Identification of Biomolecular Building Blocks by Recognition Tunneling:

Stride towards Nanopore Sequencing of Biomolecules

by

Suman Sen

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, Co-Chair
Peiming Zhang, Co-Chair
Ian R. Gould
Chad Borges

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

DNA, RNA and Protein are three pivotal biomolecules in human and other organisms, playing decisive roles in functionality, appearance, diseases development and other physiological phenomena. Hence, sequencing of these biomolecules acquires the prime interest in the scientific community. Single molecular identification of their building blocks can be done by a technique called Recognition Tunneling (RT) based on Scanning Tunneling Microscope (STM). A single layer of specially designed recognition molecule is attached to the STM electrodes, which trap the targeted molecules (DNA nucleoside monophosphates, RNA nucleoside monophosphates or amino acids) inside the STM nanogap. Depending on their different binding interactions with the recognition molecules, the analyte molecules generate stochastic signal trains accommodating their "electronic fingerprints". Signal features are used to detect the molecules using a machine learning algorithm and different molecules can be identified with significantly high accuracy. This, in turn, paves the way for rapid, economical nanopore sequencing platform, overcoming the drawbacks of Next Generation Sequencing (NGS) techniques.

To read DNA nucleotides with high accuracy in an STM tunnel junction a series of nitrogen-based heterocycles were designed and examined to check their capabilities to interact with naturally occurring DNA nucleotides by hydrogen bonding in the tunnel junction. These recognition molecules are Benzimidazole, Imidazole, Triazole and Pyrrole. Benzimidazole proved to be best among them showing DNA nucleotide classification accuracy close to 99%. Also, Imidazole reader can read an abasic monophosphate (AP), a product from depurination or depyrimidination that occurs 10,000 times per human cell per day.

In another study, I have investigated a new universal reader, 1-(2-mercaptoethyl)pyrene (Pyrene reader) based on stacking interactions, which should be more specific to the canonical DNA nucleosides. In addition, Pyrene reader showed higher DNA base-calling accuracy compare to Imidazole reader, the workhorse in our previous projects. In my other projects, various amino acids and RNA nucleoside monophosphates were also classified with significantly high accuracy using RT. Twenty naturally occurring amino acids and various RNA nucleosides (four canonical and two modified) were successfully identified. Thus, we envision nanopore sequencing biomolecules using Recognition Tunneling (RT) that should provide comprehensive betterment over current technologies in terms of time, chemical and instrumental cost and capability of de novo sequencing.

ACKNOWLEDGMENTS

I have been extremely fortunate to have Dr. Stuart Lindsay and Dr. Peiming Zhang as the advisors during my Ph.D. study. I would like to convey my heartiest reverence for their instruction, patience, and most importantly their mentorship throughout last five and half years of my research life. They provided me invaluable help for my research work through their knowledge, experience and kindness. I am overwhelmed with gratitude by all the efforts, support and guidance that I received from them during my growth as a researcher. I gratefully acknowledge all my committee members. Dr. Ian Gould and Dr. Chad Borges. Thank you for all your insightful instructions and discussions. I sincerely thank Dr. Predrag Krstic for his kind discussions and collaboration work for my projects. I would also like to mention gratefully, Dr. Brian Ashcroft and JongOne Im, for their extensive efforts on machine learning algorithms and related work, Sovan Biswas for providing me diverse synthetic molecules matching my requirements and Dr. Timothy Karcher, for his expertise and help with the XPS. It was a pleasant experience to work with Dr. Yanan Zhao, Dr. Shuai Chang, Dr. Weisi Song, Dr. Pei Pang, Dr. Subhadip Senapati, Sudipta Biswas, Saikat Manna, Dr. Padmini Krishnakumar and Dr. Parminder Kaur throughout these years. Thank you all for the collaboration, support and great ideas. The experience of working with you will be my precious memory. Also, I would like to thank Margaret Black and Michael Dodson for helping with various research and administrative obstacles. Other than my group mates, I also thank ASU for providing me the opportunity of study, teaching experience, facilities and financial support and many thanks to all my colleagues and friends with whom I worked together at ASU. Last but not the least, my gratitude to the DNA sequencing technology program of the National Human Genome Research Institute for sharing funding of my projects.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF ABBREVIATIONS

ADP	adenosine-5'-diphosphate
aq	aqueous
°C	degree Celsius
cm	centimeter
DNA	deoxyribonucleic
EDTA	ethylenedinitrilotetraacetic acid
ESI	electrospray ionization
g	gram (s)
h	hour (s)
H_2	hydrogen gas
H_2O	water
HCl	hydrochloric acid
HPLC	high pressure liquid chromatography
Hz	Hertz
L	liter
M	molar
mL	milliliter
mM	millimolar
mmol	millimole(s)

mRNA	messenger ribonucleic acid
μM	micromolar
μmol	micromole(s)
N	normal
nm	nanometer
RNA	ribonucleic acid
STM	scanning tunneling microscope
SVM	support vector machine
tRNA	transfer RNA
UV	ultraviolet
VIS	visible
XPS	x-ray photoelectron spectroscopy

CHAPTER 1

INTRODUCTION

1.1 DNA, RNA and Protein: The Central Dogma

The human body is a miraculous creation of nature. A countless number of complex and delicate processes are taking place synchronously. "The central dogma" of life narrates apparently the most decisive one among them. An enzyme, called RNA polymerase initiates the decoding process of genomic blueprint from the gene, a stretch up to several thousands of DNA base-pairs and creates a messenger-RNA or mRNA molecule with the free nucleotides from the nucleus as it slides along the DNA strand. This process is called "transcription" and the mRNA strand reciprocates the sequence of the DNA stretch. mRNA is then modified by discarding sections corresponding to non-coding DNA stretch of the gene. This edited mRNA leaves the nucleus and arrives at "ribosome" to serve as the protein manufacturing template. The mRNA code is interpreted by the ribosome and different amino acids are delivered to the ribosome by transfer-RNA or tRNA, followed by the formation of an amino acid chain according to the mRNA base array. This process is known as "translation". As the last amino acid molecule is coupled to the strand, it arranges itself into a complex spatial conformation and forms a protein molecule.

1.2 Protein

Proteins can be described as the constitutional elements of organisms and can be considered as one of the most important biomolecules for life. They play various roles such as transporters, enzymes and much more. There are 20 different naturally occurring amino acids which serve as the building blocks of all these proteins. An amino acid can be joined with a

couple of others by amide bonds and repetition results a one-dimensional chain of amino acids called peptide. The sequence of amino acids in a peptide is considered as the primary structure of a protein and controls the local folding of the chain or the secondary structure of protein. One or more long peptide chains form protein molecules and its overall folding and spatial orientation in known as the tertiary structure.

1.2.1 Protein Sequencing: Motivation

Even after successful sequencing of the complete human genome, plenty of mysteries about gene expression and subsequent protein production remain unresolved. The presence of 98% non-coding part in human genome barely contributes to the knowledge of protein biomarkers for various diseases.[1] The specific functions of non-coding DNAs are yet to be understood accurately but the effort behind sequencing this preeminent portion of the human genome seems rather impractical to resolve much smaller protein sequences or their corresponding manufacturing templates. Resolving the primary structure of proteins i.e. amino acid sequences in the peptide can serve the purpose itself and would help us to understand gene mutation, virus invading and required personalized medicine.

In addition to that, proteins have countless numbers of protein variants as a consequence of RNA splicing and post-translational modifications. Secondly, protein variants related to diseases are frequently present in extremely low concentrations. In the case of nucleic acids, low concentrations can be amplified using the polymerase chain reaction, but there is no analogous process for proteins. Hence, to develop protein biomarkers sequencing of proteins at single-molecular level is crucial.

1.3 Popular Methods

1.3.1 Edman Degradation

Edman degradation method for peptide sequencing was discovered by Pehr Edman.[2] This approach relies on the labeling of the first amino acid at the N-terminal and consequent degradation and identification through a chromatography technique. First, the peptide is exposed to phenylisothiocyanate, which reacts with the uncharged amino group of the N-terminal amino acid and a cyclic phenylthiocarbamoyl intermediate is formed. In the second step, the first N-terminal peptide bond is cleaved under a mild acidic condition and a phenylthiohydantoin (PTH) derivative of the N-terminal amino acid is formed. This PTH-

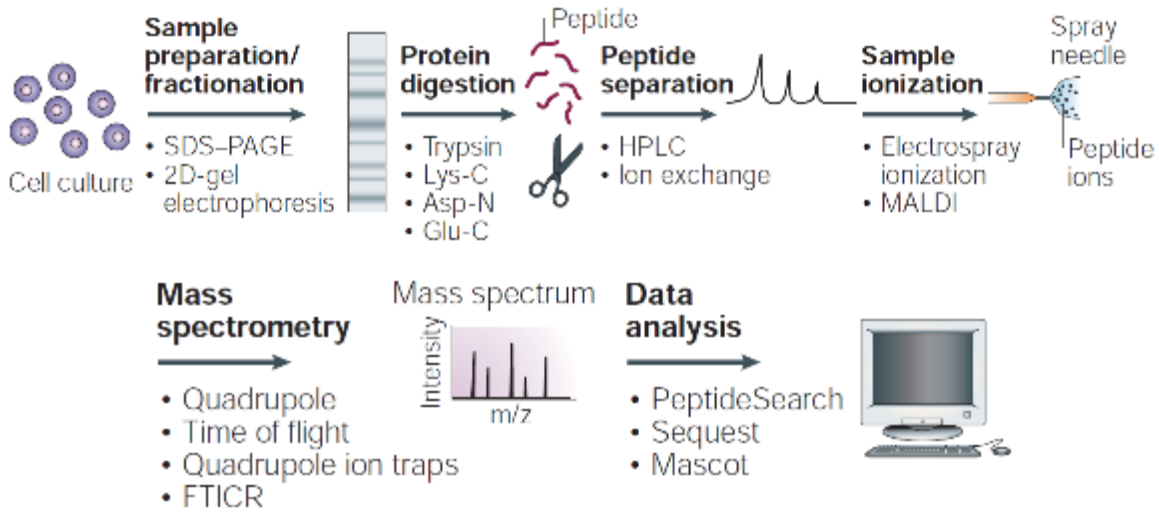

Figure 1.1. Mass-spectrometry experimental work-flow for protein/peptide sequencing

amino acid is then identified by chromatography or electrophoresis technique. Ion chromatography is the most used among them which depends on the difference in binding affinity of the ions or polar species with the ion exchange resin. After separation, the PTH-

amino acids can be characterized by UV/Visible spectroscopy, utilizing the Ninhydrin reaction, which produces a purple colored product.

1.3.2 Mass Spectrometry

Mass spectrometry based methods for protein sequencing are extremely popular in the present research environment. Ionization of target protein/peptide is the most important step and can be achieved by either electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) technique (figure 1.1[3]). Generally, protein is digested by a protease enzyme to create shorter peptide fragments and separated from each other by high pressure liquid chromatography (HPLC). Peptides are then sequentially analyzed by introducing them to the mass spectrometer directly from the HPLC column and ionizing using ESI or MALDI technique. The ionized peptide is then analyzed by the mass analyzer and compared to a database to discover its sequence. This process is repeated till the last peptide fragment is introduced into the mass spectrometer from the HPLC column.

1.3.3 Limitations

The major drawback related to Edman degradation method is the inability to sequence long polypeptides. Peptide chains longer than 50-60 (30, to be more accurate) are difficult to sequence due to incompletion of the cyclic derivative formation step and as a result, longer peptides need to be cleaved into shorter fragments, making the process much more complicated. Secondly, Edman degradation method becomes inactive in the case of peptides with chemically modified N-terminal amino acids or if the N-terminal amino acid is buried inside the protein cavity. For mass spectrometer techniques, peptide chain length could be improved compared to Edman degradation method. But, performing de novo sequencing is still challenging using mass spectrometry techniques. Besides, both methods are unable to

differentiate between enantiomeric and isobaric amino acids, which is imperative to understand the primary structure of proteins in a proper way.

1.4 The Human Genome & DNA

The genome can be regarded as the blueprint for any organism and constructed by basic building blocks known as deoxy-ribonucleic acid or DNA. It carries all the genetic and biological information of various organisms (along with some viruses) and responsible for instructions about their reproduction, development and metabolism. Long DNA strands formulate "genes", codes for different complex biological processes. All these genes are embedded in 23 pairs of "chromosomes" which reside in cell-nucleus.

1.4.1 "The DNA" and Its Structure

The answer to the question "What is DNA?" can be given as "Arguably the most important molecule in genetics and biochemistry". In 1953, James Watson, Francis Crick and Maurice Wilkins solved the complex structure of DNA[4, 5] and were awarded the Noble prize in 1962 for this monumental work. DNA is a macromolecule with a unique double helix structure made of two complementary strands.[4] Each strand is made of polymeric phosphate-deoxyribose backbone and four different kinds of nucleobases: Adenine(A), Guanine(G), Cytosine(C) and Thymine(T) as shown in figure 1[6]. A strand can also be described as a long thread of subunits called nucleotides made of a nucleobase and a phosphate-deoxyribose moiety. Both strands are anti-parallel to each other in a sense that if we follow an arrow from the 5' carbon to the 3' carbon of any deoxyribose sugar they are in opposite direction in the complementary strands. This intricate structure of DNA is maintained by two kinds of non-covalent interactions: hydrogen bonding between hydrogen bonding sites of complementary nucleobase-pairs and π-π stacking interaction between aromatic surfaces of the nucleobases.[7]

G-C and A-T are called complementary base-pairs as G(or C) only binds with C(or G) and

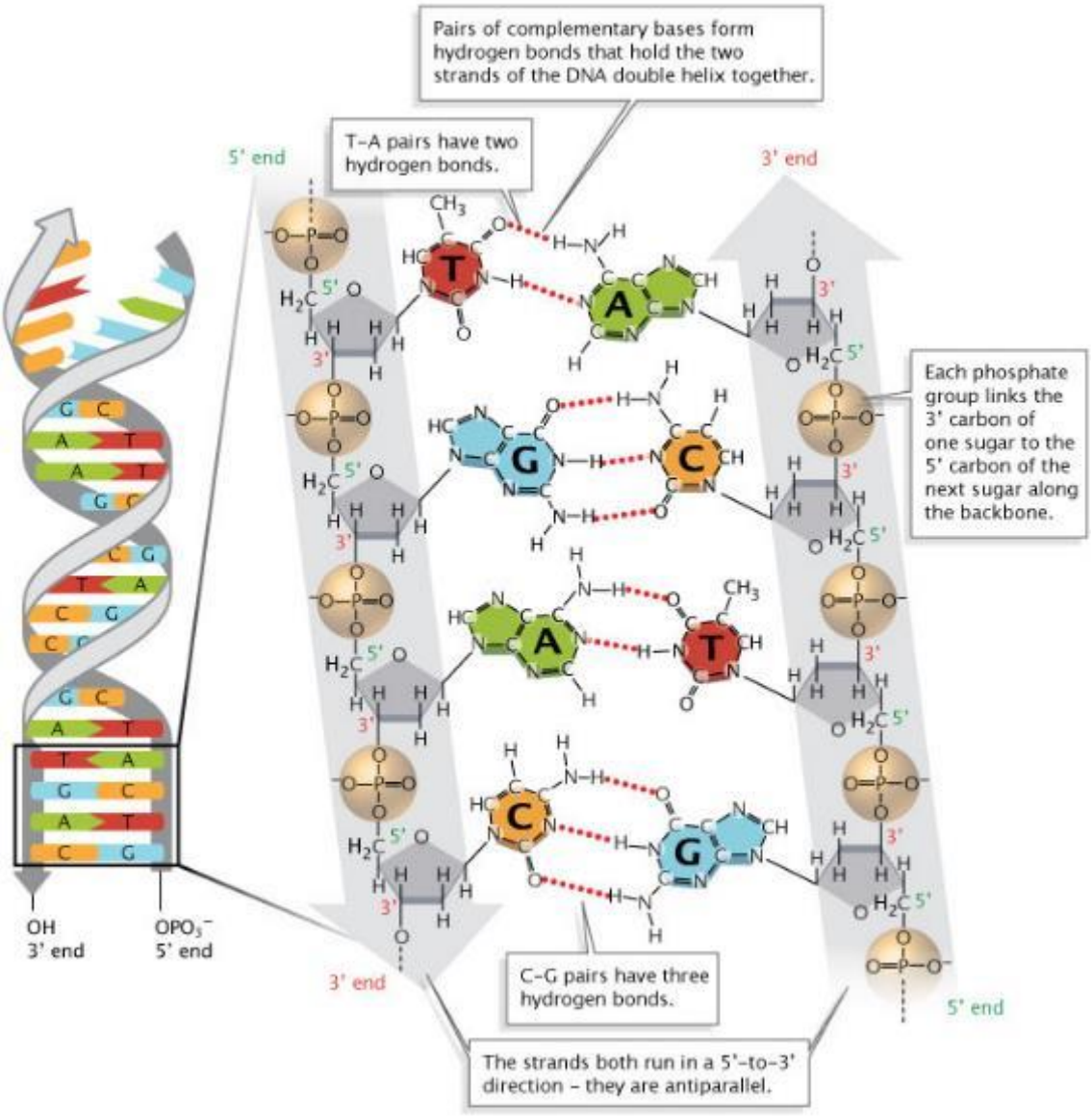

Figure 1.2. Structure of DNA double helix

A(or T) only binds with T(or A) by three and two hydrogen bonds, respectively. In a double helix structure of B-form (most common) DNA, the distance between two adjacent nucleotides is nearly 0.3nm and diameter of the helix is approximately 2nm.[8]

1.5 DNA Sequencing

DNA sequencing possesses the goal of resolving the order of nucleotides in a DNA strand. Most of the techniques determine this order in single-stranded DNA (ssDNA) rather than in double-stranded DNA (dsDNA) due to ease of the process. As we move along, I will describe some sequencing methods and it will appear that when we target a ssDNA to sequence down, we determine the sequence of the complementary strand instead, in most of the processes. But, at the end of the process, we have the complete sequence of the double helix.

1.6 DNA Sequencing: Motivation

DNA sequencing always has been a famed topic in the science community. The reason behind is its broad application in so many research fields. Here I mention some of the most popular applications of DNA sequencing.

1.6.1 Personalized Medicine

Whole Genome Sequencing (WGS) cost has declined in an unreal manner from $2.7 Billion to just several thousand dollars and the consequence is a bright future for personalized medicine. The idea is to obtain the clinically complete Human Genome data of any individual and predict the future diseases that individual is going to develop and prescribe the treatments for his/her present and predicted diseases. Recent astonishing development on genome and exome sequencing methods and increasing volume of the available database of genomic information from diverse cases of patient's genome, infectious agents and pathogen genome indicate that personalization of disease prognosis and following therapeutic is not far. [9, 10]

1.6.2 Genetic Variation in Humans and Human Diseases:

Knowledge of rare sequence variants in human germline genome or genome of an infectious pathogen can help us to understand human diseases pathogenesis, predictive development of the disease and its response towards treatment. An elaborated comparison study on the human genome that is healthy and in the phase before unfolding its death causing disease and the genome sequence variant associated with the disease is required. [11, 12]

1.6.3 Cancer Study

Carcinogenesis or Oncogenesis, the process of development of cancer is caused by genetic or epigenetic mutations of the normal healthy cells. Several top-notch, highly collaborative projects on cancer study (such as International Cancer Genome Consortium, Cancer Genome Atlas etc.) have been focusing on the sequence of genomes and exomes of different individuals, carrying various well-known categories of cancer malignancies. The idea is to recognize the mutations related to Oncogenesis. It is also necessary to differentiate genomic structure of cancer cells from non-cancerous cells. In addition, knowledge of tumor genome helps the cancer therapeutic strategies.[10]

1.6.4 Understanding the Immune System Response

Plasma cell produced Immunoglobulin (Ig) protein (or commonly known as an antibody) plays a significant role in recognizing and nullifying pathogens (various viruses and bacteria). Along with Ig, T-cell or T lymphocytes are also important in immune system due to its cell-mediated immunity. Proper monitoring of genetic rearrangements of the Ig receptors of B cells (cells of the adaptive immune system) and T-cell receptors of T cells (also part of the adaptive immune system) can immensely help in understanding the immune system response and autoimmune disorders in human. Sequencing of Ig gene rearrangement can be used as a

tool for tracking interactions between viral pathogens (e.g. HIV) and response on the adaptive immune system. This approach can also be a pathfinder against various autoimmune diseases and immunodeficiency disorders.[10, 13-15]

1.6.5 Crime Forensic

Advancement in DNA sequencing provides great help in the field of crime forensic. Sequencing of a genomic sample found in a crime spot can be matched to an existing database of criminal profiles (such as FBI's CODIS) to find any criminal suspect. Similarly, DNA samples from different criminal cases can be compared to each other to check if there is a common suspect. In addition, this tool also can help to identify any missing personal from dead-body or other related samples. Exploiting different genome variants, one can identify whether a suspect profile belongs to any specific ancestry (e.g. European, African or Asian origin). Predictions also can be made on physical characteristics such as eye color or hair color of an unknown suspect.[16] Sequencing giants are also providing special instruments particular to this field (such as Illumina's MiSeq FGx Forensic Genomics System).[17]

1.7 RNA

RNA is also a crucial biomolecule, as important as DNA and completes the bio-molecular series known as "nucleic acid" along with DNA. Unlike DNA, RNA is a single-stranded molecule and the strands are constructed with polymeric phosphate-ribose backbone and four different kinds of nucleobases: Adenine(A), Guanine(G), Cytosine(C) and Uracil(U). Though DNA does not have any variety, different types of RNAs (mRNA, t-RNA, non-coding RNA, r-RNA etc.) are present depending on various structural and functional properties. All these different RNAs found in a cell at any point of time combine together defines the "transcriptome" of that cell at that definite moment and physiological environment.

1.8 RNA Sequencing: Motivation

People have been talking about DNA sequencing for most of the time when it comes to understanding the human genome, neglecting the importance of profiling RNA structures. Though, the scenario has been changed abruptly in recent times with realizing the importance of "transcriptome" for complete knowledge of genome. During the production of RNA from DNA through transcription, the process of alternative splicing helps to generate various mRNA codes from a single gene.[18] As a consequence, several proteins can be synthesized from a single gene code. Hence, profiling all those resulting mRNAs after RNA splicing is critical. Also, single nucleotide polymorphisms (SNP) or mutations can occur during transcription, whereas post-transcriptional modifications occur beyond transcription and affect further alternation. All these phenomena require RNA sequencing to understand their powerful effects on protein production. Though RNA sample purification processes can be more complicated, still most of the popular DNA sequencing techniques are capable of sequencing RNA due to their structural similarities.

1.9 Evolution of DNA Sequencing

1.9.1 Starting Point: Sanger Sequencing

In 1977, the discovery of the Sanger Sequencing method, also named as Chain-termination method by the noble laureate Frederick Sanger and his co-workers, was nothing short of a revolution in the field of genetics and biochemistry.[19] Before being outshined by the

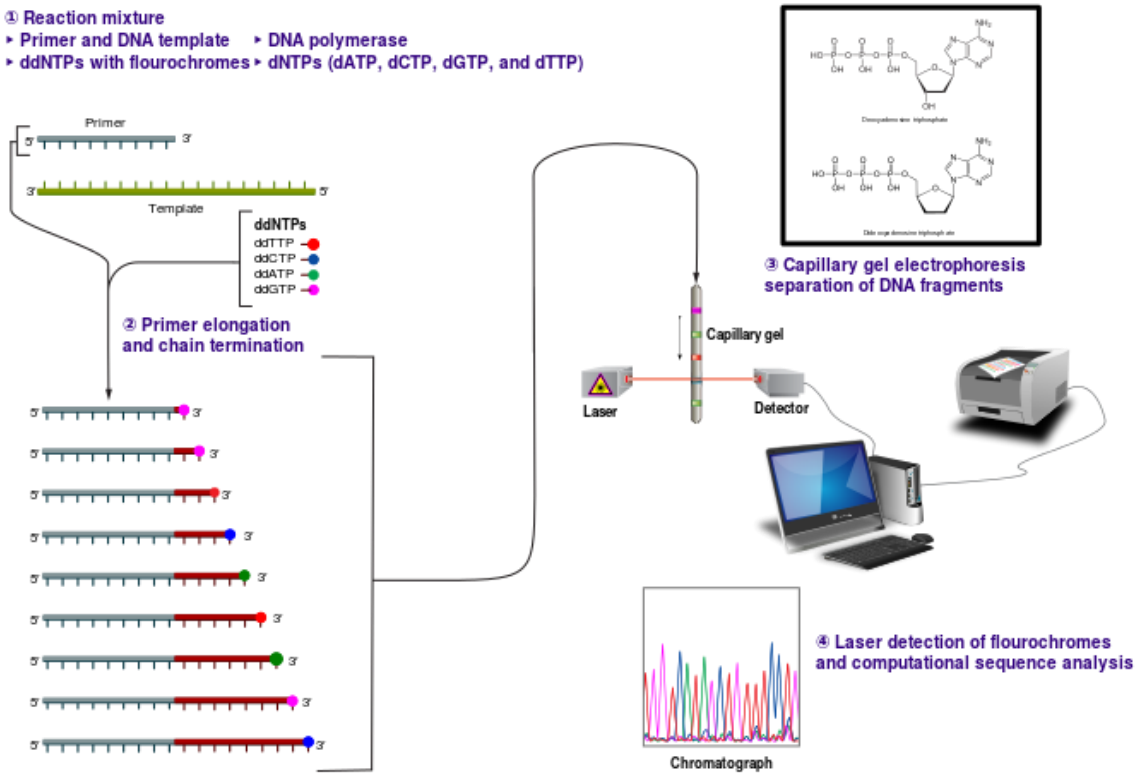

Figure1.3. Work-flow of Sanger chain-termination method

'nextgen' sequencing techniques, Sanger method was by far the standout approach for almost three decades.

Method: The method starts with denaturing dsDNA to ssDNA as its separates out to template strand and complementary strand (figure 2[20]). The template strand is combined with a primer by annealing. The primer is a short nucleic acid chain that helps DNA polymerase to start DNA replication. Next step is setting up four reaction containers, all of them having the sequencing target, the primer attached DNA strand and followed by addition of DNA polymerase and free nucleotide or dNTPs (one of them is radiolabeled with radioactive S or P), sequentially. Then only one type of ddNTPs (di-deoxy nucleotide triphosphates) is added to each container. The trick is the absence of –OH group in ddNTPs that stops DNA replication as it is unable to react with the next dNTP. Hence, terminates the

chain. Then polyacrylamide gel electrophoresis is used to run all four reaction mixtures in separate lanes of the gel, and smaller DNA fragments move faster and further from the negative end to the positive end. The presence of radiolabeled dNTPs shows various bands as an X-ray of the gel is taken. In the end, the sequence can be traced down by reading these DNA bands across the X-ray film.[19, 21] Though modern platforms use capillary gel electrophoresis instead of polyacrylamide gel (figure 2).

Drawbacks: Though it opened a new horizon in the world of biochemistry, drawbacks related to Sanger method forced it to be a back-bencher as modern day research deals with a gigantic amount of sequencing data. Pivotal challenges lie in the gel separation technique due to the inability of automation and handle many parallel separation processes, simultaneously.[22] Also, the requirement of a large amount of template DNA for each reaction and lengthy nature of the whole process were the driving force for the evolution of new sequencing strategies.[23]

1.9.2 Next Generation Sequencing

Advantages over Sanger sequencing: From the early years of the new century Sanger chain-termination method started to lose its flare as the "next-generation sequencing" (NGS) techniques challenged it with their prime advantage of massively parallel and high throughput characteristics. The amount of data originated from a single run with the NGS is enormous compared to that of Sanger sequencing. As a consequence, NGS provides very high speed and the year-long projects with Sanger sequencing can be finished in no time employing NGS. Though NGS has the issue of less accuracy on each run, a high degree of coverage makes it highly accurate after overall data assembly.[22]

Roche 454 GS FLEX: dsDNA is first fragmented in smaller strands (300-800bps) by the method called Nebulization using a Nebulizer (figure 1.4 [24]). The fragmented dsDNAs are

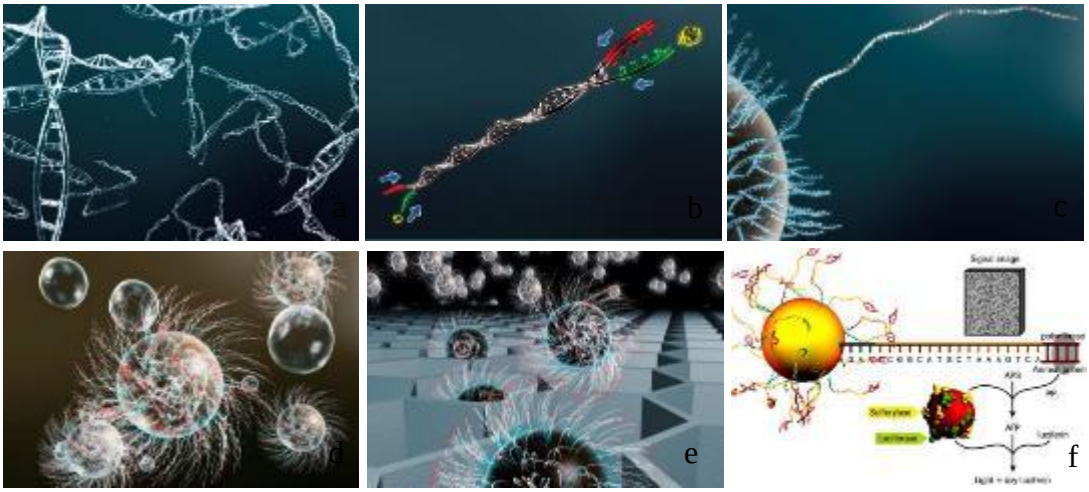

Figure1.4. Work-flow of Roche GS FLEX sequencing method

ligated with adapters and denaturized to ssDNA by heating. Attached DNA adapters help with purification, quantification, amplification and sequencing of the target DNA in a sequential manner. Each of the ligated strands is then added to a bead (by adding a large excess of beads) followed by emulsification so that each DNA fragment containing beads separate out into a small oil emulsion. Next step is the amplification of DNA fragments in each fragment-bead complexes inside oil emulsions in the presence of added primer, DNA polymerase and dNTPs. These emulsions are now called emulsion PCR micro-reactors. Amplification produces approximately 1 million new fragments on the surface of each bead. Then the emulsions are broken and these beads are loaded into pico-titer plate (PTP) where only a single bead loads inside each micro-well followed by the addition of packing beads, enzyme beads and PPiase beads. Next is the pyrosequencing step, in which a pyrophosphate unit is released when one dTNP gets incorporated into the strand during DNA strand elongation by polymerase followed by conversion of pyrophosphate into ATP by sulfurylase. This ATP is then utilized

by luciferase to oxidize luciferin and create a chemiluminescence event which is captured by detector and results consequential base identification and sequencing.[22, 25]

Drawbacks: High cost of various enzymes, lengthy sample preparation process and crosstalk between clonally amplified neighboring beads inside PTP are still challenges to overcome in this NGS technique. [25]

Illumina: Illumina's Hi Seq 2500 sequencer is currently one of the most popular platform and capable of generating 120 Gb data per rapid single run in 29 hours using dual flow-cell. [26] As shown in figure 1.5,[27] fragmented dsDNA is first ligated with double stranded adapter molecules at both terminals and then denatured to ssDNA. Next, these fragments are attached to the complementary oligo-adapter containing glass-slides of the flow-cell in a 'Bridge' orientation due to the presence of adapter at both ends of the fragments. One of the adapters serve as primers and synthesize the reverse strand in the presence of dNTPs and enzyme. Both complementary strands are then dissociated from each other and the same process is repeated extensively resulting "Bridge-amplification" of the target DNA fragment and producing millions of clonally amplified clusters. Each flow-cells contains approximately 150 million clusters, whereas each cluster contains around 1000 fragments. All the reverse strands are then washed off, leaving only forward strands, ready for sequencing. Then a primer is hybridized to adapter sequence so that 3' ends are blocked to avoid unwanted priming of the cluster fragments. Differently color-coded and fluorescence-tagged nucleotides are then added for synthesizing the complementary strands. After each addition of nucleotides, clusters are excited by lasers, identifying the base from the color code of the nucleotides and read the

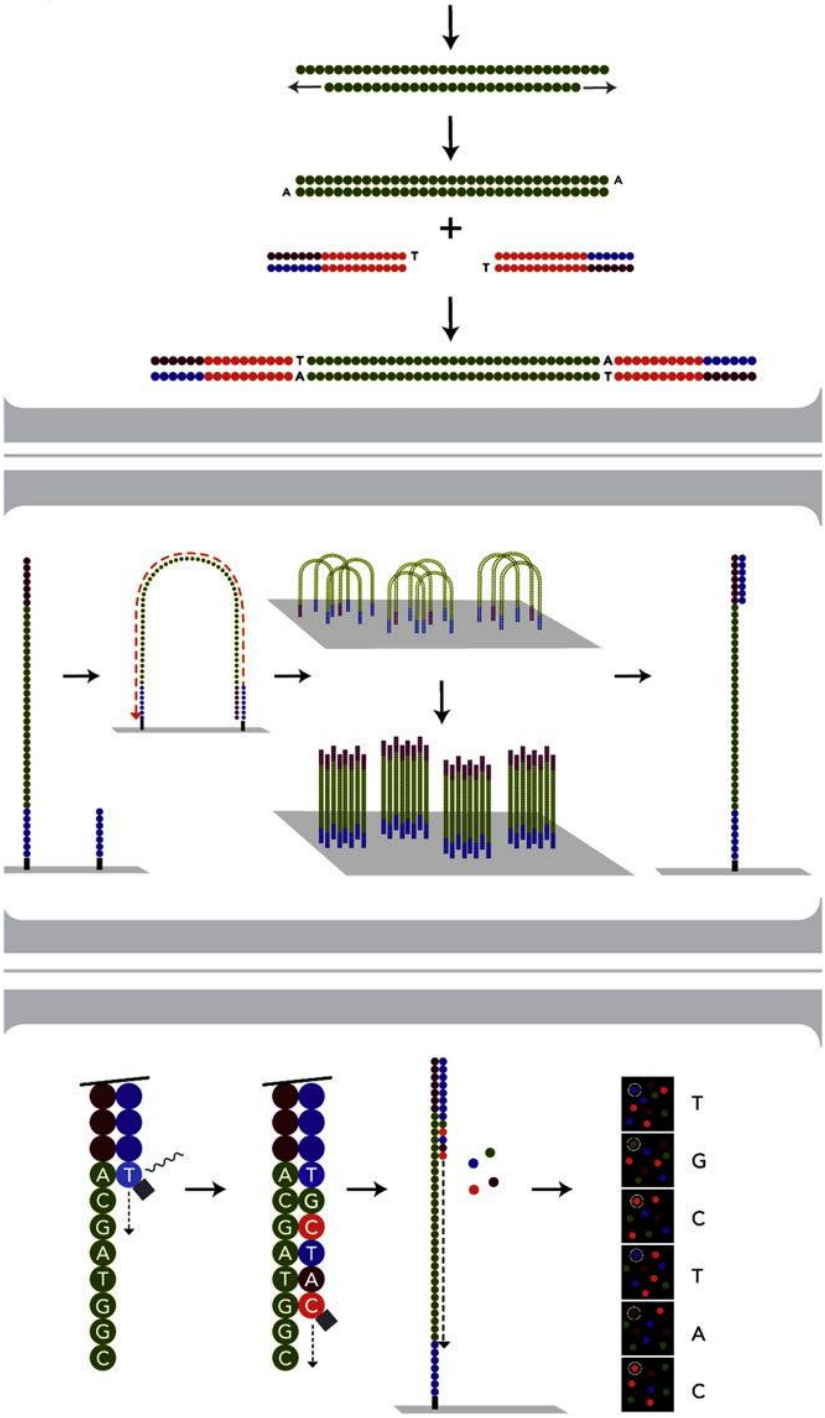

Figure 1.5. Work-flow for Sequencing by Synthesis (SBS) by Illumina

sequencing. Hence, this sequencing process is also termed as "Sequencing by Synthesis" (SBS).[27]

