Chapter 9
Single-Molecule Sequencing

Masateru Taniguchi

Abstract In this chapter, the use of single-molecule conductance for DNA sequencing, a principle target of the $1,000 Genome Project, will be discussed. Since starting the project, numerous universities and companies have attempted to develop single-molecule sequencers but have not yet demonstrated a proof of concept. A major challenge has been the fabrication of nanoelectrodes with a 1 nm gap, equal to the diameter of single-stranded DNA molecules. The breakthrough discovery of the use of tunneling currents was required to perform single-molecule electrical sequencing. This discovery led to a proof of concept using a chemically modified scanning tunneling microscope (STM) and mechanically controllable break junction (MCBJ). These single-molecule measurement technologies are now being developed for application studies.

Keywords Single molecule • Sequencing • Tunneling currents • DNA • RNA

9.1 Introduction

Many people believed that the end of the Human Genome Project in 2003 meant the dawn of personalized medicine and therapeutics based on genomic information [1–4]. However, the long time and exorbitant cost to read an entire human genome have been a significant barrier to realizing personalized medicine [5–7]. To overcome this barrier, the US National Institutes of Health (NIH) that led the Human Genome Project has founded the $10,000 and $1,000 genome projects with the goal of reading an entire human DNA sequence in 1 day for the cost of $10,000 and $1,000, respectively [8–13]. The final target of the $1,000 genome project is to develop single-molecule DNA sequencers that can identify the sequences of the four-base molecules in DNA by measuring single-molecule conductance.

First- and second-generation DNA sequencers identify base molecules via light emission by laser excitation of dye molecules that are chemically bonded to base molecules [9–11]. They require polymerase chain reaction (PCR) to amplify

M. Taniguchi (✉)
The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan
e-mail: taniguti@sanken.osaka-u.ac.jp

M. Kiguchi (ed.), *Single-Molecule Electronics*, DOI 10.1007/978-981-10-0724-8_9

sequencing templates so that sufficient material is available for generating detectable signals. Furthermore, first- and second-generation DNA sequencing technologies require fluorescent labels. In contrast, third- and fourth-generation DNA sequencing technologies directly detect single-base molecules by changes in electric current such that neither PCR amplification nor fluorescent probes are necessary [14–17]. Comparing the throughputs and total cost to determine a complete human genome shows that first-generation DNA sequencing technologies take 3 months and cost approximately $10 million, second-generation technologies take 2 months and cost approximately $0.1 million, and third- and fourth-generation technologies will take 1 day and cost approximately $1,000 [9–11]. The use of nanopore devices is expected to result in a technical leap in DNA sequencing technologies.

Nanopore-based sequencers can be classified into two categories based on changes in ionic or tunneling currents (Fig. 9.1) [16, 18]. Ionic current-based nanopores have been studied for over 30 years [16, 19]. When we observe cell surfaces, we can see innumerable holes with diameters of several nanometers. These holes, or nanopores, are what nanopore devices use to sequence DNA. They are formed by channel proteins that allow the transport of specific substances such as ions, small molecules, and DNA across the cell membrane; therefore, nanopores have the ability of molecular recognition. For example, a nanopore may only allow a DNA molecule to enter a cell when it identifies certain DNA sequences. If this single-molecule recognition ability can work on a device chip, the device would be able to identify single molecules with high precision. This idea motivated the creation of bionanopore devices. Bionanopore devices identify base molecules passing through a nanopore by detecting changes in the ionic current flowing parallel to the nanopore when a voltage is applied across the membrane [16, 18, 19]. Negatively charged, single-stranded DNA molecules flow downward through the nanopore due to electrophoresis. When no molecule passes through the nanopore, a large ionic current is able to flow through it. When a molecule enters the nanopore, the ionic current decreases as the molecular volume increases. Bionanopore devices

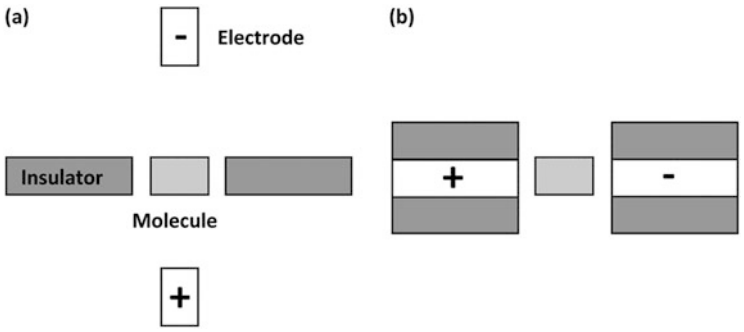

Fig. 9.1 Schematic figures of (**a**) ionic current-based nanopore and (**b**) tunneling current-based nanopores

can read DNA sequences because the devices can identify small differences in the molecular volume by observing the ionic current.

Nanopore-based sequencers that use tunneling currents are presented in Fig. 9.1b [16, 18]. Nanopores of diameter less than 10 nm are formed on a Si substrate covered with a thin Si_3N_4 film. Nanogap electrodes with spacing equal to a diameter of approximately 2 nm are assembled in a Si_3N_4 membrane. To reduce electrical noise, these electrodes are covered with a thin SiO_2 film. This nanostructure detects molecules passing through the nanopore using changes in the electric current flowing between the nanogap electrodes, not by changes in the ionic current flowing parallel to the nanopore. The current passing between the nanoelectrodes comes from a tunneling current conducted via molecules passing through the membrane. When a single DNA molecule passes through a nanopore, we can measure the tunneling current, which is potentially different for each nucleotide.

9.2 DNA Structures

The entire human genome is composed of three billion base pairs in DNA. Genomic information is obtained by sequencing DNA, where adenine always pairs with thymine with two hydrogen bonds and guanine always pairs with cytosine through three hydrogen bonds. The base molecules are chemically bonded to a sugar and phosphate group. Double-stranded DNA molecules are formed between two single-stranded DNA molecules via hydrogen bonds and can have multiple structures (Fig. 9.2). The *B*-type DNA structure has ten base pairs in a pitch, compared with the 11 base pairs in an *A*-type structure [20]. In the *B*-type DNA structure, the base pair height is 0.34 nm, and the rotation angle is 36°, compared with 0.26 nm and

