



(51) International Patent Classification:

B81B 1/00 (2006.01) C12Q 1/6825 (2018.01)

B82Y 15/00 (2011.01) G01N 27/327 (2006.01)

C12Q 1/6869 (2018.01)

(21) International Application Number:

PCT/US2020/023283

(22) International Filing Date:

18 March 2020 (18.03.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/819,783 18 March 2019 (18.03.2019) US

(71) Applicant: **BOARD OF TRUSTEES OF THE UNIVERSITY OF ARKANSAS** [US/US]; 2404 North University Avenue, Little Rock, AR 72207 (US).

(72) Inventors; and

(71) Applicants: **TUNG, Steve** [US/US]; 1784 N. Barrington Drive, Fayetteville, AR 72701 (US). **MA, Bo** [CN/US]; 1856 N. Choctaw Court, Fayetteville, AR 72701 (US).

(74) Agent: **GLASGOW, Richard, Blakely**; Wright, Lindsey & Jennings LLP, 200 West Capitol Avenue, Suite 2300, Little Rock, AR 72201 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: METHOD OF DNA BASE-CALLING FROM A NANOCHANNEL DNA SEQUENCER

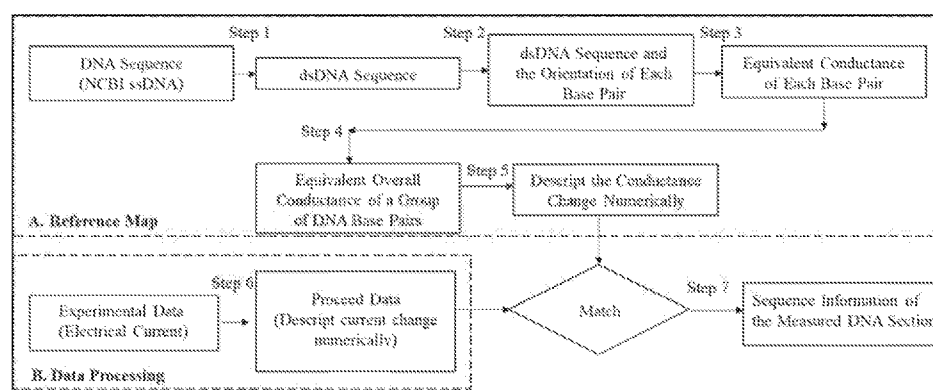


FIG. 1

(57) Abstract: A method of DNA base-calling from a nanochannel DNA sequencer. The method includes building a reference map and preparing an unknown sequence of DNA prior to the final step of data matching. The reference map includes a series of reference characters, such as numbers, that describe the change in tunneling current of a DNA strand with a known sequence. A DNA strand of unknown sequence is prepared so that the change in electrical measurement can also be described numerically. The section of match between the DNA strand of unknown sequence and the reference map is used to determine the sequence of the DNA strand.

METHOD OF DNA BASE-CALLING FROM A NANOCHANNEL DNA SEQUENCER

CROSS-REFERENCE TO RELATED APPLICATIONS

5 This application claims the benefit of U.S. Provisional Application No. 62/819,783, entitled "Method for DNA Base-Calling from a Nanochannel DNA Sequencer" and filed on March 18, 2019. The complete disclosure of said provisional application is hereby incorporated by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

10 This invention was made with government support from grant no. 1128660 awarded by the National Science Foundation and grant no. 1R21HG010055-01 awarded by the National Institute of Health. The government has certain rights in
15 the invention.

BACKGROUND ART

 Advanced DNA sequencing technologies have given great optimism to the future of public health. These technologies provide vital information to support research throughout the field of disease diagnoses, prevention, and treatment. To
20 sequence a complete human genome, which contains about 3 billion base pairs (bp), current sequencing technologies such as the Next Generation or Third Generation DNA sequencing require the DNA sample to be chopped into short segments. Through the approach of massively parallel processing, the human genome sequencing can be accomplished in weeks, however, the high cost of the
25 facility and long lead time of sequencing persists due to the short reading length of each segment.

 The world's first nanopore DNA sequencer, MinION from ONT (Oxford Nanopore Technologies) is based upon the technology of blockage current. Theoretically, when DNA is translocating through the nanopore, ionic current through
30 the nanopore is blocked by the presence of DNA. The amplitude of blockage current depends on the interaction between the DNA bases and nanopore. However, existing nanopore configurations are relatively thick (a few nanometers) and

measure the blockage current induced by multiple DNA bases instantaneously. Raw data performance assessments show that initially the ONT MinION achieved only a 60-70% sequencing accuracy because of the thickness of the nanopores.

A method of fabricating nanochannel systems for DNA sequencing and nanoparticle characterization is disclosed in U.S. Patent No. 9,718,668 (Steve Tung et. al). While the patented method made important strides in the field of DNA sequencing, the patented method fails to address one critical challenge for the application of DNA sequencing: existing technologies (like the MinION ONT, for example) are not suitable for analyzing the tunneling current measured by nanoelectrodes with a width wider than a single DNA base (about 0.3nm). Because of this issue, existing methods do not allow for the direct reading of the DNA sequence from the tunneling measurement based on its amplitude.

Furthermore, while ONT software has been developed further to increase sequencing accuracy, such software cannot be used to analyze data generated using devices such as the one described in U.S. Patent No. 9,718,668 because (a) fundamentally the measurement mechanism is different; (b) the ONT software is based on the algorithms of deep learning, which cannot be adopted to current uses for the sequencing data training step; and (c) using nanoelectrode to measure transverse current involves DNA orientation considerations that were not considered in the ONT algorithms. While the basic concept of data processing is common (i.e. to reveal the DNA sequence information based on their context), a novel DNA base-calling method for tunneling current analysis is necessary to address these challenges.

DISCLOSURE OF THE INVENTION

The present invention is directed to a method for DNA base-calling from a nanochannel DNA sequencer. Base-calling is a process that converts raw signals into readable DNA sequences. The process consists of two major tasks (building a reference map and preparing experimental data) prior to the final step of data matching. In the present invention, the reference map refers to a series of numbers built based on a standard DNA sequence to describe the change of its corresponding tunneling current. Experimental data is prepared so that the change

of electrical measurement can be described numerically. A section of match between the prepared experimental data and the reference map is used for DNA base-calling. The present invention utilizes seven sequential steps to execute these two major tasks, with mathematical models developed to accomplish the goal of each of the sequential steps. The novel DNA translocation protocol of the present invention utilizes AFM (atomic force microscope) based nanomanipulation to select and pick a single DNA molecule from a substrate surface. By moving the AFM tip in an aquaria environment, the DNA is stretched to linear during the process of DNA tunneling current measurement. This process is essential for allowing the DNA sequence to be output as the final results.

These and other objects, features, and advantages of the present invention will become better understood from a consideration of the following detailed description of the preferred embodiments and appended claims in conjunction with the drawings as described following:

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram showing an overview of the DNA base-calling process for DNA sequencing of the present invention.

Fig. 2A shows a representation of DNA translocating through single-chain nanoelectrodes and Fig. 2B shows the corresponding tunneling current.

Fig. 3A shows DNA translocating through triple-chain nanoelectrodes and Fig. 3B shows the corresponding tunneling current.

Fig. 4A shows the analytical method of matching DNA translocating with a tunneling current with the assistant of 'moving window' and Fig. 4B shows the trends associated with the tunneling current thereof.

Fig. 5A shows an example of a DNA sequence in a forward translocation direction and Fig. 5B shows the example DNA sequence in a backward translocation direction.

Fig. 6A shows the example DNA sequence of Fig. 5 converted into a double strand DNA in a forward translocation direction and Fig. 6B shows the example DNA sequence converted into a double strand DNA in a backward translocation direction.

Fig. 7 shows the effect of a DNA base pair rotating its orientation by 36 degrees per base pair and mapping the rotation as a sin wave function.

Fig. 8 shows an example of converting a section of dsDNA sequence in an array to three possible conductance arrays.

5 Figs. 9A-9B are schematics showing two extreme cases and their equivalent circuits when a DNA base pair of A-T is translocating through the gap between nanoelectrodes. Fig. 9A shows the 90° orientation and its equivalent parallel circuit and Fig. 9B shows the 0° orientation and its equivalent series circuit.

10 Fig. 10 is a table showing the conductance equation selections for corresponding base pair orientations.

Fig. 11 shows an example of dsDNA base pairs along the double helix structure changing their orientation to nanoelectrodes during the process of DNA translocation.

15 Fig. 12 is a table showing the known relative conductance of each type of nucleotide.

Fig. 13 is a schematic showing the concept of converting the conductance induced by each DNA base pair to a transverse conductance measurement.

Fig. 14 shows an example of a reference map of DNA in forward translocation direction.

20 Fig. 15A shows raw data and Fig. 15B shows processed data for the transverse current measurement when DNA translocates through a nanoelectrode at 0.1 $\mu\text{m/s}$.