Life Technology's Ion Torrent benchtop instrument adopt a simpler strategy from other massively parallel sequencing methods. High throughput methods (Illumina's HiSeq 2500,

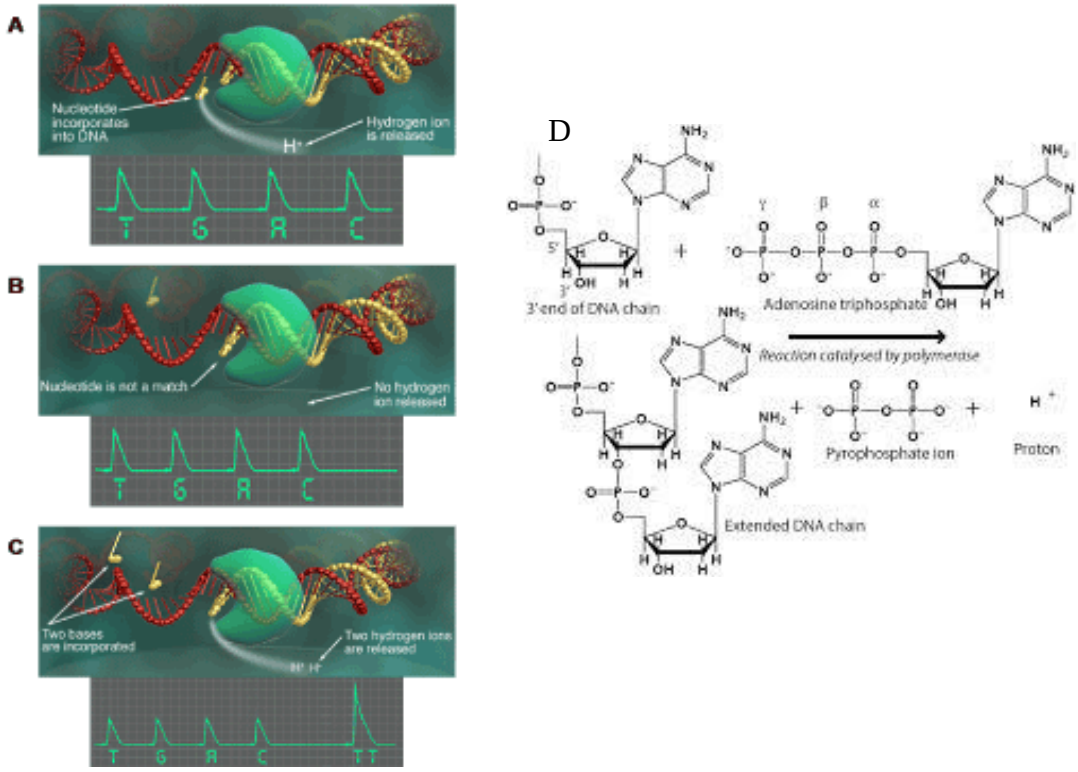

Figure 1.6. (A-C) Ion-Torrent method and (D) related chemistry

Roche 454 GS FLEX, Life technology's SOLiD, Complete Genomics etc.) are the perfect choice for larger and demanding genomic research such as cancer, personal genomics and human genome study. But in the case of small laboratories with smaller genomic projects where lower reagent cost, brief work-flow and shorter runtime are more important than high throughput, Ion Torrent platform is more popular instead of its lower throughput.[22] Ion Torrent uses a semiconductor chip containing millions of wells. The process begins with DNA fragmentation and denaturation (200-400bps long) followed by attachment of each fragment to a bead and performing clonal amplification by emulsion PCR inside micro-droplets. Each amplified bead then flows into a micro-well of the semiconductor chip. Next, the chip is

flooded with one of the four DNA nucleotides in the presence of polymerase and as soon as a nucleotide binds with the ssDNA fragment, a H^+ is released resulting change in the pH of the micro-well.[28] This pH change is converted to readable voltage change and indicates nucleotide incorporation. Hence, we have the base-call. A double change in voltage represents two same nucleotide incorporation, consecutively and no change in voltage designate a deletion site (figure 1.6).[22, 28]

1.9.3 Drawbacks of NGS and Drive for Single Molecular Sequencing

Even with all the success of NGS, it possesses an innate problem during sequencing DNA samples with long repetitive segments, a familiar characteristic of long genomes (such as the human genome) along with various bacterial genomes.[29, 30] The inability of long reads (NGS can provide reads only a few hundreds bp long), causing high error rates creating a roadblock for *de novo* genome sequencing[31] and failed to meet the four gold standards set by the National Human Genome Research Institute (NHGRI, 2004): 1) high accuracy (1 error in 10,000 bases or less), 2) very long reads, 3) high throughput and 4) cost as low as $1000 per genome. After all those techniques, nanopore-based sequencing techniques have emerged as the most potent candidates to reach the goal. The possibility of very long reads (10kb already achieved by ONT), fast and direct sequencing from freshly obtained data, the absence of complicated sample preparation and presence of sophisticated and cheap semiconductor and microfluidics device processing makes it a star contender to cover all four gold standards.

1.10 Example of Single Molecular Methods

Pacific Biosciences launched their Single Molecule Real Time (SMRT) sequencing platform on 2011. Different color-coded fluorescence tagged nucleotides are used which are able to

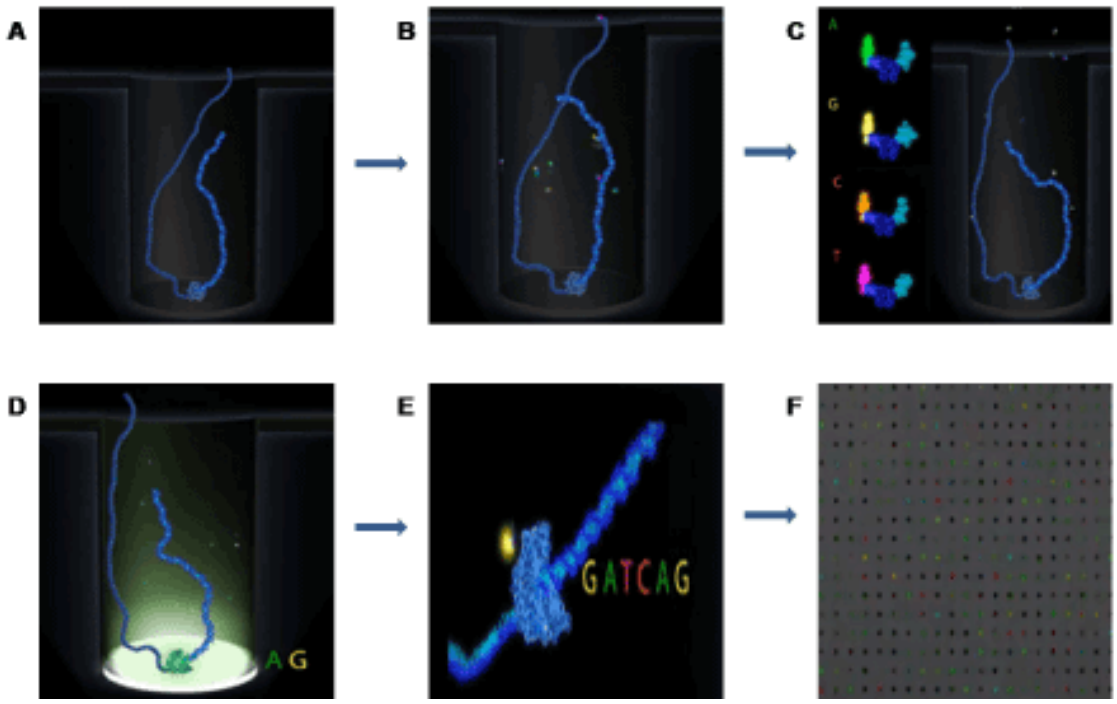

Figure 1.7. (A-F) SMRT sequencing work-flow

cleave the fluorescence tags during replication of the target ssDNA in the presence of polymerase as the tags are attached to the phosphate chains rather than to the base, unlike other fluorescence-based sequencing. When a nucleotide incorporates a fluorescence event occurs which is detected by the ZMWs (zero-mode waveguides) as shown in figure 1.7.[32] These are nanophotonic visualization cylinder with approximately a diameter of 70nm and 100nm depth.[32] As a nucleotide is incorporated by polymerase its fluorescence tag gets illuminated for several milliseconds and the detector receives a signal burst corresponding to the nucleotide. This process repeats and different nucleotides give their corresponding fluorescence signal bursts and help to get the sequence.

According to a report from Macquarie in 2013, Illumina is by far the leading company in $1.1 billion sequencing market, with a stake of 64%. Life Technologies hold the second spot by capturing 22% of the market, while Pacific Biosciences, Roche and Complete Genomics

are mostly competing for the remaining 14%.[22] Following table summarizes various sequencing specifications (such as sequencing chemistry, read length, etc.), advantages and disadvantages of the five most popular sequencing platforms.[22]

1.11 Nanopore Sequencing

Over the past decade, DNA sequencing technology has rapidly transitioned from Sanger sequencing to next-generation sequencing (NGS). NGS has become an indispensable tool for genomic medicine, making great strides in the diagnosis of diseases in clinics.[33] Since its advent, NGS has reduced the sequencing cost from about US $10 million to thousands of dollars.[34] With a state-of-the-art NGS machine, an individual human genome can be finished in a few days. Compared to Sanger sequencing (> 800 Q20 read length),[35] however, NGS has shorter read length (~ 150 bases for single end, www.illumina.com; or 200 bases, www.lifetechnologies.com) and lower raw sequencing accuracy.[36, 37] These shortcomings present challenges for use of NGS. First, NGS requires much higher sequencing coverage than the Sanger method for de novo assembly of genomes with comparable quality.[38] It generates sequencing data at a rate of 100 gigabases (Gb) per single genome for moderate coverage (~30-fold). The deluge of sequencing data requires a computing cluster or supercomputer for their analysis.[39] Secondly, short reads couldn't encapsulate long blocks of repetitive sequences, resulting in fragmented assemblies for repetitive sequences longer than the read length. Given the fact that nearly half of the human genome is filled with repeats (ranging in size from 1 - 2 bases to millions of bases),[30] a straightforward way to address the repeat issue is increasing the read length to span as many of these repeats as possible. Nonetheless, the ultimate solution is to have a method that can faithfully and continuously read the entire sequence of a chromosome from one end to another.

Current progress in nanopore sequencing has opened a new avenue to develop the sequencing technologies. A nanopore is an orifice with a nanometer diameter, which can function as a fluidic channel to conduct ions under a voltage bias. When it is embedded in a thin membrane that separates two chambers filled with conductive electrolytes, DNA molecules can electrophoretically translocate through the nanopore. Subsequently, the ionic currents would transiently be reduced because the flow of ions is blocked by DNA. [40] This is a mechanism used by a commercial product MinION for sequencing DNA by protein nanopores (www.nanoporetech.com). Since there is no theoretical limit on the length of the DNA translocation, the nanopore DNA sequencing will have the potential to solve the assembly issues related to the short reads of NGS, providing a high speed and low cost process of sequencing. However, the protein nanopore sequencing suffers from low accuracy (85%).[41] Gundlach and coworkers have demonstrated that the current blockade in a protein nanopore (Mycobacterium smegmatis porin A, referred to as MspA) is a collected event of four nucleotides (quadromer), [42] and the 256 possible quadromers produce a significant number of redundant current levels. [43] Despite many years of efforts, the nanopore has not achieved a single base resolution in DNA sequencing. Branton *et al* pointed out that "even an infinitely short channel would not achieve the required resolution" and alternative readout methods are required for the nanopore DNA sequencing.[44]

1.12 Molecular Science & Scanning Probe Techniques

After their invention in 1980's, Scanning Probe Microscopy (SPM) techniques, namely Scanning Tunneling Microscope (STM) and Atomic Force Microscopy (AFM) became exceedingly popular for imaging surfaces and understanding their topography. Compared to traditional microscopy and spectroscopy techniques, which rely on the interaction between

the sample and electron or electromagnetic radiation, these SPM methods exploit the interaction between a specific probe and the sample. Hence, they have added the benefit of gaining detailed topographic information as the probes are atomically precise and capable of interacting at the atomic level. The ability to maintain a non-contact profile during the measurement has provided these techniques the advantage of non-destructive approach with respect to the sample. Involved working principles are different from each other as STM relies on electron tunneling between probe and sample (figure 1.8[45]), whereas AFM exploits other atomic forces (such as van der Waals force) to bend its cantilever when probe interacts with the surface. As a consequence, unlike AFM, STM is limited to conducting surfaces only. But, the exponential nature of tunneling signal makes it superior in resolution. Apart from imaging and topography, analytical techniques like Scanning Tunneling Spectroscopy (STS), STM Controllable Break Junction etc. turned STM into a pathbreaker in molecular electronics research. STM has the unique feature to maintain a constant and controllable gap between probe and surface, makes it capable of acting as a biosensor. Measuring molecular conductance and exploring biomolecular interactions are routine experiments nowadays, which were mere theoretical possibilities a couple of decades back.

1.12.1 Molecular Electronics

The field of molecular electronics is devoted to the utilization of organic molecules as components of electronic devices. Though major breakthroughs in molecular electronics have

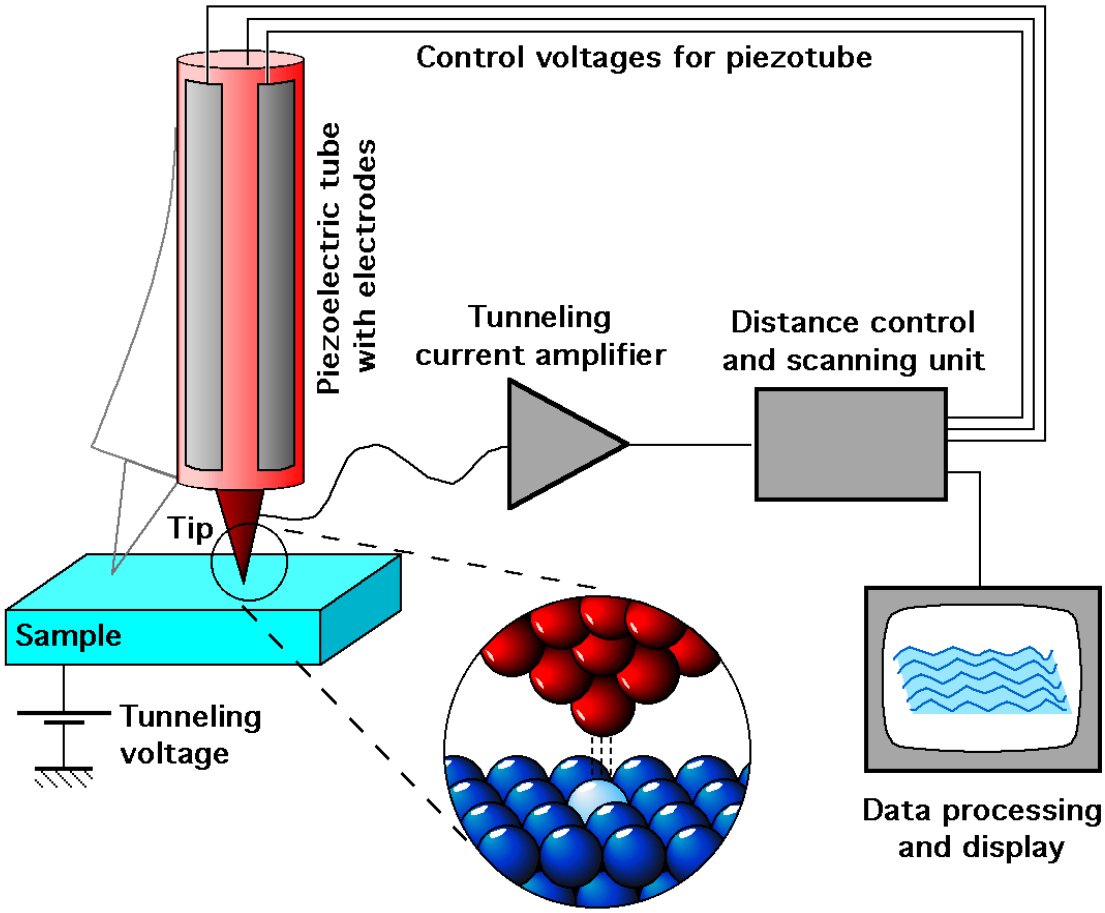

Figure 1.8. Schematic diagram and simplified instrumentation of Scanning Tunneling Microscope

been achieved in the last couple of decades, early footprints can be discovered in 1970's, instead. [46, 47] The original motivation behind molecular electronics is to replace the traditional CMOS (complementary metal oxide semiconductor) materials with organic molecules and overcome the barrier of miniaturization limit of transistors inside a computing chip. From a neutral perspective, still way to go to make it a reality. However, the latest improvements in this area such as fabricating robust, repeatable molecular junctions and reliability of well-developed "non-equilibrium Green's function" theoretical methods paved

the way of understanding the molecular properties in a nanostructure or even as a single species, resulting various new applications of this highly interdisciplinary research area.[48] Fields like molecular sensor, solar energy and thermoelectric have gained significant

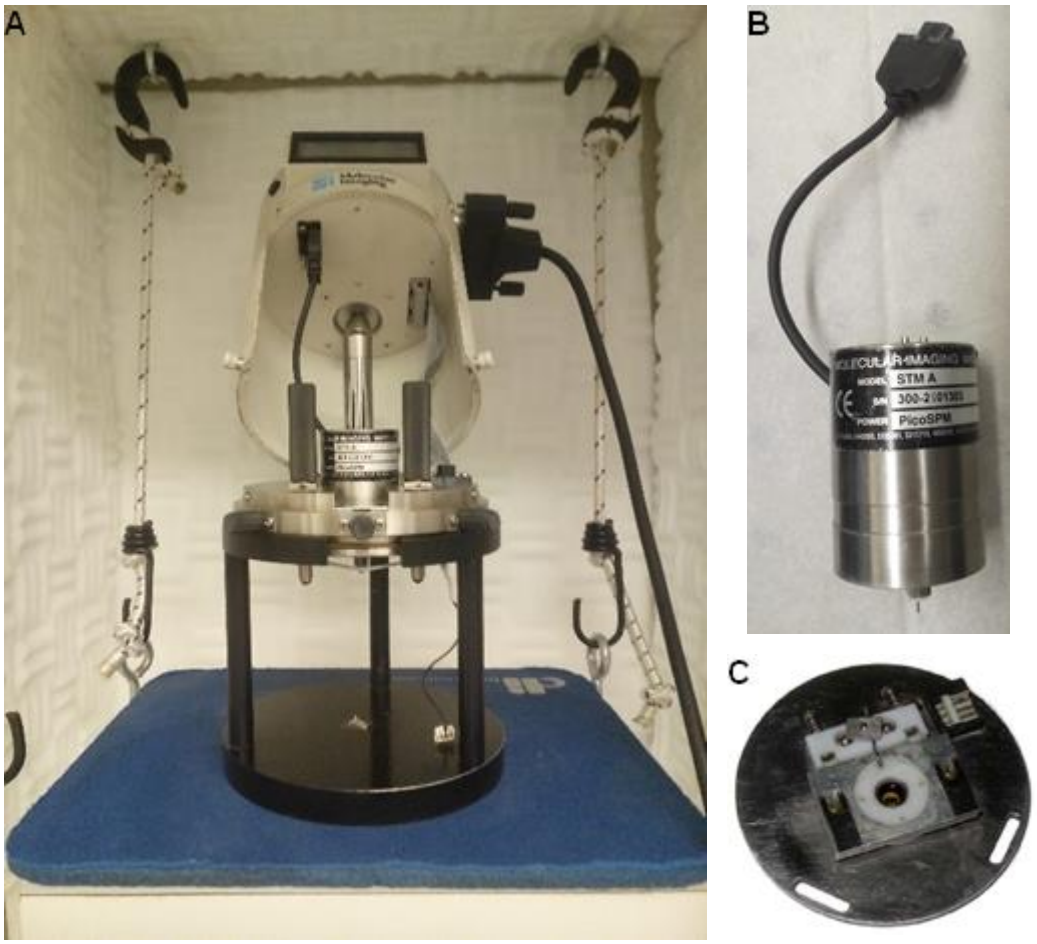

Figure 1.9. (A) STM scanner installed in STM head kept over spring stage, (B) STM scanner with connection cable and (C) STM fluid sample plate

acceleration, consequently. Required smaller size of molecules (generally, 1-100 nm) and availability of convenient synthesis methods are prime advantages of molecular electronics. This method is an excellent tool as molecules can be easily designed and engineered according to its role and desired specifications in the molecular device. Intermolecular interactions

between molecules also form stable nanostructures which is another important aspect of molecular electronics.[49]

1.12.2 Instrumentation of Scanning Tunneling Microscope

The most important parts of the STM instrumentation are the STM head (figure 1.9.A), scanner (figure 1.9.B) and fluid sample-cell (figure 1.9.C). STM controller and computer are other two important integral parts of the STM (not shown), which regulate the experiment, control the servo and record data. The scanner has a small holder in front for installing the probe and a connection cable for connecting itself with the STM head. The sample plate is connected and placed under the STM head such that the substrate is very close to the probe (~200-300 μM) and this z coordinate can be adjusted by a pair of moving magnetic screws. The scanner has piezoelectric elements that manage the movement of the probe in x, y and z directions. These movements can be controlled by preamplifiers and regulated by computer control.

1.12.3 Physical Background of Scanning Tunneling Microscope: Quantum Tunneling

Scanning Tunneling Microscope (STM) is a revolutionary invention of Gerd Binnig and Heinrich Rohrer at IBM-Zurich in 1981, which won them the Noble prize in 1986.[50] STM is based on a quantum mechanical principle called "Electron-tunneling". If a microscopic particle faces an energy barrier which has higher energy than the energy of the particle itself, then the particle will always have a finite probability to tunnel through that energy barrier. This quantum mechanical phenomenon is known as "Tunneling". In STM, tunneling of the electron is considered. Now, according to Schrodinger equation, the wavefunction of an electron to tunnel through a potential energy barrier (of energy P) is,

$$\psi(x) = Aexp\left\{-\sqrt{\frac{2m(P-E)}{\hbar^2}}x\right\}$$

where, E is the energy of the electron. (P-E) is actual barrier height for the tunneling electrons and specifically for STM it is equal to the average work-function (φ) of the metal tip and the metal substrate. m is the mass of an electron.

The probability of the electron to tunnel can be expressed as,

$$\psi(x)^2 = A^2exp\left\{-2\sqrt{\frac{2m(P-E)}{\hbar^2}}x\right\}$$

1.13 Conductance in Metal-Molecule Nanostructure

For macroscopic materials, the conductance can be expressed as,

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

So, the conductance of a material is proportional to its cross-sectional area (A) and inversely proportional to its length (L). σ is the proportionality constant, known as specific conductivity and have characteristic value for specific material.

However, for the molecular level or nano-scale conductance measurements, the previous relationship cannot be considered because of the significant difference in transport property of nano-scale systems when compared to macroscopic systems. Transport in nano-scale metal-molecule system is assumed to be an elastic tunneling process and corresponding tunneling conductance can be described by Landauer's formula. At finite bias and a series of quantum modes, the conductance will be as follows,

$$G = \frac{2e^2}{\hbar}\sum_{ij} |T_{ij}|^2$$

Where, Tij is the probability for the electron to transmit from *i*th mode of one electrode (*i.e.* tip or substrate) to the *j*th mode of the other electrode. When there is only a single atomic contact between the two electrodes (*i.e.* tip and substrate), then $\sum T_{ij} = 1$ and conductance will be equal to $\frac{2e^2}{\hbar}$, which is known as the quantum of conductance and has a value of 77.5 μS.

The tunnel current in the tunnel gap between tip and substrate is,

$$I = I_0 \exp(-\beta d)$$

or $I = G_0 V \exp(-\beta d)$

where, d is the distance between tip and substrate, I_0 is highest obtained current on atomic point contact and G_0 is the corresponding conductance. V is the applied bias between tip and substrate. Hence, the tunnel current is proportional to applied bias and decays exponentially with increasing tunnel gap distance, having β as the decay constant. β can be expressed as,

$$\beta = 2\sqrt{\frac{2m\varphi}{\hbar^2}}$$

where, φ is the energy barrier for tunneling or commonly known as the work function of the metal electrodes.

Now, if a molecular system is present in the gap, a set of molecular orbitals are generated in the gap. In such condition, Fermi level of metal electrode resides between lowest unoccupied molecular orbital (LUMO) and highest occupied molecular orbital (HOMO).

Hence, effective barrier height reduces in the presence of such molecular system in the gap. So, decay constant is modified to,

$$\beta = 2\sqrt{\frac{2m(E_{MO} - E)}{\hbar^2}}$$

As a consequence of a reduction in β value, the conductivity in the gap will increase.

1.14 Different Transport Regimes

Electron transport in metal-molecule-metal junction can occur through two different mechanisms. They are known as i) Tunneling and ii) Hopping. Dependence of electron transport on several factors such as temperature and length of the molecular system are different for these two transport regimes. [51]

1.14.1 Hopping

In longer molecular wires (≥4nm) [51], hopping is the predominant mechanism for transport. In hopping transport mechanism, charge carriers move from one electrode to the other by a series of transfer between adjacent appropriate sites within the molecular system (figure 1.10).[52] Hence, unlike tunneling, it is a multistep process and often known as multistep hopping. Hopping is a thermally activated transport process. This thermal activation process is Arrhenius type, having an activation energy (E_a) for the transport process. As a consequence, hopping has a strong dependence on temperature. Thermally activated motions of nuclei, such as bond vibration, provide favorable geometry to the molecular system to originate electronic coupling and charge transfer. Unlike tunneling, conductance through hopping transport shows proportional dependence with the inverse of length (L^{-1}) of the molecular system.[52, 53] Hopping conductance can be expressed as,

$$G \propto \frac{1}{L}\exp(-\frac{\mathrm{Ea}}{\mathrm{k}T})$$

where k is Boltzmann constant and T is absolute temperature.

1.14.2 Tunneling

On the other hand, tunneling is the transmission of an electron through a potential barrier of certain height and thickness. In tunneling transport, the particle always has some finite

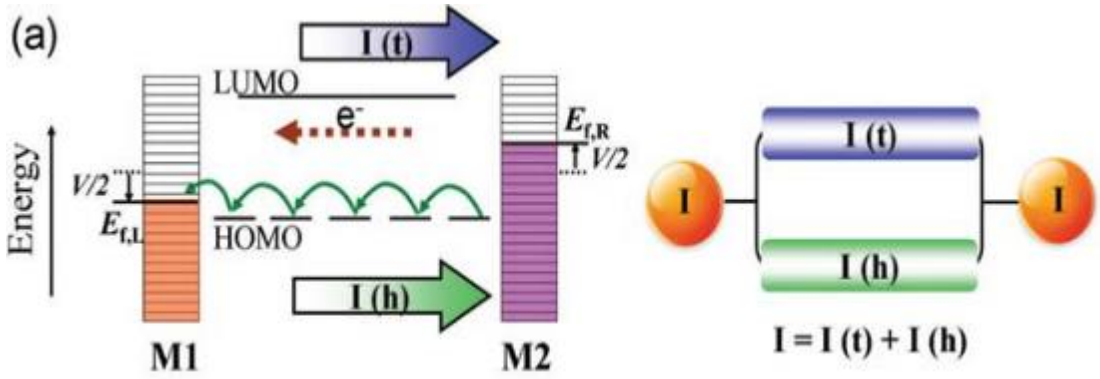

Figure 1.10. Mechanism of tunneling and hopping.

probability to transmit to the other side of the barrier. In contrast to multistep hopping, tunneling is a single step transport mechanism. Charge transport occurs through tunneling in the case of smaller molecular system (≤ 4-5nm)[51, 52] and smaller tunnel gap. Experimental studies have shown temperature invariant nature of this process.[51] Strong exponential dependence of conductance on the length of the molecular system (L) is a characteristic of tunneling transport. From the expression of tunnel conductance, exponential decrease with L can be shown as,

$$G = G_0\exp(-\beta L)$$

where β is known as the decay constant and G_0 is the quantum of conductance.

1.15 Recognition Tunneling

Ohshiro and Umezawa opened the door for a new technique for molecular recognition utilizing hydrogen bonding.[54] A gold probe, modified with a thiolated-DNA base was scanned separately over monolayers of different analytes, attached to a gold substrate. In the case of complementary base pairing between thiolated-DNA and the analyte, greater charge transfer is resulted due to stronger "mechanical adhesion" due to hydrogen bonding. This can be explained as the longer stay of the probe over the analyte which enhances the conductance. Inspired by this study, Lindsay group investigated and proved that specific DNA-base pairing can be identified by measuring the decaying tunnel current while a DNA-base modified STM tip is withdrawn from a gold substrate, functionalized by a SAM of nucleoside.[55] The stiffness of such DNA-base pairing and strength of these hydrogen bonding complexes were measured from similar tunnel junction experiments in their subsequent study. [56] Moving one step further, Lindsay group designed another decaying tunnel current experiment, where the STM tunnel junction was constituted with a DNA-base modified probe and metal substrate functionalized with a Guanidinium SAM. Then short DNA oligomers were absorbed over Guanidinium SAM by means of hydrogen bonding between Guanidinium and phosphate groups of DNA oligomers (fig 1.11)[57]. Though sensitivity of these decay current experiments were far from single base resolution (approximately 20 bases reside in the tunnel gap) yet the observed sensitivity of specific base pairing promised a highly potential sensor device for recognizing bio-analytes exploiting hydrogen-bond mediated tunneling current. Another different approach was invented by Lindsay group[58] for DNA-base pair recognition. Instead of measuring decaying tunnel current as a function of the probe and substrate separation, a fixed distance was imposed between the two electrodes and varying

tunnel current data was recorded with time. Random change in the current level was obtained as the molecular pair on the electrodes continuously bonded and released from inter-molecular hydrogen bonding interactions. As a consequence, time-current data traces appeared as "telegraph noise". This molecular junction configuration was termed as "tethered molecular pair".

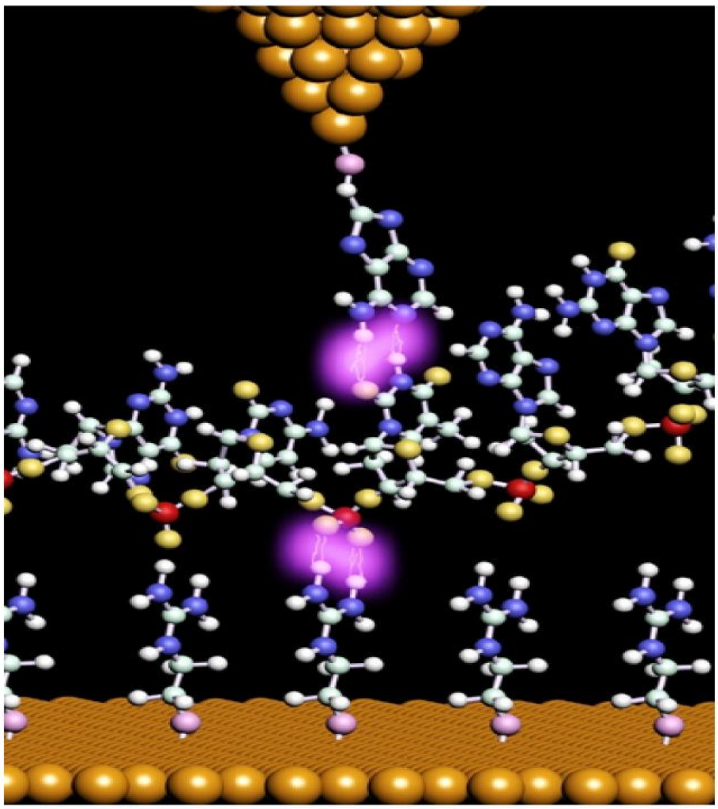

Figure 1.11. Decaying tunnel-current experiments using STM tunnel junction with three molecular members.