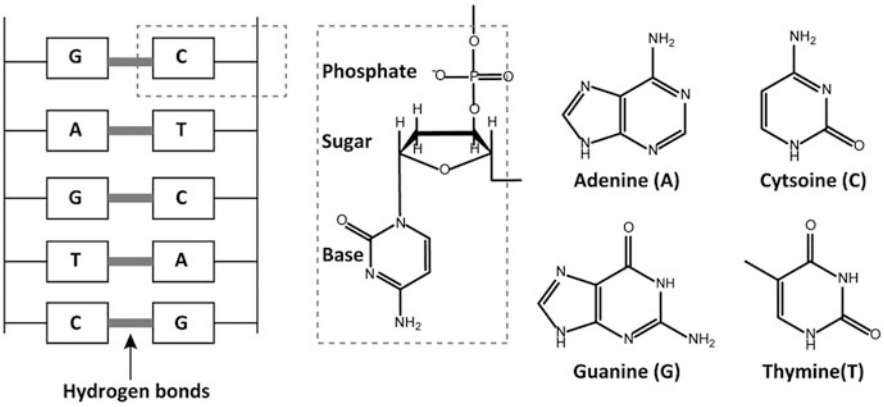

Fig. 9.2 Schematic figures of double-stranded DNA, a monomer unit, and four-base molecules

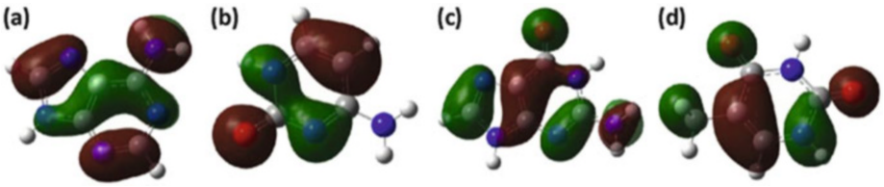

Fig. 9.3 The highest occupied molecular orbitals of (**a**) adenine, (**b**) cytosine, (**c**) guanine, (**d**) thymine

33°, respectively, observed in the *A*-type structure. In addition, DNA molecules have cations such as Na^+ to balance the negatively charged phosphate group.

Due to the electrical properties of DNA molecules, only single-stranded DNA can be used for single-molecule sequencing. The four-base molecules in DNA are approximately 1 nm and contain π-electron bonding systems. Their highest occupied molecular orbitals (HOMOs) are distributed over the whole molecule, whereas bonds in the sugar and phosphate groups are localized σ bonds (Fig. 9.3) [15]. Intermolecular π-orbital interactions between base molecules in DNA are expected because base molecules are stacked in parallel. In fact, many researchers expected that double-stranded DNA molecules could function as molecular wires and have examined their electrical conductivity. Although DNA molecules can function as insulators, semiconductors, metals, and superconductors, the best electrical description of DNA molecules is apparently that of a wide-gap semiconductor [21]. Single-molecule sequencing through conductance is difficult because of the strong interactions between base pairs in double-stranded DNA; however, the π-orbital interactions between base molecules are expected to be weak in single-stranded DNA molecules due to their more flexible molecular structures. As a result, single-molecule conductance through base molecules in single-stranded DNA is expected to originate from single-base molecules and can be potentially used for sequencing [22].

9.3 The Principle of Single-Molecule Sequencing

There are two ways to identify single-base molecules using tunneling currents (Fig. 9.4) [16, 18]. One is to identify single-base molecules using the tunneling currents generated between the STM and metal substrate that has been modified with recognition molecules. This technique, called recognition tunneling, recognizes base molecules by their hydrogen bonds [23–25]. The second method uses tunneling currents flowing between nanoelectrodes formed by the MCBJ [26–29].

The electric currents obtained using the two measurement methods come from the electron transport mechanism via the HOMO in the coherent tunneling regime where the single-molecule conductance of a base molecule can be expressed by

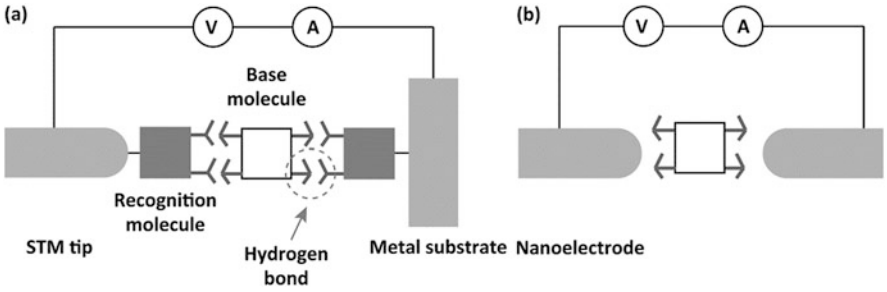

Fig. 9.4 Schematic figures of (**a**) STM and (**b**) MCBJ configurations

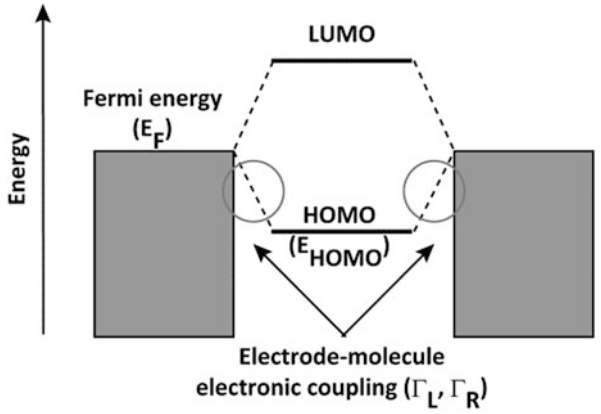

Fig. 9.5 Schematic diagram showing the molecular orbital levels, the Fermi level in the contacts, and the electrode–molecule electronic coupling

Eq. 9.1 [15]:

$$G = \frac{2e^2}{\hbar} \frac{\Gamma_L \Gamma_R}{(E_{\mathrm{HOMO}} - E_F)^2} \tag{9.1}$$

The values E_{F}, E_{HOMO}, and Γ_{L} represent the Fermi energy of the electrodes, the energy of the HOMO, and the left electrode–molecule electronic coupling strength, respectively (Fig. 9.5). Calculations based on density functional theory indicate that the HOMO energy order is guanine > adenine > cytosine > thymine > uracil [27]. Assuming that the electrode–molecule electronic coupling strengths are almost same for the four-base molecules, this order then corresponds to the order of single-molecule electrical conductance.

In an actual measurement system, electrode–molecule electronic coupling strengths strongly depend on the molecular conformation with respect to the nanoelectrodes and their tip structures, resulting in different electrode–molecule electronic coupling strengths for each base molecule. In particular, the coupling

strengths are affected by interactions between the molecules and electrodes when the molecules adsorb onto nanoelectrodes. In addition, molecular conformation is expected to be stabilized in the strong electric field between nanoelectrodes due to the dipole moments of the base molecules [22, 30–32]. Consequently, the single-molecule conductance histogram is expected to show a broad distribution over a single peak; therefore, it is difficult to expect an order in the single-molecule conductance of the base pairs from a chemically modified STM tip, because single-molecule conductance is strongly affected by the electronic structures and conformation of recognition molecules.