Fig. 16A shows an example tunneling current signal graph and Fig. 16B shows the DNA section believed to be stretched zoomed in and coded numerically.

25 Fig. 17 is a graph showing the matching result of experimental DNA sequencing data to the reference map.

Fig. 18 is a graph showing the DNA sequencing raw data and the coding result at the beginning of the signal.

30 Fig. 19 shows the format of organizing sequence data for the VISTA submission.

Fig. 20 shows the VISTA mapping result. The highlighted sequence is used to match the experimental data.

Fig. 21 is a data plot showing the experimental data (current as a function of time) divided into seven sections (A-G).

5 Fig. 22 is a graph showing the reference data matched to experimental data in four separated sections and assembly using the VISTA tool kit.

BEST MODE FOR CARRYING OUT THE INVENTION

With reference to Figs. 1-22, the present invention may be described. The present invention is directed to a method for DNA base-calling from a nanochannel
10 DNA sequencer. As noted above, base-calling requires two major tasks: reference mapping and data processing. In the present invention, these major tasks are executed in a series of seven sequential steps (as shown in Fig. 1). Steps 1-5 are useful for executing the reference mapping task and step 6 is useful for the data processing task. Step 7 calls upon the outputs of the reference mapping task and
15 data processing task in a matching step that allows for the sequence information of the measured DNA to be the final output.

Experimental Background and Method Development: Before describing the method of DNA base-calling embodied by the present invention, it should be noted that the improved DNA base-calling process of the present invention is based off
20 experimental analysis that used quantum simulation to investigate the effect various DNA base pairs have on tunneling current measurement. Such experimental process used for deriving the method of the present invention is described here for background purposes. For best results, the quantum simulation was first performed using a single chain-like Pt nanoelectrode. For simulation purposes, DNA having a
25 sequence of GCAT (top strand reading from right to left) was used as a model. This DNA model is shown in Fig. 2A. In this model, dsDNA is translocating from the left side of the nanoelectrode to the right side in total of eight steps with a 1 Å moving interval, as shown in Fig. 2A. The resulting tunnel current is summarized in Fig. 2B. As shown, the experimental results show that there is no significant difference in
30 terms of the tunneling current measurement for an A-T pair and T-A pair. Likewise, there is no significant difference for a G-C pair and a C-G pair. Based on this

conclusion, the DNA sequence with four base pairs (A-T pair, T-A pair, C-G pair, and G-C pair) could theoretically be equivalent to a sequence with two base pairs (A-T pair, and C-G pair). Results also show that G-C and C-G pairs are generally more conductive than A-T and T-A pairs.

5 A further simulation with the goal of analyzing tunneling current without single base resolution was performed. The same DNA model as the first simulation was used and the width of the nanoelectrodes were increased from single chain to triple chain as shown in Fig. 3A. Simulation results, shown in Fig. 3B, demonstrate that the relationship between the normalized tunneling current and the DNA base pairs
10 are not clear. Thus, it was determined that no DNA sequencing data can be directly 'read' based on the tunneling current amplitude. Thus, it was determined that the non-single base sequencing resolution requires an additional step of data processing for DNA base calling. A concept of a "sliding window" was developed to describe the moving of DNA through the detecting range of nanoelectrodes. To do
15 so, the DNA model was assigned with a box that had a length equivalent to a 3 bp long DNA segment. Based on model configuration, the 3 bp length was determined based on the detection range of the nanoelectrodes (4.2 Å). To trace the change, DNA base pairs composition inside of the sliding window during the DNA translocation were changed and the process of DNA translocation was assembled
20 as shown in Fig. 4A. From the DNA Position #1 to DNA Position #3, the A-T pair came into the detection range and increased the tunneling current. From the DNA Position #3 to Position #6, the leading pair, the G-C pair, was swapped out by the T-A pair. Because T-A pair was shown to be less conductive than the G-C pair, the tunneling current was expected to decrease slightly. Then, from DNA position #6 to
25 position #8, only two DNA base pairs were left in the detection range and a rapid decline in tunneling current was expected, as shown in Fig. 4B. The moving DNA strand therefore can be correlated with the tunneling current change rather than the amplitude of the tunneling current. Based on the simulation results, with the assistance of a 'moving window', the change of tunneling current showed a clear
30 correlation to the sequence of tested DNA. Since the ultra slow DNA translocation speed could make the tunneling current sequence-dependent, the base call method

of the present invention (described below) was developed to analyze the experimental data for DNA sequencing.

DNA Base-Calling Process: The process begins with a single-stranded DNA (ssDNA) sequence. At step 1, the ssDNA is converted into a double-stranded DNA (dsDNA) sequence based on the ssDNA sequence. A dedicated notation is used to describe the base pair information along the DNA strand. For example, step 1 may use the basic DNA base pairing principle (A to T and C to G) to complement the double stranded DNA (dsDNA). It may be seen then, that an ssDNA sequence having the sequence shown in the forward translocation direction Fig. 5A, for example, will, in the backward translocation direction, have the sequence shown in Fig. 5B. Based off basic DNA base pair principles, the ssDNA sequence above may be converted into a dsDNA sequence having the sequence shown in Fig. 6A in the forward translocation direction and the sequence shown in Fig. 6B in the backward translocation direction. For exemplary purposes, the notation 'S' may represent a 'C-G' pair and the notation 'W' represents an 'A-T' pair. In one embodiment, programming (such as MATLAB) may be used to convert the reference ssDNA sequence into the appropriate dsDNA sequence using arrays and matrices to appropriately execute base pairing.

When a translocating DNA strand hits a nanoelectrode gap, each DNA base pair will interact with the nanoelectrodes in a particular orientation. When a dsDNA is translocating through a pair of patterned nanoelectrodes, polarization of the DNA base pair and the direction of tunneling current will vary depending on the base pair's particular orientation. Theoretically, during DNA translocation, the orientation of each base pair is determined by its position along the DNA double helix structure. dsDNA has a helix structure and each base pair twists at an angle of 36 degrees. Thus, a complete 360 degree turn is achieved every 10 DNA base pairs. The effect of this orientation change can be represented by a sine wave as shown in Fig. 7. The horizontal axis is the position of DNA base pair along the double helix structure, and the vertical axis is the value of the sine wave function.

For experimental purposes in developing the invention, and based on the fact that every 10 base pairs completes a 360 degree turn, experimental base pairs were

described by one of three orientations (0°, 36°, and 72°). To simplify the analysis even further, orientations of 0°, 36°, and 72° were approximated to 0°, 45°, and 90° to accommodate the theory of using equivalent circuits. These approximated orientations are shown in Fig. 7. These three orientations (O_1 , O_2 , and O_3) are
 5 represented by the following equations, where x represents the index:

$$\begin{aligned} O_1(x) &= \text{abs}(\sin(36 * x)) \\ O_2(x) &= \text{abs}[\sin(36 * (x + 1))] \\ O_3(x) &= \text{abs}[\sin(36 * (x + 2))] \end{aligned}$$

10 Using this process, the orientation of each of the DNA base pairs can be successfully described numerically to describe its periodical property. It should be noted, of course, that in actual practice of the invention described herein the orientation of the first DNA base pair determines the orientation of all of the base pairs that follow. It may be noted, then, that the simplistic experimental view of the
 15 base-pairs of 0°, 36°, and 72° (approximated to 0°, 45°, and 90°) may no longer apply. Instead, in practice, the orientation of each DNA base pair along the strand is determined using the orientation of the leading pair, and the orientation of the leading pair can fall anywhere in the range of 0° to 360°. If the leading pair is 0°, for example, the second pair will be 36°. If, however, the leading pair is 1.5°, the
 20 second pair will be 37.5°. At step 2, a matrix is built and contains the dsDNA sequence in the first row and the orientation of each corresponding base pair in the second row. Using the dsDNA sequence of base pairs and the orientation of each base pair, matrices can be established. For example, for a short piece of ssDNA with a sequence of GCGTA, a dsDNA sequence of SSSWW may be determined based
 25 off of basic base pairing principles (as described above). Assuming the orientations of the first base pair is one of 0°, 36°, and 72° (which are approximated to 0°, 45°, and 90°), three rows of a matrix or three individual matrices may be generated (An example is shown in Fig. 8). The different matrices (DSDNA O_1 , O_2 , and O_3) represent the three possible configurations of DNA and nanoelectrodes. In each
 30 matrix, the first line shows the sequence of the dsDNA (in either forward or backward moving direction) and the second line contains the orientation information

of each DNA base pair with a same vertical index. For ease of describing the invention, the matrices in Fig. 8 show arrows indicating orientation, but in computational processing, the second row of each matrix contains numbers obtained through the sine wave function as described above. It should be noted, of course, that the particular sequence described herein (GCGTA) is exemplary only and the present invention may be useful for analysis of any sequence. Likewise, while for purposes of describing the invention the same simplistic assumptions for base pair orientation are presented, in practice, the orientation of the first base pair (which may be anywhere from 0 to 360 degrees) is used to determine the true orientation of the remaining base pairs.

