Gotfredsen, and Klaus Bock (2000). Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chem. Rev.* **100**, 4589-4614.

[87] Nancy Leymarie and Joseph Zaia (2012). Effective Use of Mass Spectrometry for Glycan and Glycopeptide Structural Analysis, *Anal. Chem.* **84**, 3040-3048.

[88] Gabe Nagy and Nicola L. B. Pohl (2015). Monosaccharide Identification as a First Step toward de Novo Carbohydrate Sequencing: Mass Spectrometry Strategy for the Identification and Differentiation of Diastereomeric and Enantiomeric Pentose Isomers, *Anal. Chem.* **87**, 4566-4571.

[89] Shuo Huang, Shuai Chang, Jin He, Peiming Zhang, Feng Liang, Michael Tuchband, Shengqing Li, and Stuart Lindsay (2010). Recognition Tunneling Measurement of the Conductance of DNA Bases Embedded in Self-Assembled Monolayers, *J. Phys. Chem.* **114**, 20443-20448.

[90] Shuo Huang, Jin He, Shuai Chang, Peiming Zhang, Feng Liang, Shengqin Li, Michael Tuchband, Alexander Fuhrmann, Robert Ros and Stuart Lindsay (2010). Identifying single bases in a DNA oligomer with electron tunneling, *Nature Nanotechnology* **5**, 868-873.

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

[92] https://web.stanford.edu/group/murmann_group/cgi-bin/mediawiki/index.php/Aldo_Pe%C3%B1a_Perez

[93] Ulman, A. (1991). An Introduction to Ultrathin Organic Films: From Langmuir--Blodgett to Self-Assembly. (Academic Press).

[94] Shuai Chang, Jin He, Peiming Zhang, Brett Gyarfas, and Stuart Lindsay (2011). Gap Distance and Interactions in a Molecular Tunnel Junction, *J. Am. Chem. Soc.* **133**, 14267-14269.

[95] Birendra N. Pramanik, Peter L. Bartner, Urooj A. Mirza, Yan-Hui Liu and Ashit K. Ganguly (1998). Electrospray Ionization Mass Spectrometry for the Study of Non-covalent Complexes: an Emerging Technology, *J. Mass Spectrom.* **33**, 911-920.

[96] http://scikit-learn.org/stable/index.html

[97] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg (2008). Top 10 algorithms in data mining, *Knowl. Inf. Syst.* **14**, 1-37.

[98] Matthew J.Cracknell (2014). AnyaM.Reading, Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, *Computers & Geosciences* **63**, 22-33.

[99] Jack V. Tu (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes., *J. Clin. Epidemiol* **49**, 1225-1231.

[100] Corinna Cortes and Vladimir Vapnik (1995). Support-Vector Networks, *Machine Learning* **20**, 276-297.

[101] Alexandros Karatzoglou, David Meyer, and Kurt Hornik (2006). Support Vector Machines in R, Journal of Statistical Software 15.

[102] http://stackoverflow.com/questions/4629505/svm-hard-or-soft-margins

[103] Daniel B. Werz, René Ranzinger, Stephan Herget, Alexander Adibekian,Claus-Wilhelm von der Lieth, and Peter H. Seeberger (2007). Exploring the Structural Diversity of Mammalian Carbohydrates ("Glycospace") by Statistical Databank Analysis. *ACS Chem. Biol.* 2, 685–691.

[104] Alexander Adibekian, Pierre Stallforth, Marie-Lyn Hecht, Daniel B. Werz, Pascal Gagneuxd and Peter H. Seeberger (2011). Comparative bioinformatics analysis of the mammalian and bacterial glycomes. *Chem. Sci.* **2**, 337-344.

[105] P. Both1, A. P. Green1, C. J. Gray1, R. Sardzık, J. Voglmeir, C. Fontana, M. Austeri, M. Rejzek, D. Richardson, R. A. Field, G. Widmalm, S. L. Flitsch and C. E. Eyers (2014). Discrimination of epimeric glycans and glycopeptides using IM-MS and its potential for carbohydrate sequencing. *Nat. Chem.* **6**, 65-74.

[106] Pei Pang, Brian Alan Ashcroft, Weisi Song, Peiming Zhang, Sovan Biswas, Quan Qing, Jialing Yang, Robert J. Nemanich, Jingwei Bai, Joshua T. Smith, Kathleen Reuter, Venkat S. K. Balagurusamy, Yann Astier, Gustavo Stolovitzky, and Stuart Lindsay (2014). Fixed Gap Tunnel Junction for Reading DNA Nucleotides. *ACS Nano* **8**, 11994–12003.

[107] Daniel Branton, et al. (2008). Nanopore Sequencing. *Nat. Biotechnol.* **26**, 1146-1153.

[108] Joseph R. Bishop, Manuela Schuksz & Jeffrey D. Esko (2007). Heparan sulphate proteoglycans fine-tune mammalian physiology, *Nature* **446**, 1030-1037.