In ideal nanopore devices, a single DNA molecule passes through a nanopore at a constant velocity [17, 33]. Similarly, a single DNA molecule is expected to pass through a nanoelectrode gap in one direction and ideally with a constant speed for greater accuracy. Moreover, the throughput has to be higher when single-molecule electrical sequencing technology is applied to a practical sequencer. In other words, single DNA molecules have to enter a nanoelectrode gap at high frequencies. Consequently, identifying single-base molecules and controlling the translocation speed are key for the realization of higher accuracy and throughput. This chapter will only focus on methods for identifying single-base molecules using single-molecule conductance measurements [17, 33].

9.4 Measurement and Analysis Methods

The protocols to measure tunneling currents of single-base molecules and single DNA molecules are almost the same for STM-based [23–25] and MCBJ-based [26–29] methods. First, in the STM-based method, an STM tip and a metal substrate are modified with recognition molecules. Second, the distance between the STM, substrate, and spacing between nanoelectrodes are fixed by measuring tunneling currents in solutions containing analytes. Then, the gap spacing is calculated on the basis of tunneling currents. The gap spacing is kept constant via feedback control using piezo devices. In actual experiments, measured electric currents result from tunneling and ionic currents because the sample solutions contain ionic buffers. To reduce ionic currents, an STM tip and nanoelectrodes formed by the MCBJ are covered with insulating materials. After fixing the gap spacing, current–time profiles are measured at a sampling rate of less than 250 KHz by applying a bias voltage of 0.1–0.7 V.

When current–time profiles of single-base and DNA molecules are measured, spikelike signals characterized by the maximum current (I_P) and duration (t_d) are observed (Fig. 9.6a) [23–29]. Single-molecule conductance histograms are formed by several 100–1,000 points of I_p data. Generally, single-molecule conductance histograms of the four-base molecules show single peaks with broad distributions [23–29]. Large overlaps of the peaks seem to prevent identifying single-base molecules based on single-molecule conductance (Fig. 9.6b). Let us go back to the method for sequencing DNA using optical measurements.

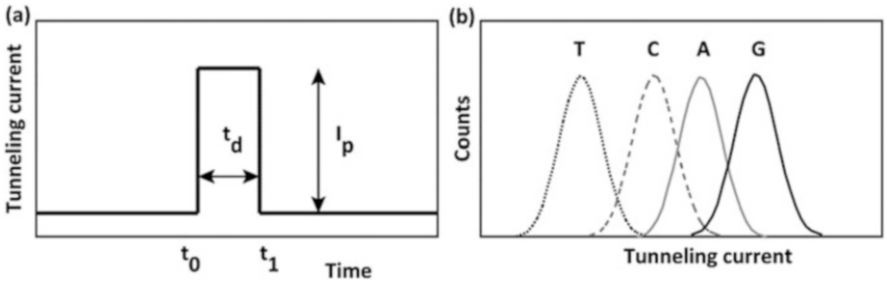

Fig. 9.6 (**a**) Tunneling current–time profile characterized by the maximum current I_p and duration of the current t_d and (**b**) schematic figure of tunneling current histograms of four-base molecules

Optical DNA sequencers can identify four-base molecules using four types of dye molecules whose absorption and fluorescent spectra overlap, similar to the conductance histograms [34]. Thus, before processing the spectra, they comprise the sum of the four spectra; however, after processing, all four-base molecules can be identified. The individual spectrum of the four-base molecules corresponds to the probability density function of wavelength $f_x(\omega)$. In the simplest case, the spectrum intensity at wavelength ω in the measured spectra can be expressed using the probability density function of the four-base molecules:

$$c_A f_A(\omega) + c_C f_C(\omega) + c_G f_G(\omega) + c_T f_T(\omega), \tag{9.2}$$

where C_x (X = A, C, G, and T) denotes the coefficient at ω. For example, the measured spectral intensity at a particular wavelength is obtained by the sum of adenine (80 %), cytosine (10 %), guanine (8 %), and thymine (2 %). In this case, the spectral intensity at this wavelength is assigned to adenine.

Similarly, the measured electric currents can be represented as the sum of electric currents multiplied by probability density functions, corresponding to the single-molecule conductance histograms of the four-base molecules: $G_x(g)$ (X = A, C, G, and T). Therefore, in the simplest model, the measured single-molecule conductance can be expressed by Eq. 9.3:

$$a_A G_A(g) + a_C G_C(g) + a_G G_G(g) + a_T G_T(g), \tag{9.3}$$

where a_X denotes the coefficient at the measured single-molecule conductance g. Consequently, small distributions of the probability density functions for single-molecule conductance allow us to sequence DNA with high precision, similar to an optical spectrum. However, peak wavelengths in the fluorescent spectrum are fixed, whereas peak positions in single-molecule conductance histograms alter due to changes in electrode–molecule electronic coupling strengths.

The current–time profiles measured for single-stranded DNA molecules are analyzed using the following procedure [27]. Based on the Eq. 9.3, the measured

conductance (G) is assumed to be formed from the sum of the single-molecule conductance of base molecules comprising single-stranded DNA. The ratio of the molecular conductance of each base molecule the overall measured G must be determined to assign the measured G to single base. First, single-molecule conductance histograms are constructed from all data points. In an effort to obtain single-molecule conductance histograms of 4 base molecules, Gaussian functions are fitted to conductance histograms. Second, the current–time profiles are transformed into probability–time profiles. The probabilities of the four-base molecules are calculated by inserting probability–time profiles into probability functions of the four-base molecules. Third, the probabilities of individual-base molecules between t_0 and t_1 are integrated using probability–time profiles. Finally, the most likely base molecules are assigned on the basis of the highest integrated probabilities.

However, a significant problem specific to single-molecule DNA analysis remains. Existing DNA sequencers analyze data using machine learning, i.e., a hidden Markov model [34] instead of a simple linear model. Both linear and hidden Markov models assume that the measured values of single-base molecules are independent events. In optical measurements, the observed spectra are independent because the optical methods measure the fluorescence spectrum of dye molecules connected to the terminus of DNA molecules, not the bases themselves. However, in single-molecule identification via tunneling currents, the measured conductance of single-base molecules is not independent because the conductance of single-base molecules can be affected by neighboring base molecules through chemical bonds. In an effort to overcome this issue, single-molecule science should be fused with information technology.