At step 3, the DNA base pair information and its orientation are combined and an equivalent conductance for each base pair is generated. To calculate the equivalent conductance for each base pair, equivalent circuits are used to refer to corresponding base pair orientations to the nanoelectrodes, as shown in Figs. 9A-B. The top and bottom limit of conductance due to the DNA translocation can be quantified by setting the orientation of a DNA base pair to 90° and 0°, respectively. These two orientations are equivalent to a parallel and series circuit as shown in Fig. 9A and 9B, respectively (using a base pair of A-T as an example).

Based on the conductance equations of the parallel and series circuit, the equivalent conductance of each base pair may be calculated given the base pair's orientation to the nanoelectrodes using the following equations:

$$G_p = \frac{1}{R_1} + \frac{1}{R_2}$$

$$G_s = \frac{1}{R_1 + R_2}$$

The equations provided above (where G_p refers to the conductance of an equivalent parallel circuit and G_s refers to the conductance of an equivalent series circuit) are used to calculate the equivalent conductance of each base pair based on the relationships shown in Fig. 9 and represented in Fig. 10. In the above equations, $1/R_1$ is the conductance of one of the DNA bases and $1/R_2$ is that of the other DNA base in the base pair.

To better understand the relationship between the position of a DNA base pairs to their orientation, consider the following example using a dsDNA with the sequence of GCGTAC, where the first base pair is assumed to be in the 90 degree orientation. As noted above, the notations S and W may be used to refer to particular base pairs (G-C and A-T, respectively). Subscripts may be used to indicate the position of the base pair along the double helix structure. When this DNA section was translocated through the nanoelectrodes, the double helix structure twists the orientation of the DNA base pairs in steps of 36° for each (rounded to 45°) as previously described, and as shown in FIG. 11. The equivalent conductance of S₁, S₃, W₄, and S₆ are determined by using the parallel (G_p) and series (G_s) equations above according to their orientations. In contrast, the equivalent conductance of S₂ and W₅ are determined by first calculating the conductance of that base pair by using both the parallel and series equations, and then taking a mean of those two values. These data processing steps are repeated for each DNA base pair along the sequence of reference DNA for reference map construction. Once the orientation and sequence information is combined and the calculations of conductance are complete the DNA sequence (for example, GCGTA) is converted into three possible conductance arrays: $\sigma_1(x)$, $\sigma_2(x)$, and $\sigma_3(x)$ by using the conductance number of each nucleotides listed in Fig. 12. In practice, conductance arrays will correspond to each possible orientation sequence for each translocation direction.

At step 4, the system conductance is produced using the conductance of each DNA base pair. The system conductance is defined as the conductance that should be theoretically detected by the nanoelectrodes. Due to the width of the nanoelectrode detection range, the system conductance may be calculated by combining multiple DNA base pairs simultaneously. The number of DNA base pairs that should be included in this calculation is determined by the 'window' size as described previously. The equivalent system conductance is determined by combining the conductance of each DNA base pair based on the physical properties of the experimental setup to simulate the measured tunneling current. The

conductance arrays generated through step 3 consist of the conductance of each individual DNA base pair.

In practice, each instantaneous tunneling measurement is composed by the tunneling effect of multiple DNA base pairs due to the large width of the

5 nanoelectrodes. Fig. 13 demonstrates an additional step required to convert the conductance listed in any array into a format that represents the instantaneous measured conductance due to the measure of conductance being attributed to multiple DNA base pairs. When a section of dsDNA translocates through the gap of nanoelectrodes, as shown in in Fig. 13, DNA base pairs from i to j contribute to the
10 conductance measurement simultaneously. Based on its location and the distance to the center of detection range, a electron transmission probability function, T , can be used to determine the contribution of each DNA base pair. The measured conductance, $\Delta\sigma$, through a section of dsDNA in the transverse direction is calculated by following the equation:

$$15 \quad \Delta\sigma(x) = C + \sum_i^j (\sigma_x * T_x)$$

where, C is the background baseline shift, σ_x is the conductance of each DNA base pair, T_x is the transmission probability based on the location of the DNA base pair

$$T_x = \frac{\hbar^2 v^2}{\hbar^2 v^2 + U^2}$$

where, $\hbar$ is the reduced Planck number, v is the applied potential bias, and U is an
20 evaluating number in the range from 0 to 1 for describing the alignment position. $U = 0$ when the DNA base pair is in the middle of the nanoelectrodes where the transmission probability $T_x = 1$. After this step, the conductance of each DNA base pair, stored in arrays of $\sigma_t(x)$, were converted to measurement conductance, stored in arrays of $\Delta\sigma_t(x)$.

25 At step 5, dedicated numbers are used to describe the system conductance change numerically. After this step, the reference map is ready to be used. In this final step of reference map construction, the theoretically established measurement conductance arrays $\Delta\sigma_t(x)$ are used. In order to find a match between experimental

data and theoretical data, the change of amplitude rather than the absolute value of the amplitude must be used. To accommodate the computer processing requirement, the change of the theoretical data must be described numerically using the following equations:

$$\Psi_t(x) = \begin{cases} 4; & \text{if } \Delta\sigma(x+1) - \Delta\sigma(x) > 0; \\ 2; & \text{if } \Delta\sigma(x+1) - \Delta\sigma(x) = 0 \\ 0; & \text{if } \Delta\sigma(x+1) - \Delta\sigma(x) < 0; \end{cases}$$

where, the array $\Delta\sigma(x)$ is the measured conductance based on a group of DNA base pairs appearing in the nanoelectrodes detection range. After this process, in the reference map, the change of the system conductance due to the translocating DNA is expressed numerically without a physical vector. In this way, each time the DNA moved forward one base pair distance, the measurement conductance change of increase, decrease, and flat were represented by the number of 4, 0, and 2, respectively. For a DNA translocating through the gap of sensing nanoelectrodes, a series of numbers is generated to describe the change of measured conductance due to this translocation event. Once this process is repeated on all conductance arrays, the reference maps are prepared and ready to be used. An example reference map is shown in Fig. 14.

Step 6 is the process of experimental preparation to interpret experimental electrical current change numerically in the same way as the reference map described. That is, experimental data processing follows the same principle by converting data to a series of numbers that represent the change of tunneling current. To do this, a section of experimental data where the DNA is believed to be stretched is selected for analysis. In one embodiment, it may necessary to process the raw data to reduce noise level. For example, noise level may be reduced using a 3rd-order Butterworth LPF with a cutoff frequency of 45 Hz or a Keithley 6485 with a sampling frequency of 1000 Hz. It is contemplated that various equipment or working conditions may be used as known in the art for reducing noise level of the data, and that the particular frequency and other parameters should be modified according to the particular equipment used. In any event, the noise reduction used in this step is only for the purpose of finding the stretched DNA sections. Data before and after processing is shown, for example, in Fig. 15.

After data processing, the DNA tunneling current is plotted. An example is shown in Fig. 16A. Based on the DNA translocation speed, the data is divided into small sections to obtain the average in each of these sections. For example, if the DNA is translocating at 0.1 $\mu\text{m/s}$, the experimental data is dissected into small sections with the length of 3 ms for each section. This is based on the fact that the distance between the two neighbor DNA base pairs is 0.3 nm. With a 3 ms interval, theoretically, there will be 1 DNA base pair in each data section. The average is then taken in each section, and the conductance change from one section to another can be numbered by following the same principle. The numbering and its matching results are shown in Fig. 16B as an example. As the example shows, the coded sequence (also referred to in the application as a listing) could be written as [040222220240244242].

In the seventh step, the processed experimental data is used to find a match on the reference map. The obtained result indicates the position of that matching, which is used to retrieve the sequence information from the standard DNA sequence database. Thus, the developed DNA base calling method identifies the sequenced DNA by conducting a match study between the experimental data and theoretical reference maps (as shown in Fig. 17). In the example above, the coded sequence of [040222220240244242] is used to find a match in the matrices. DNA base calling is then accomplished by finding the corresponding DNA base information by reverse tracing the matrix of measurable conductance ($\Delta\sigma$) to the original DNA matrix. After this process, the position of the DNA sequenced in this experiment is located. As shown in Fig. 17, by tracing the reference map, sequence information of [ACTGCCCCTGCTTTCTTC] is located.

In order to carry out the method, it may be seen that the following must be known: (a) the target DNA, (b) the DNA translocation speed, and (c) the width of the nanoelectrodes. Instead of directly 'reading' DNA sequence through challenging the current fabrication technology limitation to have a sub-0.3 nm wide sensing nanoelectrodes, the method of the present invention significantly reduces the cost of sequencing for applications where DNA identification is desired. It may be seen that the method of the present invention may be directly used by or embedded in a deep

learning algorithm to work with sophisticated mathematical models for further analysis.