In the very next year (2010), another study by Chang et al.[59] adopted somewhat different strategy and introduced free DNA nucleosides inside a tunnel junction, where both, STM

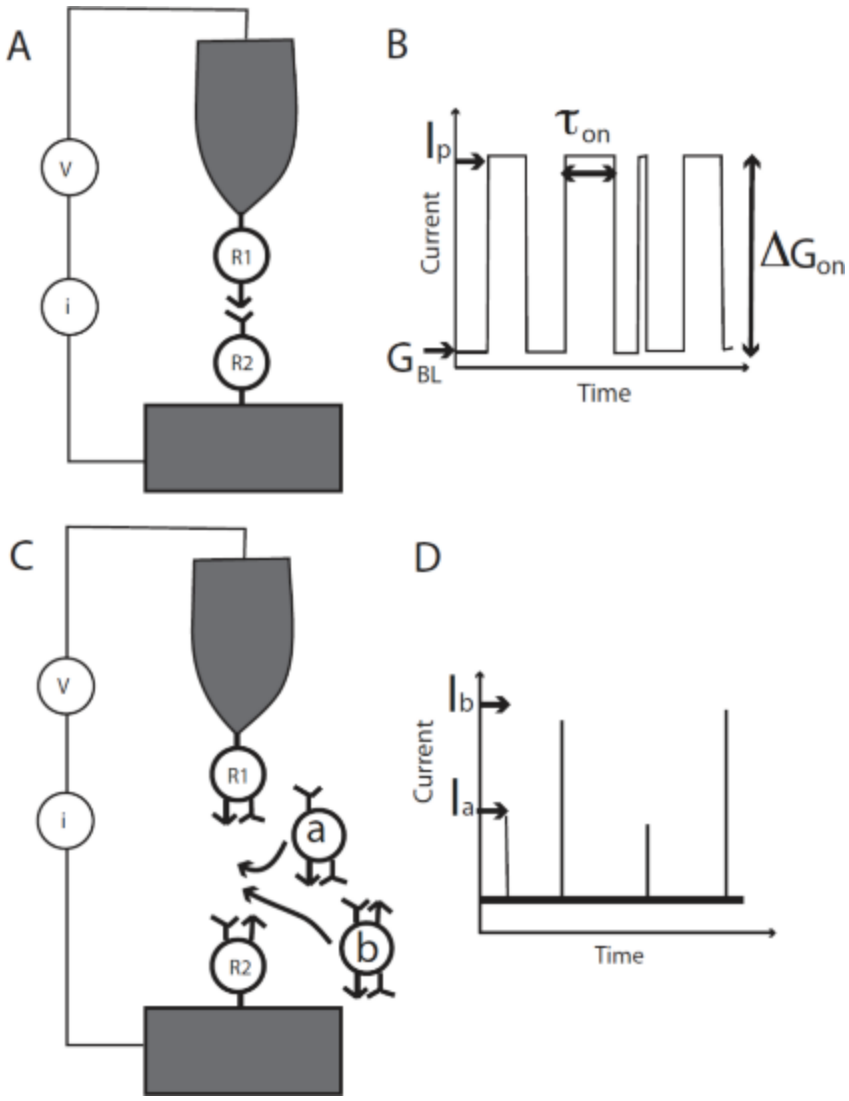

Figure 1.12. Different categories of Recognition Tunneling and characteristic "telegraph noise"

probe and substrate are functionalized with a universal "recognition molecule" or "reader molecule". This junction set-up was defined as "free analyte configuration". Then current-time traces were recorded instead of measuring decay current. This technique provided the advantage of using a "universal reader molecule" for all analytes unlike requirement of analyte-specific functionalization of the probe or the substrate. A clean, spikeless tunnel current baseline was obtained in the absence of any analyte inside the junction, while injection of DNA

base solution generates a train of stochastic tunnel current spikes. Henceforth, Recognition Tunneling could be categorized into a couple of classes, namely "tethered molecular pair" and "free analyte" as described above. Figure 1.12[60] shows the simplified models for these two experimental configurations. In both the cases, a constant baseline conductance (G_{BL}) is maintained to attain an approximately fixed gap between the functionalized probe and substrate. The current level increases and corresponds to "on conductance" (G_{ON}) as soon as the molecular bridge forms. It is quite obvious that an optimized gap between the probe and substrate is one of the important prerequisites of these experiments if not the most. Chang et al.[59] reported a gap distance of approximately 2.5 nm (fig 1.12) for their tunnel junction with benzamide reader modified gold electrodes. Though it is intrinsically hard to maintain perfectly fixed gap throughout the experiment, yet sustaining constant baseline conductance serves the purpose, adequately. Among the two, "free analyte" approach gained superior attention due its universal nature i.e. same tunnel junction could be used for different analytes. This characteristic creates a possibility to sequence heteropolymers of biomolecules (ssDNAs, peptides, etc.) if they could be dragged between such tunnel gap, slowly enough to read each single building blocks from their characteristic current spikes. Figure 1.13 describes the physical picture of these molecular junctions. The conductance of a pure metal tunnel junction (without any molecular present) possessing a gap separation L can be approximated as, [60]

$$G \approx G_0 \exp\left(-1.02\sqrt{\phi}L\right) = G_0 \exp(-\beta L)$$

Here ϕ is the work function which is equal to the difference (V – E_F). V is the potential

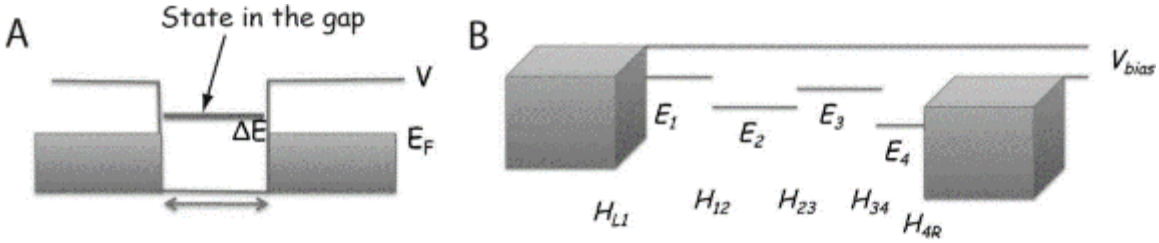

Figure 1.13. (A-B) Energy description of metal-molecule junction involved in Recognition Tunneling

barrier facing by the electrons inside the metal and E_F is the Fermi energy of the metal electrodes. β is known as the electronic decay length and express the decrease in conductance with increasing gap distance. In this situation, there is no available state in between the electrodes to assist electron transport. But, new energy levels (the state in bold line is an example, figure 1.13.A)[60] are generated with the introduction of a molecular bridge spanning the gap. As a consequence, the barrier encountered by the tunneling electron is reduced from ϕ to ΔE and conductance is modified to,

$$G' \approx G_0 \exp\left(-1.02\sqrt{\Delta E}L\right) = G_0 \exp(-\beta' L)$$

β' is the electronic decay length of the bridged metal-molecular junction, which itself can be properly expressed with the help of a linear combination of atomic orbitals coming from atomic components (E_1, E_2, E_3 etc. in figure 1.13.B) of the molecule and hopping matrix elements of neighboring orbitals. In figure 1.13.B, H_{12}, H_{23}, H_{34} etc. describe the interactions among adjacent atomic orbital of the molecule, whereas H_{L1} and H_{4R} define the interactions between left and right metal electrodes and their neighboring atomic contact orbitals. Hence,

the elastic transmission (assuming the non-interacting nature of the electrons) of the metal-molecule junction can be calculated using the Green's function:

$$G_{L,R}(E) \sim \Sigma \frac{\langle L|\psi_n\rangle\langle\psi_n|R\rangle}{E - E_n - i\partial}$$

Here ψ_n is the n^{th} state of the metal-molecule junction. These are basically reformed molecular orbitals including the interaction with the metal electrodes, commonly termed as probe and substrate. E_n is the eigenenergy of the n^{th} state and ∂ is a quantity that tends to zero, whereas L and R states with energy E within the left and right metal electrodes, respectively. Now, the conductance of this metal-molecule junction can be determined by following expression,

$$\mathrm{G} = \frac{2e^2}{\hbar} \int_{-eV/2}^{eV/2} T(E)dE$$

Where T(E) is the transmission function near the metal Fermi level and V is the applied bias voltage between the electrodes.

CHAPTER 2

FROM GOLD ELECTRODE TO PALLADIUM ELECTRODE

2.1 Disadvantages with Gold Tunnel Junctions

In most of the research experiments related to electron tunneling and molecular conductance measurements, gold has been the metal of choice to the scientific community. Whether it is in mechanical break junction experiments[61], in the repetitive formation of molecular junctions[62] or measuring single molecular conductance employing self-assemble molecular junctions[63] , gold always has been the preferred electrode material.

2.1.1 Plastic Deformation

Two crucial properties of gold have made it so popular. Firstly, thiol self-assembly on Au is a very straightforward process with enough understanding of the details[64]. Also, Au has high plastic deformation that leads to specific features in break junction measurements[65]. But, some other studies prefer tunnel junctions with less or ideally no plastic deformation during the experiments.[60] [66] These experiments require fixed gap in the tunnel junction which might be an issue with Au tunnel junctions. Instead, plastic deformation can lead to a geometrically unstable junction.

2.1.2 Diffusion of Gold into Silicon

Gold diffuses reasonably fast into silicon as interstitial impurities[67, 68] creating structural defects in the silicon crystal. These interstitial gold atoms then occupy substitutional sites by any of the two controversially competing pathways: "Frank-Turnbull" mechanism[69] or "kick-off" mechanism.[70] As a consequence, an undesirable electronic defect is generated in

the silicon crystal of the semiconductor device, known as "deep level traps". [71, 72] These defects trap the charge carriers and affects their velocity within the semiconductor. In addition, the non-radiative lifetime of the carriers is shortened by the traps leading to a detrimental effect on semiconductor device quality. But, the diffusivity of palladium into silicon is 20 to 50 (depending on temperature) times lower to that of gold[73]. Hence, compare to gold, palladium is much more compatible with CMOS (complementary metal oxide semiconductor) devices.

2.2 Conductance

A higher level of tunnel conductance is always desirable for molecular electronics study as it provides better signal to noise ratio for all these measurements. In a theoretical finding, Lawson *et al.* [74] showed that molecular junctions consist of a pair of metal electrodes bridged with phenyl-dithiol exhibit higher conductance in case of palladium compare to that of gold. Hence, palladium might be a better choice to obtain larger current signals from recognition tunneling measurements.

2.3 Fabrication of Palladium Substrate

Electron beam evaporated palladium thin films (200 nm thick) were used for all recognition tunneling experiments. All the films were made in ASU CSSER cleanroom using the Lesker PVD75 Electron Beam Evaporator (Lesker# 3). Circular silicon wafers (10 cm diameter) were purchased from Silicon Quest International and 99.99% pure palladium and titanium metal targets were bought from Kurt J. Lesker Company. Prior to use, silicon wafers were cleaned with hydrofluoric acid. Over the silicon wafer, a thin Ti adhesive layer (5nm thick) was deposited at a deposition rate of 0.02 nm/sec. Pd layer was deposited over that Ti adhesive

layer maintaining a deposition rate of 0.1 nm/sec. Small squares of 1 cm × 1 cm were cut before any measurement. STM images show that the surface was not very smooth and small grains are apparent (figure 2.1). [75]

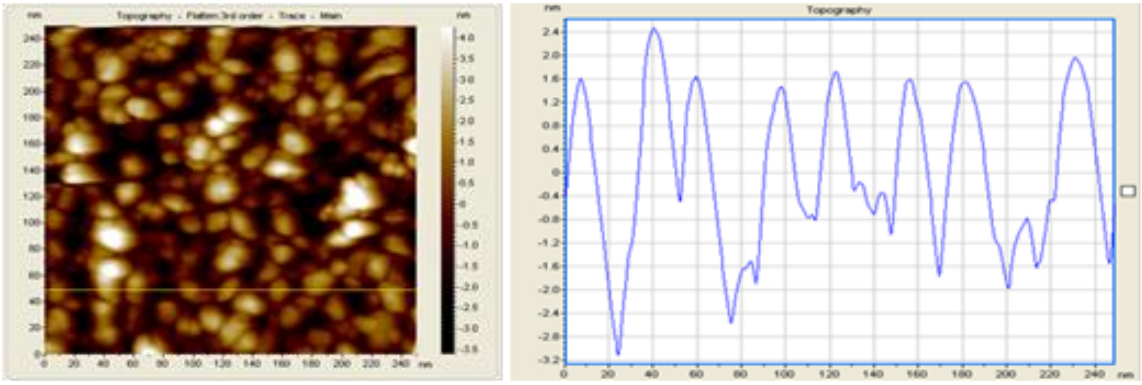

Figure 2.1. STM image of the bare palladium substrate. The Images was taken using a Scanning Tunneling Microscope (Agilent) interconnected with a computer system. Movement of the tip and set up of tunnel parameters (set point and bias) were controlled by PicoSPM software. Several 100nm × 100nm and 250nm × 250nm square areas were scanned for acquiring images of bare palladium substrate. Pt-Ir (90:10) tips were used and scanning was done in air. Scan rate was 3 lines/sec during image acquisition.

2.4 Palladium Probe Preparation

Palladium probes were made from Pd wires with 0.25 mm diameter (from California Fine Wires) after electrochemical etching in con. HCl and Ethanol mixture (1:1) followed by partial insulation with polyethylene. Probes with leakage current > 1 pA were discarded and the rest were modified with reader molecule and used in RT measurements.

2.4.1 Probe Etching

Etching profile is really crucial. Etched tips should not be too long and triangular kind of shape at the top part of the tip is preferable. Also, we need a sharp enough tip for the

experiment. But, etching an extremely sharp tip frequently combines a long profile with the sharpness. Also, there is a chance of over exposure after coating for too sharp tips. On the other hand, if we try to make the profile broader and shorter there is a good chance of having a blunt tip, which may have a problem of a bad experiment along with a chance of over-coating. Sacrificing the sharpness to some extent is preferable to attain a broader and shorter etching profile as problems like bad imaging, over-coating is rare compared to the problem of leakage.

For the 1st step of etching, a value of 35 kHz was used in the signal generator. A voltage of 30 V and current range of 280-350 mA were fixed. The 1st step is to basically obtain a roughly long and sharp shape to start with the 2nd step.

The 2nd step is very subtle and more important. We can achieve the desired shape or profile in this step. Firstly, a couple of dips (fast dipping and taking out) in the solution keeping the voltage at ~20V was done and the probe was checked under the microscope. If the profile is still long another couple of dips may be required. To make it sharper, subtle lowering of voltage (~12-15V) is needed. We do not need to be concerned about the current during the dipping of tips in the 2nd step etching as it is too subtle to control. If the tip is blunt after many trials, we have to start over with the 1st step etching. Hence, it a trial and error process in the 2nd step rather than having a strict protocol to follow.

2.4.2 Probe Coating

The coating is much easier compared to etching. The traveling distance of the tip during coating (i.e. length of the coated portion) was maintained in the range ~7-7.5 mm. Coating temperature is a variable parameter which may change with the humidity or local temperature of the lab and probably with the profile of the top part of the etched tips. A temperature range

of 210-225°C can be used as a can be considered as the starting point. For a well-coated tip, the protrusion should as small as it can be. If there is no apparent protrusion under the microscope, the tip may be over-coated or it still can be a good tip after checking in STM. At the same time, tips with very small obvious exposure under the microscope are really good too. So, while checking under the microscope there is always a boundary of uncertainty.

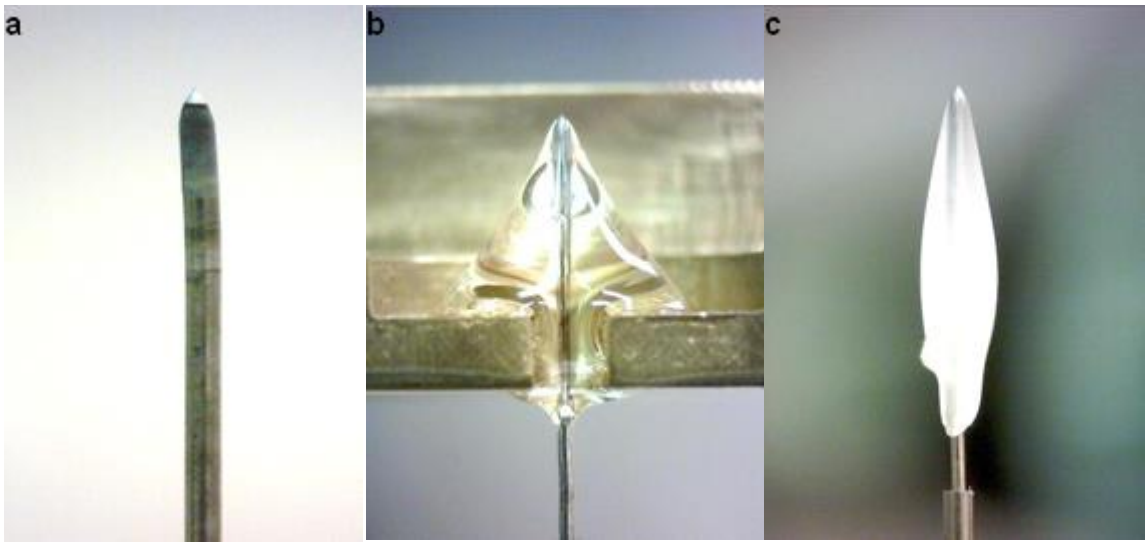

Figure 2.2. Preparation of STM probe (a) after electrochemical etching, (b) during polyethylene insulation and (c) after insulation

Hence, it is preferable to check the leakage for both, tips with no obvious exposure and those with very small protrusion before functionalizing them with the reader molecule. If they have leakage the same tip can be made useable after few more trials with the coating if it appears that the etching profile did not have any problem. Lastly, because this fabrication process (etching and coating) is not totally automated, required adjustments can vary a little bit from person to person.

2.5 Tunneling Transport Through Water Molecules: Water Signals with Gold Electrodes

In their latest work with gold tunnel junction and 4(5)-(2-mercaptoethyl)-1H-imidazole-2-carboxamide as the reader molecule, Chang et al. found that some considerable

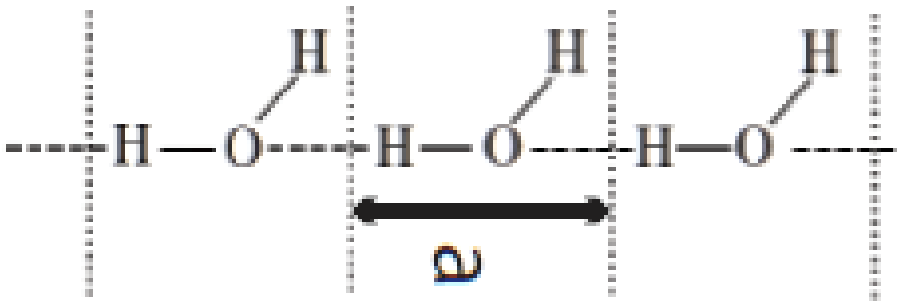

Figure 2.3. Chain structure of water

signals were obtained in absence of any nucleotide at 0.5 V bias and 6pA set-point (corresponding to 12pS tunnel gap conductance) of tunnel junction during control experiments with phosphate buffer only.[76] These current spikes were generated by tunneling through small layer structured water molecules (figure 2.3).[60] Though these water signals, on average, are shorter than the current spikes generated from the DNA bases yet distribution of water spikes shows a considerable overlap with spike distributions of DNA bases.[76] These signals were completely vanished when gap conductance was reduced to 8pS (0.5V bias & 4pA set-point). But, no considerable current signal was generated by nucleotides after their introduction in the tunnel junction, at this tunneling condition (0.5V bias & 4pA set-point).

Electron tunneling through water molecules has been proved by several works on this topic. Tunneling through layer structure of water molecules (figure 2.4) was shown by Zinn and Porter[77] by their tunneling measurement through a thin water layer residing between a pair of spherical mercury electrodes.[77] More studies[78, 79] were able to show evidence of

electron tunneling through a small number of water molecules in tunneling experiments using STM. Boussaad *et al.* found that tunneling current through water medium between a pair of fixed nano-electrodes switches between two discrete levels. According to them, tunneling can occur via local states of water layer and that causes the switching of tunnel current to a higher

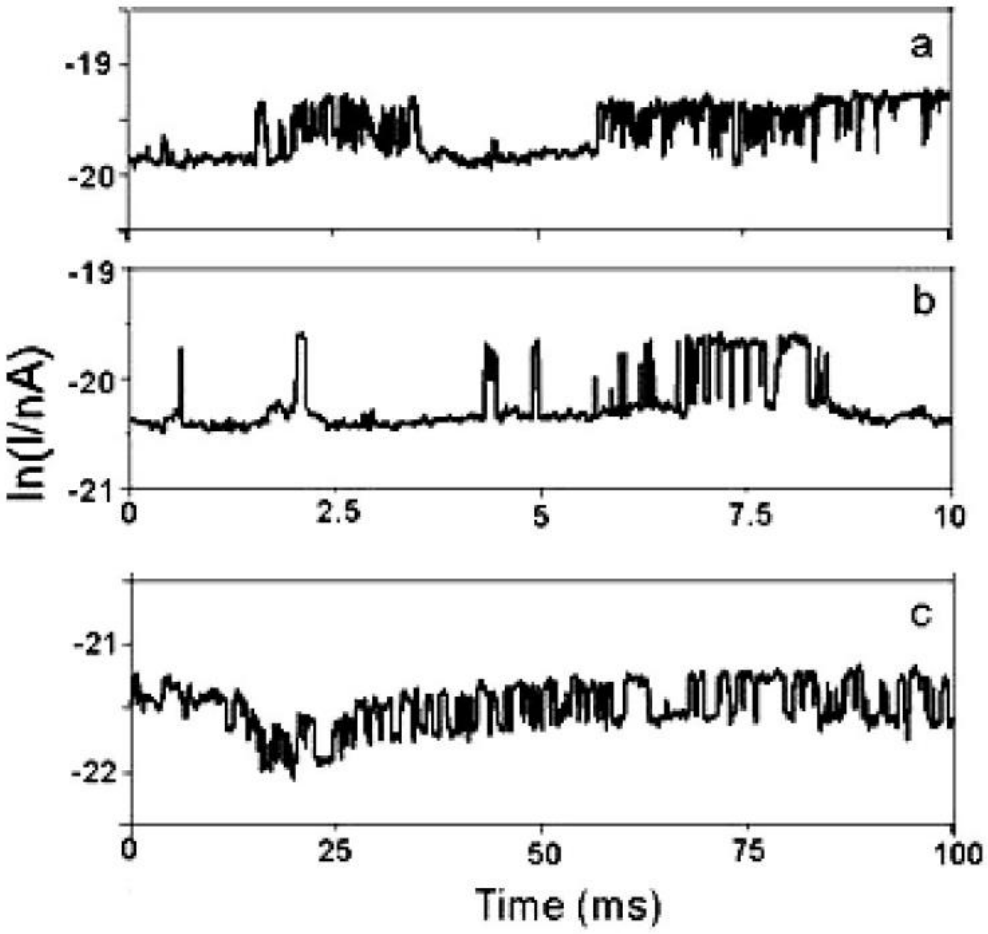

Figure 2.4. Two level switching of tunneling current in a metal-water-metal

level. [62] The assumption is that the local states of layer structured water reside close enough to the Fermi level of metal nano-electrodes and make tunneling through water feasible. Tunneling measurements in chloroform showed no such switching of tunnel current providing strong statement in favor of tunneling through the water.[62] Now, electron

tunneling through a definite layer or chain structure of water molecules is possible if intermolecular motion of water can be neglected on the time scale of fast tunneling process. Then tunneling occurs for several distinct configurations of water layer structure which can be called "frozen configurations".[80] In typical STM experiment, the order of tunneling time is ~ 10^{-16} s. Hence, intermolecular solvent motion and effect of intramolecular solvent dynamics are indeed negligible on the time scale of tunneling process. Consequently, static or frozen water configurations can be assumed during tunneling through water layer structure.[80]

CHAPTER 3

SURFACE CHARACTERIZATION OF SELF-ASSEMBLED MONOLAYERS OF VARIOUS RECOGNITION MOLECULES

All different reader molecules were synthesized in our lab by my project collaborators and before performing any recognition tunneling experiment, palladium substrate and partially insulated probes were needed to be functionalized. It was imperative to optimize the functionalization condition for different readers and characterize the formation of self-assembled monolayer using various surface characterization techniques. I used FTIR (Fourier Transform Infra-Red) spectroscopy, Water contact angle measurement, Thickness measurement by Ellipsometry, XPS (X-ray Photoelectron Spectroscopy) characterization and ARXPS (Angle-resolved X-ray Photoelectron Spectroscopy) thickness measurement. It should be mentioned that all surface characterization experiments were performed only on modified metal substrates and modified probes were not suitable for these measurements due to their minuscule dimensions.

3.1 Formation of Self-Assembled Monolayers

Electron beam evaporated palladium substrates (as described in chapter 2) were cut into small square pieces of 1 cm^2 × 1 cm^2 dimension and used as the substrates for self-assembled monolayers.

3.1.1 Imidazole Reader

0.5 mM ethanolic solution was prepared in properly degassed ethanol. A small piece of palladium substrate was then hydrogen flamed for approximately 30 seconds and immersed in ~2 ml freshly prepared solution of Imidazole reader. After 16-18 hours, palladium substrate

was taken out of the solution and washed thoroughly with plenty of ethanol to ensure that the SAM is devoid of any physical absorption over it. The substrate is then blown dry with argon flow and used for experiments, immediately.

3.1.2 Triazole Reader & Pyrrole Reader

Exactly similar functionalization condition was used to form SAMs of Triazole reader and Pyrrole reader.

3.1.3 Benzimidazole Reader

Following the same protocol did not work well in case of Benzimidazole reader as measured thickness was quite higher than the expected value, suggesting a large extent of physical absorption. The concentration of the solution and deposition time were optimized and 10 μM solution concentration and 10-12 hours deposition period was proved to be fruitful to achieve proper monolayer.

3.1.4 Pyrene Reader

100μM ethanolic solution of reader molecule was prepared and Pyrrolidine was added to the solution for thioacetate deprotection (1:100 molar ratio for Pyrene reader and Pyrrolidine). After a couple of minutes, Palladium probes and substrates were dipped into this free thiol solution and kept for 6-8 hours for modification and washed thoroughly with ethanol followed by nitrogen or argon flow before any experiment.

3.1.5 2-Phenylethane Thiol

A palladium substrate was immersed in an ethanolic (degassed) solution of 2-Phenylethane thiol (50 μM) for 2.5 h, washed thoroughly with ethanol, dried under a nitrogen flow, and used immediately.

3.2 Fourier Transform Infrared Spectroscopy (FTIR)

3.2.1 Introduction

Infrared spectroscopy has been an attractive technique for the material scientists for a long period of time for characterization of organic, inorganic compounds, polymers, self-assembled

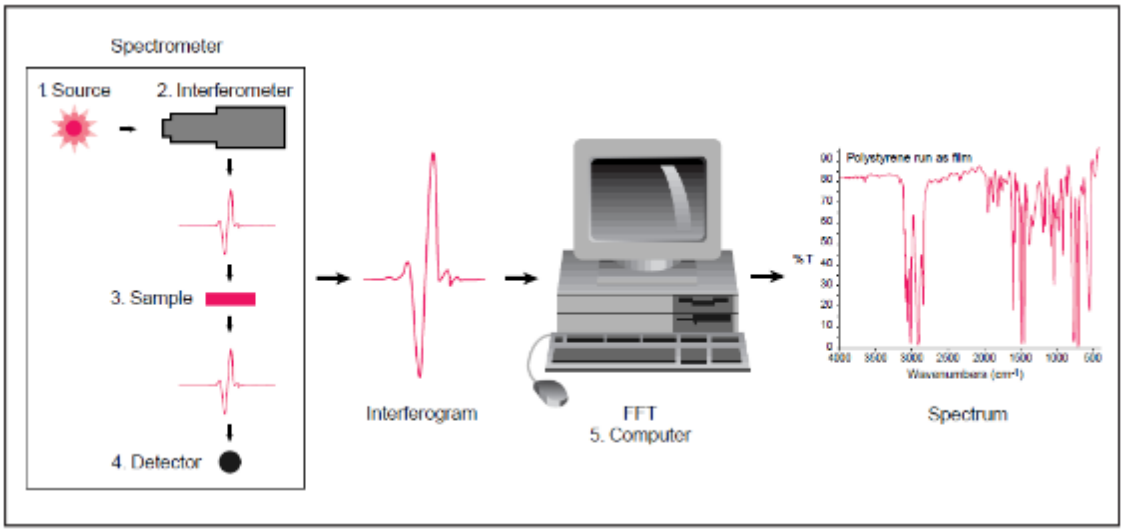

Figure 3.1. Simplified schematic of components of FTIR instrument

monolayers and so on. But, for any sample to be analyzed by this technique has to be IR active, which means the molecule must possess at least one vibrational motion that can change the dipole moment of the molecule in order to absorb IR radiation. These vibrational motions are called IR active modes. When a sample is irradiated by infrared radiation, different frequency of that radiation is absorbed by the sample depending upon the presence of different IR active modes in the sample. These IR active modes represent different structural moieties or functional groups within the sample and a fingerprint of those functional groups can be detected by their IR absorbance.

3.2.2 Working principle & Instrumentation

Originally, dispersive infrared spectroscopy technique was used to irradiate different IR frequency separately and record corresponding IR absorbance, making it a slow process. Fourier Transform Infrared (FTIR) spectroscopy is an ingenious improvement over the ancestor technique as the entire frequency range of the IR radiation can be focused on the sample and resulting absorbance can be recorded, simultaneously. This makes FTIR far more advantageous regarding the speed and sensitivity. Modern FTIR instrument is composed of four vital components. IR radiation source, interferometer, detector and digital analysis setup. [81] IR radiation is emitted from a water-cooled IR source and enters into the interferometer which employs a beamsplitter to divide the original beam into two parts. These two beams again couple with each other at the beamsplitter, after one of them reflects from a fixed mirror and the other beam reflects from another movable mirror that can change its distance (order of few millimeters) from the beamsplitter. As a consequence, an optical path difference is generated among these beams, resulting interference between them. The signal of the recombined beam that comes out of the interferometer after the interference is called an "interferogram". This is represented as a plot of beam intensity (as vertical axis) versus optical path difference or OPD (as horizontal axis) and is a time domain spectrum as OPD depends on scan-time and movable mirror velocity. An interferogram in FTIR is associated with recombined beams from all the frequency components, simultaneously.[81] As this IR beam is channelized through the sample different frequency components are absorbed by the sample and the residual beam is transmitted to the detector. An interferogram is then obtained from the detector response carrying all the spectral information in the time domain. In the digital analysis setup, this interferogram is converted into a frequency domain IR spectra employing a mathematical process called Fourier Transformation. Typical IR spectrum is represented as

wavenumber (reciprocal of wavelength) as the horizontal axis and corresponding absorbance or percentage transmittance as the vertical axis. Transmittance (T) is defined as the ratio of transmitted beam intensity (I) and incident beam intensity (I_0), whereas, Absorbance (A) is expressed as the logarithm of the reciprocal of transmittance.

$$A = \log_{10}(1/T) = -\log_{10} T = -\log_{10}(I/I_0)$$

3.2.3 Experimental

FTIR spectrum of the pure powder sample of reader molecule and SAM over palladium substrate were recorded and compared to test the functionalization. A Nicolet 6700 FT-IR

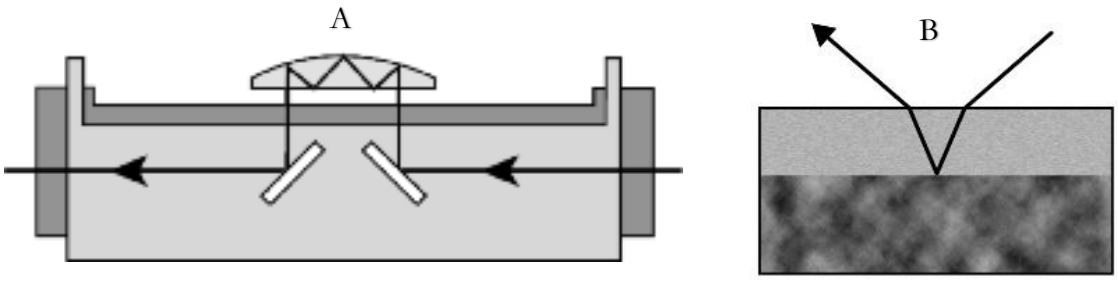

Figure 3.2. IR beam path for (a) single bounce ATR and (b) Specular reflectance

instrument from Thermo Electron Corporation, combined with an MCT detector was used for the purpose. But different accessories were required for pure powder and organic thin film samples due to their obvious difference in physical properties.

Smart Orbit

Smart Orbit, a single reflection diamond ATR (Attenuated Total Reflection) accessory was used to record spectrum for pure powder samples. An ATR crystal is made of highly dense material and has a high value of refractive index. IR beam with an incident angle larger than the critical angle gives total internal reflection and generates an evanescent wave that comes into the contact with the sample (figure 3.2.A)[82]. The sample absorbs energy from the

evanescent wave according to the presence of active vibrational modes present in it and results in attenuation of energy in the wave. This attenuated IR beam is then directed to the detector and interferogram is generated, followed by its conversion to FTIR spectrum.[83] Befor the collection of the powder sample data, a background run was performed each time after cleaning the sample stage properly with methanol. Then approximately 0.5 mg sample was mounted on the sample stage and spectrum was recorded with 128 scans at 4 cm^{-1} resolution.

Smart SAGA

FTIR of SAM samples were obtained using SAGA (Specular Apertured Grazing Angle) accessory which relies on the principle of "specular reflectance". In this technique, a high angle of incidence is maintained (80° with respect to surface normal) to create a long traveling path for the IR beam through the sample, (figure 3.2.B)[82] resulting very high sensitivity.[83] Use of integrated germanium polarizer also boosts the sensitivity as it reduces the S-polarized light which is generally responsible for diminishing sensitivity of spectral data. During spectrum recording, SAM samples were laid upside down over a circular pore with a diameter of 8 mm, which acts as the active specimen area for interaction with the IR beam. A clean bare palladium substrate (from the same batch that we used for SAM preparation) was used to collect background spectrum, prior to measuring any SAM sample. 256 scans and 4 cm^{-1} resolution was maintained during each spectrum collection. OMNIC program was used for smoothing and baseline correction of the recorded spectrum.

3.2.4 Result

Imidazole

The vibration around 3300 cm^{-1} is assigned to N-H stretching, 1690 cm^{-1} to C=O stretching and ~2920 cm^{-1} to aliphatic C-H stretching from methylene group.

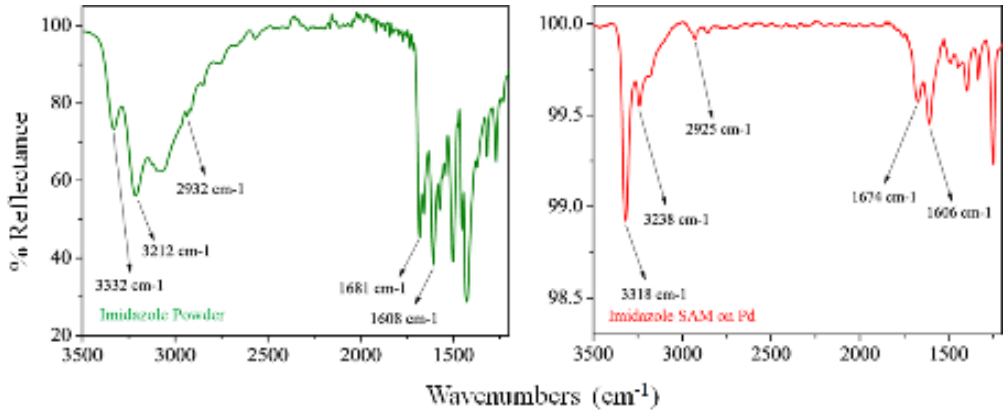

Figure 3.3. FTIR spectrum of Imidazole (a) powder sample and (b) SAM on Pd

Benzimidazole

The vibration around 3300 cm-1 is assigned to N-H stretching and 1690 cm-1 to C=O stretching.