9.5 Single-Molecule Identification of Base Molecules

Single-molecule identification of base molecules has been demonstrated using STM and MCBJ methods. Lindsay et al. demonstrated single-molecule resolution of single DNA bases via hydrogen-bond-mediated tunneling currents using a nucleobase-functionalized Au STM probe and nucleoside-functionalized Au substrate in organic solvents [23–25]. This quantum sequencing approach has shown single-base resolution using tunneling currents generated between the Au STM probe and the Au substrate functionalized with 4-mercaptobenzoic acid (Fig. 9.7). This enabled the differentiation of single-molecule conductance in the order of deoxyadenosine > deoxycytidine > deoxyguanosine > thymidine [25]. Furthermore, this approach allowed identification of single-base molecules of DNA with a conductance order of dCMP > dGMP ∼ dmCMP > dAMP, although there were no detectable signals for dT [23]. In addition, the order of peak conductance was $d(C)_5$ > $d(mC)_5$ ∼ $d(A)_5$ for oligomers. Probe drift prevented single-molecule electrical sequencing of d(ACACA) and d(CmCCmCC) [23]; however, two conductance

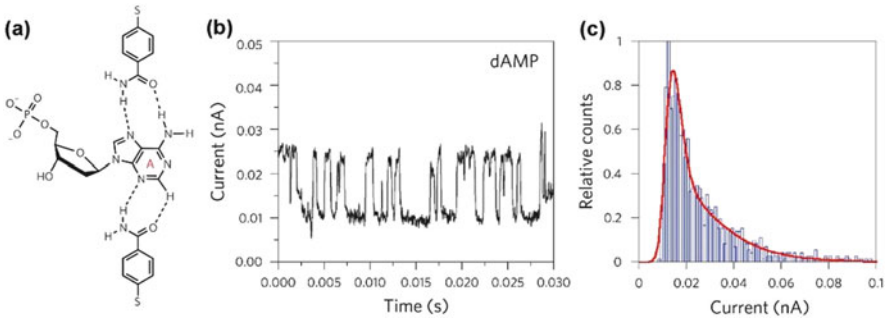

Fig. 9.7 (**a**) Schematic of recognition and base molecules, (**b**) current–time profile of dAMP, and (**c**) current histogram of dAMP (Reprinted with permission from Macmillan Publishers Ltd: ref. [23], copyright 2010)

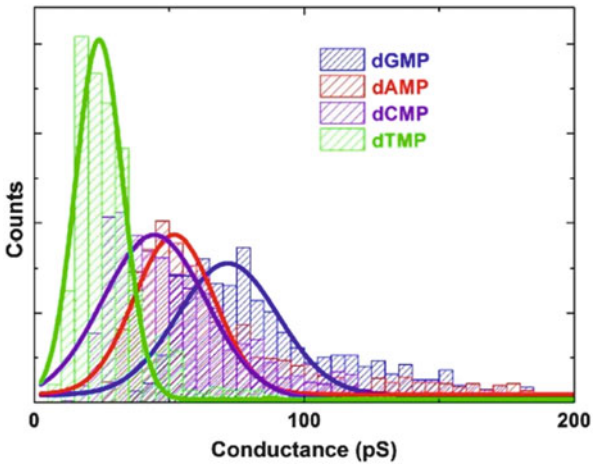

Fig. 9.8 Single-molecule conductance histograms of four-base molecules. Ref. [27]

plateaus corresponding to each of the two-base molecules were observed in current–time profiles.

In the MCBJ method, when current–time profiles are measured in aqueous solutions containing base molecules, spikelike signals are observed [26, 27, 29]. The corresponding single-molecule conductance histograms of the four-base molecules show single peaks (Fig. 9.8). The peak conductance order is guanine > adenine > cytosine > thymine [27]. Note that this experimental order agrees with HOMO energies obtained from quantum chemical calculations.

If the electron transport mechanism in single-base molecules is coherent tunneling, then methylated dCMP and 8-oxo-dGMP can be identified because their HOMO energies differ from those of other base molecules. Methylated dCMP is the most extensively studied epigenetic mark because of its direct relevance to human

health and disease. 8-Oxo-dGMP is known as a biologically important marker for assessing oxidative stress in humans, such as damage caused by smoking, relevant to aging and diseases. However, existing DNA sequencers cannot identify chemically modified base molecules without chemical modifications and pretreatments. Hence, identifying two chemically modified base molecules is very important. Theoretical calculations show that the HOMO energies of dCMP and methylated dCMP are −6.1 and −6.0 eV, respectively, whereas those of dGMP and 8-oxo-dGMP are −5.7 and −5.6 eV, respectively. Assuming that the electrode–molecule electronic coupling strengths are the same in the original and chemically modified base molecules, single-molecule conductance orders are expected to be methylated dCMP > dCMP and 8-oxo-dGMP > dGMP, consistent with experimental results [29].

9.6 DNA Sequencing

Although STM and MCBJ methods have attempted to sequence DNA, only the MCBJ method has achieved DNA sequencing [27]. Here, the cases of TGT and GTG are presented. Spikelike signals were obtained for DNA oligomers as well as single-base molecules. High and low bands show single-molecule conductance from guanine and thymine, respectively (Fig. 9.9). As clearly shown in the single-molecule conductance histogram, three peaks corresponding to baseline, guanine, and thymine are obtained from the current–time profiles of single-base molecules. Next, using the above analysis method, current–time profiles are analyzed to

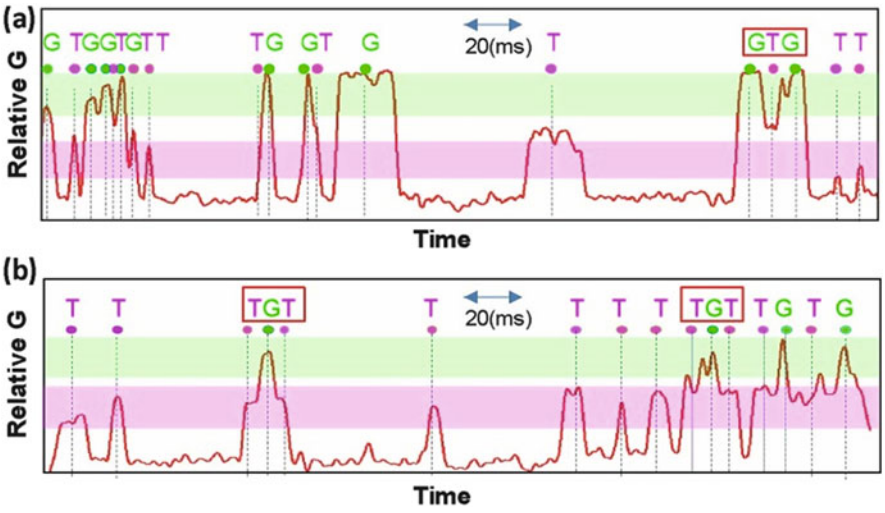

Fig. 9.9 Relative single-molecule conductance–time profiles of GTG and TGT. Ref. [27]

determine the sequences of the base molecules in DNA. In one case, two lower plateaus of thymine were observed in the first and third positions, with guanine forming the middle higher plateau. In the other case, guanine was observed to form the first and third higher plateaus, with thymine in the lower center position. This result demonstrates that the MCBJ method can achieve electrical sequencing of a single DNA molecule. Under the present experimental conditions, not only GTG and TGT but also G, T, TG, and GT can be identified due to the stochastic traps of single DNA molecules.