Using Experimental Data to Estimate the Accuracy of the DNA Base-Calling Process: As described below, additional experimental data was employed to

5 describe the method of DNA base-calling accuracy determination. A long DNA sequence raw data from a piece of λ DNA is partially coded and plotted in Fig. 18. The coding process was conducted based on the signal coding and code matching method described above. Based on the DNA translocation speed, the beginning of this serial data was coded within a time interval of 3 ms. The goal of this first step
10 data coding is to determine the length of the first section, noted as section A, with no mismatching. Then, referring back to the coding matrix described above, the sequencing of section A is found as [CCACGCGGGATGA].

To determine the gene information for the sequenced section of the λ DNA starting with the sequencing of [CCACGCGGGATGA], the DNA mapping techniques
15 were carried out using the VISTA tool. The DNA sequence information of the section A was managed into a text file in the format shown in Fig. 19. The name of the sequence was defined by a text string starting with a '>' mark. After the success of the file submission, the mapping result was returned and shown in Fig. 20. The highlighted VISTA mapping result suggests the direction of coding of the rest of the
20 raw data shown in Fig. 18.

The rest of the data shown in Fig. 18 was coded based on the suggestion of the VISTA mapping result, and the coding result is shown in Fig. 21. The redundant sequence found in sections of B, D, F was determined because it repeated the data found in sections either before or after it. Section C can only be correctly coded if
25 using the time interval of 2.6 ms. Compared to the 3 ms time interval used in section A, the section C is physically less stretched. Some of the numbers of coding result in section C represent errors. For each error, the number on top of the error code represents the corrections suggested by the VISTA mapping results. As shown in Fig. 21 (starting from left to right), a "2" has been corrected to a "4," a "4" has been
30 corrected to a "2," a "4" has been corrected to a "0," and a "2" has been corrected to a "0." As a result, the data section A contains the DNA sequence of

[CCACGCGGGATGA], and the section C contains the sequence of [ACCTGTGGCATTGTGCTGCCGGAAC] after the correction.

In order to successfully code sections E and G, the time interval of 1 ms has to be employed. For this particular group of experimental data, the 1 ms time interval reaches to the limit of the DAQ system used for data collection which only has a 1000 Hz maximum sampling frequency and causes the raw data to be unsuitable for further analysis. Therefore, the sequence data in sections E and G will not be counted for determining the base-calling accuracy.

In summary, in this particular group of DNA sequencing data, a total length of 40 base pair DNA was successfully processed using the disclosed base-calling method with 4 errors, which suggests a 90.47% local accuracy using the equation:

$$\varepsilon = 1 - \frac{\delta}{N}$$

where, ε is simply the accuracy, δ is the count of errors, and N is the total number of DNA sequence embedded in the raw data. The DNA sections E and G were not counted as a successful processing result due to the limited sampling frequency. In section E, there were 21 base pairs during the time of 21 ms. Similarly, in section G, there were 14 base pairs DNA packed in a time duration of 13 ms.

In a macro scale, the success rate of the DNA sequencing result was low by giving a total of 36 base pairs correctly read out from a 75 bp DNA segment. It roughly gives the global accuracy of 48%. Though 48% is not a significant number, it still shows the potential when considering the 65% raw accuracy of Oxford Nanopore MinION that has been developed for a decade.

The improvement for using this disclosed DNA base-calling accuracy can be achieved from two major perspectives. The most obvious way for achieving higher base-calling accuracy is to use an advanced DAQ system with a higher sampling frequency. In this particular example, the major fall back of the global accuracy is due to the limit of the DAQ system. The other improvement can be realized through the dimension reduction of the sensing element to improve the signal to noise ratio. The data used in this study as an example was measured using a 100 nm wide

nanoelectrodes. The changing of the conductance caused by the translocating DNA was described above using the equation of:

$$\Delta\sigma(x) = C + \sum_i^j (\sigma_x * T_x)$$

With a 1 nm width reduction of the nanoelectrodes, the $\Delta\sigma$ is reduced by ~1% in average. The change of the signal to noise ratio can be described using the following equation:

$$\Delta\alpha \cong \frac{\Delta\sigma(x)}{\Delta\sigma'(x)} - 1$$

where the $\Delta\alpha$ is the improvement of the signal to noise ratio and $\Delta\sigma'(x)$ is the overall conductance measured by nanoelectrodes with a reduced width. Based on the equations, reducing the width of nanoelectrodes from the current 100 nm to 50 nm will double the signal to noise ratio. The connect between the improved signal to noise ratio and the overall DNA base-calling accuracy enhancement is still under investigation.

The present invention has been described with reference to certain preferred and alternative embodiments that are intended to be exemplary only and not limiting to the full scope of the present invention.

WE CLAIM:

1. A method of DNA-base calling, comprising the steps of:
 - (a) building a reference map comprising reference characters
corresponding to a change in conductance measured as a known sequence of
5 double-stranded DNA is translocated through nanoelectrodes of a DNA sequencer;
 - (b) determining conductance of an unknown sequence of double-stranded
DNA measured as said unknown sequence of double-stranded DNA is translated
through said nanoelectrodes of said DNA sequencer
 - (c) determining changes in said conductance between adjacent sections
10 of said unknown sequence of double-stranded DNA;
 - (d) assigning reference characters corresponding to said changes in
conductance to create a listing;
 - (e) matching said listing to said reference map; and
 - (f) determining a sequence of said unknown sequence of double-stranded
15 DNA based on said matching of said listing to said reference map.
2. The method of claim 1, wherein said reference characters of said reference
map and said reference characters of said listing are numerals.
3. The method of claim 1, wherein said step of building a reference map
comprises the steps of:
 - 20 (a) converting a known sequence of single-stranded DNA to said double-
stranded DNA comprising a plurality of base pairs;
 - (b) determining orientations of said plurality of base pairs relative to a first
base pair of said plurality of base pairs;
 - (c) calculating an equivalent conductance of each of said plurality of base
25 pairs of said double-stranded DNA based on said orientations of said plurality of
base pairs;
 - (d) calculating system conductances of adjacent sections of said double-
stranded DNA; and
 - (e) assigning said reference characters of said reference map
30 corresponding to changes in said system conductances in adjacent sections of said
double-stranded DNA.

4. The method of claim 2, further comprising the step of building a matrix comprising said known sequence of said double-stranded DNA and said orientation of said plurality of base pairs.
5. The method of claim 2, wherein said step of calculating an equivalent conductance of each of said plurality of base pairs of said double-stranded DNA comprises the step of selecting a formula based on said orientations of said plurality of base pairs.
6. The method of claim 2, wherein said system conductances are equal to a sum of said equivalent conductance of each of said plurality of base pairs within a detection range of said nanoelectrodes.
7. The method of claim 6, wherein said detection range of said nanoelectrodes is determined by a width of said nanoelectrodes.
8. The method of claim 1, wherein prior to step (b) noise reduction is performed on said unknown sequence of double-stranded DNA.
9. The method of claim 1, wherein prior to step (c) said conductance of said unknown sequence of double-stranded DNA is plotted.