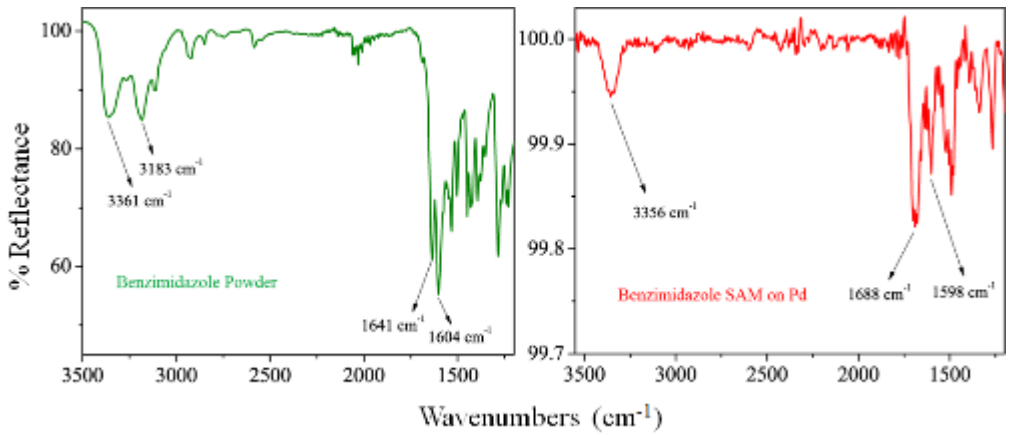

Figure 3.4. FTIR spectrum of Benzimidazole (a) powder sample and (b) SAM on Pd

Triazole

The vibration around 3400 cm-1 is assigned to N-H stretching, 1690 cm-1 C=O stretching and ~2920 cm-1 aliphatic C-H stretching from methylene group.

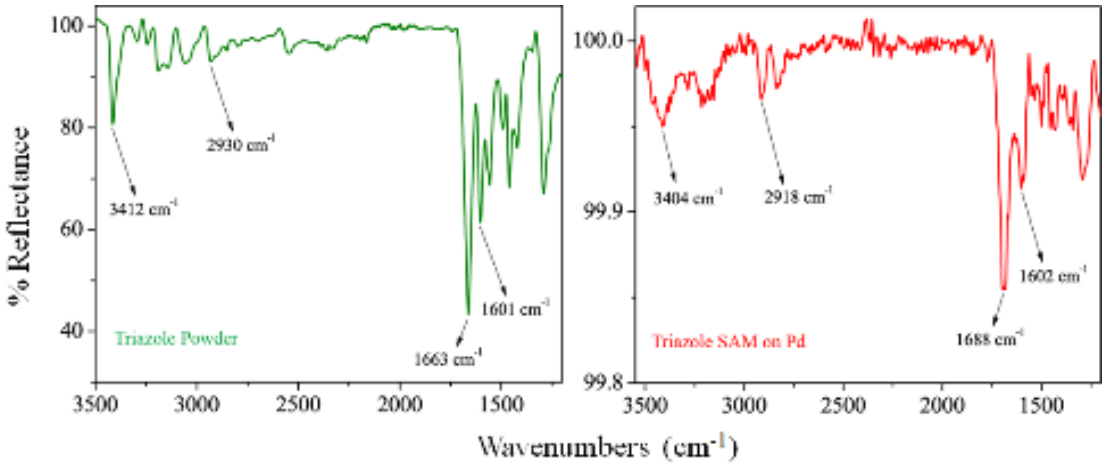

Figure 3.5. FTIR spectrum of Triazole (a) powder sample and (b) SAM on Pd

Pyrrole

The vibration around 1690 cm-1 is assigned to C=O stretching and ~2925 cm-1 to aliphatic C-H stretching from methylene group, and 2560 cm-1 S-H stretching, which is absent in SAM spectrum.

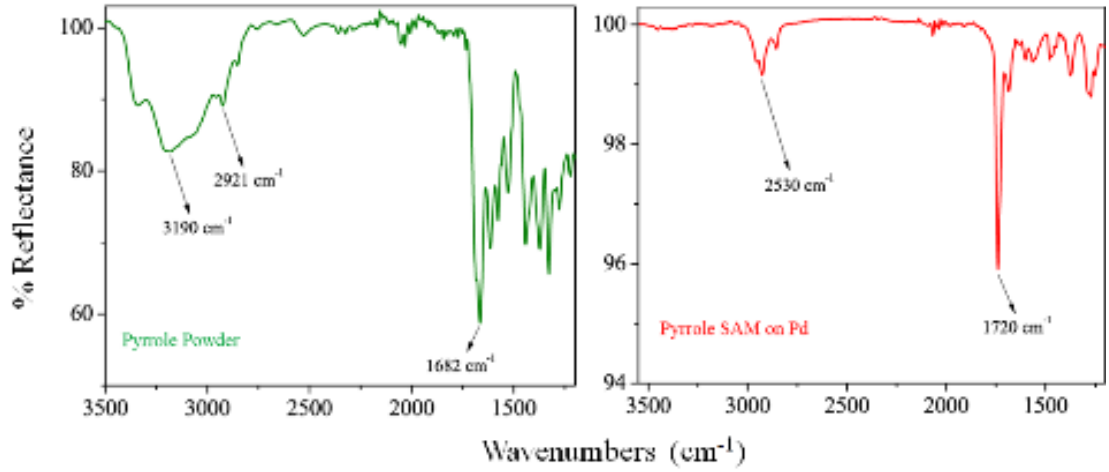

Figure 3.6. FTIR spectrum of Pyrrole (a) powder sample and (b) SAM on Pd

Pyrene

The vibration around 3040 cm-1 is assigned to aromatic C-H stretching, 1600 cm-1 aromatic

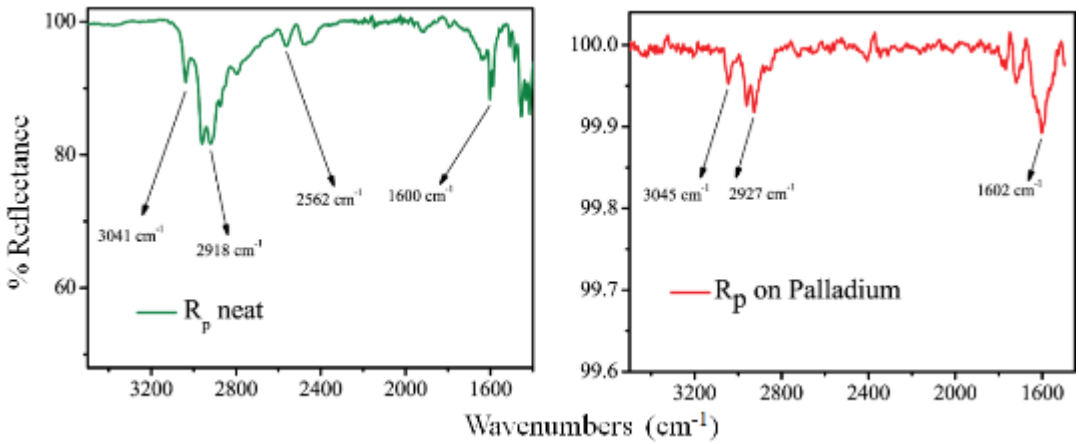

Figure 3.7. FTIR spectrum of Pyrene (a) powder sample and (b) SAM on Pd

C-C stretching, ~2925 cm-1 aliphatic C-H stretching from methylene group, and 2564 cm-1 S-H stretching, which is absent in SAM spectrum.

2-Phenylethane Thiol

The vibration around 3020 cm-1 is assigned to aromatic C-H stretching, ~1600 cm-1

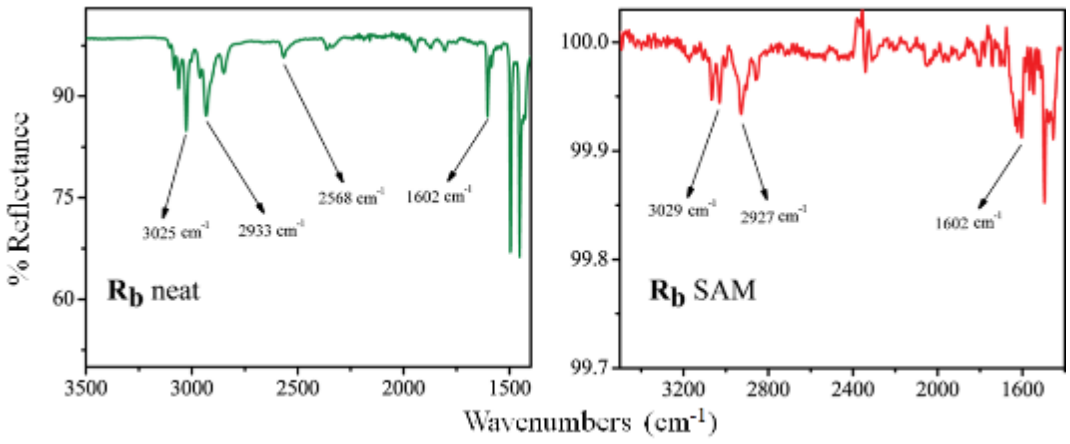

Figure 3.8. FTIR spectrum of 2-Phenylethane thiol (a) powder sample and (b)SAM on Pd

aromatic C-C stretching, ~2930 cm-1 aliphatic C-H stretching from methylene group, and 2568 cm-1 S-H stretching, which is absent in the SAM spectrum.

3.3 Thickness: Ellipsometry

Ellipsometry is a popular technique that measures optical constants and thickness of thin films with thickness ranging from several microns to sub-nanometer. It is a non-contact and

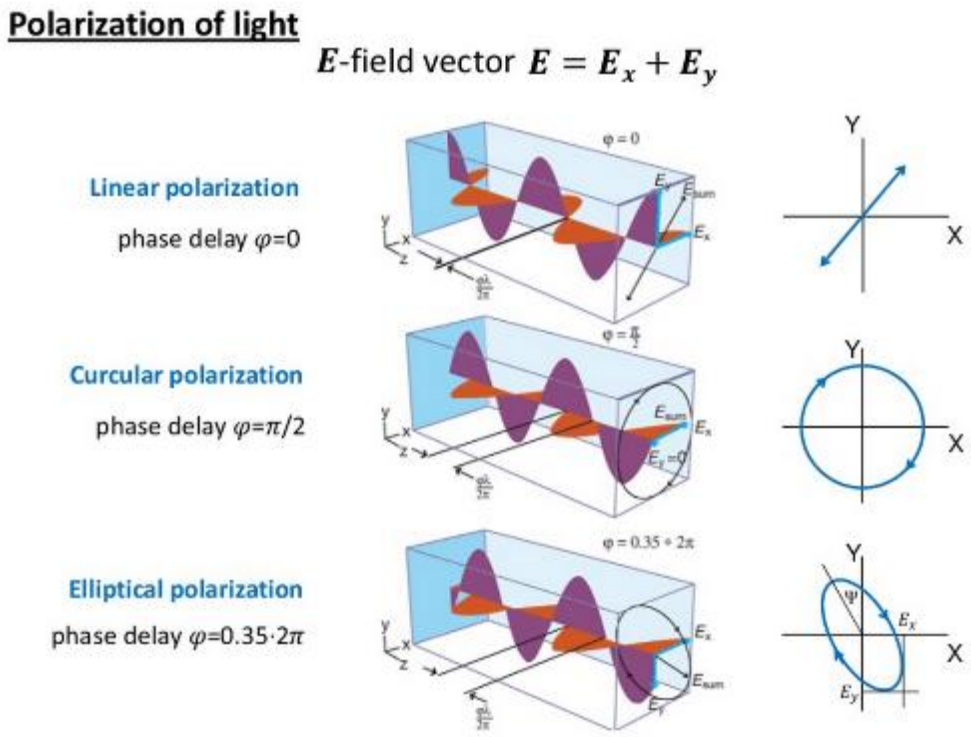

Figure 3.9. Polarization of light.

non-destructive method and as a result, can be applied to various kinds of samples like metals, semiconductor, dielectrics, polymers, organic films and so on. The related working principle relies on detecting the shift of polarization of polarized light after its reflection and/or refraction from the sample. The polarization change is interpreted by a couple of variables known as amplitude ratio (ψ) and phase difference (Δ). Hence, other surface properties such

as roughness, crystallinity, doping concentration, etc. that can affect optical response can also be evaluated by Ellipsometry.

3.3.1 Introduction & Working principle

Light is a combination of electric and magnetic waves fluctuating perpendicular to each other

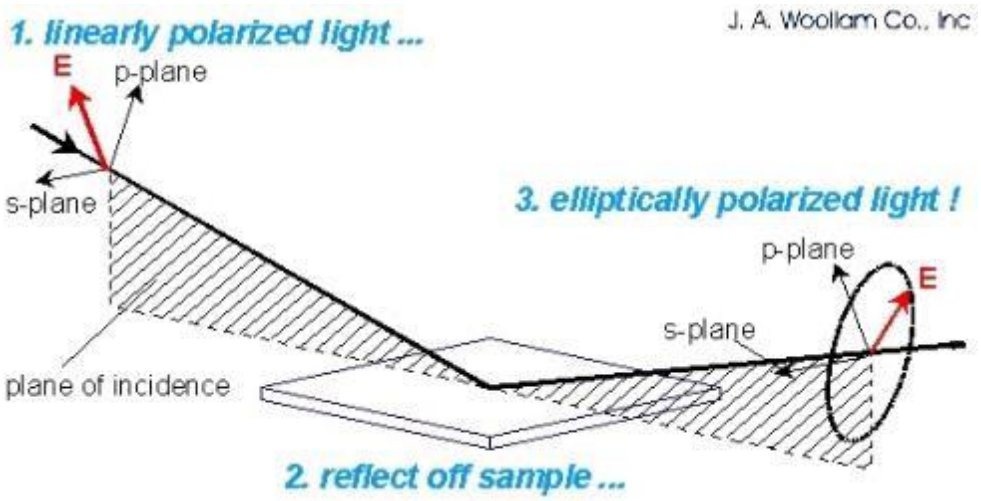

Figure 3.10. Propagation and polarization of light wave during Ellipsometry.

and also perpendicular to the direction of propagation of light. But, the polarization state of light is dictated by the electric field only. If the electric field vector oscillates in a straight line with respect to time on a plane that is perpendicular to the direction of propagation of light, we call it a linearly polarized light. An electric wave can also be added to another electric wave that is perpendicular to the original wave, generating a new resultant wave with new polarization state. When an electric wave is joined in-phase ($n\pi$) by another wave, a linear polarization state is obtained with a shift in orientation. In the case of two equal amplitude waves merging completely out-of-phase ($n\pi/2$) a circular polarization results. But, for two waves with different amplitudes and/or arbitrary phase difference (in between $n\pi$ and $n\pi/2$) we obtain an elliptical polarization. This is the resultant polarization state that is obtained in

Ellipsometer when the linearly polarized incident light is transformed after reflection and transmittance from the sample. Hence, the technique is named as Ellipsometry. The

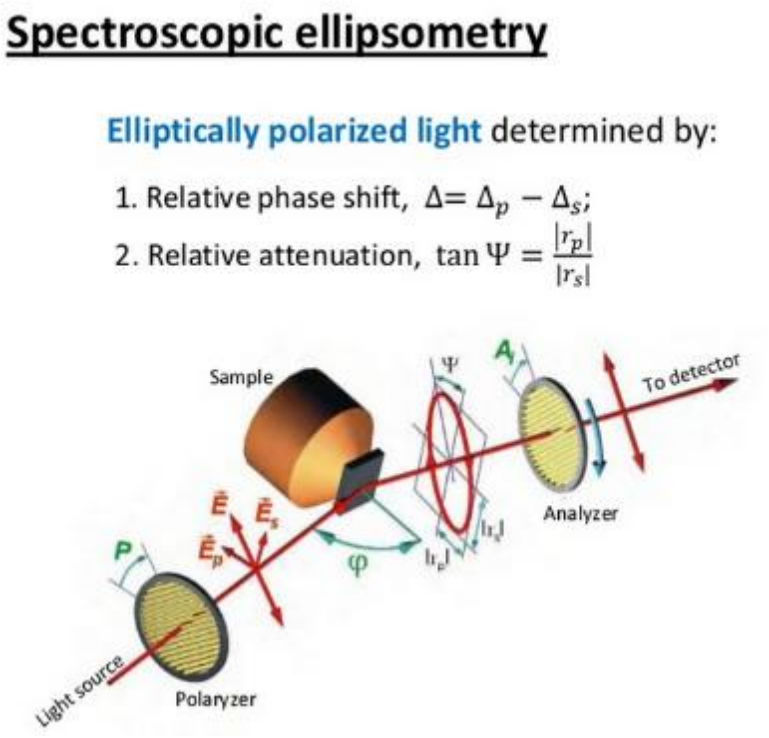

Figure 3.11. Simplified schematic of Spectroscopic Ellisometry.

component of light that oscillates in the plane of incident is called p-polarized light and the component oscillating perpendicular is termed as s-polarized light. When light incident on the sample these two components are in phase and overall polarization state is linear. But, this two components reflect differently from the sample and results elliptical polarization (figure 3.9).[84] This difference is represented as complex reflection ratio.

$$\frac{\mathrm{Rp}}{\mathrm{Rs}} = \tan\psi \exp(i\Delta)$$

$\tan\psi$ represents the ratio of the modulus and Δ dictates the change in phase of p- and s-polarized light (figure 3.10[85] & 3.11[84]).

A light source with a wavelength ranging from UV to IR can be used. Randomly polarized light comes out of the source and pass through a rotating polarizer that allows waves with p- and s-polarized states only. After reflection from the sample and consequent transition to elliptical polarization, the light beam again passes through a rotating analyzer which again separates linearly polarized waves having different orientations. The orientations of linearly polarized waves depend on the instantaneous orientation of the rotating analyzer. The light coming out of the analyzer is then collected by the detector and electrical signal is generated corresponding to amplitude ratio (ψ) and phase difference (Δ). ψ and Δ are functions of wavelength. Hence, they should always be reported along with the specific wavelength of light in use. These measured values of ψ and Δ are then used by an inbuilt model to determine important physical properties like thickness, optical constants, surface roughness, etc. Optical models are constructed by expected thickness of the film, complex refractive index, $n` = n + ik$ (n is the refractive index and k is the extinction coefficient) and complex dielectric constant, $\varepsilon` = \varepsilon 1 + \varepsilon 2$ ($n`^2 = \varepsilon`$). ψ and Δ are then calculated theoretically employing the model and compared with the experimentally obtained data. Regression analysis is used to fit the data changing the parameters of the optical model and the process is continued until a good fit is obtained.

3.3.2 Experimental

LSE STOKES Ellipsometer (GAERTNER Scientific Corporation) was used for measuring the thickness of SAMs of different reader molecules on palladium substrates. Prior to the modification, bare metal substrates were cleaned by hydrogen flame annealing, followed by the determination of ellipsometric parameters (n & k), which were required for measurement of SAM thickness. A HeNe laser with a characteristic wavelength of 632.8 nm and incident

angle of 70° was maintained in the instrument. The refractive index of the organic SAM layer was set to 1.46 during thickness measurement. For each substrate, 5-7 measurements were

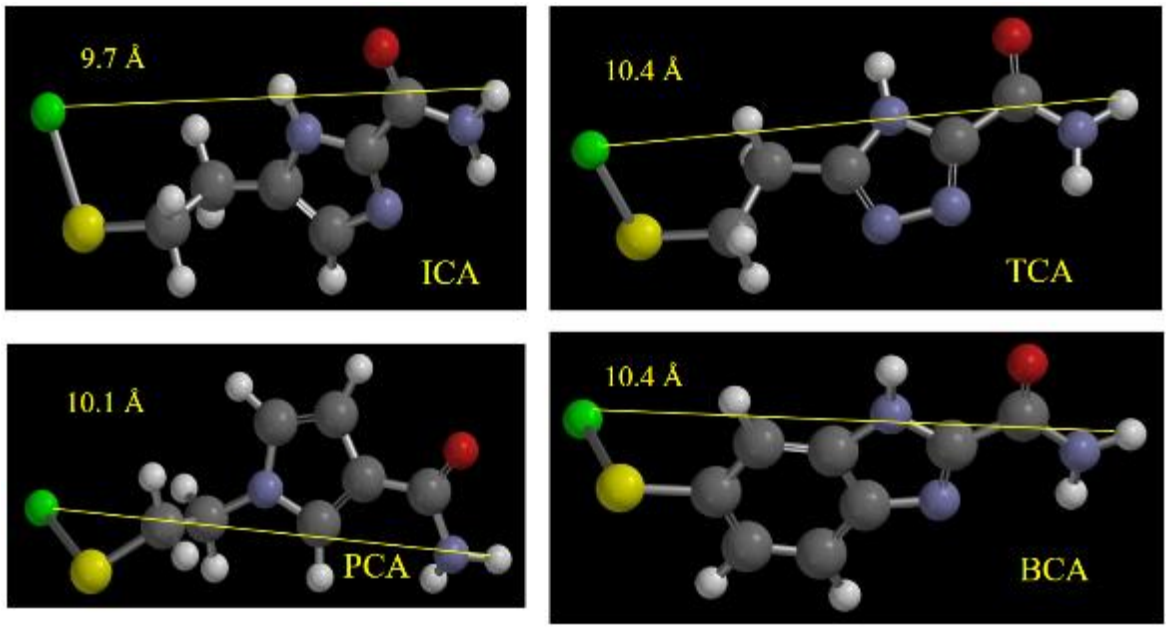

Figure 3.12. Calculated length of four recognition molecules (ICA=Imidazole, TCA=Triazole, PCA=Pyrrole & BCA=Benzimidazole) by SPARTAN

taken at different locations on the substrate and values of the thickness were averaged out.

3.3.3 Result

Experimentally obtained thickness of different monolayers are summarized in table 3.1. and molecular lengths of reader molecules in their free optimized state were calculated using SPARTAN and given in figure 3.14. for a preliminary comparison (though both must be somewhat different from each other for various reasons).

Table 3.1. Measured thickness of SAMs formed by different reader molecules using Ellipsometry

Name of the Reader	Thickness by Ellipsometry (Å)

Imidazole	7.6 ± 1.2
Benzimidazole	8.1 ± 1.2
Triazole	8.4 ± 1.3
Pyrrole	9.1 ± 0.8
Pyrene	8.6 ± 0.6
2-Phenylethane Thiol	5.6 ± 0.9

3.4 X-Ray Photoelectron Spectroscopy

3.4.5 Introduction & Working Principle

In X-ray Photoelectron Spectroscopy (XPS), a beam of X-ray is utilized to irradiate a particular material and from the spectroscopic characteristic of the emitted photoelectrons, chemical identity and elemental electronic states of the sample is determined. XPS is also known as Electron Spectroscopy for Chemical Analysis (ESCA) due to its ability to resolve elemental composition and empirical chemical formula. XPS is a very popular surface analysis technique due to its sensitivity depth, which is limited to only about 10 nm from the top of the surface. A very high vacuum (10^{-9} torr) is maintained which results exceedingly long (30-40 km) mean free path of emitted photoelectrons so that the electrons can travel to the detector without any inelastic collisions.[86] The process starts with the incident of a photon with energy hν on the sample and consequent ejection of electrons from the core-shell of atoms. These electrons, known as photo-electrons, are removed to the vacuum level with a certain amount of kinetic energy. Now, as a consequence of hole formation in the core level, a higher level electron with higher energy comes down to lower energy core level and acquire

the hole with releasing the excess energy. This excess energy is often absorbed by some higher level electrons and emitted to vacuum as Auger-electrons. Hence, for Auger-electrons its kinetic energy depends only on the binding energy difference of the involved levels and

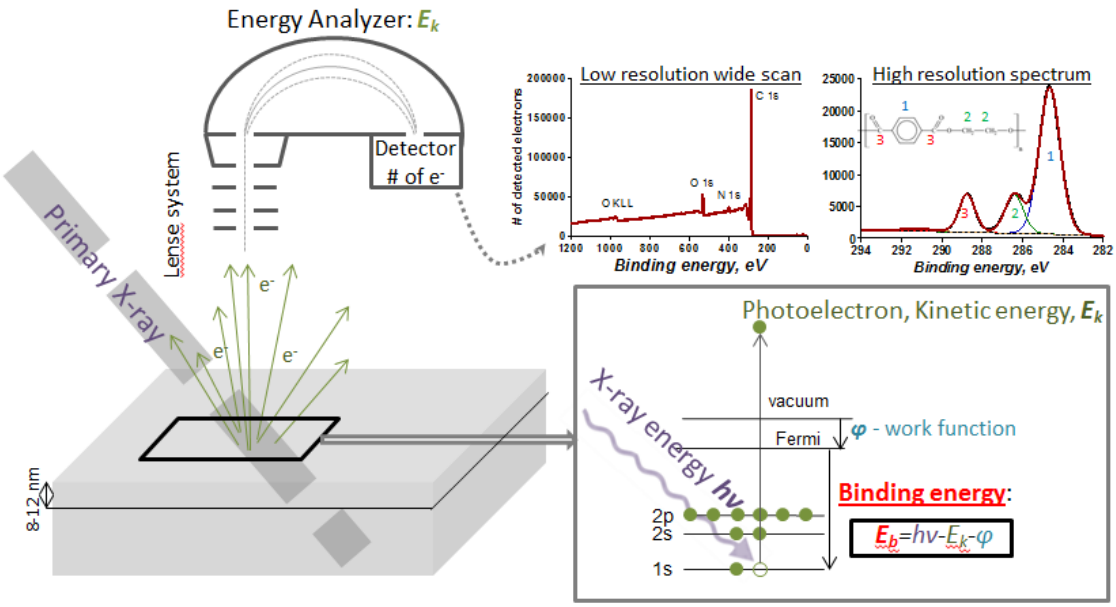

Figure 3.13. Working principle, instrumentation and sample XPS spectra

independent of applied X-ray photon energy. But, in the case of photo-electrons, kinetic energy (KE) depends on the energy of the incident photon, binding energy (BE) of core level and also on the work function of the sample (φ). So, kinetic energy of photoelectrons can be expressed as

$$KE(photoelectron) = h\nu - BE(core\ level) - \phi(sample)$$

Therefore, having the knowledge of hν, φ and BE elemental detection can be done from the value of kinetic energy of the photo-electrons (figure 3.13) [87]. A monochromatic X-ray beam is required, otherwise, from a definite core level, photo-electrons with a variety of kinetic energy will reach the detector and assignment of elements or electronic states will be difficult.

Surface sensitivity of XPS can be indicated by a property called "Sampling Depth". It is defined as the depth from which 95% of all photo-electrons are scattered as they reach the surface and information can be obtained from these electrons. Sampling depth depends on the mean-free-path of incident X-ray radiation (λ) and equals to 3λ for a definite XPS instrument. AlKα is the most popular X-ray source and has λ value in the range of 1-3.5 nm. Hence, for this condition sampling depth is in the range 3-10nm.

3.4.6 Instrumentation

The X-ray beam is generated from a monochromatic source and irradiated on the sample. Then the ejected photo-electrons are collected by electron energy analyzer system and channelized to the electron-detector. The analyzer is a concentric hemispherical analyzer that is maintained at a certain bias to capture electrons with definite energy range.

3.4.7 Experimental

VG ESCALAB 220i-XL photoelectron spectrometer with Al-Kα radiation (15keV) at 6 x 10E-10 mbar base pressure was used to record X-ray photoelectron spectra of both powder samples and SAMs on Pd substrates. C(1s), Pd(3d), N(1s) and S(2p) core level spectra were recorded at a pass energy of 20 eV and wide scan spectra were obtained at a pass energy of 150 eV. For determination of monolayer thickness, Angle-resolved XPS (ARXPS) was applied, where the thickness was measured three times on the same sample at three different angles of the incident energy beam. Tanuma Powell Penn 2M model was used for thickness calculation. Density data of different universal readers was required for thickness measurement and was calculated using Chemsketch program.

3.4.8 Result

Table 3.2. Expected and measured elemental ratio of SAMs formed by different reader molecules using XPS

Reader	Atom	% in XPS	Ratio in XPS	Expected ratio
Imidazole	S	5.17%	1	1
	C	26.77%	5.2	6
	N	12.22%	2.4	3
Benzimidazole	S	1.51	1	1
	C	25.7	17.02	9
	N	5.53	3.67	2
Triazole	S	5.52	1	1
	C	26.90	4.87	5
	N	20.71	3.75	4
Pyrrole	S	4.22	1	1
	C	30.93	7.33	7
	N	11.67	2.76	2

3.9 Angle-Resolved X-Ray Photoemission Spectroscopy (ARXPS)

ARXPS is a depth profiling technique that is popularly used to determine the thickness of organic self-assembled monolayers on metal substrates. In this process, emitted photoelectrons are collected at various angles to the sample surface. Therefore, electrons are

detected coming from a different depth of the sample (figure 3.14).[86] The technique undergoes by stepwise tilting of the sample with respect to the kinetic energy analyzer. Surface sensitivity increases with increasing tilt angle with respect to the surface normal. ARXPS is advantageous over the sputtering process for thickness measurement due to its non-destructive nature.

3.9.1 Result

Thickness values obtained from ARXPS are summarized in table 3.3. and calculated density of the reader molecules using CHEMSKETCH, required for ARXPS study, are listed in table

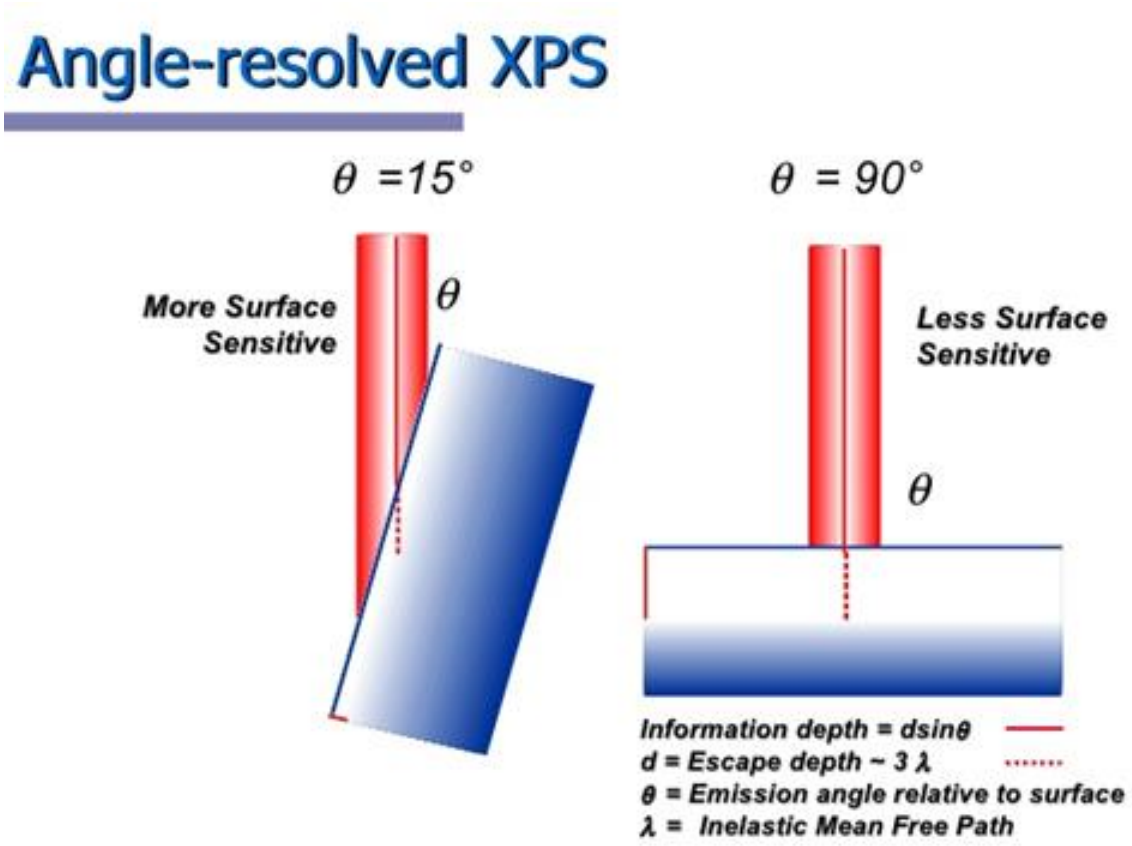

Figure 3.14. Working principle of ARXPS

3.4. ARXPS thickness values are little higher compare to thickness obtained from

Ellipsometry. This can be attributed to the calculated density data, which may not be very accurate due to the use of a basic level program.

Table 3.3. Measured thickness of SAMs formed by different reader molecules using Angle Resolved XPS

Name of the Reader	Thickness by ARXPS (Å)
Imidazole	10.5 ± 1.3
Benzimidazole	10.2 ± 1.9
Triazole	12.6 ± 0.7
Pyrrole	12.3 ± 1.8
Pyrene	7.8 ± 0.4

Table 3.4. Calculated density of different pure reader samples, required for thickness measurements of SAMs using Angle-Resolved XPS

Reader	Density (gm/cc) [calculated using Chemsketch]
Imidazole	1.35 ± 0.06
Benzimidazole	1.53 ± 0.06
Triazole	1.44 ± 0.06
Pyrrole	1.30 ± 0.10
Pyrene	1.26 ± 0.06

3.10 Water Contact Angle Measurement

Contact angle measurement is a very popular technique to characterize surface coating. The value of water contact angle on a surface indicates the hydrophilicity or hydrophobicity of the material.

3.10.1 Introduction & Working Principle

Liquid contact angle (θc) for a liquid droplet on a solid surface is defined as the angle

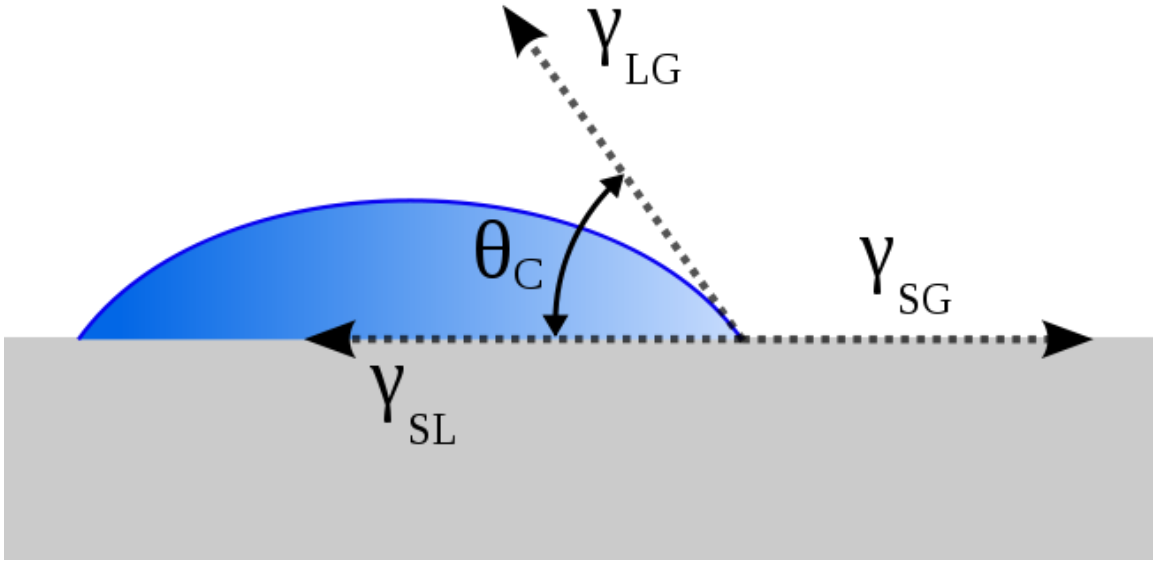

Figure 3.15. Definition of Contact Angle

between the solid-liquid interface and liquid-vapor interface. It also can be described as the angle between the solid surface and the tangent drawn through the liquid from the contact point of the solid, liquid and vapor phase together (figure 3.15). For a liquid droplet, thermodynamics between the three phases can be explained by Young equation as follows,

$$\gamma_{SG} = \gamma_{LG} \cos\theta c + \gamma_{SL}$$

where, γ_{SG} = Interfacial tension between solid and vapor, γ_{LG} = Interfacial tension between liquid and vapor, γ_{SL} = Interfacial tension between solid and liquid figure 3.15).[88] Generally, contact angles are measured by computer program from a two-dimensional image recorded by a CCD camera. In a common contact angle measurement instrument, all these components are integrated together with the movable (required for focusing by the camera) sample stage and LED light (for illuminating the droplet).