In the present study, the movement of single DNA molecules is attributed to Brownian motion. When a single TGT DNA sequence approaches the nanoelectrodes and returns, electric signals corresponding to a single thymine molecule are obtained. When a single TGT DNA sequence passes through the nanoelectrodes, electric signals corresponding to TGT are obtained. Similarly, electric signals corresponding to a guanine molecule were obtained when a single TGT DNA sequence approached the electrodes and returned.

9.7 RNA Sequencing

RNA is formed from adenine, cytosine, guanine, and uracil, which have two OH groups in a sugar group [20]. Recent studies reveal that microRNA (miRNA) is a small RNA molecule that has approximately 22 base molecules and regulates gene expression at the translational level. MiRNA expression profiling of human tumors has identified the signature associated with diagnosis, staging, progression, prognosis, and response to treatment; however, existing DNA sequencers cannot directly identify sequences of miRNA base molecules [35]. Determination of such sequences is based on transcription of RNA to DNA, i.e., reverse transcription, and transcription errors are expected to result in low accuracy. Instead, single-molecule sequencing technologies can directly sequence RNA, eliminating transcription errors and the need for pretreatments related to reverse transcription.

In addition to DNA, single-molecule conductance histograms of the four-base molecules of RNA are determined using the MCBJ method. Experimental results show that the conductance order for single-molecule RNA is guanine > adenine > cytosine > uracil, corresponding to their HOMO energies. In addition, an RNA molecule of UGAGGUA is sequenced using the MCBJ and analysis methods. Here, three typical current–time profiles are shown (Fig. 9.10). Similar to the current–time profiles of DNA, those of miRNA showed electric currents corresponding to each base molecule. UGAGG is obtained from the upper current–time and U, G, A, and AU in the middle current–time. A long fragmented sequence UGAGGUA was obtained in the lowest profile. Here, 35 fragment sequences were used to assemble the sequence contigs. Sequence contig 1 was assembled as UGA using 13 fragment sequences. Similarly, contigs 2, 3, and 4 were assembled as GAGG, AGGUA, and AGGU, respectively, using the 17-, 9-, and 13-base fragment sequences. The four contigs were assembled as UGAGGUA.

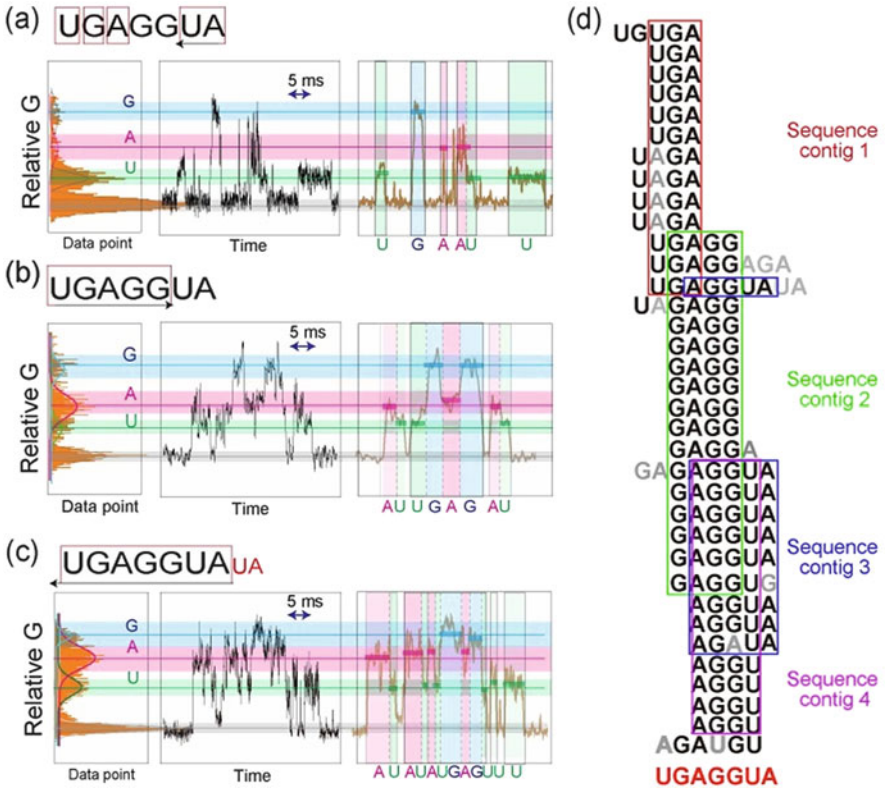

Fig. 9.10 Relative single-molecule conductance–time profiles of (**a**) UGA, (**b**) UGAGG, (**c**) UGAGGUA, and (**d**) the 35 fragmented sequences used for resequencing the microRNA of UGAGGUA. Ref. [27]

Single-molecule conductance can be expressed in Eq. 9.1, where electrode–molecule electronic coupling shows a single molecular conformation with respect to nanoelectrodes and their tip structures. Comparing current–time measurements of single-base molecules with single-base molecules in DNA, the degree of freedom in the latter is lower due to restriction of chemical bonds. In fact, the full width at half maximum of single-molecule conductance of single-base molecules for DNA and RNA is lower because the smaller degree of freedom makes variation in electrode–molecule interactions smaller [27].

9.8 Peptide Sequencing

Peptides are biomolecules formed by 20 types of amino acids and have many biological functions [36–42]. For example, a peptide formed by 29 amino acids exhibits a hormone action to raise blood sugar levels, whereas another peptide

formed by seven amino acids performs antibacterial action. Based on knowledge of different peptide functions, natural peptides can act as biomarkers of physical conditions and diseases, whereas artificial peptides can be drug discovery targets [37, 38, 41, 43–45].