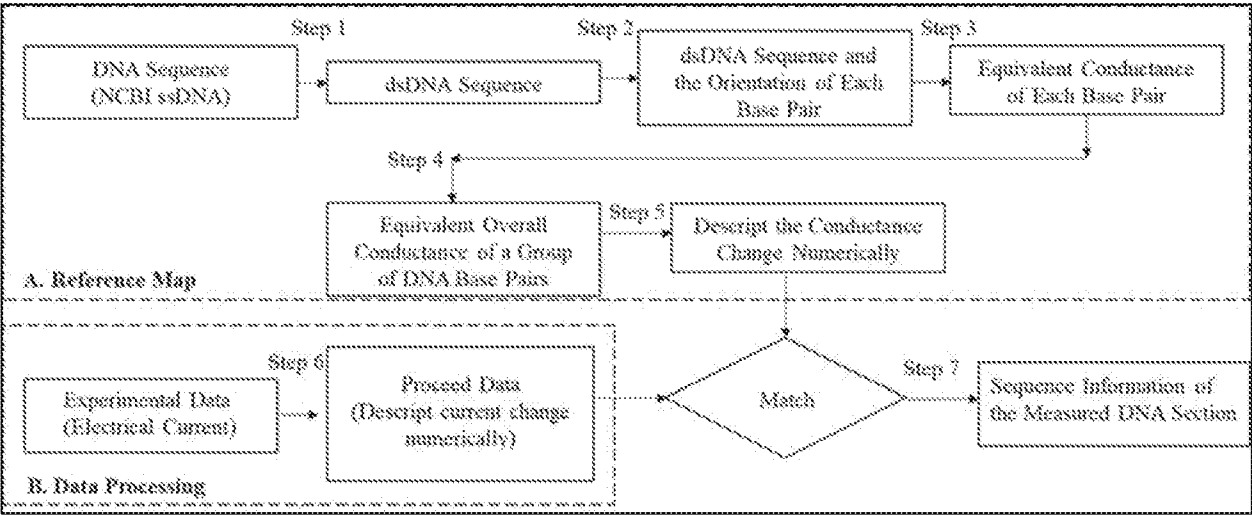


FIG. 1

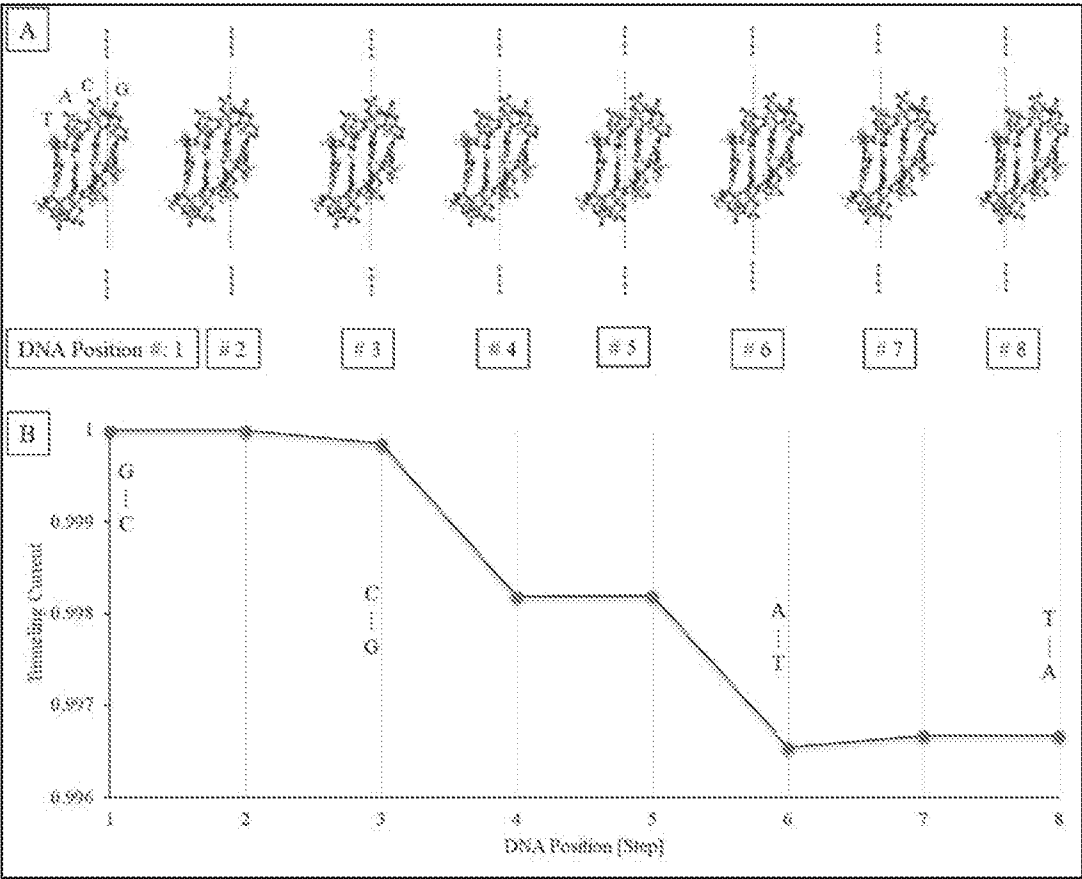


FIG. 2A

FIG. 2B

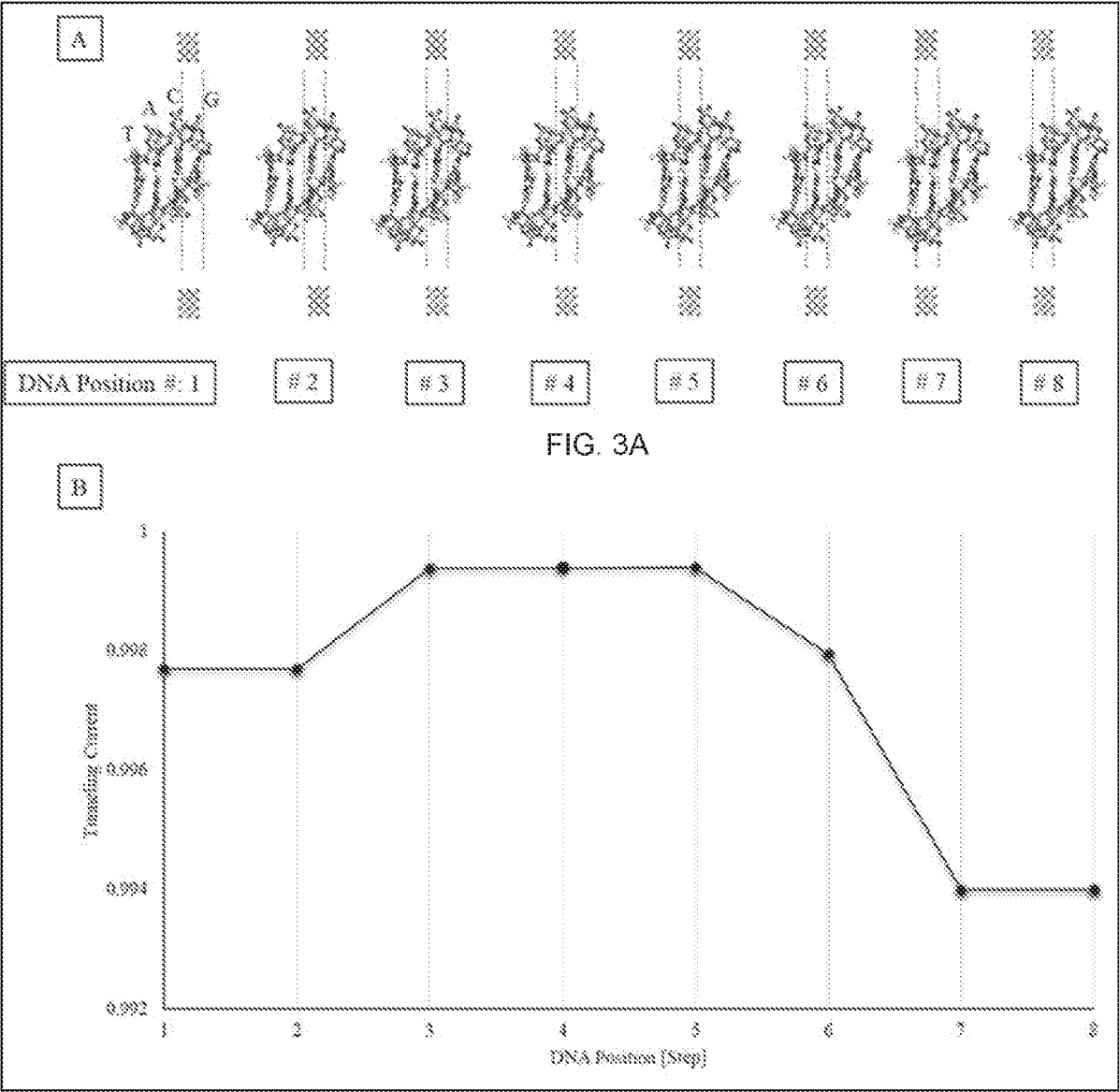


FIG. 3A

FIG. 3B

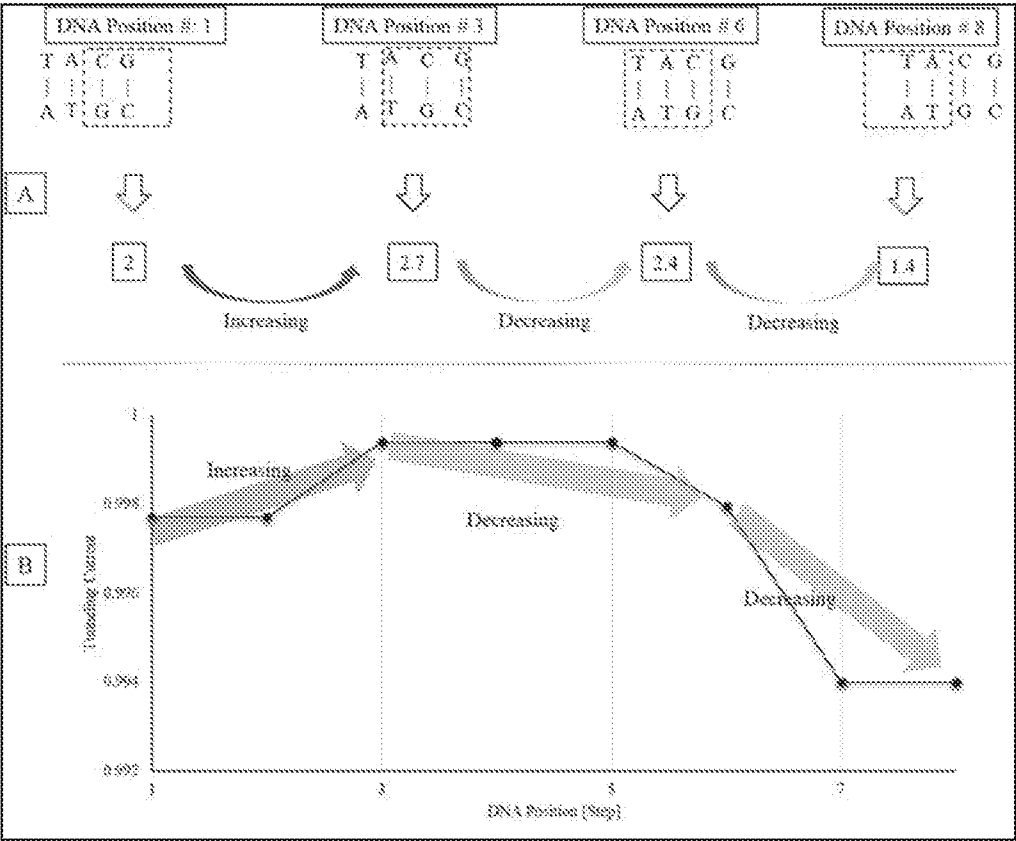


FIG. 4A

FIG. 4B

A

DNAf(x)	GGGCGGCGACCTCGCGGGTTT...TCCGGTGATCCGACAGGTTACG
---------	--

B

DNAb(x)	GCATTGGACAGCCTAGTGGCCT...TTTTGGGCGCTCCAGCGGCGGG
---------	---

FIG. 5A

FIG. 5B

FIG. 5

A	
DSDNAI(x)	SSSSSSSSWSSWSSSSSSWWWS...WWSSSSWSWSSSWSWSSWWWS