3.10.2 Experimental

Water contact angles were measured for hydrogen annealed palladium substrate prior to SAM formation and for functionalized substrate after SAM formation using an Easydrop Drop Shape Analysis System (KRÙSS GmbH, Hamburg). The volume of each water droplet for static contact angle measurements was 1 μL. 5-6 measurements were taken at different locations of the each functionalized and unfunctionalized palladium substrates.

3.10.3 Result

All contact angle values are listed in table 3.5. Bare hydrogen flame annealed palladium substrate showed a contact angle value of 8°-10°. The SAMs with amide group and azole moiety are hydrophilic and give a value in the range 30°-40°, whereas Pyrene and 2-PET SAMs are hydrophobic, as expected due to the presence of only hydrocarbon functionalities.

Table 3.5. Values of water contact angle of SAMs formed by different reader molecules

Name of the Reader	Water Contact Angle (°)
Imidazole	33.1 ± 5.1

Benzimidazole	39.3 ± 4.2
Triazole	35.7 ± 4.0
Pyrrole	40.3 ± 2.8
Pyrene	67.8 ± 4.5
2-Phenylethane Thiol	79.5 ± 4.1

CHAPTER 4

RECOGNITION TUNNELING OF AMINO ACIDS EMPLOYING HYDROGEN BONDING

4.1 Introduction

In this project, we showed that single amino acids can be identified by trapping the molecules between two electrodes that are coated with a single layer of sensor molecules, then measuring the electron tunneling current across the Scanning Tunneling Microscope

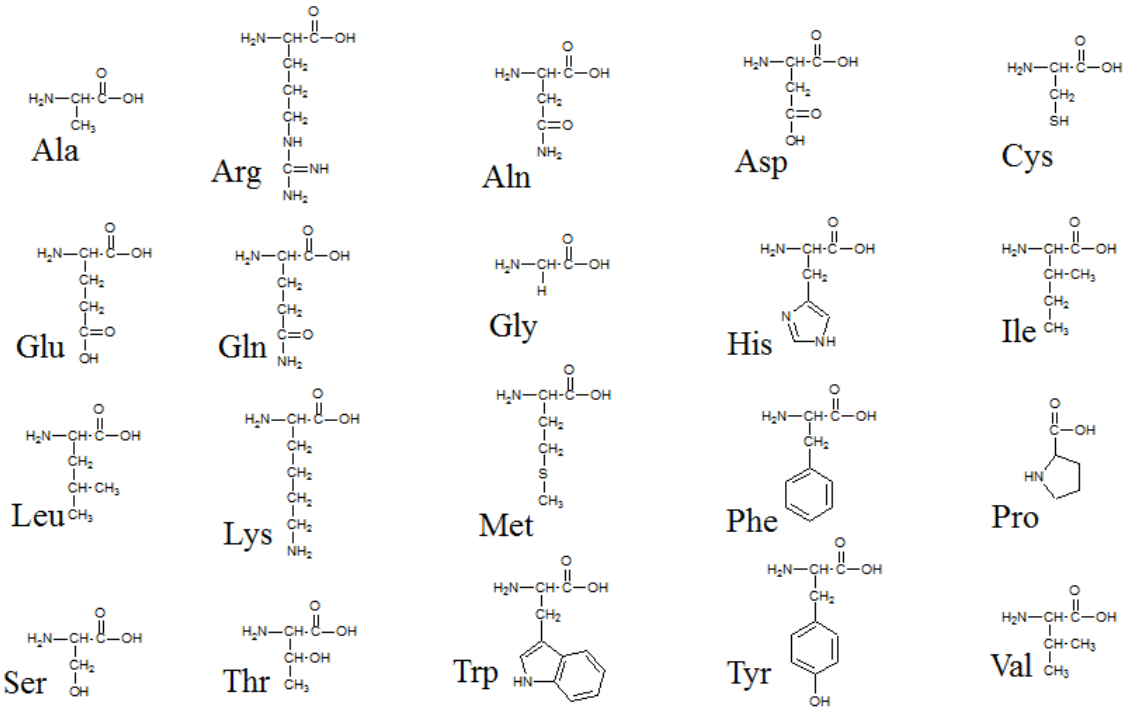

Figure 4.1. Naturally occuring amino acids and their three letter codes

junction. A given molecule can bind in more than one ways in the junction and we, therefore, use a machine-learning algorithm to distinguish between the sets of electronic 'fingerprints'

associated with each binding motif. This technique is known as "recognition tunneling" (described in chapter 1) and can distinguish two isobaric amino acids, a pair of D and L enantiomers and a methylated amino acid from its non-methylated analogue. Along with

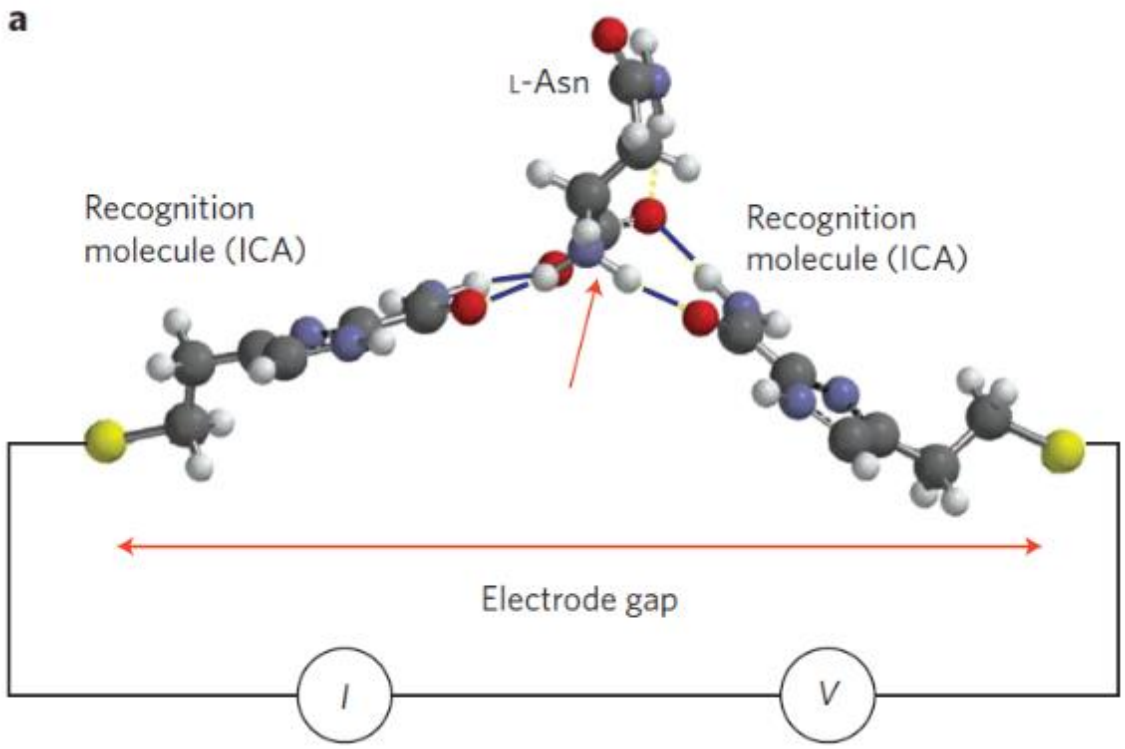

Figure 4.2. Optimized binding orientation of L-Asn in the STM nano-gap

different binding motifs, different dipole moment vector is also responsible for generating 'electronic fingerprints' (figure 4.2.) for the corresponding analyte. Short peptide chains also can be identified.[89] The results suggest that direct electronic sequencing of single proteins could be possible by sequentially measuring the products of exopeptidase digestion, or by using a molecular motor to pull proteins through a tunnel junction integrated with a nanopore.

4.2 Experimental

4.2.1 Preparation of Analytical Solutions

Amino acids were obtained from Sigma Aldrich (98% purity) and dissolved in 1 mM

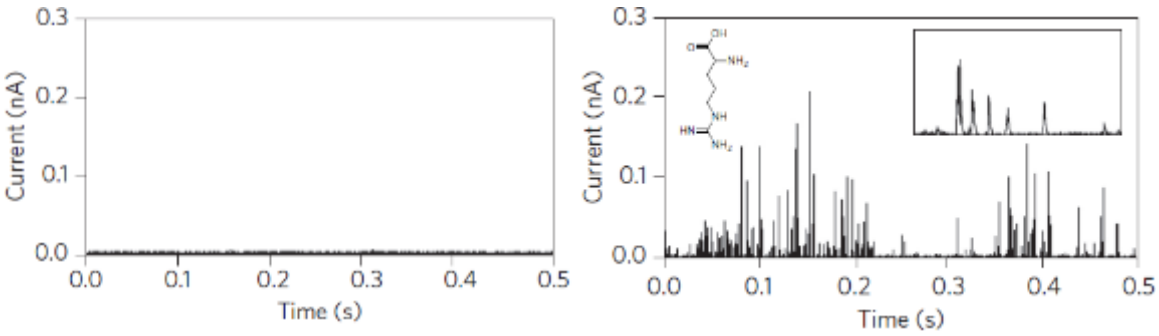

Figure 4.3. Typical signal trace from (a) control and (b) Arg tunneling experiment

phosphate buffer (pH 7.4), made using water from a Milli-Q system with specific resistance of 18 milliohm-cm and total organic carbon contamination below 5 ppb. Peptides were obtained from CPC Scientific and solutions prepared as for the amino acids.[89]

4.2.2 RT Experiment

We used two different PicoSPMs (Agilent Technologies) equipped with custom LabView interfaces for data acquisition. Tunnel current was sampled at 50 kHz. The -3 dB bandwidth of the current-to-voltage converter was 7 kHz, but useful signals were obtained out to the Nyquist limit of 25 kHz after correction for the instrumental response. The liquid cells were cleaned in Piranha (note: these solutions are potentially explosive and must be handled with extreme care) and rinsed with Milli-Q water and ethanol. The current set point was set to 4 pA with 0.5 V bias applied (probe positive, as this results in less leakage) and the probe approached with integral and proportional gains set to 1.0. The surface was scanned to ensure that the grain structure of the Pd was clearly visible. The microscope was left to stabilize for

at least 2 h before signals were recorded, and the integral and proportional gains were then reduced to 0.1. The control (1.0 mM phosphate buffer at pH 7.4) was run before an amino-acid solution was measured. Recordings were distorted by movement of the Z transducer during runs in which a series of high-amplitude spikes were recorded, but this artifact was common to all analytes and incorporated into the training of the SVM. We used different batches of substrates and probes for each run, usually recording four runs for each analyte. We also alternated measurements between different instruments. In this way, the influence of small changes in experimental conditions could be removed from the final analysis.[89]

4.3 SVM Analysis

Complete SVM data analysis process is described in chapter 8.

4.4 Result & Discussion

Figure 4.3.a. shows tunnel current-time trace is clean with only phosphate buffer in the

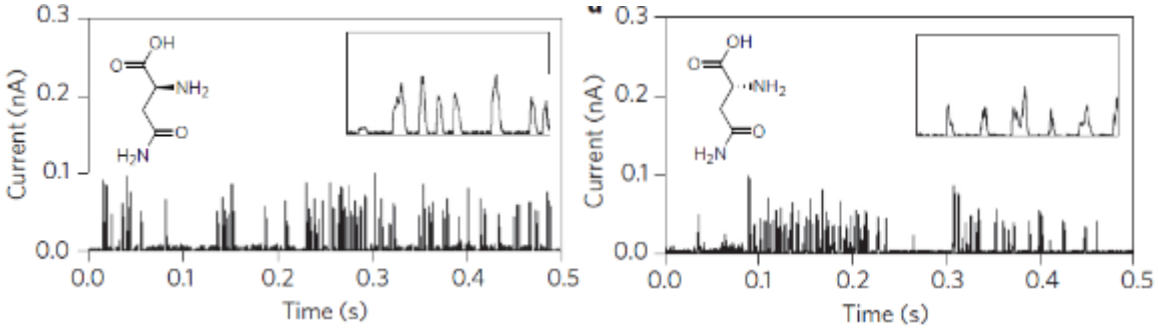

Figure 4.4. Typical signal trace from (a) L-Asn and (b) D-Asn

tunnel junction and after adding amino acid solution (Arg in figure 4.3.b.) in the junction ample of tunnel current spikes are generated. Following the same experimental procedure, control experiments are performed before measuring tunneling data for each amino acid. Typical tunneling current vs time data traces are showed for L-Asn (figure 4.4.a.), D-Asn (figure 4.4.b.),

Gly (figure 4.5.a.), m-Gly (figure 4.4.b.), Leu (figure 4.6.a.) and Ile (figure 4.4.b.).[89] For complete data analysis by SVM, four data sets of each of these above mentioned amino acids

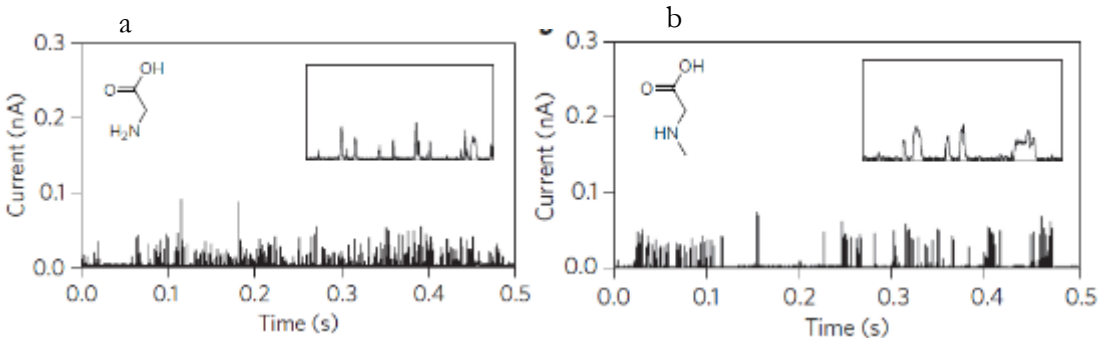

Figure 4.5. Typical signal trace from (a) Gly and (b) m-Gly

are obtained. In addition, all twenty naturally occurring amino acids (figure 4.1.) and multiple small peptides have been tested to obtain recognition tunneling signals. I have contributed in the probe preparation, substrate fabrication and the recognition tunneling experiments of multiple amino acids. All twenty naturally amino acids showed recognition tunneling signals, though tryptophan and tyrosine required higher current set-point (6 pA & 10 pA, respectively,

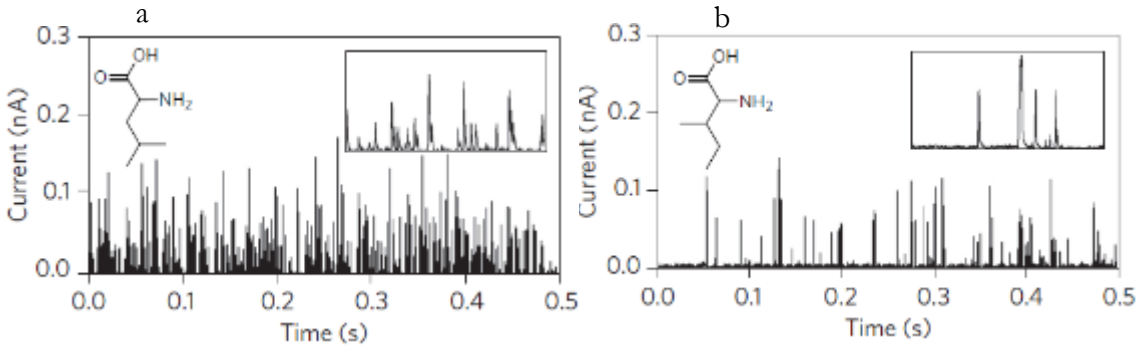

Figure 4.6. Typical signal trace from (a) Leu and (b) Ile

instead of 4pA).

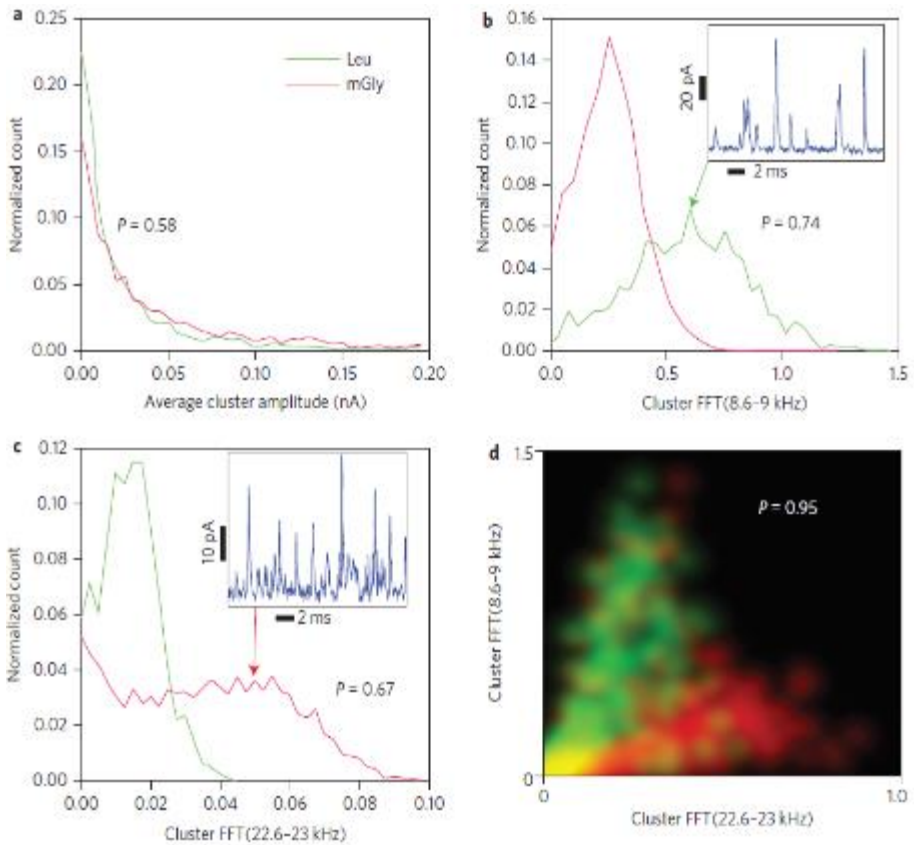

Figure 4.7. Signal features identify analytes. a, Peak amplitudes are exponentially distributed so provide little discrimination. Assigning the larger spikes to mGly (red curve) yields an accuracy (P=0.58) only slightly better than random (0.5). b,c, Particular Fourier components of the clusters show more separation, producing 74% (b) and 67% (c) accuracies if called solely on the more probable value of the feature. 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 as a function of the two FFT feature values. The colour scale shows mGly data points as red and Leu points 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.

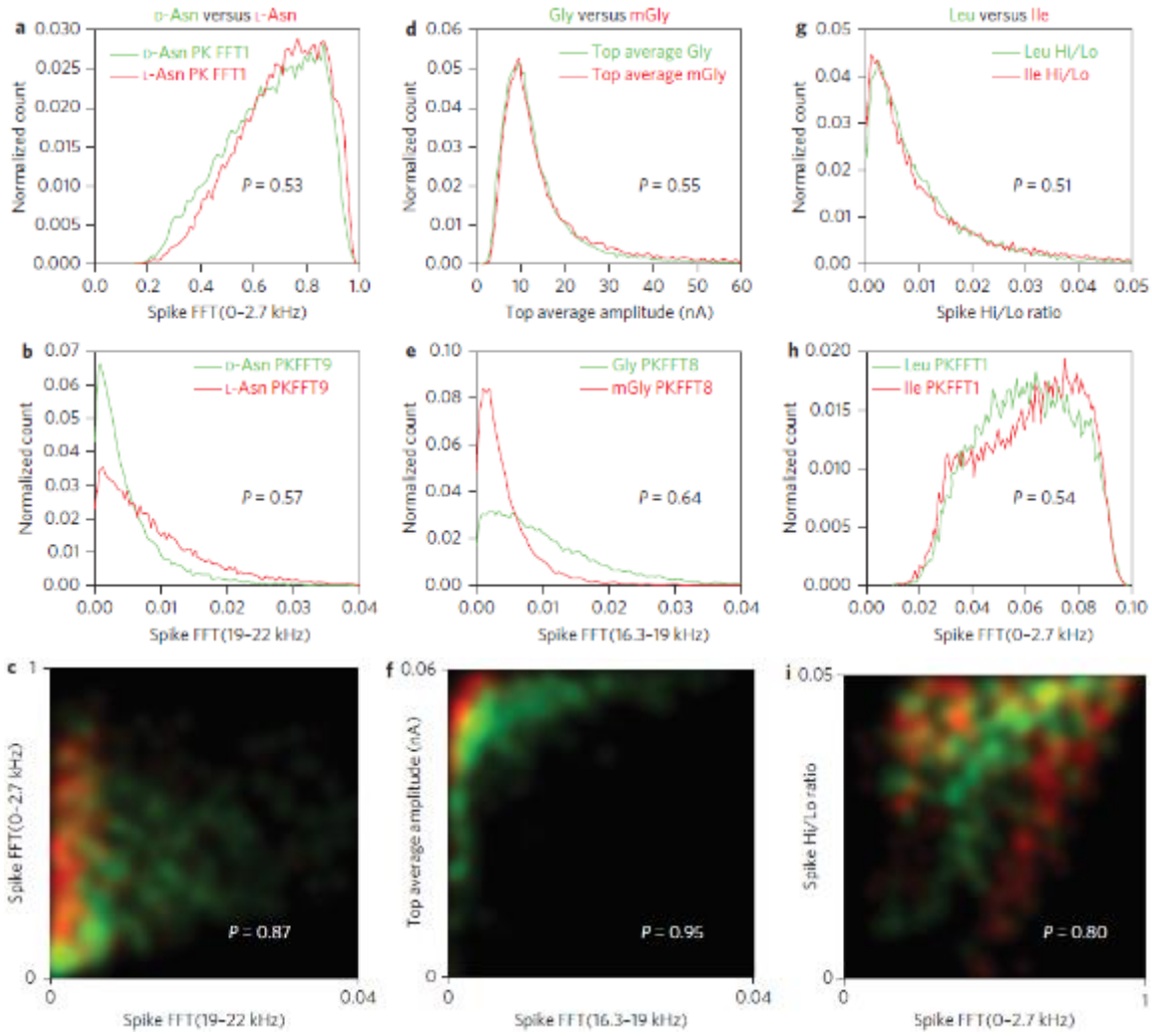

Figure 4.8. Closely related pairs of analytes can be significantly separated (>80%) using just two signal features together. All data are for pure solutions of one analyte. a–i, Chiral enantiomers D-Asn and L-Asn (a–c), Gly and mGly (d–f) and the isobaric isomers Leu and Ile (g–i) are quite well separated 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). The two-dimensional maps plot probability densities for the analyte pairs 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

We tried to differentiate the amino acids from their corresponding peak amplitudes, but the

distributions are so much overlapped that significant level of identification of individual amino acids appears to be difficult. Figure 4.7.a. represents average cluster amplitude distribution of m-Gly and Leu, indicating minimum amino acid identity can be revealed from this. But, as

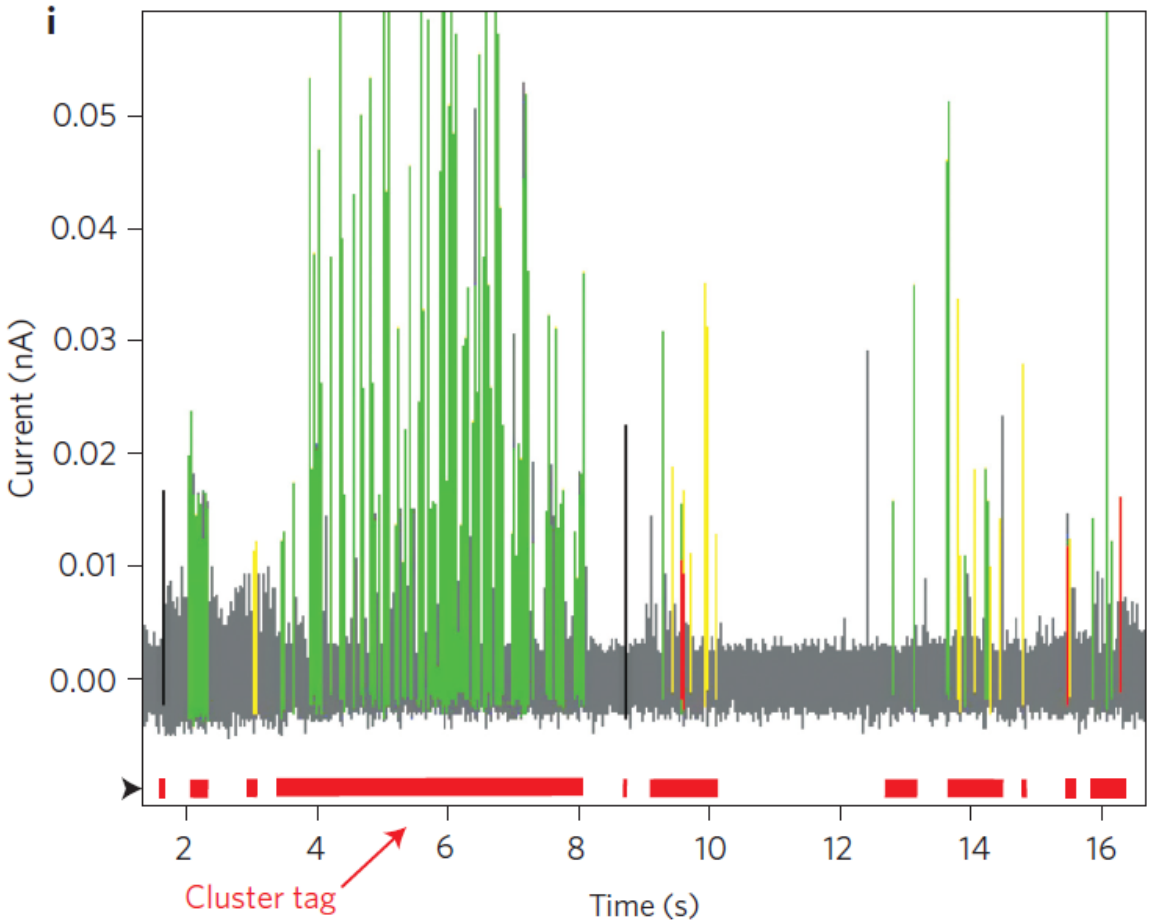

Figure 4.9. Signal trace for Arg, colour-coded according to the peak assignments made by a machine learning algorithm (green, correct; red, wrong call; black, 'water peak'; yellow, common to all amino acids). The red bars at the bottom mark signal clusters generated by a particular single-molecule binding event

we consider a couple of frequency components (spectral domain features) of the same cluster data (figure 4.7.b. & 4.7.c.) moderate level (~0.7) of analyte differentiation probability can be achieved. Now, as we combine these pair of one-dimensional histograms and plot a two-dimensional histogram a dramatically high data separation probability is obtained. Figure 4.7.d.

shows m-Gly and Leu can be categorized with a significantly high probability of 0.95 where m-Gly and Leu data points clustered in the red and green colored areas on the diagram. The yellow patches represent overlapped data of the two amino acids. This is a proof of Cover's theorem, which states that this kind of data separability improves in higher dimensions. Similar two-dimensional analysis to classify between L-Asn and D-Asn (enantiomeric pair), Leu and

Table 4.1. Percentage accuracies of all seven amino acids

Table 1 | Accuracy with which any one of seven pure analytes is identified from the total pool of data taken from all seven pure samples using 52 signal features together.

Number of spikes	Arg	D-Asn	L-Asn	Gly	Ile	Leu	mGly
1	95.14	94.99	96.99	97.24	96.87	94.36	96.45
3	98.77	99.67	99.99	99.67	99.99	99.55	99.99
5	99.99	99.99	99.99	99.99	99.99	99.99	99.99

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

Ile (isobaric pair) and Gly and m-Gly (methylated analogue) produce minimum data classification probability of 0.8 (figure 4.8.). [89]

Instead of only two, when more signal features (say N) are used analyte separability reaches close to 100% in an N-1 dimensional hyperplane. Similarly, more than two amino acids (seven, in our case) are classified as it was done for a pair of amino acids. In this case, any one of the amino acids is classified individually from rest of the amino acids using a common data pool.

Signal spikes common to different analytes and rare spikes from control experiments (water signals) are identified as noise spikes during training the SVM and excluded during analysis. This obviously increases the accuracy of separation between the analytes. Figure 4.9. explains this in a color-coded sample tunnel current vs. time data train of Arg. SVM accurately calls the green spikes as Arg but give a wrong call (call it other than Arg) on the red spikes. Black and yellow spikes are identified as water signal and common noise spikes from all amino acids, respectively. The red bars under the signal train indicate different clusters corresponding to

different binding events. Table 4.1. summarized the individual calling accuracies achieved after SVM analysis for seven amino acids that were previously mentioned. Separation accuracies in the order of 96-97% are obtained when single signal spikes have been used for the analysis. But, if majority voting is used for any three or five signal spikes from different signal clusters,

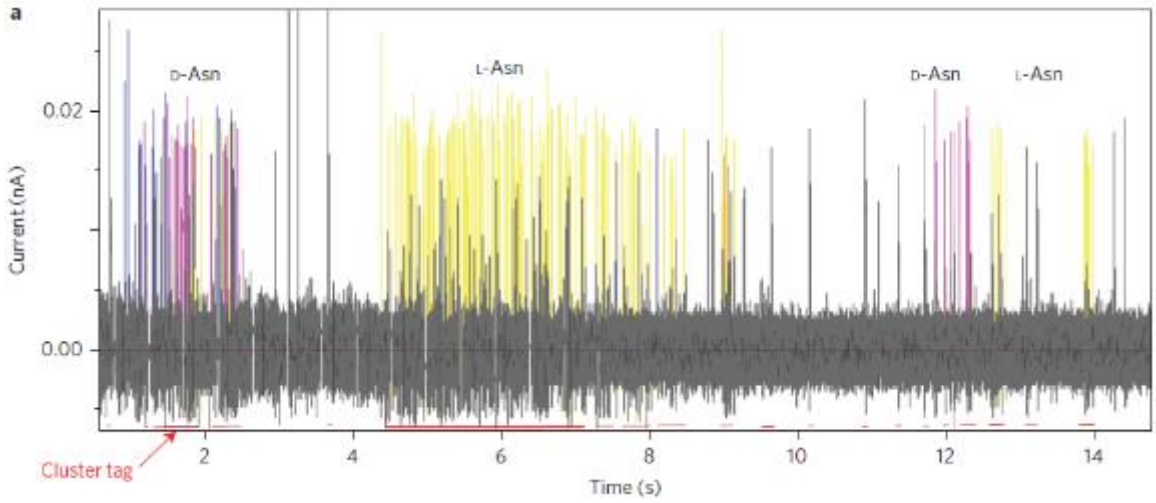

Figure 4.10. Data train obtained from a L-Asn and D-Asn mixture and analyzed by SVM

calling accuracies reaches extremely close to 100%.

Mixtures of analytes are tested with recognition tunneling, also. Data for different concentration ratio of L-Asn and D-Asn mixtures are recorded and analyzed by SVM. It assigns all spikes in a cluster to only one amino acid and proves that each cluster corresponds to a single amino acid binding event inside the tunnel gap (figure 4.10.).[89] Though accurate analyte ratio could not be measured by SVM as the generation of clusters mostly depends on the diffusion of the amino acid molecules inside the nanogap, which may not be homogeneous for different amino acids. Hence, the number of clusters is not simply proportional to the number of analyte molecules in the system. Besides, different analyte has different binding strength and interaction with the recognition molecule. As a consequence, stronger binding is

expected to produce longer cluster and generates more spikes than that in a short-lived cluster from a weaker binding. So, the number of spikes do not indicate the amino acid ratio either. [89]

4.5 Conclusion

Recognition tunneling (RT) can be used as a single molecular spectroscopy for recognizing different molecular classes. Its ability to differentiate between isobaric, enantiomeric molecules presents a great advantage over some other popular molecular recognition techniques (for example mass spectroscopy). All twenty naturally occurring amino acids generates signals in RT experiments. Also, isobaric amino acids and enantiomeric amino acids can be differentiated from each other with extremely high accuracy. All these provides a huge promise for single-molecule nanopore sequencer for peptide analysis employing RT.