Sequences of amino acids are very significant to determine peptide functions, although peptides do not perform these functions until one or more amino acids are chemically modified. This chemical modification is called posttranslational modification and enables proteins to exhibit biological functions [46–51]. The phosphorylation of tyrosine is a well-known posttranslational modification that is closely related to cancers and metabolic diseases [46, 51]. Since there are no methods to amplify and directly sequence peptides, preparing peptide samples is time-consuming and expensive. In these cases, single-molecule sequencing technologies are expected to reduce time and cost because sample preparation is not necessary.

As well as all four-base molecules and chemical modified base molecules of DNA and RNA, amino acid molecules and chemical modified amino acid were identified via tunneling currents using the STM tip modified with recognition molecule[52] and MCBJ [28] (Fig. 9.11). In the STM measurements, amino acid molecules are captured via hydrogen bonds formed between amino acids and recognition molecules (Fig. 9.12) [52]. When recognition tunneling currents were measured in buffer solutions of individual amino acid molecules, spikelike signals were observed, similar to the DNA and RNA measurements discussed above. However, recognition tunneling current histograms show exponential decay and the overlaps of the wide range. This is why distinctions between different amino acid molecules are difficult. A machine learning algorithm called support vector machine (SVM) [53–58] was introduced in an effort to overcome this issue. SVM was used to distinguish amino acid molecules into two classes by obtaining large recognition tunneling current–time data of two amino acid molecules and learning from these data parameters. Of course, using SVM to distinguish the signals strongly depends on the signal features of the two chosen amino acid molecules. Two specific signal features (top average amplitude of recognition tunneling currents and Fourier component) of the tunneling current distributions were selected and were able to distinguish glycine (Gly) and methylglycine (mGly) with a probability of $P = 0.95$. This is in stark contrast to the probability of $P = 0.55$ obtained using only recognition tunneling currents [52]. This method was able to identify five amino acids, one chemically modified amino acid, and one pair of enantiomers. In addition, the single-molecule method shows the potential to determine the mix ratio of two amino acid molecules in a solution formed by the two molecules. Because SVM algorithms are binary classifiers in general, the representation of time series data and what feature parameters to select are the key for realizing to identify all 20 amino acid molecules.

The conductance–time profiles of the 20 amino acids are measured using the MCBJ to obtain single-molecule conductivity [28]. Unlike the base molecules of DNA and RNA, there are differences in the molecular size of amino acid molecules. Therefore, single-molecule conductance measurements were performed

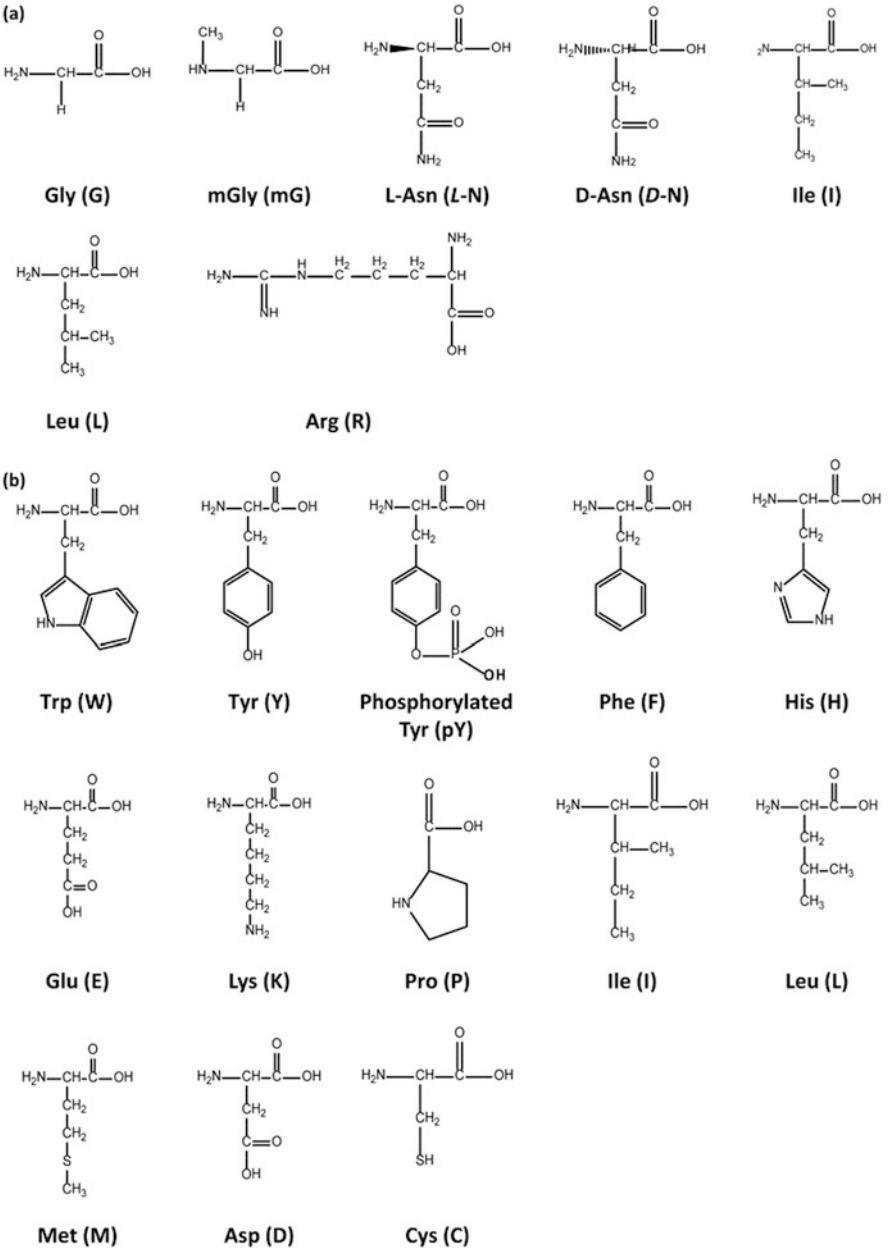

Fig. 9.11 Amino acids identified via (**a**) recognition tunneling and (**b**) tunneling currents

using 0.7- and 0.5-nm-nanogap electrodes. Similar to the base molecules of DNA and RNA, the histograms corresponding to different amino acids show single easily distinguishable peaks. When the 0.7-nm-nanogap electrodes are used, nine