B	
DSDNAb(x)	SSWWSSWSWSSSWWSWSSSSWW...SWWWSSSSSSSWSSWSSSSSSSS
S=C-G pair; W=A-T pair	

FIG. 6A

FIG. 6B

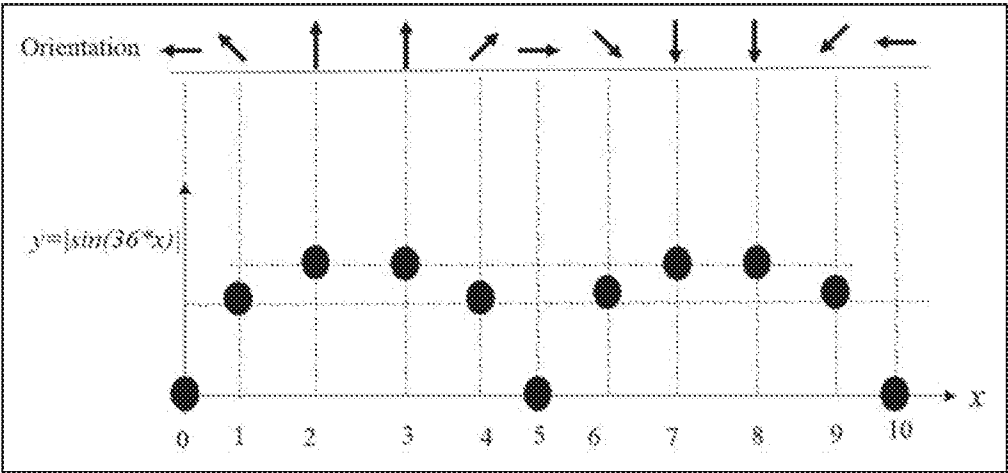


FIG. 7

5/13

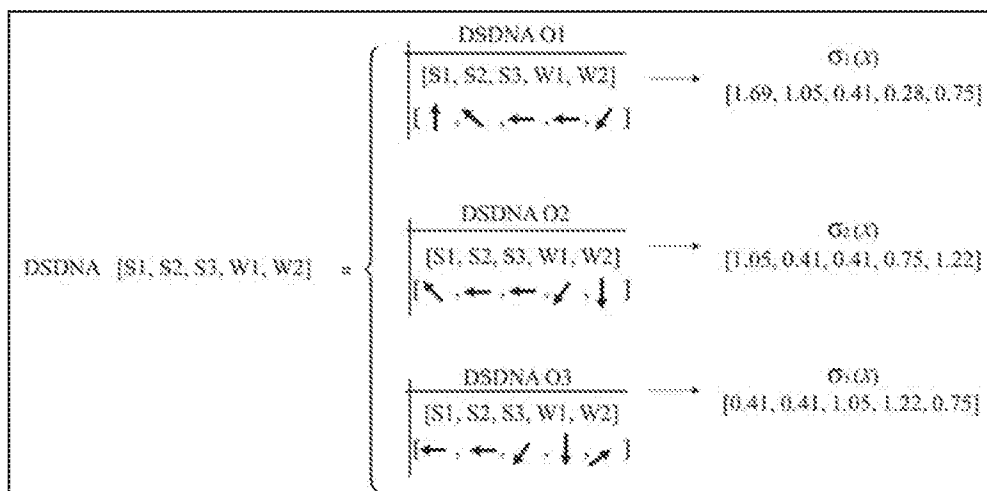


FIG. 8

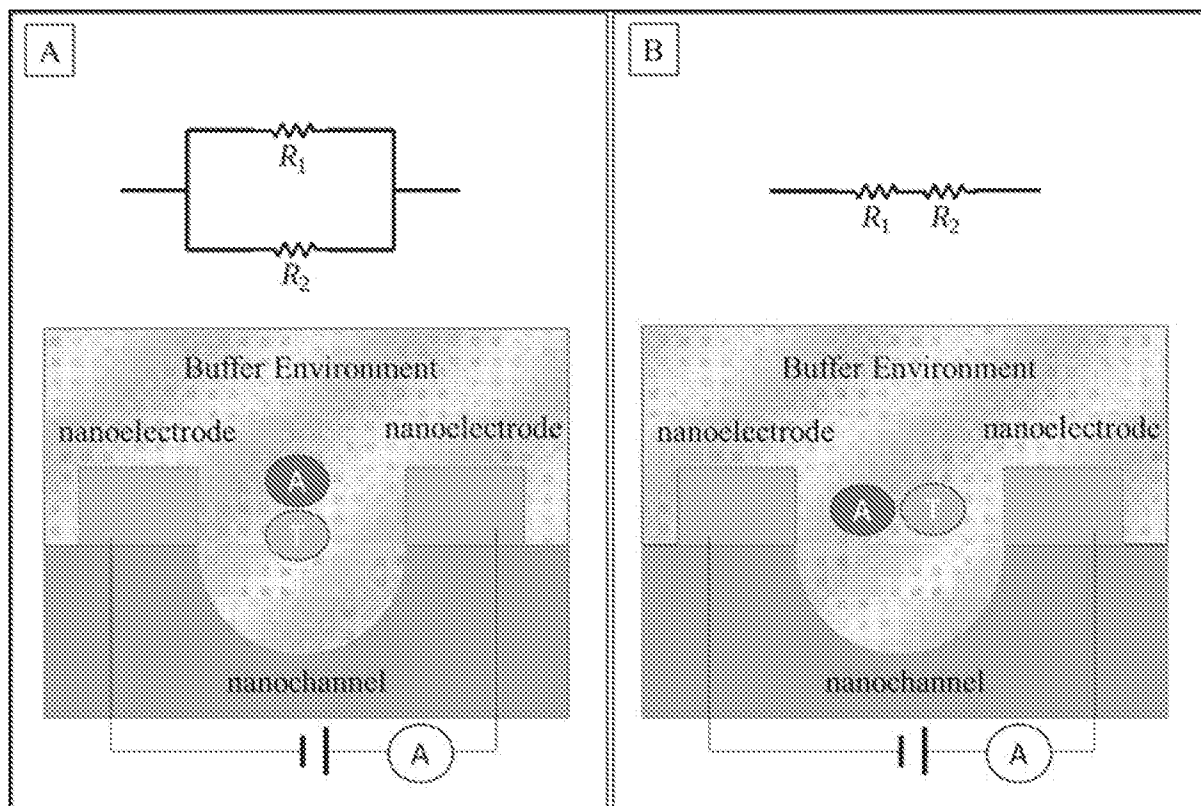


FIG. 9A

FIG. 9B

6/13

DNA Base Pair Orientations	Selected Conductance Equation
$\uparrow$ or $\downarrow$	G_p
$\leftarrow$ or $\rightarrow$	G_s
$\swarrow$ or $\searrow$	Mean (G_p and G_s)