CHAPTER 5

RECOGNITION TUNNELING OF DNA NUCLEOSIDE MONOPHOSPHATES EMPLOYING HYDROGEN BONDING

5.1 Introduction

To read DNA bases more accurately, we have synthesized a series of nitrogen-based heterocycles to examine their capabilities to interact with naturally occurring DNA nucleobases by hydrogen bonding in nanogaps. These recognition molecules are Benzimidazole (Bi), Imidazole (I), Triazole (T) and Pyrrole (P). The chemistry of complex formation in aqueous solution was studied by electrospray ionization mass spectrometry (ESI-MS). The study shows strong 1:1 complex and weak 2:1 complex between one and two reading molecules respectively with one DNA base. All these reader molecules are able to recognize four DNA monophosphates with an accuracy over 90%, except Pyrrole reader. Benzimidazole reader shows the highest base calling accuracy compared to other three molecules. [90]

Electron tunneling between two electrodes spaced by a nanometers gap has been proposed as a mechanism to read DNA bases, [91], and much progress has been made in this field. [92] When a single stranded DNA pass through an electrode nanogap embedded in a solid-state nanopore, the interactions of an individual DNA base with the electrodes should result in changes in the tunneling currents (illustrated in Figure 5.1.A). The tunneling current is highly sensitive to changes in distance (~ an order of magnitude per 0.1 nm) and the distance between two adjacent bases in a single stranded DNA is ~ 4 Å. So, the tunneling measurement should provide a highly spatial resolved method for readout of DNA sequences. It has been

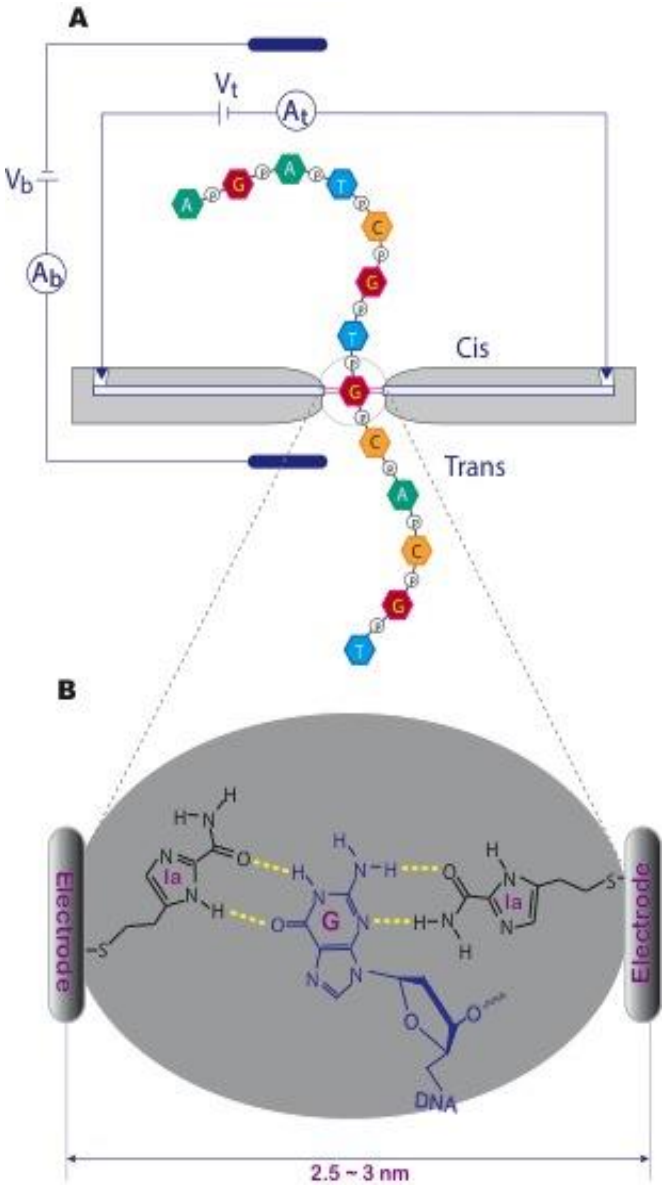

Figure 5.1. Cartoon illustrating (A) A tunneling device embedded in a nanopore to read DNA bases when they sequentially translocate through the nanopore; (B) Recognition interactions in the nano-gap where read molecules (universal reader) attached to the electrodes catch a DNA base by forming a hydrogen bonding complex and escalate electronic signals.

reported that nucleoside monophosphates and oligonucleotides can generate tunneling signals in a 0.8 nm nanogap,[93] but they overlapped one another. We rationalize that when two

electrodes are functionalized with recognition molecules that can capture a DNA base to form a tunneling junction (Figure 5.1.B) [90], it allows the electrons to pass through the nanogap more efficiently and create electrical signals related to the particular DNA base. All of the DNA bases have their different chemical structures so that they can form distinguishable tunneling junctions, resulting in unique signatures for their identification. We call this method sequencing by recognition tunneling.[60] This approach should increase chemical sensitivity and specificity of a tunneling nanogap in the detection of individual DNA bases. For the recognition tunneling, a recognition molecule that can form complexes with the naturally occurring DNA base in the nanogap is critical. Previously, we designed 1-*H*-Imidazole-2-carboxamide (**I**) as a universal reader.[94] It can form hydrogen-bonded triplets with DNA bases in the nanogap when attached to electrodes through a short carbon chain, as illustrate in figure 5.1.B. Our data show that **I** can recognize all of the DNA bases, but it only reach ~ 80% accuracy for each DNA base on average. [76] That leaves room for us to improve chemically the molecular recognition. In this project, we put our efforts in developing new reader molecules to increase the reading accuracy. [90]

5.2 Molecular Principles for Design of Universal Readers.

Our initial work on developing a universal reader began with benzamide (B) (Figure 5.2) [90] because the majority of hydrogen bonding motifs existing in DNA bases are a form of donor and acceptor alternation so that the amide group is a suitable moiety for the recognition interactions. Our data show that the benzamide moiety reads DNA base A, C, G, and methylated C but T. This prompted us to design and synthesize the first generation of reader molecule, 1H-imidazole-2-carboxamide (I), that contains more hydrogen bonding sites and a flexible linker, and we found out that it recognized all the DNA bases in nanogaps. Our data

indicates that the imidazole molecule has reached ~ 80% accuracy in identification of the five DNA bases, much higher than the random sampling (~ 20%). Thus, the imidazole-2-carboxamide provides us a framework on which we can build a true universal reader. We have

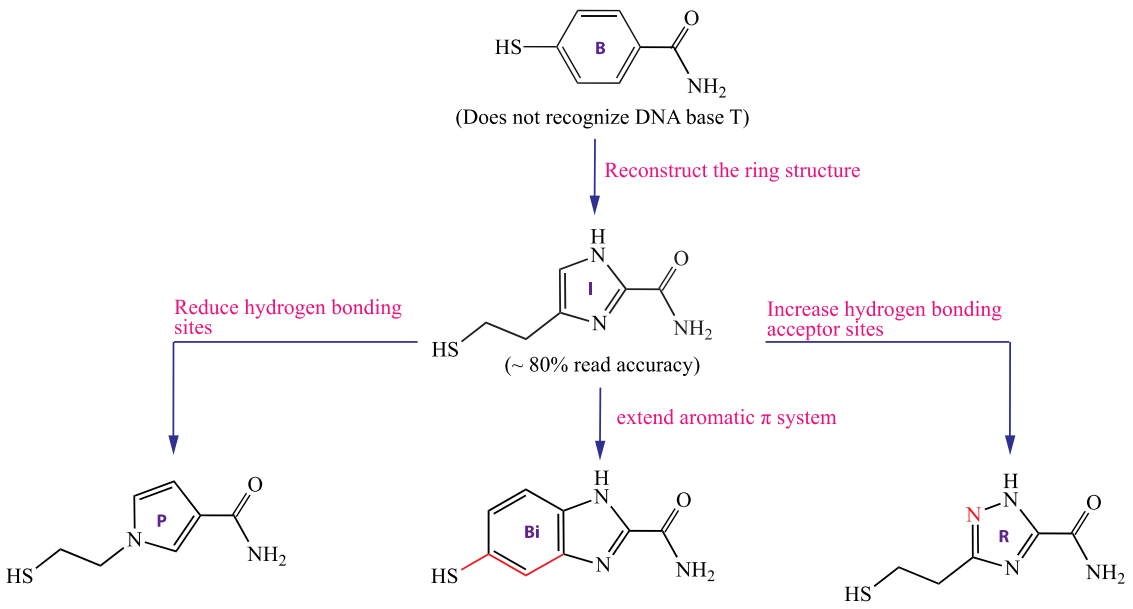

Figure 5.2. Different Universal reader candidates derived from the imidazole-2-carboxamide molecule

created three candidates of the universal reader by fine-tuning the chemical structure of the imidazole ring (Figure 5.2). [90] First one is 1-(2-Mercaptoethyl)-1H-pyrrole-3-carboxamide (P), the pyrrole ring of which has higher π electron densities on the aromatic carbons than the imidazole ring. The second one is 5-mercapto-1H-benzo[d]imidazole-2-carboxamide (Bi) that extends the π system of the imidazole ring and is more rigid, and third one is 3-(2-mercaptoethyl)-1H-1,2,4-triazole-5-carboxamide (R) that has one more hydrogen bonding sites than imidazole. By studying these molecules, we should have more insights into effects of chemical structures on recognition of DNA bases. It should be noted that each of these molecules is connected with a thiol function either through a two-carbon chain or itsequivalent

in length (see Bi in Figure 2 where the thiol is placed at a position of two carbon-carbon bonds away from the imidazole ring as drawn in red) for their attachment to electrodes.

5.3 Experimental

5.3.1 Preparation of Analytical Solutions

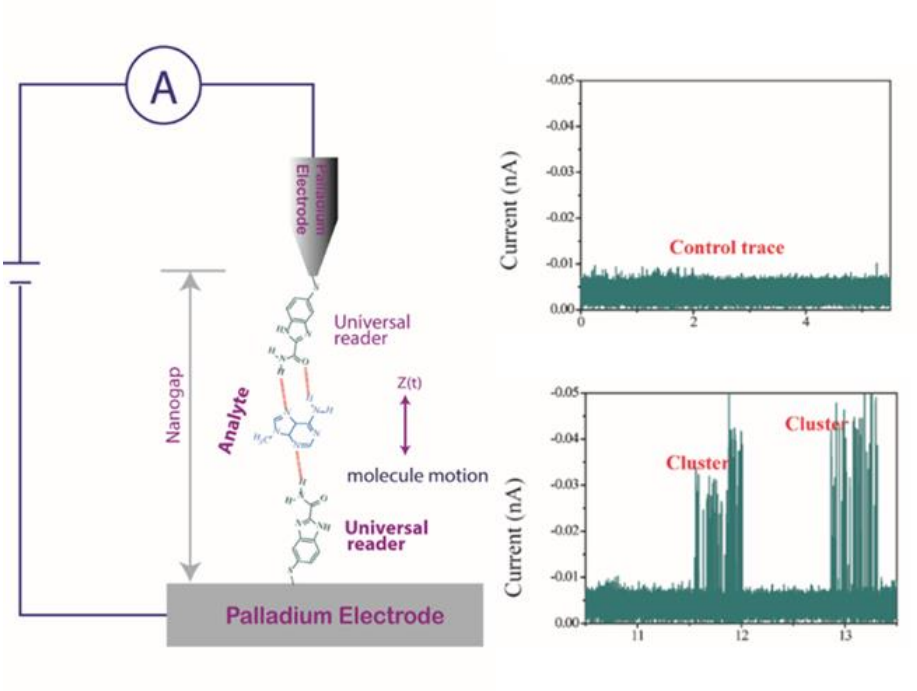

Figure 5.3. Schematic of Recognition tunneling typical control trace, signal cluster and signal spike

DNA monophosphates were obtained from Sigma Aldrich as sodium salts and dissolved in 1 mM phosphate buffer (pH 7.4), made using water from a Milli-Q system with total organic carbon contamination below 5 ppb.

5.3.2 RT Experiment

In a typical RT experiment, the measurement followed a process of mounting the functionalized Pd-STM probe and Pd-substrate to a PicoSPM scanning tunneling microscope, stabilizing the tunnel junction in a phosphate buffer (1.0 mM, 7.4 pH) until a clean baseline

was generated (~ 2 h), introducing an analyte solution (typically 100 μM in 1.0 mM phosphate buffer, pH 7.4) to the liquid cell, and collecting current recordings under a predefined tip-

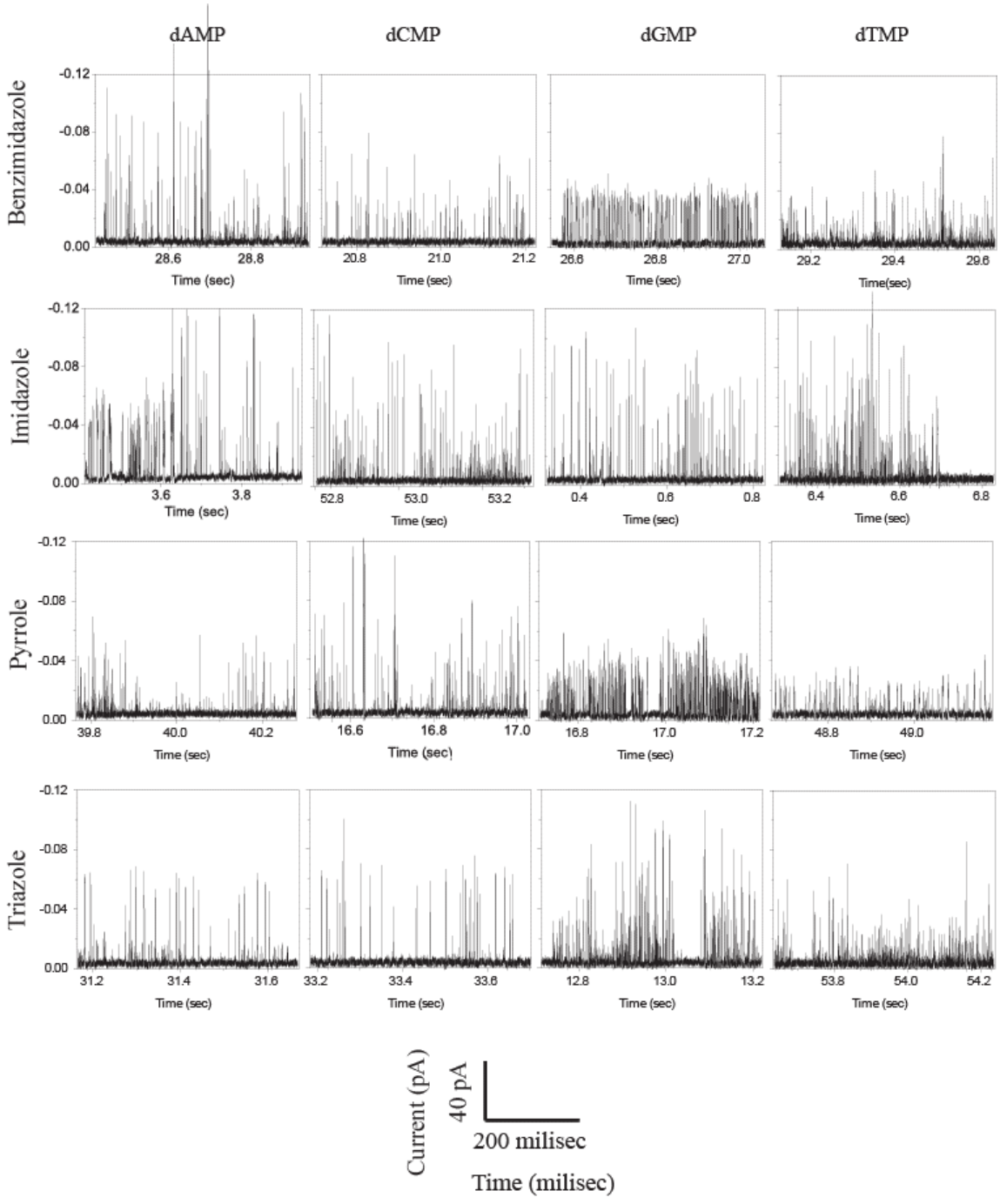

Figure 5.4. Signal clusters for different DNA monophosphates with different universal readers at 4 pA set-point current and 500 mV probe bias

substrate bias. Four naturally occurring DNA nucleoside monophosphates (dAMP, dCMP, dGMP and dTMP) were used as analytes. For each analyte, four separate experiments were run with freshly made probes, substrates, and samples. We used different batches of substrates and probes for each run, usually recording four runs for each analyte. We also alternated measurements between different instruments. In this way, the influence of small changes in experimental conditions could be removed from the final analysis.

5.4 Result & Discussion

We have used Scanning Tunneling Microscope (STM) to create the nanogaps for studies of recognition tunneling (RT). In a typical RT experiment, for example, a tunneling current was set at 4 pA with a voltage bias of 0.5 V, which corresponded to a nanogap of ~ 2.4 nm distance. [95] Most of the readers gave better DNA monophosphate separation at 4pA set point compare to 2pA set point in STM. For all of the readers with 2pA set point, the level of separation is around 90%, which goes up to around 95% or more in case of 4pA set point.

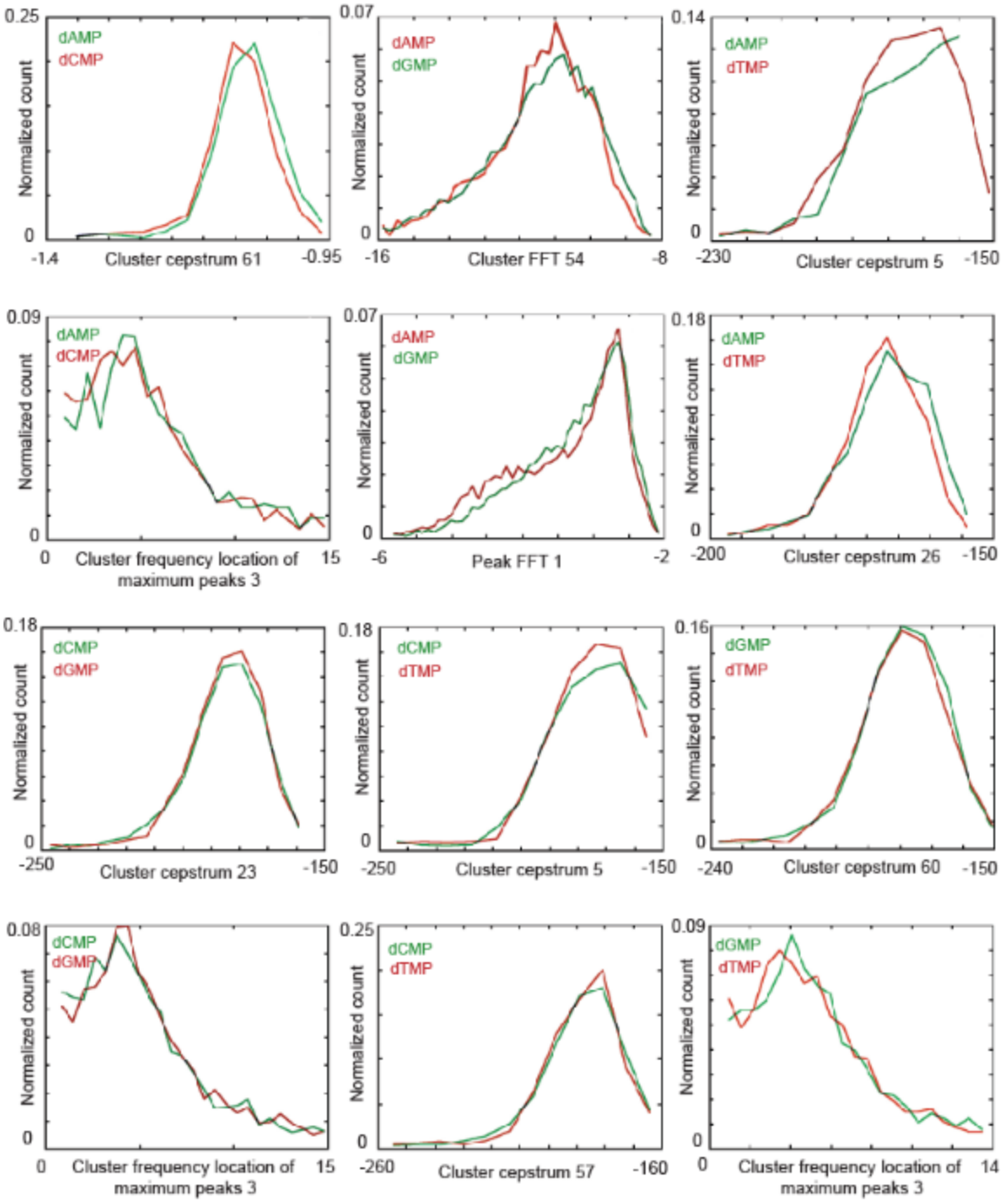

Figure 5.5. Different 1D histograms used to plot the best 2D histograms obtained for Imidazole reader to compare any two DNA monophosphates

This information concludes that a nanogap with 4pA set point gives optimum dimension. We propose that gap at 4 pA provides bridging conformations that are much different from each

other. Whereas at 2 pA the gap is so big that the bridging interactions with different DNA monophosphates have less differentiating features from each other. Four data sets for each DNA monophosphates were obtained at both 2 pA and 4 pA current set-point. But, I will

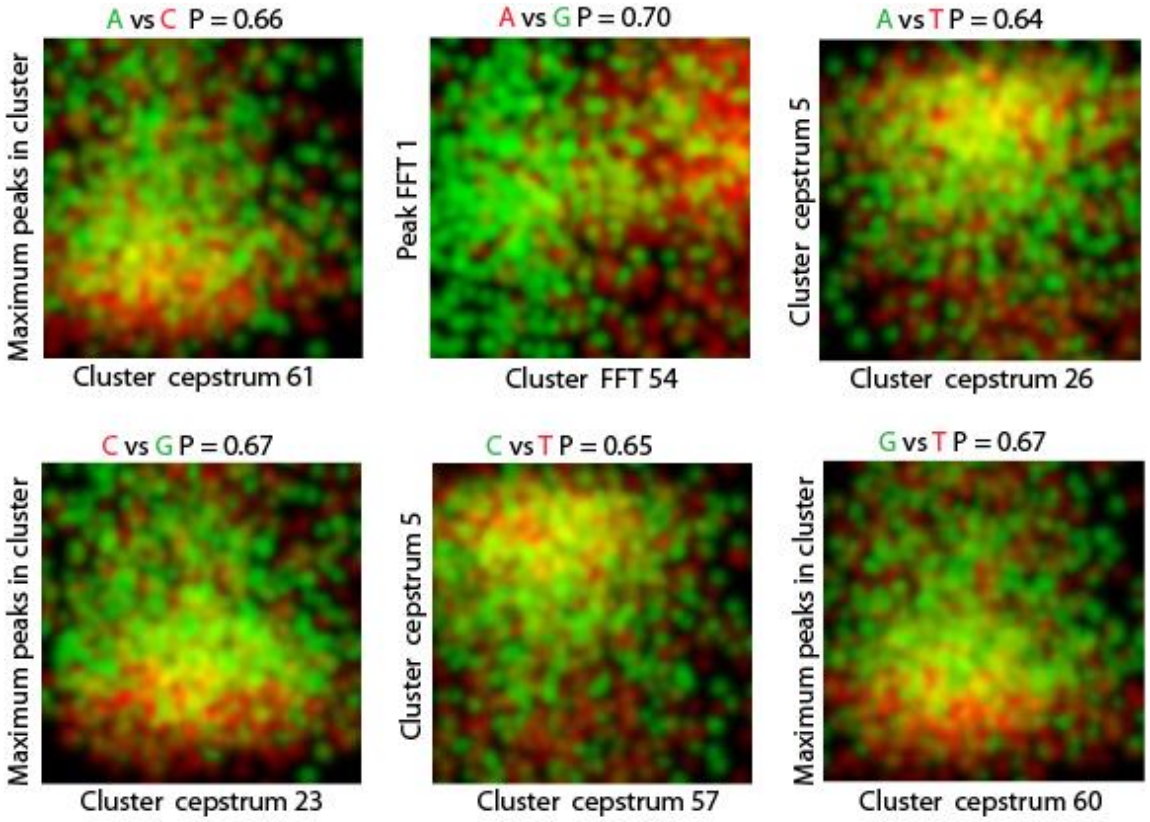

Figure 5.6. Highest obtained accuracies from 2D histogram for Imidazole Reader

describe mostly the result obtained at 4 pA as this is a proven better condition for RT experiments, providing better DNA recognition. Typical traces of data train of DNA monophosphate RT experiments with different reader molecules at 4pA current set-point in shown in figure 5.4. [90]

The servo on setting of the STM during RT experiment was a proven disadvantage for our DNA monophosphate separation purpose on the basis of tunnel current amplitude of different DNA monophosphates. The servo helps the instrument to maintain a fixed passage

of tunnel current between the source and drain. As a consequence, the nanogap in STM junction keeps changing its dimension as a function of different bridging conformation of reader molecules and analytes. Hence, tunnel current produced by different DNA monophosphates give a broad distribution, eventually.

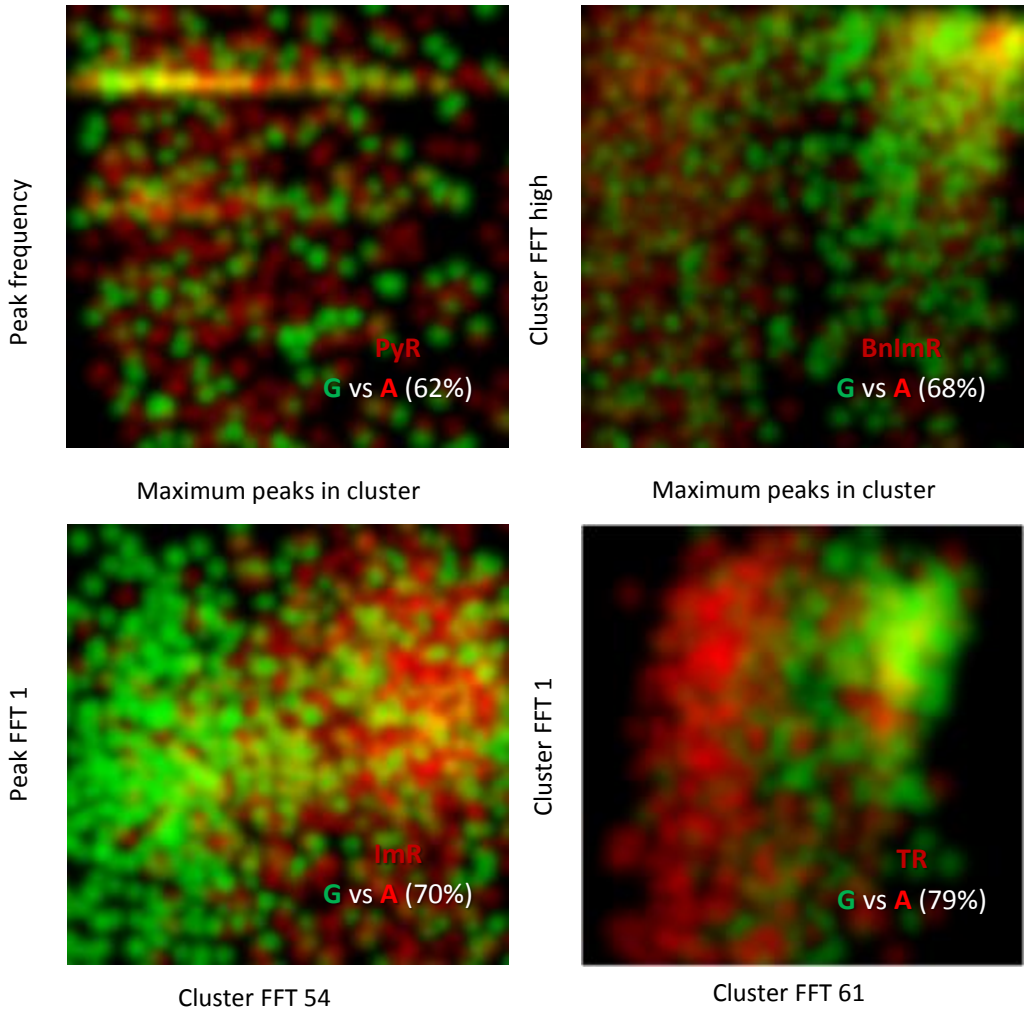

Figure 5.7. Highest obtained accuracies from 2D histogram for different Readers

Support Vector Machine (SVM) used various features associated with the spikes and clusters of different analytes to perform a classification. Various signal features were used for this purpose. FFT, cepstrum (Cepstrum is defined as inverse Fourier transformation of the log-magnitude of the Fourier spectrum) component of spike amplitude and so on. The goal is

to separate different analytes from each other with a high accuracy. We compared the base separation achieved by different universal readers at 4 pA set. Figure 5.5 summarizes different

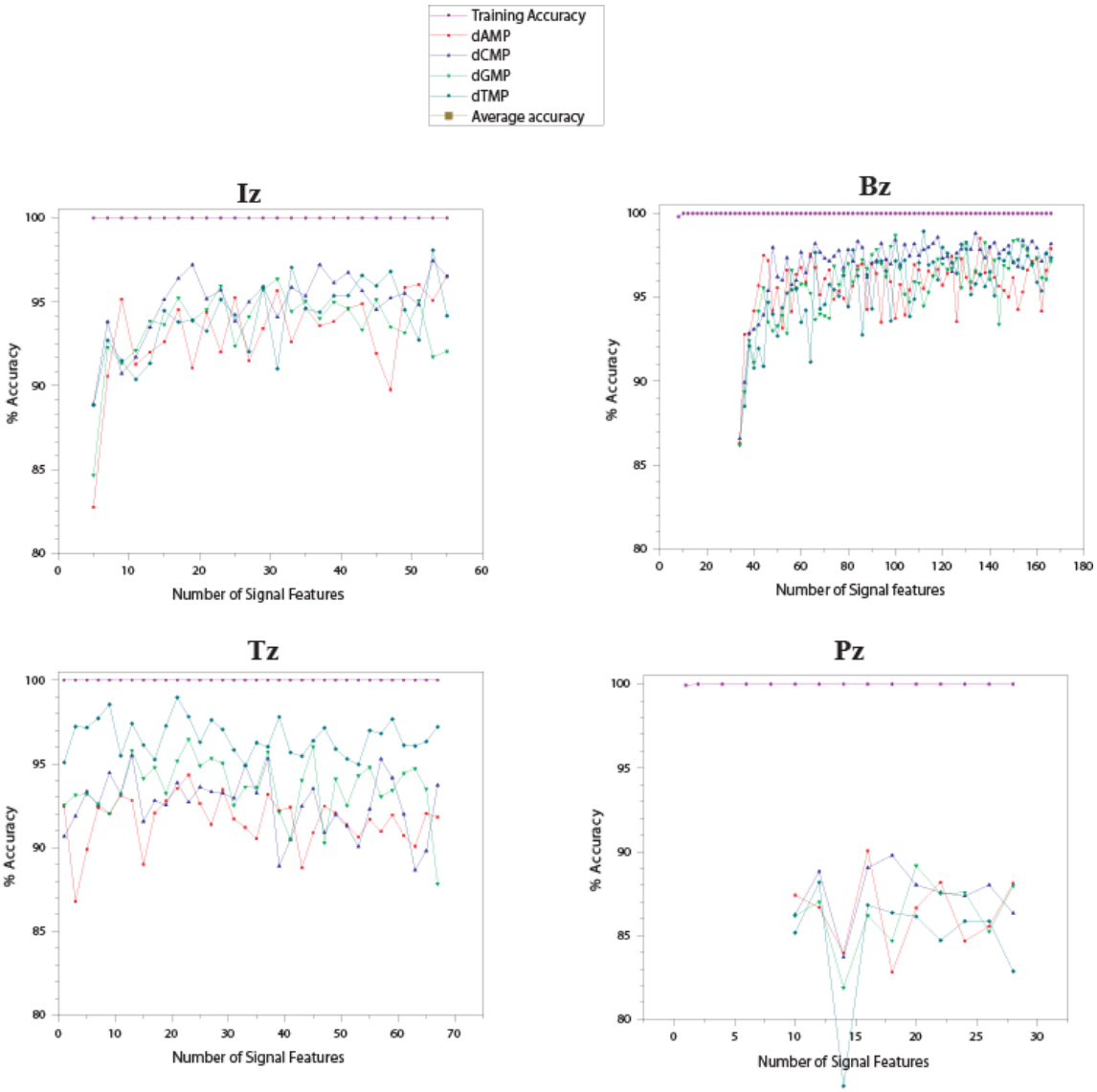

Figure 5.8. Number of signal features introduced vs training accuracy, individual DNA base calling accuracy and average DNA base calling accuracy plot of different Readers

one-dimensional histograms used to plot the best two-dimensional histograms obtained for Imidazole reader to compare any two DNA monophosphates. As it can be seen that for any two DNA monophosphates they are mostly overlapped, giving a separation probability just

over random (0.5). Hence, the two-dimensional histograms are not well separated either, but separation probability increases to ~0.7, which is much better compare to using single signal feature. Two-dimensional histograms of other readers can be obtained similarly from their corresponding one-dimensional plots. This is an illustration of covers theorem showing that accuracy in pattern recognition increases with number of dimensions, which represents the number of signal features in our SVM analysis process. In two-dimensional separation, different reader molecules perform adequately, showing a moderate separation probability of ~0.7. Triazole reader gives a better performance compare to others producing a value of ~0.8 (figure 5.7). [90]

We compared the base separation achieved by different universal readers at 4pA set point with the help of more signal features. Benzimidazolezole reader clearly stands out as the best option to call all of the nucleoside monophosphates with high accuracy, whereas Pyrrole reader could not reach even 90% calling accuracy (89%). This could be explained by the lack of hydrogen bonding sites on Pyrrole ring. Triazole and Imidazole achieved similar values of calling accuracies, suggesting increment in the number of ring nitrogen atom from 2 to 3 does not affect the binding interactions to a large extent. Lastly, the highest calling accuracy obtained by Benzimidazole could be related to the expected higher conductance of the elongated aromatic system compares to other reader molecules. Another explanation could be drawn from the less number of orientations of Benzimidazole reader inside the tunnel gap, which might help to attain distinct binding conformation of the DNA monophosphates in the junction.

Figure 5.8 represents changes in number of signal features introduced vs training accuracy, individual DNA base calling accuracy and average DNA base calling accuracy for different

recognition molecules. We found in our two-dimensional analysis, Triazole performs better compare to other three reader molecules. Number of signal features vs. accuracy plot verifies that as Triazole obtains high accuracy with a lower number of features, whereas Benzimidazole

Table 5.1. Calling Accuracies for Different Readers

	dAMP	dCMP	dGMP	dTMP	Average
Benzimidazole	98.5	98.8	98.7	98.9	98.7
Imidazole	96.5	97.4	96.4	98.1	97.1
Pyrrole	90.1	89.8	89.2	88.2	89.3
Triazole	94.3	95.5	96.5	99.0	96.6

and Imidazole require more number of features to attain comparable higher accuracy.