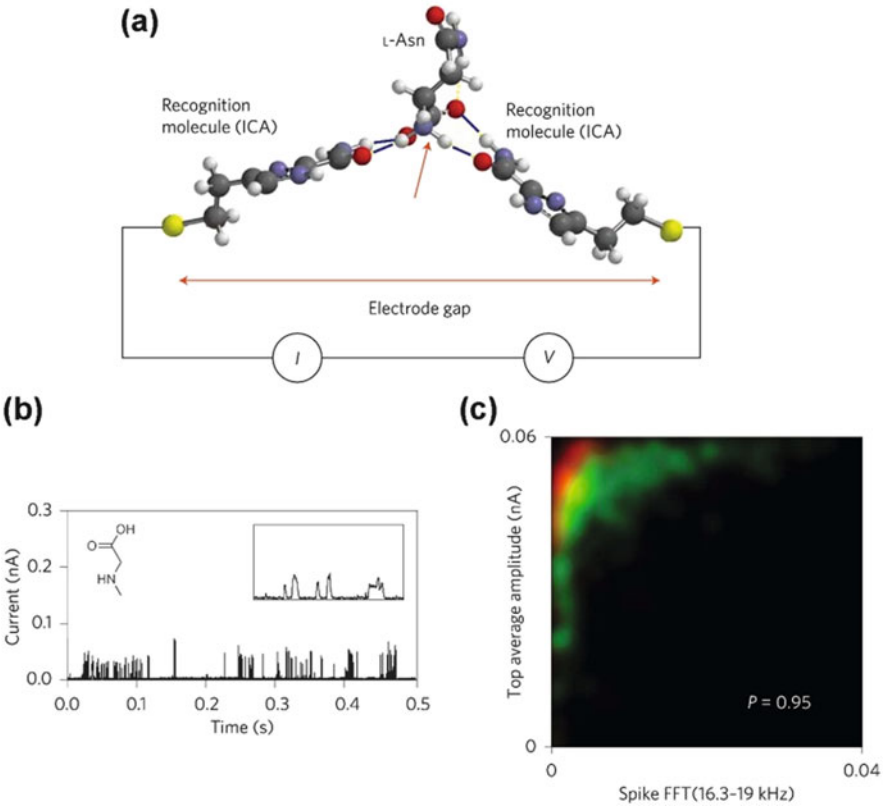

Fig. 9.12 (**a**) Schematic of recognition and amino acid molecules, (**b**) current–time profile of *N*-methylglycine, and (**c**) two-dimensional plot of probability density as a function of the FFT feature value and top average amplitude (Reprinted with permission from Macmillan Publishers Ltd: ref. [52], copyright 2014)

amino acids can be identified. The nine conductance histograms obtained from conductance–time measurements using the 0.5-nm-nanogap electrodes show single peaks, similar to those observed when using the 0.7-nm-nanogap electrodes. As a result, 12 amino acid molecules could be identified via tunneling currents using the two different nanogap electrodes. Notably, tyrosine and phosphotyrosine could be distinguished by tunneling currents. Estimating the single-molecule conductance order of amino acid molecules using their HOMO energies is difficult because the amino acid molecules are not π-electron systems.

Next, using the 0.7-nm-nanogap electrodes, electrical measurements of the original peptide and the phosphorylated peptide were performed [28]. Based on the single-molecule conductance measurements of individual amino acids and the above analysis method, the three measured peaks were assigned to tyrosine (Y), phenylalanine (F), and phosphotyrosine (pY) (Fig. 9.13a). In this analysis, other

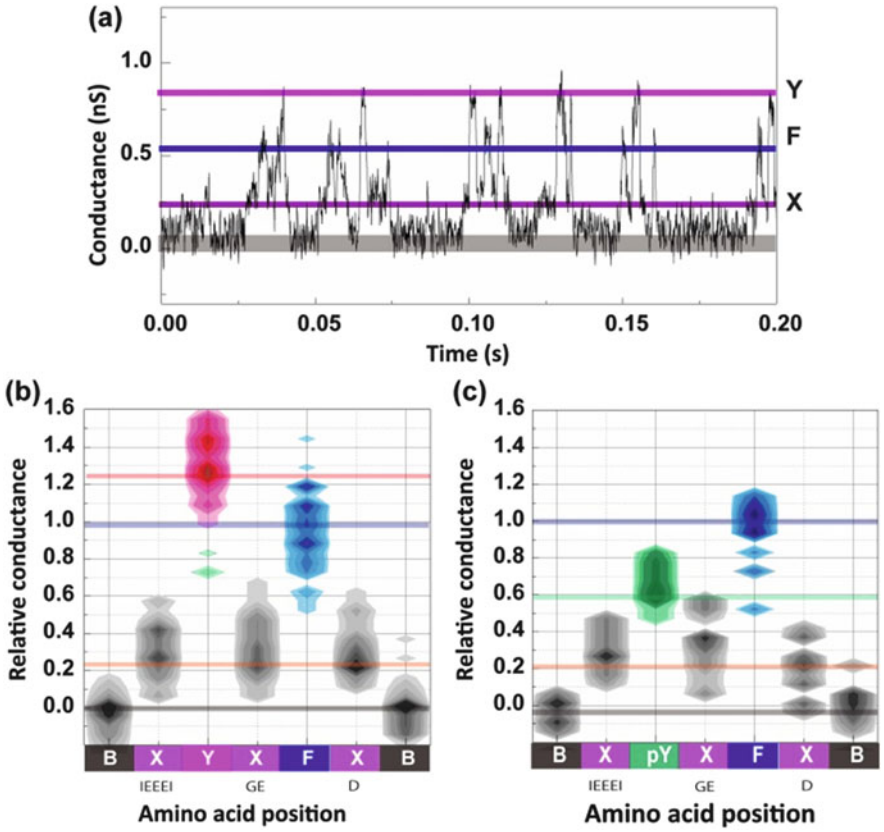

Fig. 9.13 Conductance–time profile of (**a**) a peptide, (**b**) contour map of original peptide, and (**c**) chemical modified peptide. Ref. [28]

amino acids were assigned to a generic group X. The partial sequence of amino acids that comprises the peptides can be read from their conductance–time profiles. The intensities of the plotted data show the signal numbers as percentages of the total. Based on their conductance intensities, the most likely amino acid sequence can be determined. The left (Fig. 9.13b) and right (Fig. 9.13c) figures correspond to the original peptide and phosphorylated peptide, respectively. These results demonstrate that peptides can be partially sequenced and posttranslational modifications can be detected by conductance profile analysis.

A current–time signal contains information of a fragmented sequence. The original and chemically modified peptides have one tyrosine and phosphotyrosine, respectively. Therefore, the number of electric signals of tyrosine and phosphotyrosine is expected to be the same as in the original and chemically modified peptides in solution. This indicates the possibility of quantitative analysis where sequences of amino acids in peptides and their proportion in a mixture can

be simultaneously determined. When current–time profiles of a solution with a ratio of the original/chemically modified peptide = 1/5 are measured, the ratio of fragmented sequences containing the original and chemically modified peptides is expected to be 1:5. Experimental results obtain a ratio of 1:4.3 [28]. The experimental results confirm that the current–time profile analysis can be used to effectively determine the ratio of original to chemical modified peptides.