FIG. 10

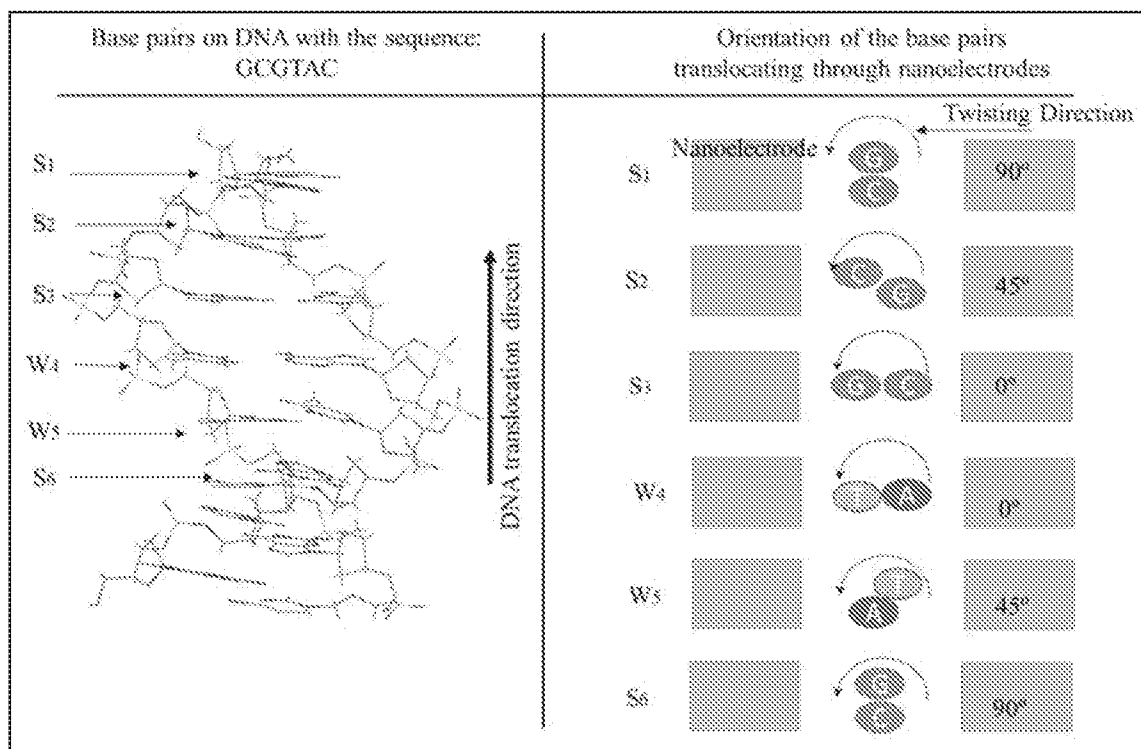


FIG. 11

Nucleotides	Relative Conductance (G)
G	1 ± 0.25
C	0.69 ± 0.25
A	0.77 ± 0.20
T	0.45 ± 0.12

FIG. 12

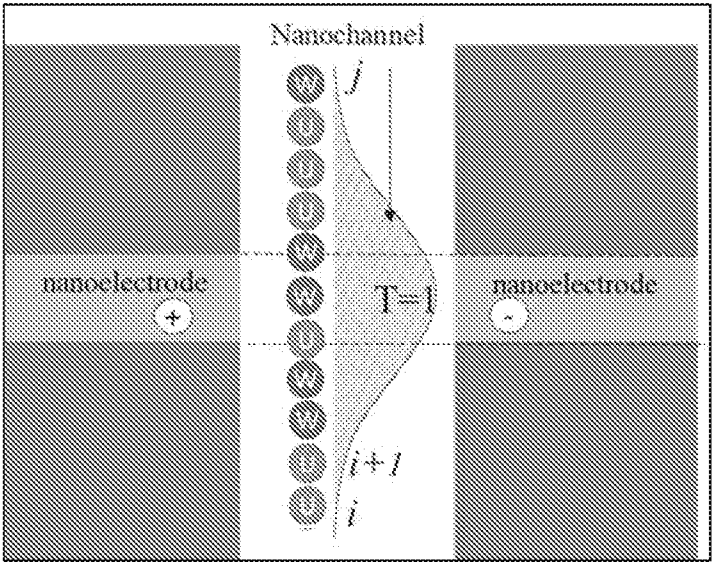


FIG. 13

$\Psi_1(x)$	20000202202400222222440202244442244.....24242044200224204224442222422244
-------------	--