[109] Ashwani Khurana, Peng Liu, Pasquale Mellone, Laura Lorenzon, Bruno Vincenzi, Kaustubh Datta, Bo Yang, Robert J. Linhardt, Wilma Lingle, Jeremy Chien, Alfonso Baldi, and Viji Shridhar (2011). HSulf-1 Modulates FGF2- and Hypoxia-Mediated Migration and Invasion of Breast Cancer Cells, *Cancer Res.* **71**, 2152-2161.

[110] Jinping Lai, Jeremy Chien, Julie Staub, Rajeswari Avula, Eddie L. Greene, Tori A. Matthews, David I. Smith, Scott H. Kaufmann, Lewis R. Roberts, and Viji Shridhar (2003). Loss of HSulf-1 Up-regulates Heparin-binding Growth Factor Signaling in Cancer, *J. Bio. Chem.* **278**, 23107-23117.

[111] Jie Li, Min-Li Mo, Zhao Chen, Jie Yang, Qiu-Shi Li, Dian-Jun Wang, Hui Zhang, Ying-Jiang Ye, Jun-Pu Xu, Hai-Long Li, Fang Zhang and Hai-Meng Zhou (2011). HSulf-1 inhibits cell proliferation and invasion in human gastric cancer, *Cancer Science* **102**, 1815-1821.

[112] Keishi Narita, Jeremy Chien, Sally A. Mullany, Julie Staub, Xiang Qian, Wilma L. Lingle, and Viji Shridhar (2007). Loss of HSulf-1 Expression Enhances Autocrine Signaling Mediated by Amphiregulin in Breast Cancer, *J. Bio. Chem.* **282**, 14413-14420.

[113] Allison H. Bartlett and Pyong Woo Park (2010). Proteoglycans in host–pathogen interactions: molecular mechanisms and therapeutic implications, *Expert Reviews in Molecular Medicine* **12**.

[114] Luigi Bergamaschini1, Emanuela Rossi, Carlo Vergani, and Maria Grazia De Simoni (2009). Alzheimer’s Disease: Another Target for Heparin Therapy, *The Scientific World Journal* **9**, 891-908.

[115] Qing Ma1, Umberto Cornelli, Israel Hanin, Walter P. Jeske, Robert J. Linhardt, Jeanine M. Walenga, Jawed Fareed and John M. Lee (2007). Heparin Oligosaccharides as Potential Therapeutic Agents in Senile Dementia, *Current Pharaceutical Design* **13**, 1607-1616.

[116] Elazabeth Pennisi (2014). DNA Sequencers Still Waiting for The Nanopore Revolution, *Science* **343**, 829-830.

[117] Abdelghani Oukhaled, Laurent Bacri, Manuela Pastoriza-Gallego, Jean-Michel Betton, and Juan Pelta (2012). Sensing Proteins through Nanopores: Fundamental to Applications, *ACS Chem. Biol.* **7**, 1935-1949.

[118] David Rodriguez-Larrea and Hagan Bayley (2013). Multistep protein unfolding during nanopore Translocation, *Nature Nanotechnology* **8**, 288-295.

[119] Masateru Taniguchi (2015). Selective Multidetection Using Nanopores, *Anal. Chem.* **87**, 188-199.

[120] Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Andrew Adey, Ian C Nova, Jonathan M Craig, Kyle W Langford, Jenny Mae Samson, Riza Daza, Kenji Doering, Jay Shendure and Jens H Gundlach (2014). Decoding long nanopore sequencing reads of natural DNA, *Nature Biotechnology* **32**, 829-834.

[121] S. Garaj, W. Hubbard, A. Reina, J. Kong, D. Branton and J. A. Golovchenko (2010). Graphene as a subnanometre trans-electrode Membrane, *Nature* **467**, 190-194.

[122] Lisen Kullman, Mathias Winterhalter, and Sergey M. Bezrukov (2002). Transport of Maltodextrins through Maltoporin: A Single-Channel Study, *Biophysical Journal* **82**, 803-812.

[123] Li-Qun Gu, Stephen Cheley, and Hagan Bayley (2003). Electroosmotic enhancement of the binding of a neutral molecule to a transmembrane pore, *PNAS* **100**, 15498-15503.

[124] Luciana R. Teixeira, Petr G. Merzlyak, Angela Valeva, and Oleg V. Krasilnikov (2009). Interaction of Heparins and Dextran Sulfates with a Mesoscopic Protein Nanopore, *Biophyscial Journal* **97**, 2894-2903.

[125] Laurent Bacri, Abdelghani Oukhaled, Erven Hémon, Fenseth Banzouzi Bassafoula, Loïc Auvray, Régis Daniel (2011). Discrimination of neutral oligosaccharides through a nanopore, *Biochemical and Biophysical Research Communications* **412**, 561-564.

[126] Takemasa, M., Fujita, M. and Maeda, M. (2012). Glycan Analysis Using a Solid State Nanopore. in *Nanopores for Bioanalytical Applications. Proceedings of the Interantional Conference* (eds. Edel, J. & Albrecht, T.) 89-92 (RSC Publishing, UK.

[127] Sudipta Biswas, Weisi Song, Chad Borges, Stuart Lindsay, and Peiming Zhang (2015). Click Addition of a DNA Thread to the N‑Termini of Peptides for Their Translocation through Solid-State Nanopores, *ACS Nano* **9**, 9652-9664.