9.9 Perspective

Tunneling currents can detect small differences in electronic structures of single molecules. Single-molecule electrical sequencing using tunneling currents can identify sequences of base molecules in DNA and RNA and sequences of amino acids in peptides. In addition, unlike existing DNA sequencing techniques, single-molecule sequencing methods can identify base and amino acid molecules, many of which are very important disease markers. Single-molecule sequencing methods may allow us to realize quantitative analysis where sequences of amino acids and the proportion of peptides in a mixture can be determined. This single-molecule quantitative analysis can be applied to DNA and RNA, particularly miRNA.

The development of methods for controlling the speed of single molecules is key for obtaining high accuracy, high throughput, and the length of the sequences that are read [17, 33]. An ideal method is to flow single DNA, RNA, and peptide molecules in one direction at a constant speed. Many research groups are currently working on developing single-molecule speed control technologies [17, 33]; however, technical leaps are required because the hydrodynamics of single molecules in solutions cannot be explained by conventional hydrodynamics.

It is found that information science plays a significant role in distinguishing base and amino acid molecules. Nanoscience and nanotechnology studies based on physics and chemistry require an understanding of the origins of the measurement data, and it is often difficult to completely understand and control the phenomena. For example, large overlaps of distributions of single-molecule conductance appear to be a big barrier for identifying single molecules. However, information science allows us to stochastically identify phenomenon using large data. Hybridization of nanotechnology and information science will lead to remarkable innovations in single-molecule science and technologies.

Acknowledgment This work is supported by KAKENHI Grant No. 26220603 for financial support.

References

1. Levy S et al (2007) Plos Biol 5:2113
2. Wheeler DA et al (2008) Nature 452:872
3. Manolio TA, Brooks LD, Collins FS (2008) J Clin Invest 118:1590
4. Lander ES et al (2001) Nature 409:860
5. Feng SH, Jacobsen SE, Reik W (2010) Science 330:622
6. Xu X et al (2011) Nature 475:189
7. Chen JF et al (2013) Nat Commun 4:1595
8. Schloss JA (2008) Nat Biotechnol 26:1113
9. Mardis ER (2011) Nature 470:198
10. Kircher M, Kelso J (2010) Bioessays 32:524
11. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ (2012) Nat Rev Microbiol 10:599
12. Metzker ML (2010) Nat Rev Genet 11:31
13. Holt RA, Jones SJM (2008) Genome Res 18:839
14. Rothberg JM et al (2011) Nature 475:348
15. Zwolak M, Di Ventra M (2008) Rev Mod Phys 80:141
16. Branton D et al (2008) Nat Biotechnol 26:1146
17. Venkatesan BM, Bashir R (2011) Nat Nanotechnol 6:615
18. Taniguchi M (2015) Anal Chem 87:188
19. Dekker C (2007) Nat Nanotechnol 2:209
20. Saenger W (1984) Principles of nucleic acid structure. Springer, New York
21. Taniguchi M, Kawai T (2006) Physica E 33:1
22. Zwolak M, Di Ventra M (2005) Nano Lett 5:421
23. Huang S et al (2010) Nat Nanotechnol 5:868
24. Chang S, He J, Kibel A, Lee M, Sankey O, Zhang P, Lindsay S (2009) Nat Nanotechnol 4:297
25. Chang SA, Huang S, He J, Liang F, Zhang PM, Li SQ, Chen X, Sankey O, Lindsay S (2010) Nano Lett 10:1070
26. Tsutsui M, Taniguchi M, Yokota K, Kawai T (2010) Nat Nanotechnol 5:286
27. Ohshiro T, Matsubara K, Tsutsui M, Furuhashi M, Taniguchi M, Kawai T (2012) Sci Rep 2:501
28. Ohshiro T, Tsutsui M, Yokota K, Furuhashi M, Taniguchi M, Kawai T (2014) Nat Nanotechnol 9:835
29. Tsutsui M, Matsubara K, Ohshiro T, Furuhashi M, Taniguchi M, Kawai T (2011) J Am Chem Soc 133:9124
30. Krems M, Zwolak M, Pershin YV, Di Ventra M (2009) Biophys J 97:1990
31. Lagerqvist J, Zwolak M, Di Ventra M (2006) Nano Lett 6:779
32. Lagerqvist J, Zwolak M, Di Ventra M (2007) Biophys J 93:2384
33. Yokota K, Tsutsui M, Taniguchi M (2014) Rsc Adv 4:15886
34. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University Press, Cambridge
35. Kim VN, Han J, Siomi MC (2009) Nat Rev Mol Cell Biol 10:126
36. Ganz T (2003) Nat Rev Immunol 3:710
37. Zasloff M (2002) Nature 415:389
38. Lewis RJ, Garcia ML (2003) Nat Rev Drug Discov 2:790
39. Peschel A, Sahl HG (2006) Nat Rev Microbiol 4:529
40. Haass C, Selkoe DJ (2007) Nat Rev Mol Cell Biol 8:101
41. Hancock REW, Sahl HG (2006) Nat Biotechnol 24:1551
42. Brogden KA (2005) Nat Rev Microbiol 3:238
43. Hruby VJ (2002) Nat Rev Drug Discov 1:847
44. Purcell AW, McCluskey J, Rossjohn J (2007) Nat Rev Drug Discov 6:404
45. Fjell CD, Hiss JA, Hancock REW, Schneider G (2012) Nat Rev Drug Discov 11:37

46. Mann M, Jensen ON (2003) Nat Biotechnol 21:255
47. Bode AM, Dong ZG (2004) Nat Rev Cancer 4:793
48. Westermann S, Weber K (2003) Nat Rev Mol Cell Biol 4:938
49. Gallego M, Virshup DM (2007) Nat Rev Mol Cell Biol 8:139
50. Walsh G, Jefferis R (2006) Nat Biotechnol 24:1241
51. Witze ES, Old WM, Resing KA, Ahn NG (2007) Nat Methods 4:798
52. Zhao YA et al (2014) Nat Nanotechnol 9:466
53. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. Acm Trans Intell Syst Technol 2:27
54. Burges CJC (1998) Data Min Knowl Discov 2:121
55. Hsu CW, Lin CJ (2002) IEEE Trans Neural Netw 13:415
56. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Mach Learn 46:389
57. Suykens JAK, Vandewalle J (1999) Neural Process Lett 9:293
58. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D (2000) Proc Natl Acad Sci U S A 97:262