5.5 Conclusion

Eventually, Benzimidazole attains higher DNA base calling accuracy (average 98.7%) compare to Imidazole (average 97.1 %) and Triazole (average 96.6%). On the other hand, calling accuracy for Pyrrole reader never improved over 90% (average 89.3%). We need to mention that SVM analysis for different readers utilized different numbers of survived signal features (after feature correlation analysis). This depends on characteristics of water spikes, common spikes and RT spikes obtained from different readers and discussed in more detail in chapter 8. Individual and average DNA base calling accuracies are summarized in table 5.1. [90]

CHAPTER 6

RECOGNITION TUNNELING OF DNA NUCLEOSIDE MONOPHOSPHATES EMPLOYING AROMATIC STACKING INTERACTION

We demonstrate experimental evidence of strong π-π aromatic stacking interaction between DNA nucleotides and a couple of sensor molecules, deliberately designed for the purpose and attached to the electrodes (Pd probe and substrate) of a Scanning Tunneling Microscope (STM). Non-equilibrium Green Function (NEGF) calculations of energy-optimized sandwiched π-π stacking structure of sensor molecule-nucleobase-sensor molecule complex (including slabs of Pd electrodes) complemented the experimental findings, nicely. Using the stochastic tunneling current spikes, obtained from STM experiments different nucleotides, we are able to detect and distinguish all four DNA nucleotides with a very high level of accuracy (98%) by employing a machine learning algorithm, called Support Vector Machine (SVM). Exploiting only a couple of signal features of the current spikes, two nucleotides are distinguished with 90% accuracy. Furthermore, the result shows that, sensor molecules with small aromatic rings (such as only benzene ring) are not suitable for strong stacking interaction with nucleotides. [96]

6.1 Introduction

Several highly sophisticated techniques have been utilized so far to move towards the target of cheaper, faster and convenient genome sequencing[97]. After the early days of Sanger sequencing[98], amplification-based sequencing employing polymerase chain reactions have acquired huge popularity due to high throughput and low-cost factors.[31, 97, 99-101] But,

these methods face a massive roadblock for samples having repetitive DNA sequences, which is a common phenomenon in the case of large genomes (e.g. human genome) and for a verity

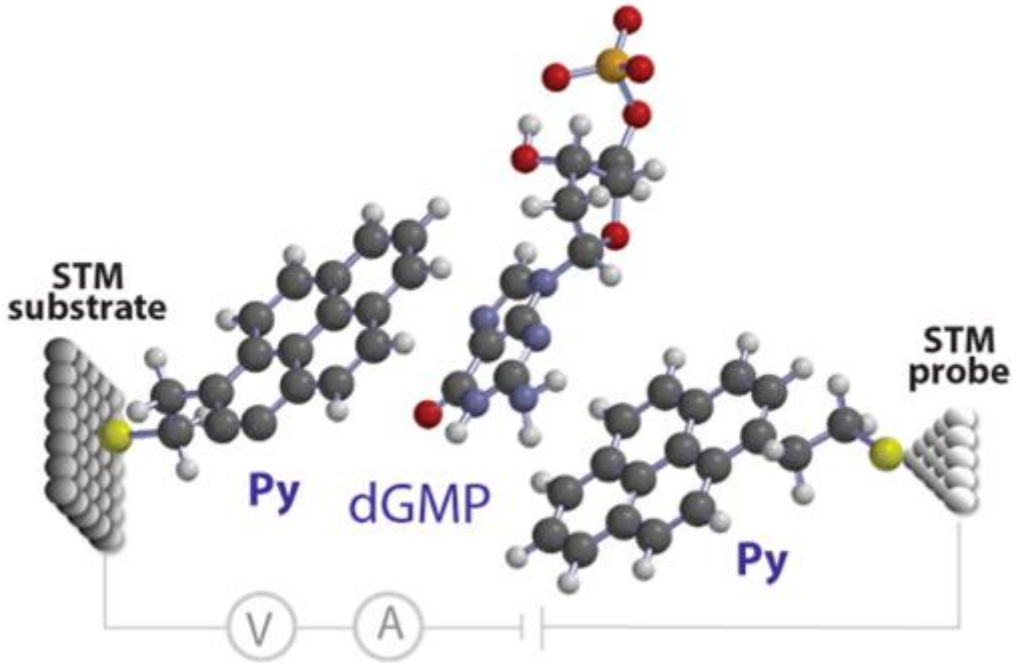

Figure 6.1. Schematic diagram of the experimental set-up in a Scanning Tunneling Microscope (STM), with Pyrene modified probe and substrate. A dGMP molecule is stacked in between the Pyrene molecules of the probe and substrate.

of bacterial genomes.[29, 30] The inability of long reads, the requirement of the amplification step (can provide reads only few hundreds bp long), causing high error rates and challenging *de novo* genome sequencing.[31] Recently commercialized single molecule real-time sequencing (SMRT by Pacific Biosciences), without the need of any polymerase chain reactions for amplification, shows the capability of very long reads (10,000 bp, even longer) with fast sequencing rate and low cost per base sequence. Still this technique possesses challenges to improve on its notably high error rate (~13%) and very high instrumentation cost. [37]

After all those methods, nanopore-based sequencing techniques have emerged as the most potent candidates to reach the goal.[102, 103] Capability of very long reads (10kb already achieved by ONT), fast and direct sequencing from freshly obtained data, an absence of complicated sample preparation and presence of sophisticated and cheap semiconductor and microfluidics device processing make nanopore-sequencing a star contender.[102-106]

Studies have shown that, transverse tunnel conductance of analytes (e.g. nucleotides, amino acids) depends exceedingly on geometrical factors (e.g. orientation) [89, 107] instead of their intrinsic electronic features. Likewise, theoretical studies on ssDNA translocation through graphene nanopores demonstrate that ion current blockade generated by DNA translocation can be nucleotide specific but excessively dependent on the orientation of the nucleotides[108], which is immensely hard to control due to high speed of translocation events.

Successful experimental studies on molecular conductance measurements using break junction with STM[109-111] and theoretical studies on DNA sequencing exploiting nucleobase-aromatic system π-π stacking encouraged our sensor molecule designing and tunnel junction measurements. Theoretical studies on interactions between nucleobase and aromatic systems like Naphthalene, Boron Nitride and Graphene boosted our interest.[112-114] Simulations showed the preference of parallel stacked orientation of the aromatic surfaces[115, 116] and from our Non-equilibrium Green Function (NEGF) calculations we obtained similar sandwiched π-π stacking structure of three-membered sensor molecule-nucleobase-sensor molecule complex.

Though, a plethora of fascinating theoretical studies on stacking interaction between nucleotides and several different aromatic systems (Graphene nanoribbon, naphthalene,

Boron nitride, etc.) including their probable applications for DNA sequencing are available, a number of experimental studies in this area are still scarce. This study can play an important role to encourage towards such experimental exploration.

In this study, we recognize the DNA nucleotides through π-π stacking interaction with a pair of aromatic sensor molecules in an STM nanogap (figure 6.1)[96] and distinguish the nucleotides with a high level of accuracy, using SVM for signal decoding. Similar tunnel junction can be made on a nanopore device, possessing a couple of tunneling electrodes, functionalized with the same sensor molecules (Pyrene Reader), through which ssDNA can be pulled through and sequence down. The high translocation speed of ssDNA through solid state nanopres is a noted obstacle for nanopore sequencing. To counter this problem, a reliable strategy can be the use of non-covalent adhesive force between nucleobase and different aromatic molecules. Hence, stacking among the nucleotides and our sensor molecules is notably strong and may serve significantly to slower translocation.[96]

6.2 Experimental

6.2.1 Preparation of Analytical Solutions

DNA nucleotides (monophosphate sodium salt or neutral compound) were obtained from Sigma Aldrich and dissolved in 1 mM phosphate buffer (pH 7.4), made using water from a Milli-Q system with specific resistance of 18 MΩ-cm and total organic carbon contamination below 5 ppb.

6.2.2 RT Experiment

In a typical RT experiment, the measurement followed a process of mounting the

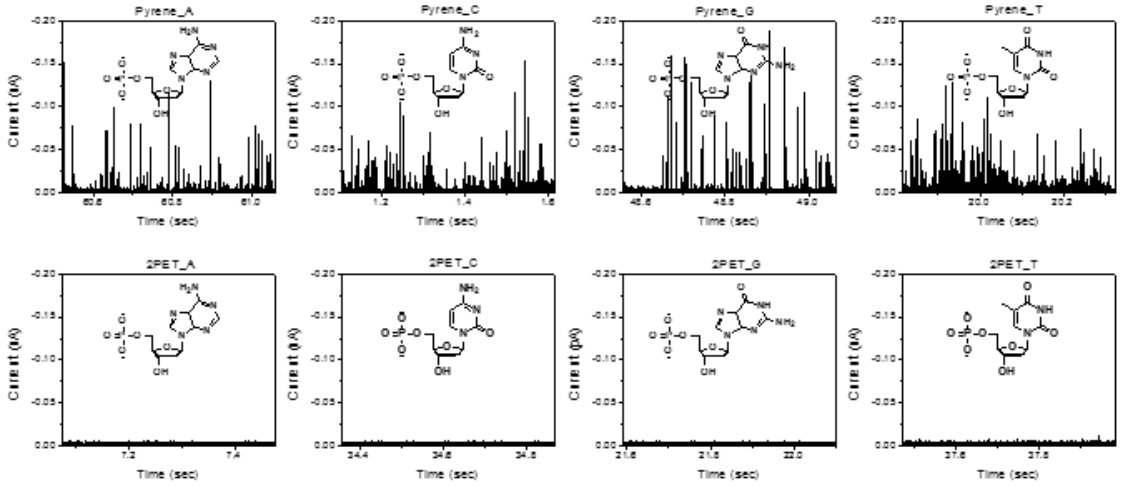

Figure 6.2. (a) Current-time traces obtained after adding DNA nucleotides in a STM tunnel junction with 2-PET (2-Phenylethenethiol) modified Palladium probe and Palladium substrate. (b) Current-time traces obtained after adding DNA nucleotides in a STM tunnel junction with 2-PET modified Palladium probe and Pyrene reader modified Palladium substrate.

functionalized Pd-STM probe and Pd-substrate to a PicoSPM scanning tunneling microscope,

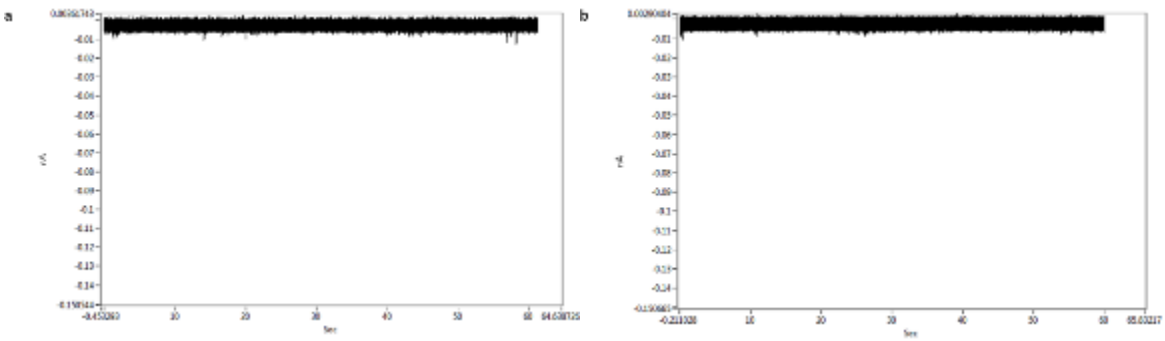

Figure 6.3. Current-time traces obtained after adding 100 μM solution of (a) Abasic DNA nucleotide and (b) D-Glucose in tunnel junction

stabilizing the tunnel junction in a phosphate buffer (1.0 mM, 7.4 pH) until a clean baseline was generated (~ 2 h), introducing an analyte solution (typically 100 μM in 1.0 mM phosphate buffer, pH 7.4) to the liquid cell, and collecting current recordings under a predefined tip-

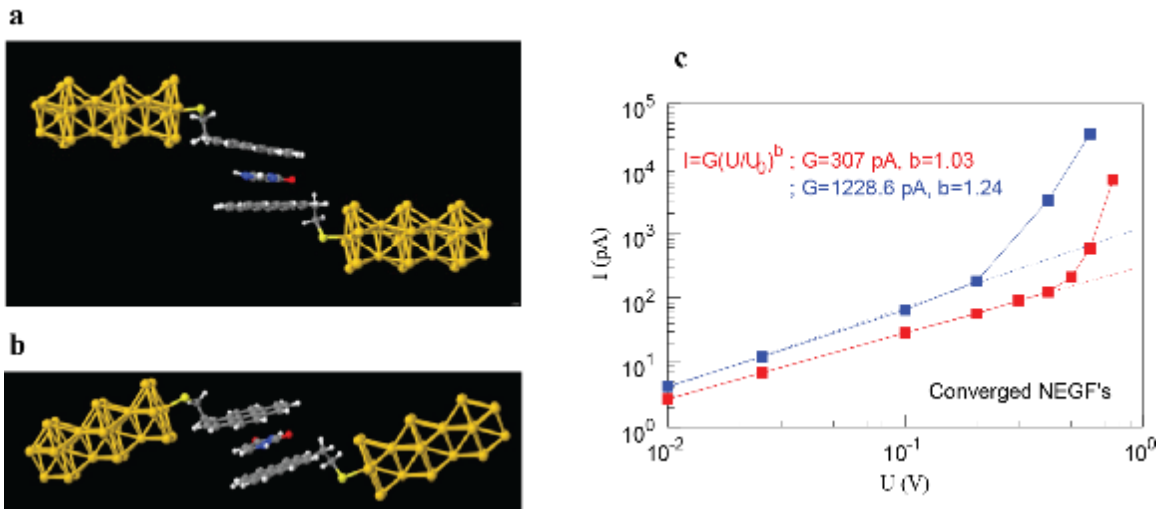

Figure 6.4. (a) Energy-minimized structure of the pi-pi stacked complex between two Pyrene Readers (attached to Palladium electrodes) and Guanine nucleo-base. (b) Energy optimized structure of the pi-pi stacked complex between two Pyrene Readers (attached to Palladium electrodes/metal slabs) and Thymine nucleo-base. (c) Theoretically observed current vs bias plots for complex (a) [blue] and complex (b) [red].

substrate bias. Four naturally occurring DNA nucleoside monophosphates (dAMP, dCMP, dGMP and dTMP) as well as two sugar molecules—deoxyribose-5-monophosphate and D-glucose—were used as analytes. For each analyte, four separate experiments were run with freshly made probes, substrates, and samples.[96]

6.3 Result and Discussion

We carried out the measurements using a setup as shown in Figure 6.1. We switched to Palladium electrodes from Gold electrodes as a reason of higher conductance and rigidity of

Palladium, resulting bigger gap distance in the tunnel junction and easier fabrication of the probes.[74, 75] Only PB buffer solution (1mM, 7.4 pH) generates overall clean baseline in the STM nano-junction of Palladium probe and substrate, functionalized with Pyrene Reader.

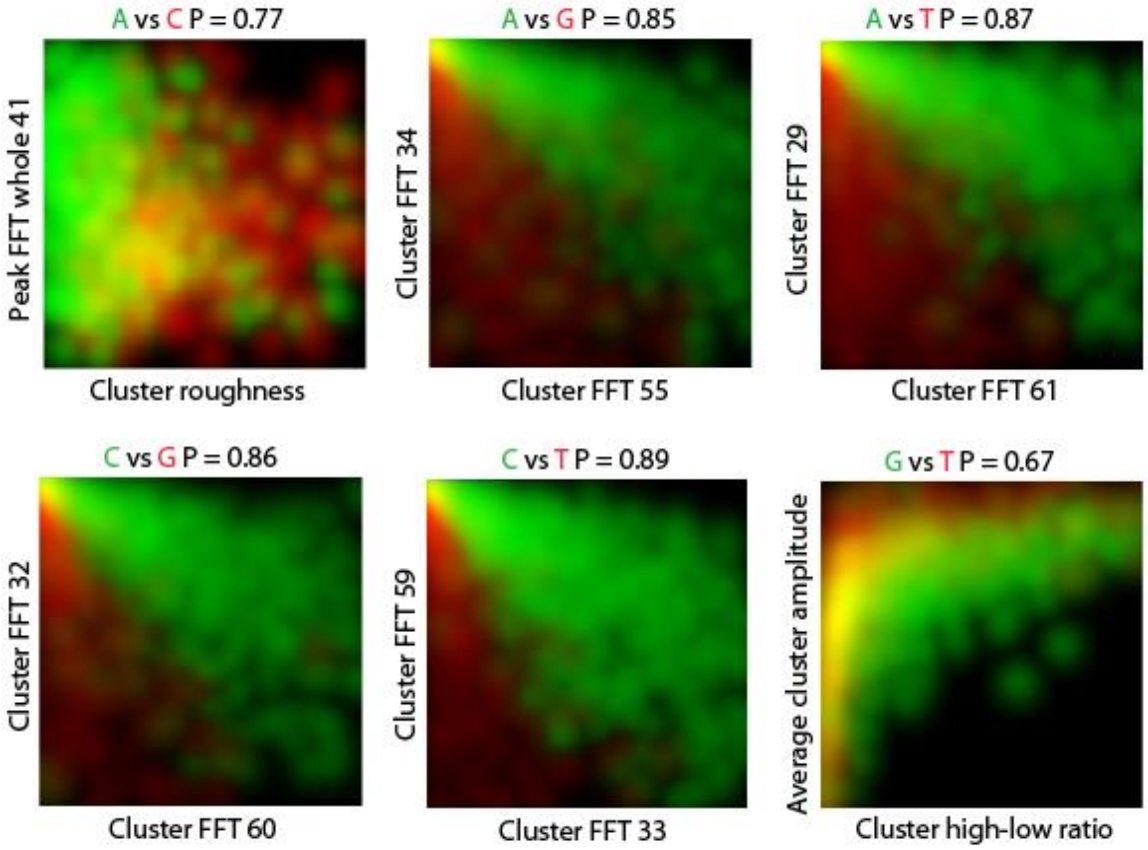

Figure 6.5. (a-g) Two dimensional histograms representing extent of separation between any two of the four DNA nucleotides, using only two signal feature/ parameters. (a) dAMP & dCMP, (b) dAMP & dGMP, (c) dAMP & dTMP, (d) dCMP & dGMP, (e)dCMP & dTMP, (f) dGMP & dTMP The percentage of separation (P) for each pair of nucleotides combinations are showed on the lower right corner of the corresponding plot.

Whereas, when we added all four DNA monophosphate solutions (in PB buffer) they give ample spikes, mostly part of spike-clusters. Sample spike-clusters for all four nucleotides are showed in figure 6.2(a). A three-membered aromatic stacking complex (Pyrene-Nucleotide-

Table 6.1. Result summary of all different recognition tunneling experiments done with different tip and substrate modification for various analytes

Analyte	Tip modification	Substrate modification	RT Signals
dGMP	Pyrene	Pyrene	Yes
	Benzene	Pyrene	Yes
	Benzene	Benzene	No
dAMP	Pyrene	Pyrene	Yes
	Benzene	Pyrene	Yes
	Benzene	Benzene	No
dCMP	Pyrene	Pyrene	Yes
	Benzene	Pyrene	No
	Benzene	Benzene	No
dTMP	Pyrene	Pyrene	Yes
	Benzene	Pyrene	No
	Benzene	Benzene	No
Abasic monophosphate	Pyrene	Pyrene	No
D-(+)-glucose	Pyrene	Pyrene	No

Pyrene) is believed to be formed and the resulting tunneling current through this non-covalent aromatic system gives rise to the current spikes. A substantial number of spikes are obtained with amplitude over 100pA, reciprocating our theoretically obtained value of several hundreds

of pico-amps for optimized Pyrene-Base (guanine)-Pyrene complex. Though the amplitude range is extremely broad probably due to the exponential dependence of tunnel current on

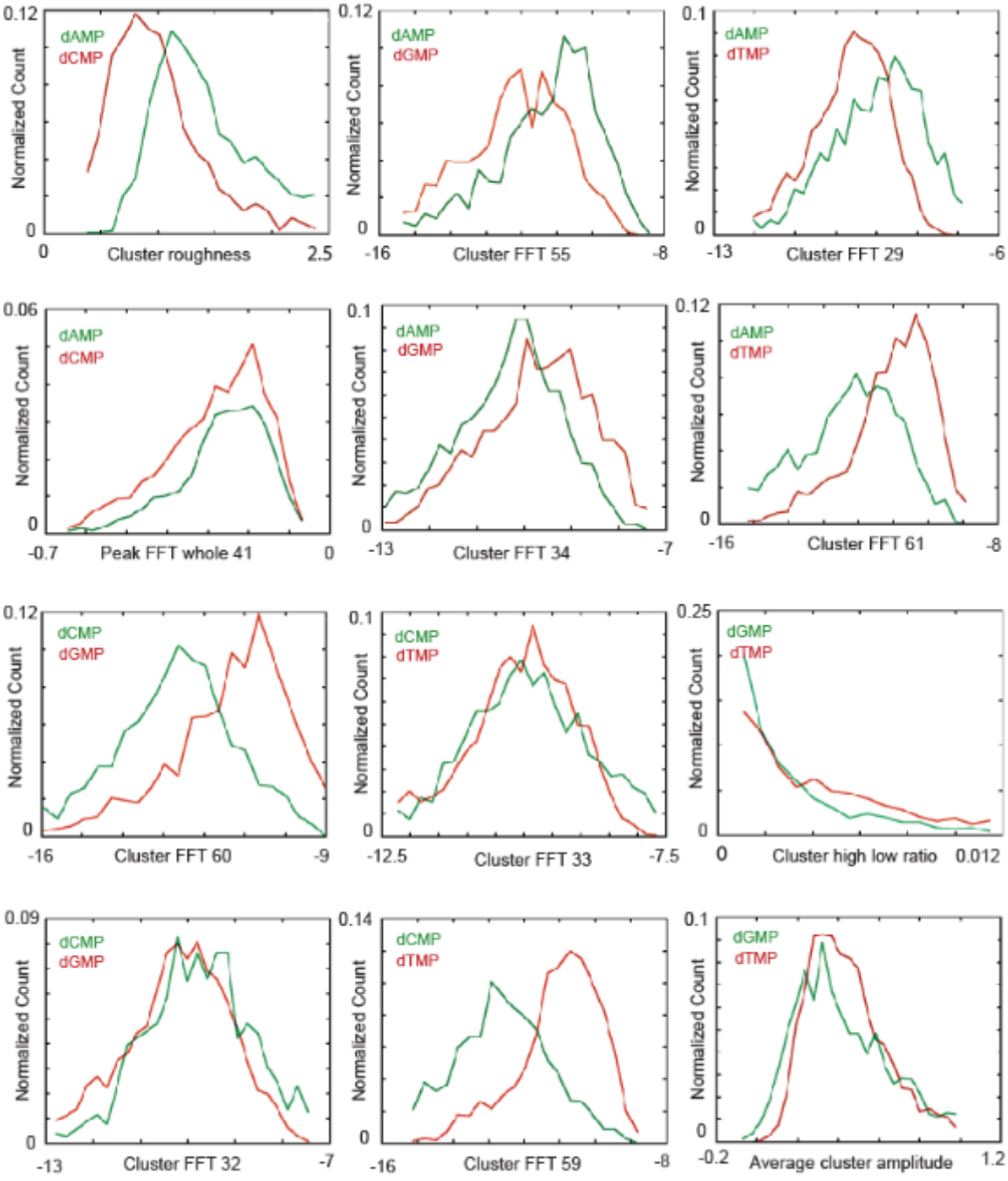

Figure 6.6. 1D histograms used to generate the 2D histograms for different pair of DNA nucleotides.

spatial position. When both the electrodes are functionalized with 2-Phenylethanethiol (which is implied as Benzene Reader), tunneling spikes are absent [figure 6.2.b]. This can be attributed to the smaller aromatic surface of Benzene, which is unable to provide strong enough π-π stacking with the nucleobases. Study by Grimme *et al.* [115] also reveals that aromatic rings as large as having 10-15 carbon atoms show significantly strong interaction. Analytes devoid of any aromatic moiety have been analyzed. Abasic DNA monophosphate and D-(+)-glucose fail to show such tunneling current spike-clusters even with Pyrene modified probe and substrates (figure 6.3 & 6.4, respectively) [96], indicating π-π stacking specific sensing property of Pyrene reader, which can be utilized to differentiate nucleotides from sugars and non-aromatic amino acids. Summary of all tunneling experiments with different analytes along with different set of probe and substrate modification is provided in Table1. [96]

Though obtained spike amplitudes were not specific to the different DNA monophosphates as corresponding amplitude distributions were closely overlaid, similar to the findings from our previous study on amino acids. [89] According to Zhao *et al.* [89], only tunnel current measurement through nucleotides may not be sufficient to discriminate them from each other as the orientation factor overwhelms the electronic-structural property of nucleotides. Hence, we use a complex machine learning algorithm, called Support Vector Machine (SVM) for nucleotide classification purpose as it uses a handful of other signal features rather than only considering tunnel current amplitude. In our recent work on amino acids, we demonstrated the great capability of SVM to discriminate between analytes with high calling accuracy. In the current work, we exploited 261 signal features related to the single spikes and clusters. All signal features are listed in chapter 8.

It has been shown by Zhang *et al.* [107] that among all possible configurations for aromatic stacked complexes parallel situated aromatic rings show the highest stability. Similar configurations were obtained for our theoretically studied energy-minimized structures (figure 6.4.a-b) [96] of such pi-pi stacked aromatic complexes. We used only nucleobases for the theoretical study instead of the complete nucleotides. Also, sugar and phosphate moieties do not have any significant contribution to stacking interaction. A distance of 0.7 nm was maintained between the aromatic surfaces of two Pyrene molecules. Complex with Guanine showed higher conductance compare to that of Thymine from the corresponding theoretical study on current vs. bias (figure 6.4.c).[96] Now, the theoretically obtained energy-minimized structures are the most probable ones. But, it is quite obvious to have a lot of different yet similar configurations that should be really close in energy to the most stable configurations. Hence, we didn't expect to get much difference between the distributions of peak average amplitude of dGMP and dTMP. But as we plotted these distributions after control noise filtration by SVM, we found a noticeable difference between the distribution peaks.

We use only any two of the 261 signal features to generate two-dimensional histograms and check the extent to which any of the two analytes can be differentiated from each other. Every possible combination of two nucleotides (total six) was analyzed (figure 6.5). All these plots are created by combining a couple of one-dimensional histograms, each of which corresponds to one of the various signal features. The green and the red portion of the plot correspond to the non-overlapped region of the well-separated analytes, whereas the yellow (and yellowish) part shows the overlapped data points from both analytes. The black region of the plots is devoid of data points from any of the analytes. We use a non-linear algorithm for calculating the percentage of separation (P) in each case. A value of P = 0.5 (or 50%) corresponds to a

perfectly random case with no effective separation. As evident from (e) almost 90% separation

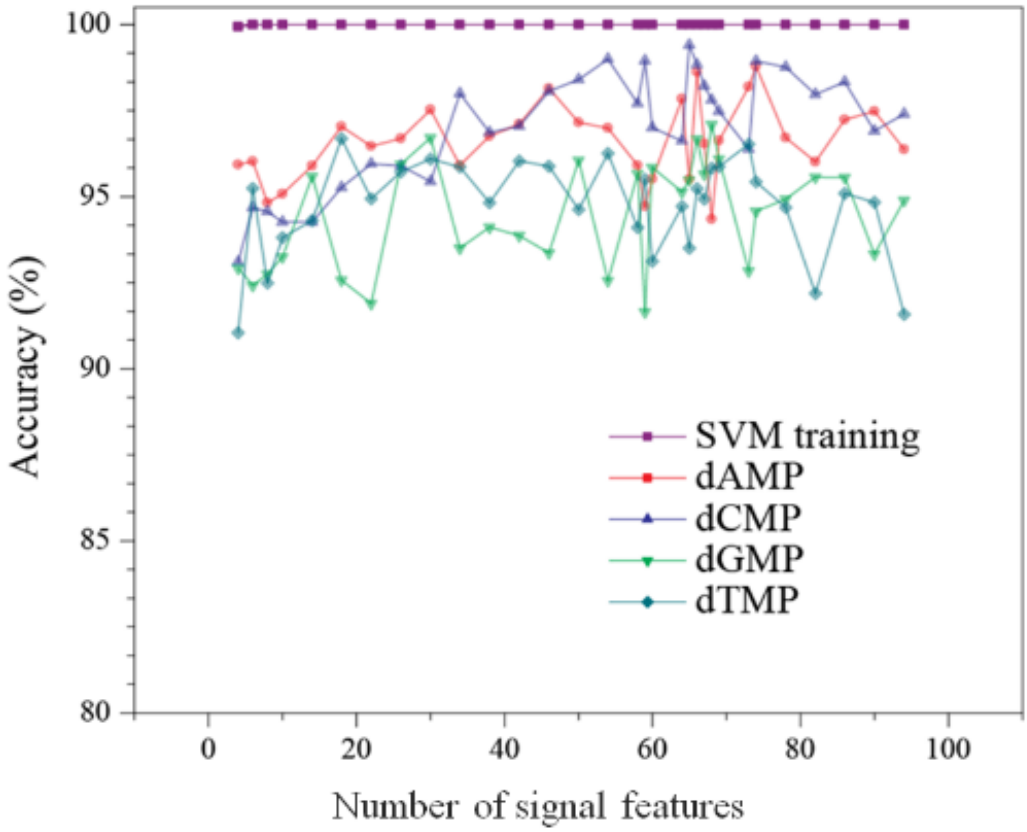

Figure 6.7. Plot for obtained nucleotide calling accuracy vs number of used parameters/signal features

between dCMP and dTMP was achieved using only couple of signal features. Whereas, the pair dGMP & dTMP showed least (67%) separated area on the histogram (f). The one-dimensional histograms for the used signal features are shown in figure 6.6. [96]

From using simply two signal features and discriminating only two analytes we proceed to distinguish all four nucleotides employing more signal features. As we introduce more and more signal features for the SVM analysis, it increases the base-calling accuracy. Average base-calling accuracy is plotted against the number of signal features (or number of parameters used in SVM) in figure 6.7. A very high calling accuracy value of 98.0% can be achieved.

Use of different probes and substrates in each experiment causes micro-scale geometrical changes to one tunnel junction going from another. This creates minute differences in shape

Table 6.2. Highest accuracy (%) that can be achieved with different readers for determining individual DNA nucleotides by RT

	dAMP	dCMP	dGMP	dTMP	Average
Pyrene	98.8	99.4	97.1	96.7	98.0
Imidazole	96.5	97.4	96.4	98.1	97.1

and characteristics of tunnel spikes and clusters obtained from different experiments even with the same analyte due to slight changes in the bonding (stacking) geometry. Hence, we perform predictive analysis to check the reproducibility of signal features for different experiments. SVM is trained with all the data sets of all four nucleotides, except one data set of any of the nucleotides which is going to be tested. Then the trained SVM is used to test predictive accuracy of the testing data set. For all nucleotides obtained 'most likely' predictive accuracy value is over 80%, whereas maximum predictive accuracy value for individual data set for different nucleotides went over 95%. This is encouraging considering the value would be 25% in case of perfect randomness. [96]

6.4 Conclusion

Using our current STM study of single DNA monophosphates, we can propose a system which demonstrates strong π-π stacking between the nucleotides and sensor molecules. This study does not solely depend on amplitude distribution of tunnel current for nucleotide discrimination and able to achieve 98.0% accurate nucleotide-calling. Also, it can increase the DNA translocation time by means of stacking interaction through the solid state nanopore,

coupled with a pair of atomically flat metallic electrodes modified with sensor molecules to measure tunneling current signatures from nucleotides.

CHAPTER 7

RECOGNITION TUNNELING OF RNA NUCLEOSIDE MONOPHOSPHATES

7.1. Introduction

In the recent years, RNA sequencing has become a compulsory and complementary research field for better interpretation of genomics. There are various types of RNAs (such as

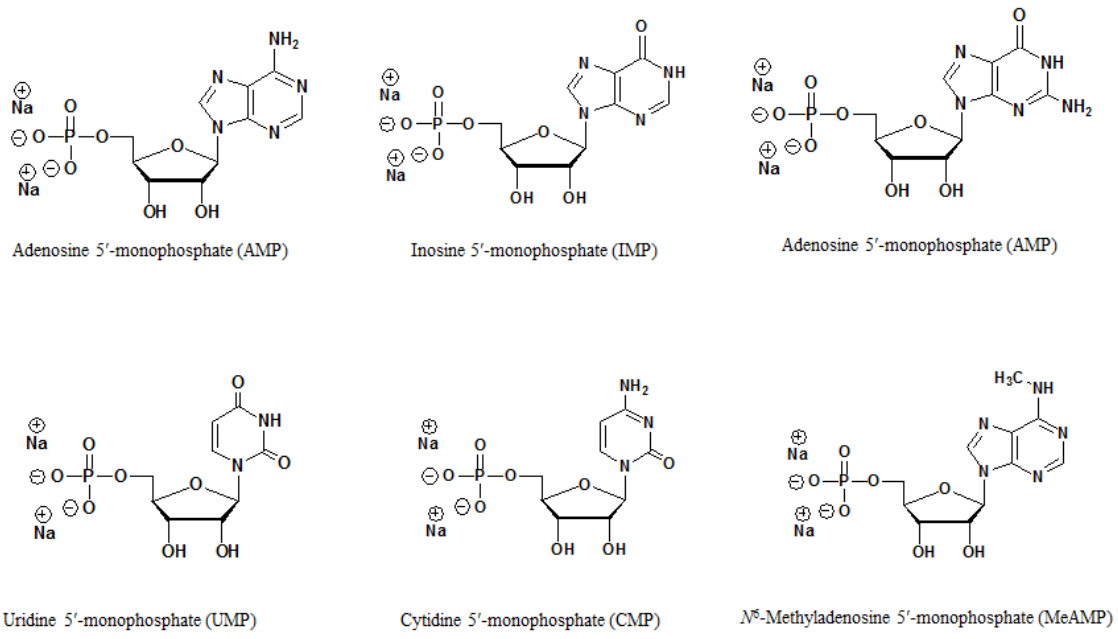

Figure 7.1. Naturally occurring and modified RNA nucleotide monophosphates used for recognition tunneling experiments

mRNA, tRNA, non-coding RNA, etc.) in a developed organisms like the human, controlling different processes inside a cell. Among them, mRNA or messenger RNA can be considered as the most important species in the mechanism of gene expression and protein production. Understanding RNA splicing, single nucleotide polymorphisms (SNPs) and post-

transcriptional modifications are of prime importance to predict physiological developments and diseases of an organism.[18] Most of the conventional NGS methods are used for

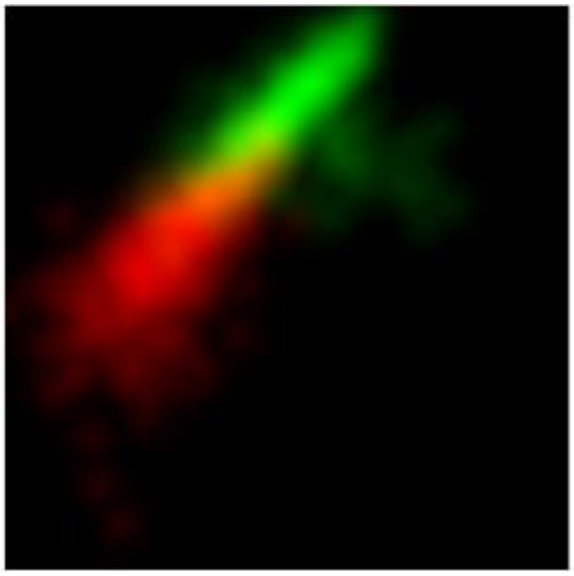

Figure 7.2. Example of a typical two Dimensional histogram with a pair of analytes [AMP (red) vs CMP (green)]

RNA sequencing and possesses similar drawbacks that are faced in case of DNA sequencing. Hence, we targeted to identify RNA building blocks by recognition tunneling technique with an ultimate goal of their nanopore sequencing.

7.2. Experimental

7.2.1. Preparation of Analytical Solutions

RNA monophosphates were obtained from Sigma Aldrich as sodium salts and dissolved in 1 mM phosphate buffer (pH 7.4), made using water from a Milli-Q system with a specific resistance of 18 MΩ-cm and total organic carbon contamination below 5 ppb.

7.2.2. RT Experiment

In a typical RT experiment, the measurement followed a process of mounting the functionalized Pd-STM probe and Pd-substrate to a PicoSPM scanning tunneling microscope, stabilizing the tunnel junction in a phosphate buffer (1.0 mM, 7.4 pH) until a clean baseline was generated (~ 2 h), introducing an analyte solution (typically 100 μM in 1.0 mM phosphate buffer, pH 7.4) to the liquid cell, and collecting current recordings under a predefined tip-

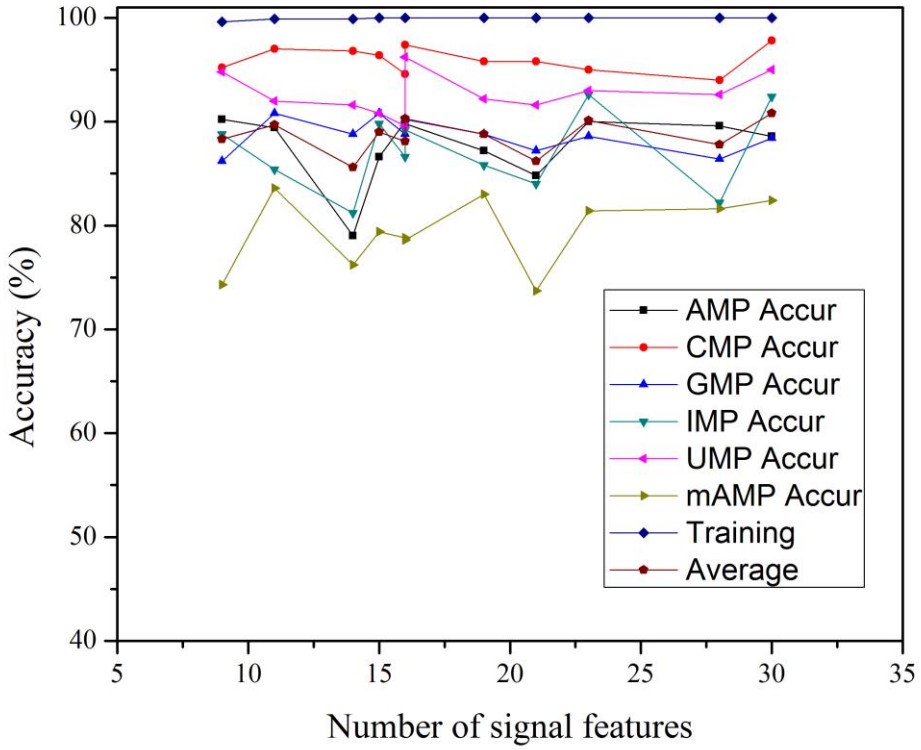

Figure 7.3. Plot for obtained RNA nucleotide calling accuracy vs number of signal features used

substrate bias. Four naturally occurring RNA nucleoside monophosphates (AMP, CMP, GMP and UMP) and two modified RNA nucleoside monophosphates (IMP and N-MeAMP) were used as analytes. For each analyte, four separate experiments were run with freshly made

probes, substrates, and samples. We used different batches of substrates and probes for each run, usually recording four runs for each analyte.

7.3. Result and Discussion

We have focused on six different ribonucleotide monophosphates (figure 7.1), four

Table 7.1. Naturally occurring and modified RNA nucleotide monophosphates used for recognition tunneling experiments

Analyte	**SVM Accuracy (%)** (highest value obtained for individual analyte)
Adenosine 5'-monophosphate (AMP)	90.2
Inosine 5'-monophosphate (IMP)	92.6
Guanosine 5'-monophosphate (GMP)	90.8
Uridine 5'-monophosphate (UMP)	96.2
Cytidine 5'-monophosphate (CMP)	97.8
N^6-Methyladenosine 5'-monophosphate (MeAMP)	83.6
Average	91.9

naturally occurring RNA building blocks (AMP, GMP, CMP & UMP) and a couple of modified ribonucleotide monophosphates (IMP & N-methyl AMP). Inosine monophosphate is an important molecule in metabolism and N-methyl adenosine is known as an activator of glycogen phosphorylase b. [117] All these six analytes showed recognition tunneling signals at 4 pA tunnel current set-point and 500 mV probe bias with Imidazole reader as the recognition molecule.

SVM analysis shows that two-dimensional separation could be achieved with high accuracy as presented in figure 7.2. Here AMP (red point) and CMP (green points) are separated with 93% accuracy. Though with increasing number of analytes this RNA calling accuracy tends to lower down (figure 7.3) and the individual recognition level for N-methyl AMP is markedly low (83.6%). The overall RNA separation accuracy could reach only ~92%, which is significantly low compared to the accuracy obtained with DNA monophosphates (~97%) with the same recognition molecule at same experimental condition (4 pA current set-point and 500 mV probe bias). The presence of an extra hydroxyl group (ribose sugar instead of de-oxy ribose) should possess some effect on molecular orientation in the tunnel gap, but it is difficult to predict the exact role of the hydroxyl group in differentiating the binding motifs of the nucleoside monophosphates. Table 1. summarizes the highest RNA calling accuracies of individual analytes and the average value.

7.4 Conclusion

Work in this project is still going on and a couple of more modified ribonucleotides are yet to be examined. One of them is 8-oxoguanine monophosphate, which is closely related to carcinogenesis. Still, it can be concluded that ribonucleotides can be identified by recognition tunneling technique and an average RNA calling accuracy of over 90% (91.9% to be precise) can be achieved, promising the possibility of single molecular RNA sequencing using nanopore device.

CHAPTER 8

DATA ANALYSIS: SUPPORT VECTOR MACHINE

8.1 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm for classification, outlier detection and regression, commonly used in bioinformatics, pattern recognition and many more related fields. For classification task (which is our matter of interest), using a small set of training data SVM builds up an optimized hyperplane (in one or higher dimensional space) or series of hyperplanes and using them categorizes new testing data points.

8.1.1 Theoretical Background

Figure 8.1.A presents an example of two classes of data. One is the empty circles and the other one is a class of black dots. As we can see, there are multiple boundary lines (H_1, H_2 and H_3) that can separate the two classes from each other. But, are all these solution boundaries equally good? To formalize the idea of the betterment of the solution a "margin" is introduced, which can be explained as the width of the band around the solution (or decision) boundary without any training data-points residing inside the band. Training data-points lying on the edge of the margin supports the decision boundary and are called the support vectors. This is elaborated in figure 8.1.B, where H_3 line (showing in black) proves to be the optimal solution boundary possessing the widest margin.

Suppose we have only two classes (y_i) of data which are linearly separable in a two-dimensional plane and a line can separate them from each other. The same idea corresponds to an N-1 dimensional separating hyperplane when each data point has N number of features

or dimensions. The separating line or one-dimensional hyperplane H_3 can be mathematically described as (figure 8.1.B),

$$w \cdot x - b = 0$$

w is the unit vector normal to the hyperplane and b is an arbitrary number. Similarly, we can define the planes constructing the margins of this hyperplane,

$$w \cdot x_i - b \geq 1 \text{ for black dots}$$

$$\mathrm{w} \cdot \mathrm{x_i} - \mathrm{b} \leq -1 \text{ for black circles}$$

The width of the margin can be derived from above relations and is equal to $2/\|\mathrm{w}\|$. The optimized hyperplane should have the widest margin. Hence, maximizing $2/\|\mathrm{w}\|$ or minimizing $\|\mathrm{w}\|$ or $\|\mathrm{w}\|^2/2$, maintaining the previous constraint of the margin (showed in the following relation) we can find the best solution for the hyperplane.

$$\mathrm{y_i}(\mathrm{w} \cdot \mathrm{x_i} - \mathrm{b}) - 1 = 0$$

Now, this problem can be constructed as a Lagrangian formulation

$$L = \frac{1}{2}\|\mathrm{w}\|^2 - \sum \alpha_\mathrm{i}[\mathrm{y_i}(\mathrm{w} \cdot \mathrm{x_i} - \mathrm{b}) - 1]$$

where α is Lagrange multiplier. After solving this optimization problem, the solution hyperplane can be obtained as a linear combination of the training support vectors,

$$\mathrm{w} = \sum \alpha_\mathrm{i} \mathrm{y_i} \mathrm{x_i}$$

More theoretical details could be found from renowned work of Vladimir Vapnik, a famous Russian mathematician.[118] [119]

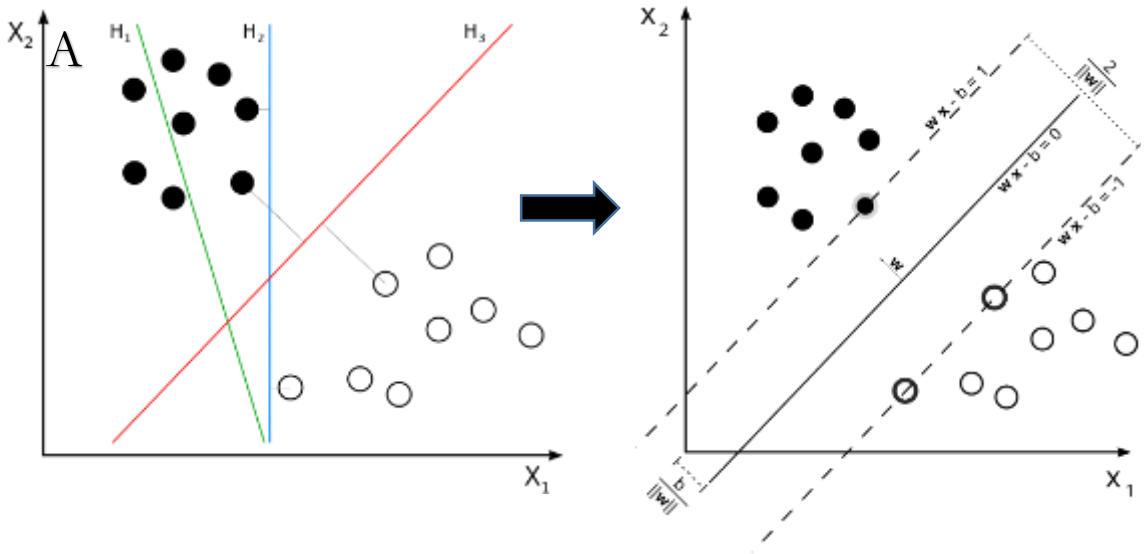

Figure 8.1. (A-B) Binary classification by support vector machine to find the best seperating boundary

Two classes of data can sometimes be linearly non-separable. This can happen with data

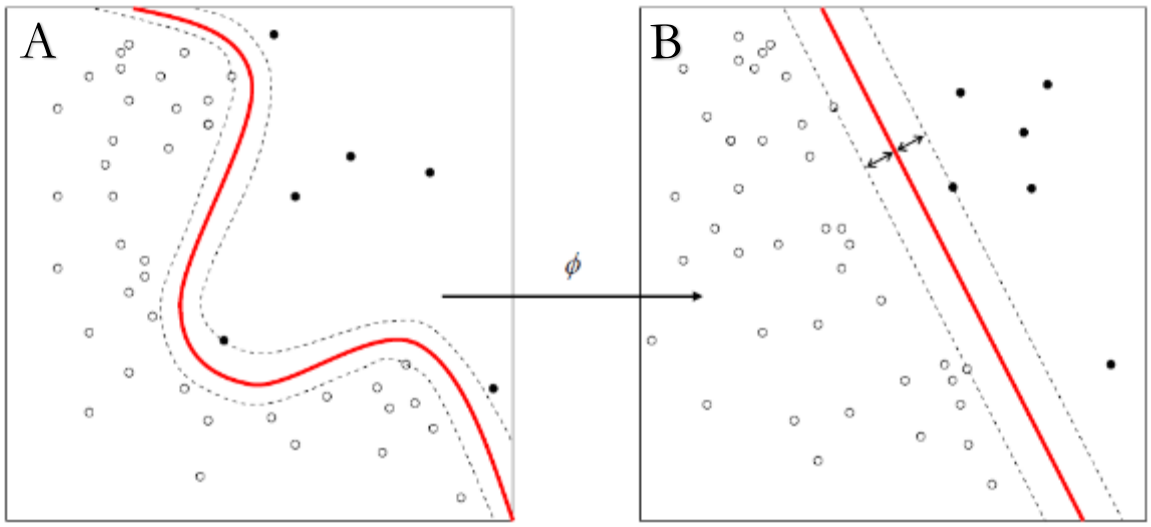

Figure 8.2. (A-B) Kernel transformation of non-linear data classes

that has some amount of noise, which is perfectly expected while dealing with practical experimental data sets. A common scenario is the presence of few data points of one class inside a dense cluster of the other class. A straight line can hardly provide a well-supported

decision boundary in such case. This has been exemplified in figure 8.2.A where no straight line can separate all the black dots from the empty circles. Hence, these two data classes are linearly non-separable in this space. A method of mathematical convenience called Kernel machine is used to transform the data into a different space where the classes are separable (figure 8.2.B). This transformation is often indicated with a function Φ(x) (figure 8.2).

8.2 Data Analysis Process

We used the kernel-mode SVM available from https://github.com/vjethava/svm-theta. [89] Each spike above 15 pA in amplitude was characterized using the features listed in table 8.1. JongOne Im and Brian Ashcroft carried out most of the part of SVM analysis related to different projects.

Table 8.1. List and description of all signal features used for Support Vector Machine analysis

	Feature Name	**Feature Description**
Primary Features	P_max Amplitude	Maximum amplitude of the peak
	P_average Amplitude	Average current of the peak
	P_top Average	Average of the peak above half maximum
	P_peak Width	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_average Amplitude	Average amplitude of all cluster peaks
	C_top Average	Average amplitude of all peaks above half maximum
	C_cluster Width	Cluster time length in millisecond
	C_roughness	Standard deviation of whole cluster signal
	C_max Amplitude	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 - 9	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)

The shape of each spike was characterized by constructing a fast Fourier transform (FFT). The resulting Fourier amplitude distribution was then downsampled using linear interpolation into nine bins of equal frequency intervals from zero to 25 kHz. FFT amplitudes (before downsampling) were averaged across three equally spaced frequency intervals (0–2.7 kHz, 8.4–11.1 kHz and 22.3–25 kHz), and these averages were used as additional features, as was the ratio of the highest to lowest FFT bins useful.

Clusters contain additional information. They were identified with a Gaussian broadening algorithm. The peaks used to locate the clusters were subject to a 15 pA threshold, but once a cluster was identified, all the data in it were used for the analysis, so amplitudes down to the baseline were included. We also developed a series of features to describe these clusters. These

included the spike frequency within a cluster, as well as the Fourier spectrum of the whole cluster (deconvolved for instrumental response by spectral division). Clusters contain many more data points than individual spikes, so the downsampling of the FFT was much finer, with a total of 61 bins used (each one corresponding to 25 kHz/61 or 410 Hz in width). Cepstrum amplitudes were calculated from the Fourier transform of the power spectrum, downsampling again to 61 frequency bins.

So as not to bias the analysis towards features with bigger numerical values and ranges, we rescaled all features as follows. The distribution of each signal feature was measured for any one analyte (one of the DNA nucleotides, RNA nucleotides or amino acids). The scale factor and additive constant required to move the mean of the distribution to zero and the standard deviation to 1.0 were calculated. Feature values for all the other analytes were remapped using the same linear transformation.[89]

Feature selection was performed in three stages. First, those features that showed too much linear correlation were removed. All the data from the entire pool were used to generate a correlation matrix where correlations are shown by off diagonal elements. Trial and error resulted in rejecting all feature combinations for which correlation coefficient $\geq$ 0.7. We chose one feature from each overly correlated set to represent the set in the next stage of analysis.[89]

Second, a comparison was performed for each feature for its variation over repeated experiments on the same analyte versus the variation between the different analytes. One dimensional histograms of all feature values were compiled for each experimental run for a given analyte. The absolute values of the differences between the normalized histograms were

accumulated to give an 'in-group' fluctuation. The same procedure was carried out for all possible pairs of analytes to give an 'outgroup' measure of fluctuation. Parameters were then ranked by the magnitude of the ratio of out-group to in-group fluctuation and the worst parameters were dropped. Finally, the usefulness of the remaining features was evaluated by determining the identification accuracy obtained with a randomly selected group of them. A tree search was used to maximize the efficiency of this process. After this process, a definite number of signal features are survived and those were used for the SVM classification analysis of the analyte series. Different analyte series (such as DNA nucleotides, RNA nucleotides and amino acids) and different recognition molecules (Imidazole, Benzimidazole, Pyrrole, Triazole and Pyrene) possesses different binding motifs and dipole moment vectors. As a consequence, signal features varies differently for all these diverse data types and correlation analysis of the signal features shows unique results for all these varied data sets. For example, SVM analysis for Benzimidazole reader data sets survived with more signal features compared to that of Imidazole reader data sets.

A subset of known data is used to find the support vectors that best partition the known data and thus train the SVM. Generally, a small fraction of data is taken from each and every data-set for this training. Many combinations of this training data is chosen randomly and SVM was trained several times to check the reproducibility. Data from subsequent analyses are then identified according to on which side of the partition they reside. The theory that I described previously has considered only two classes of data. But, our projects most often required analysis of multiple analytes and multiclass SVM was built. The strategy was to build another binary classifier that separates one analyte from the rest (one versus rest) and repeating the same analysis for each analyte.[89]

Full details of the SVM (written in Matlab) can be found in a download of the data analysis code available from https://svmsignalanalysis.codeplex.com/.

REFERENCES

1. Elgar, G. and T. Vavouri, *Tuning in to the signals: noncoding sequence conservation in vertebrate genomes.* Trends Genet., 2008. **24**(7): p. 344-352.

2. Edman, P., *Method for determination of the amino acid sequence in peptides.* Acta Chem. Scand., 1950. **4**: p. 283-93.

3. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.

4. Watson, J.D. and F.H.C. Crick, *A Structure for Deoxyribose Nucleic Acid.* Nature, 1953. **171**: p. 737-738.

5. Wilkins, M.H.F., A.R. Stokes, and H.R. Wilson, *Molecular Structure of Deoxypentose Nucleic Acids.* Nature, 1953. **171**: p. 738-740.

6. Whitney, J. 2011: Science-Art.com.

7. Yakovchuk, P., E. Protozanova, and M.D. Frank-Kamenetskii, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Res, 2006. **34**(2): p. 564-74.

8. Richmond, T.J. and C.A. Davey, *The structure of DNA in the nucleosome core.* Nature, 2003. **423**: p. 145-150.

9. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

10. Boyd, S.D., *Diagnostic applications of high-throughput DNA sequencing.* Annu Rev Pathol, 2013. **8**: p. 381-410.

11. Gilissen, C., et al., *Disease gene identification strategies for exome sequencing.* Eur J Hum Genet, 2012. **20**(5): p. 490-7.

12. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease.* Annu Rev Med, 2012. **63**: p. 35-61.

13. Wu, Y.C., et al., *High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations.* Blood, 2010. **116**(7): p. 1070-8.

14. Venturi, V., et al., *A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing.* J Immunol, 2011. **186**(7): p. 4285-94.

15. Wu, X.*e.a.*, *Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing.* Science, 2011. **333**: p. 1593-1602.

16. Koboldt, D. *Next-gen Sequencing for DNA Forensics*. Mass Genomics 2014 10/17/2014 [cited 2015 07/06/2015]; Available from: http://massgenomics.org/2014/10/next-gen-sequencing-dna-forensics.html.

17. Illumina. *The Next Generation of DNA Analysis for Criminal Justice*. 2015 [cited 2015 07/06/2015]; Available from: http://www.illumina.com/company/news-center/feature-articles/the-next-generation-of-dna-analysis-for-criminal-justice.html.

18. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

19. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc. Nati. Acad. Sci. USA, 1977. **74**: p. 5463-5467.

20. Diagram of Sanger dideoxy sequencing. (Courtesy of Wikipedia and Estevez, J.) https://www.promegaconnections.com/wp-content/uploads/2013/11/673px-sanger-sequencing11.png

21. Sanger, F., *DETERMINATION OF NUCLEOTIDE SEQUENCES IN DNA.* Noble Lecture, 1980. **1980**: p. 431-447.

22. Fei, Y., *DNA Sequencing, Sanger and Next-Generation Sequencing.*, in *Applications of Molecular genetics in Personalized Medicine*, K. Ulucan, Editor. 2014, OMICS Group eBooks. p. 1-10.

23. EMBL-EBI. *Improvements on the previous technology*. Train Online 07/10/2015]; Available from: http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/improveme.

24. Roche. *Products - Tehcnology: 454 Life Sciences, a Roche Company.* 07/09/2015]; Company Website]. Available from: http://454.com/products/technology.asp.

25. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

26. Illumina. *Hi Seq 2500 specifications.* 07/11/2015]; Available from: http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.html.

27. Tucker, T., M. Marra, and J.M. Friedman, *Massively parallel sequencing: the next big thing in genetic medicine.* Am J Hum Genet, 2009. **85**(2): p. 142-54.

28. Lindsay, S., *Biochemistry and semiconductor electronics--the next big hit for silicon?* J Phys Condens Matter, 2012. **24**(16): p. 164201.

29. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.

30. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2012. **13**(1): p. 36-46.

31. Ozsolak, F., *Third-generation sequencing techniques and applications to drug discovery.* Expert Opin Drug Discov, 2012. **7**(3): p. 231-43.

32. Foquet, M., et al., *Improved fabrication of zero-mode waveguides for single-molecule detection.* Journal of Applied Physics, 2008. **103**(3): p. 034301.

33. McCarthy, J.J., H.L. McLeod, and G.S. Ginsburg, *Genomic Medicine: A Decade of Successes, Challenges, and Opportunities.* Science Translational Medicine, 2013. **5**: p. 189sr4.

34. Hayden, R.C., *The $1000 genome.* Nature, 2014. **507**: p. 294-295.

35. Metzker, M.L., *Emerging technologies in DNA sequencing.* Genome Res, 2005. **15**(12): p. 1767-76.

36. Chan, E.Y., *Next-Generaion Sequencing Methods: Impact of Sequencing Accuracy on SNP discovery*, in *Single Nucleotide Polymorphism. Method in Molecular Biology*, A.A. Komar, Editor. 2009. p. 95-111.

37. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.

38. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nat Rev Genet, 2014. **15**(2): p. 121-32.

39. Puckelwartz, M.J., et al., *Supercomputing for the parallelization of whole genome analysis.* Bioinformatics, 2014. **30**(11): p. 1508-13.

40. Kasianowicz, J.J., et al., *Characterization of individual polynucleotide molecules using a membrane channel.* Proc. Natl. Acad. Sci. USA, 1996. **93**: p. 13770–13773.

41. Jain, M., et al., *Improved data analysis for the MinION nanopore sequencer.* Nat Methods, 2015. **12**(4): p. 351-6.

42. Manrao, E.A., et al., *Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase.* Nat Biotechnol, 2012. **30**(4): p. 349-353.

43. Laszlo, A.H., et al., *Decoding long nanopore sequencing reads of natural DNA.* Nature Biotechnology, 2014. **32**(8): p. 829-833.

44. Branton, D., et al., *The potential and challenges of nanopore sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1146-53.

45. Wikipedia. *Scanning Tunneling Microscope*. 07/29/2015]; Available from: https://en.wikipedia.org/wiki/Scanning_tunneling_microscope.

46. Mann, B. and H. Kuhn, *Tunneling through fatty acid salt monolayers.* Journal of Applied Physics 1971. **42**: p. 4398-405.

47. Aviram, A. and M.A. Ratner, *Molecular rectifiers.* Chemical Physics Letters 1974. **29**: p. 277-83.

48. Editorial, *Does molecular electronics compute?* Nat Nanotechnol, 2013. **8**(6): p. 377.

49. Ratner, M.A., *Introducing molecular electronics.* Materials Today, 2002. **5**(2): p. 20-27.

50. Binning, G. and H. Rohrer, *Scanning Tunneling Microscopy.* Surface Science, 1983. **126**: p. 236-244.

51. Luo, L., S.H. Choi, and C.D. Frisbie, *Probing Hopping Conduction in Conjugated Molecular Wires Connected to Metal Electrodes†.* Chemistry of Materials, 2011. **23**(3): p. 631-645.

52. Liu, H., et al., *Can the transition from tunneling to hopping in molecular junctions be predicted by theoretical calculation?* J Comput Chem, 2011. **32**(8): p. 1687-93.

53. Hines, T., et al., *Transition from Tunneling to Hopping in Single Molecular Junctions by Measuring Length and Temperature Dependence.* JACS, 2010. **132**: p. 11658–11664.

54. Ohshiro, T. and Y. Umezawa, *Complementary base-pair-facilitated electron tunneling for electrically pinpointing complementary nucleobases.* Proc Natl Acad Sci U S A, 2006. **103**(1): p. 10-4.

55. He, J., et al., *Identification of DNA Basepairing via tunnel current decay.* Nano Letters, 2007. **7**: p. 3854-3858.

56. Chang, S., et al., *Tunnelling readout of hydrogen-bonding-based recognition.* Nat Nanotechnol, 2009. **4**(5): p. 297-301.

57. He, J., et al., *A hydrogen-bonded electron-tunneling circuit reads the base composition of unmodified DNA.* Nanotechnology, 2009. **20**(7): p. 075102.

58. Chang, S., et al., *Tunnel conductance of Watson-Crick nucleoside-base pairs from telegraph noise.* Nanotechnology, 2009. **20**(18): p. 185102.

59. Chang, S., et al., *Electronic signatures of all four DNA nucleosides in a tunneling gap.* Nano Lett, 2010. **10**(3): p. 1070-5.

60. Lindsay, S., et al., *Recognition tunneling.* Nanotechnology, 2010. **21**(26): p. 262001.

61. Reed, M.A., et al., *Conductance of a molecular junction.* Science, 1997. **278**: p. 252-4.

62. Boussaad, S., et al., *Discrete tunneling current fluctuations in metal–water–metal tunnel junctions.* The Journal of Chemical Physics, 2003. **118**(19): p. 8891.

63. Cui, X.D., et al., *Reproducible measurement of single molecule conductivity.* Science, 2001. **294**: p. 571-4.

64. Wink, T., et al., *Self-assembled Monolayers for Biosensors.* The Analyst, 1997. **122**(4): p. 43R-50R.

65. Van Ruitenbeek, J.M., *Experiments on conductance at the atomic scale*, in *Quantum Mesoscopic Phenomena and Mesoscopic Devices in Microelectronics*. 2000, Springer: Berlin. p. 35-50.

66. Haiss, W., et al., *Measurement of single molecule conductivity using the spontaneous formation of molecular wires.* Physical Chemistry Chemical Physics, 2004. **6**(17): p. 4330.

67. Sprokel, G.J. and J.M. Fairfield, *Diffusion of Gold into Silicon Crystals.* Journal of the Electrochemical Society, 1965. **112**: p. 200-203.

68. Mathiot, D., *Gold, self-, and dopant diffusion in silicon.* Physical Review B, 1992. **45**(23): p. 13345-13355.

69. Frank, F.C. and D. Turnbull, *Mechanism of diffusion of copper in germanium.* Phys. Rev., 1956. **104**: p. 617-18.

70. Goesele, U., W. Frank, and A. Seeger, *Mechanism and kinetics of the diffusion of gold in silicon.* Applied Physics, 1980. **23**: p. 361-8.

71. Cappelletti, M.A., A.P. Cédola, and E.L. Peltzer y Blancá, *Theoretical study of neutron effects on PIN photodiodes with deep-trap levels.* Semiconductor Science and Technology, 2009. **24**(10): p. 105023.

72. Tavendale, A.J. and S.J. Pearton, *Deep level, quenched-in defects in silicon doped with gold, silver, iron, copper or nickel.* J. Phys.: Condens. Matter, 1983. **16**: p. 1665–73.

73. Vicente, J., et al., *In-Diffusion and Annealing Kinetics of Palladium in Silicon.* J. Electrochem. Soc., 1993. **140**: p. 868-70.

74. Lawson, J.W. and C.W. Bauschlicher, *Transport in molecular junctions with different metallic contacts.* Physical Review B, 2006. **74**(12).

75. Chang, S., et al., *Palladium electrodes for molecular tunnel junctions.* Nanotechnology, 2012. **23**(42): p. 425202.

76. Chang, S., et al., *Chemical recognition and binding kinetics in a functionalized tunnel junction.* Nanotechnology, 2012. **23**(23): p. 235101.

77. Porter, J.D. and A.S. Zinn, *Ordering of Liquid Water at Metal Surfaces in Tunnel Junction Devices.* J. Phys. Chem., 1993. **97**: p. 1190-1203

78. Sumetskii, M., A.A. Kornyshev, and U. Stimming, *Adatom diffusion characteristics from STM noise: theory.* Surface Science, 1994. **23**: p. 307-309.

79. Sumetskii, M., *Exponentially narrow current dip for resonant-tunneling structure of three quantum dots.* Physical Review B: Condensed Matter and Materials Physics, 1993. **48**: p. 14288-90.

80. Mosyak, A. and A. Nitzan, *Numerical simulations of electron tunneling in water.* J. Chem. Phys., 1996. **104**: p. 1549-1559.

81. *Introduction to Fourier Transform Infrared Spectroscopy.* 2002, Thermo Nicolet Corporation.

82. Accessories, T.S.S., *sampling tools for Thermo Scientific Nicolet FT-IR Systems*, T. Scientific, Editor.

83. Accessories, T.S.S., *sampling tools for Thermo Scientific Nicolet FT-IR Systems*, T. Scientific, Editor.

84. Co., J.A.W. *Ellipsometry Tutorial.* 08/11/2015]; Available from: http://www.jawoollam.com/tutorial_1.html.

85. *Spectroscopic ellipsometry.* 08/11/2015]; Available from: http://www.slideshare.net/nirupam12/spectroscopic-ellipsometry/14.

86. *XPS.* 08/22/2015]; Available from: http://www.slideshare.net/cynon/xps-3529469.

87. *VACUUM ADVENTURES.* 08/22/2015]; Available from: https://kartyush.files.wordpress.com/2012/05/xpsprinciple.png.

88. Wikipedia. 01/07/2016]; Available from: https://en.wikipedia.org/wiki/Contact_angle#/media/File:Contact_angle.svg.

89. Zhao, Y., et al., *Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling.* Nat Nanotechnol, 2014. **9**(6): p. 466-73.

90. Biswas, S., et al., *Hydrogen-Bonding Universal Readers for Identification of DNA Bases in Nanogaps by Recognition Tunneling with High Accuracy.* Manuscript in Preparation

91. Lagerqvist, J., M. Zwolak, and M.D. Ventra, *Fast DNA Sequencing via Transverse Electronic Transport.* Nano Lett, 2006. **6**(4): p. 779-782.

92. Yokota, K., M. Tsutsui, and M. Taniguchi, *Electrode-embedded nanopores for label-free single-molecule sequencing by electric currents.* RSC Advances, 2014. **4**(31): p. 15886.

93. Tsutsui, M., et al., *Single-molecule sensing electrode embedded in-plane nanopore.* Sci Rep, 2011. **1**: p. 46.

94. Liang, F., et al., *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, 2012. **18**(19): p. 5998-6007.

95. Chang, S., et al., *Gap distance and interactions in a molecular tunnel junction.* J Am Chem Soc, 2011. **133**(36): p. 14267-9.

96. Sen, S., et al., *The pi-pi Stacking as an Alternative to Hydrogen Bonding for Identification of DNA Bases in Nanogaps by Recognition Tunneling* Ready for Submission.

97. Ku, C.-S. and D.H. Roukos, *From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic madicine.* Expert Rev. Med. Devices, 2013. **10**: p. 1-6.

98. Sanger, F. and A. Coulson, *A rapid method for determining sequences in DNA by primed synthesis.* Journal of molecular biology, 1975. **94**: p. 441–448.

99. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.

100. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

101. Drmanac, R.*e.a.*, *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.* Science, 2010. **327**: p. 78-81.

102. Venkatesan, B.M. and R. Bashir, *Nanopore sensors for nucleic acid analysis.* Nat Nanotechnol, 2011. **6**(10): p. 615-24.

103. Feng, Y., et al., *Nanopore-based fourth-generation DNA sequencing technology.* Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 4-16.

104. Maitra, R.D., J. Kim, and W.B. Dunbar, *Recent advances in nanopore sequencing.* Electrophoresis, 2012. **33**(23): p. 3418-28.

105. Haque, F., et al., *Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA.* Nano Today, 2013. **8**(1): p. 56-74.

106. Wang, Y., Q. Yang, and Z. Wang, *The evolution of nanopore sequencing.* Front Genet, 2014. **5**: p. 449.

107. Zhang, X.G., et al., *First-principles transversal DNA conductance deconstructed.* Biophys J, 2006. **91**(1): p. L04-6.

108. Wells, D.B., et al., *Assessing graphene nanopores for sequencing DNA.* Nano Lett, 2012. **12**(8): p. 4117-23.

109. Wu, S., et al., *Molecular junctions based on aromatic coupling.* Nat Nanotechnol, 2008. **3**(9): p. 569-74.

110. Martın, S., et al., *Identifying Diversity in Nanoscale Electrical Break Junctions.* JACS, 2010. **132**: p. 9157–9164.

111. Schneebeli, S.T., et al., *Single-molecule conductance through multiple pi-pi-stacked benzene rings determined with direct electrode-to-benzene ring connections.* J Am Chem Soc, 2011. **133**(7): p. 2136-9.

112. Min, S.K., et al., *Fast DNA sequencing with a graphene-based nanochannel device.* Nat Nanotechnol, 2011. **6**(3): p. 162-5.

113. Cho, Y., et al., *Noncovalent Interactions of DNA Bases with Naphthalene and Graphene.* Journal of Chemical Theory and Computation, 2013. **9**(4): p. 2090-2096.

114. Lee, J.-H., et al., *Physisorption of DNA Nucleobases onh-BN and Graphene: vdW-Corrected DFT Calculations.* The Journal of Physical Chemistry C, 2013. **117**(26): p. 13435-13441.

115. Grimme, S., *Do special noncovalent pi-pi stacking interactions really exist?* Angew Chem Int Ed Engl, 2008. **47**(18): p. 3430-4.

116. Solomon, G.C., et al., *The Chameleonic Nature of Electron Transport through π-Stacked Systems.* JACS, 2010. **132**: p. 7887–7889.

117. Morange, M., et al., *AMP analogs: their function in the activation of glycogen phosphorylase b.* Eur. J. Biochem., 1976. **65**(2): p. 553-63.

118. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**: p. 273.

119. Boser, B.E., I.M. Guyon, and V.N. Vapnik, *A training algorithm for optimal margin classifiers.* Proceedings of the fifth annual workshop on Computational learning theory, 1992. **92**: p. 144.

APPENDIX A

COPYRIGHT & PERMISSION

Order Completed

Thank you very much for your order.

This is a License Agreement between Suman Sen ("You") and Nature Publishing Group ("Nature Publishing Group"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

Get the printable license.

License Number	3832111410469
License date	Mar 18, 2016
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Biotechnology
Licensed content title	The potential and challenges of nanopore sequencing
Licensed content author	Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner et al.
Licensed content date	Oct 9, 2008
Type of Use	reuse in a dissertation / thesis
Volume number	26
Issue number	10
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	As a strand of DNA emerges from a nanopore, a 'phosphate grabber' on one functionalized electrode and a 'base reader' on the other electrode form hydrogen bonds (light blue ovals) to complete a transverse electrical circuit through each nucleotide as it is translocated through the nanopore.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Identification of Bio-molecules Using Recognition Tunneling: Stride Towards Nanopore Sequencing
Expected completion date	May 2016
Estimated size (number of pages)	150