FIG. 14

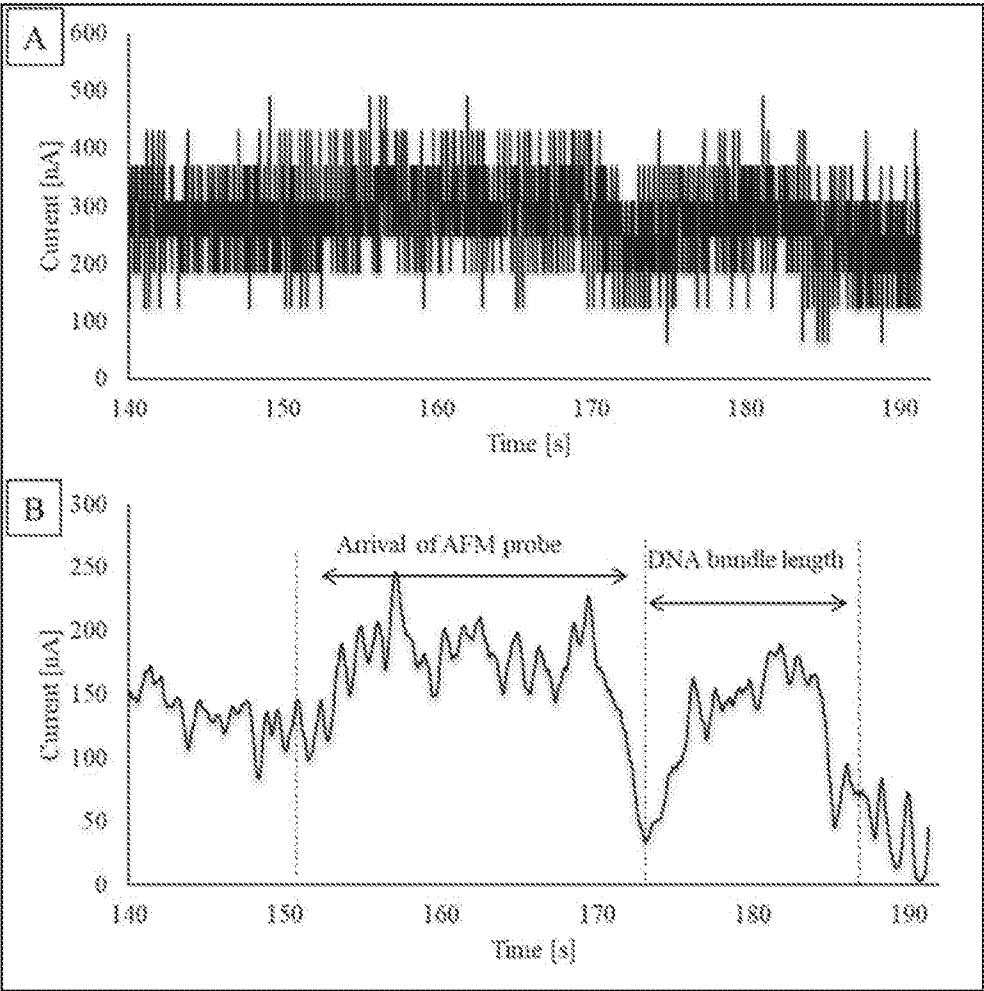


FIG. 15A

FIG. 15B

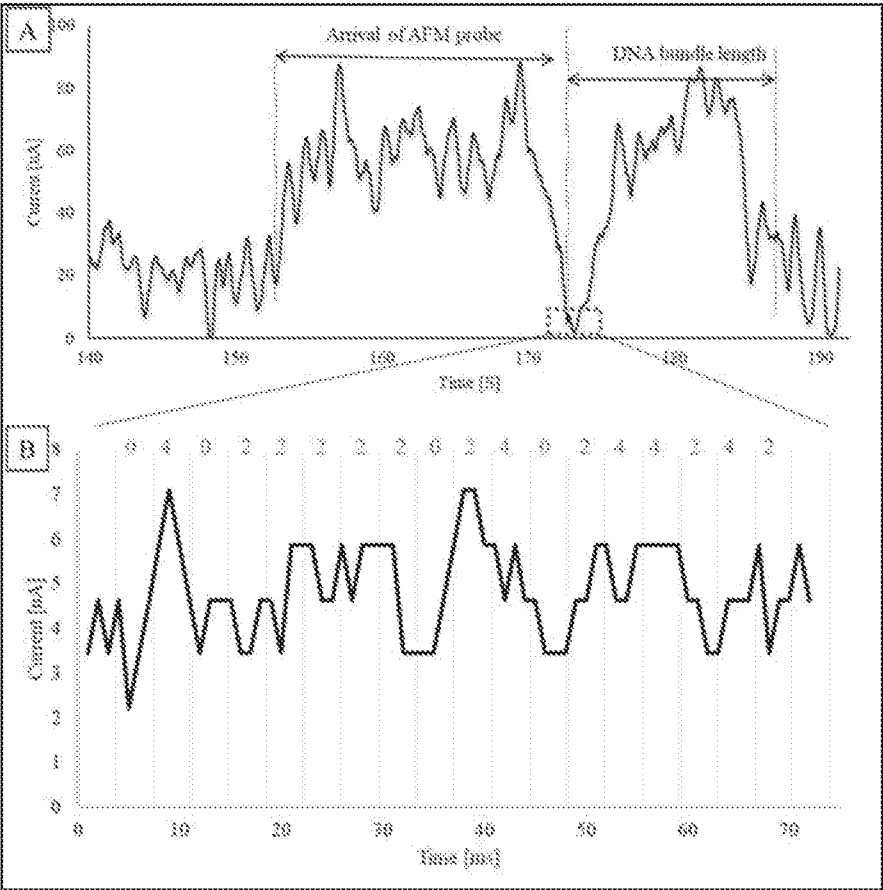


FIG. 16A

FIG. 16B

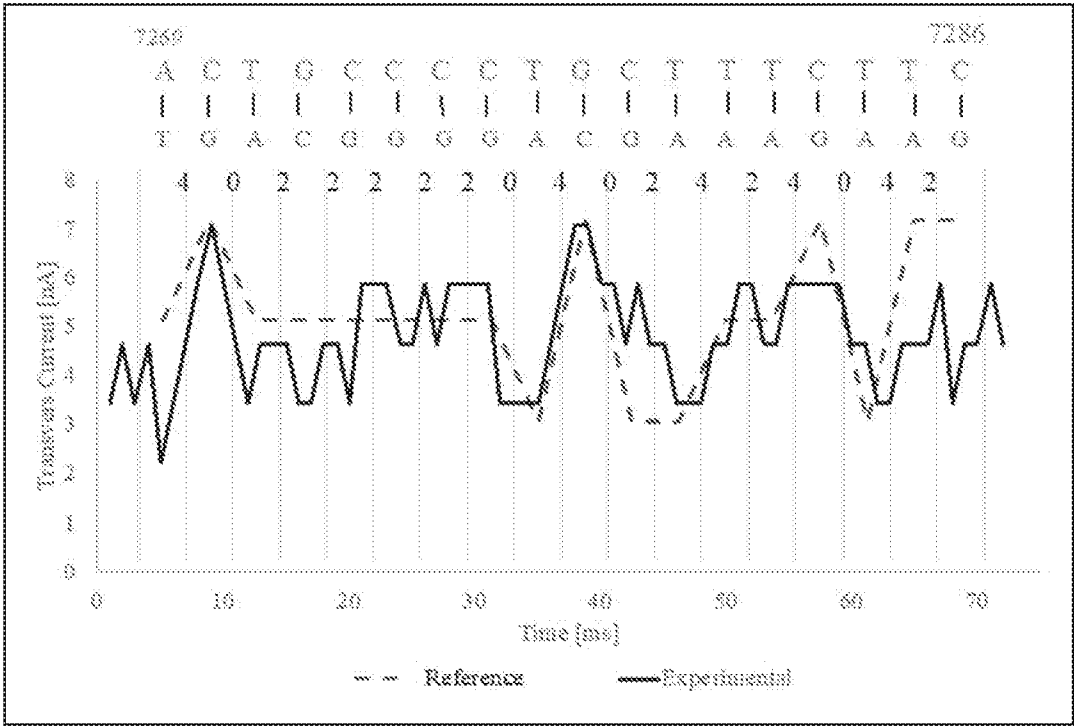


FIG. 17

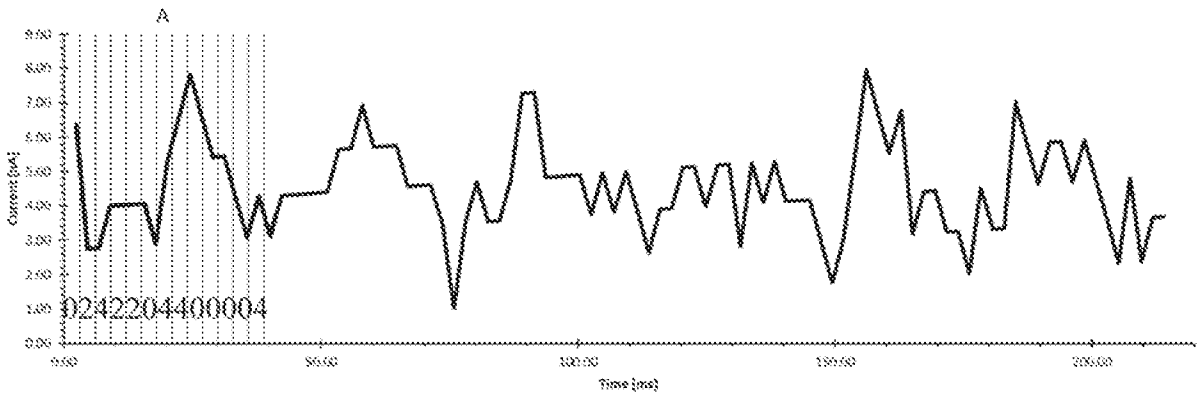


FIG. 18



FIG. 19



FIG. 20

12/13

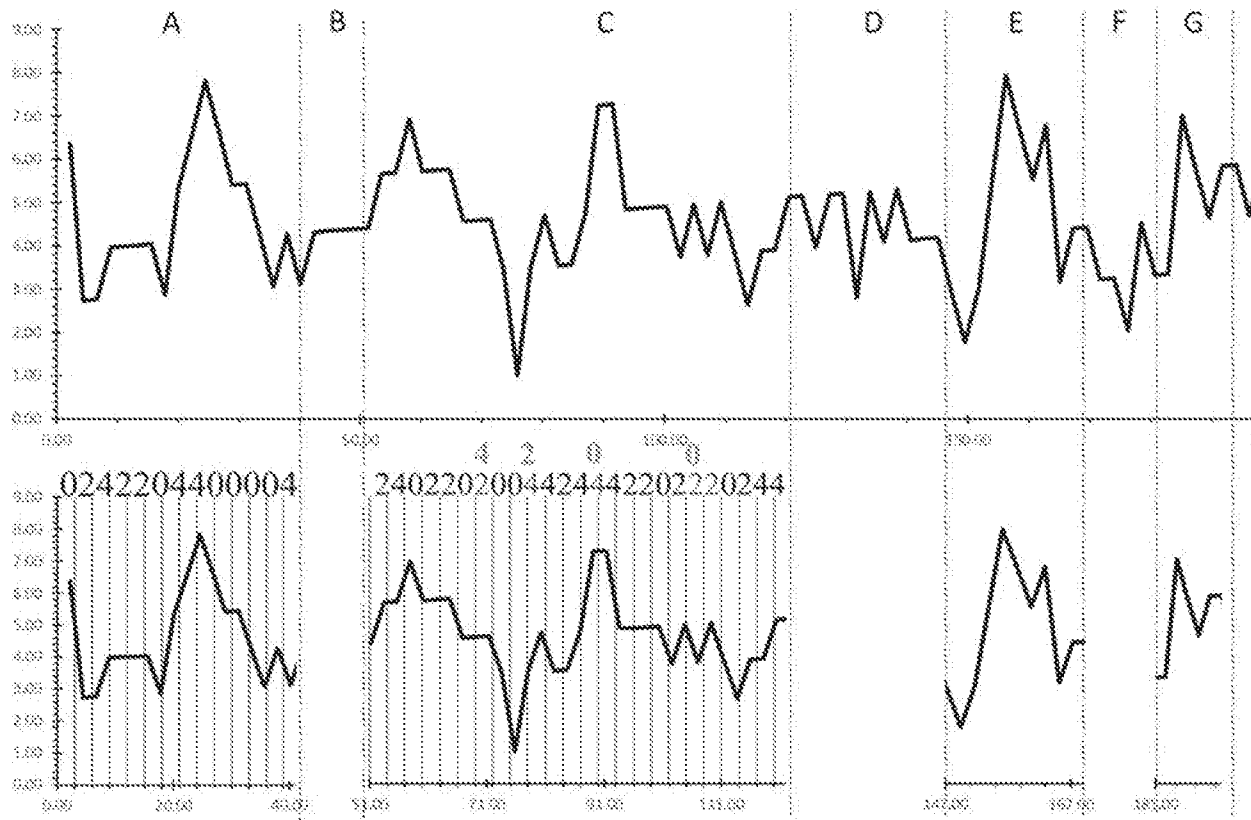


FIG. 21

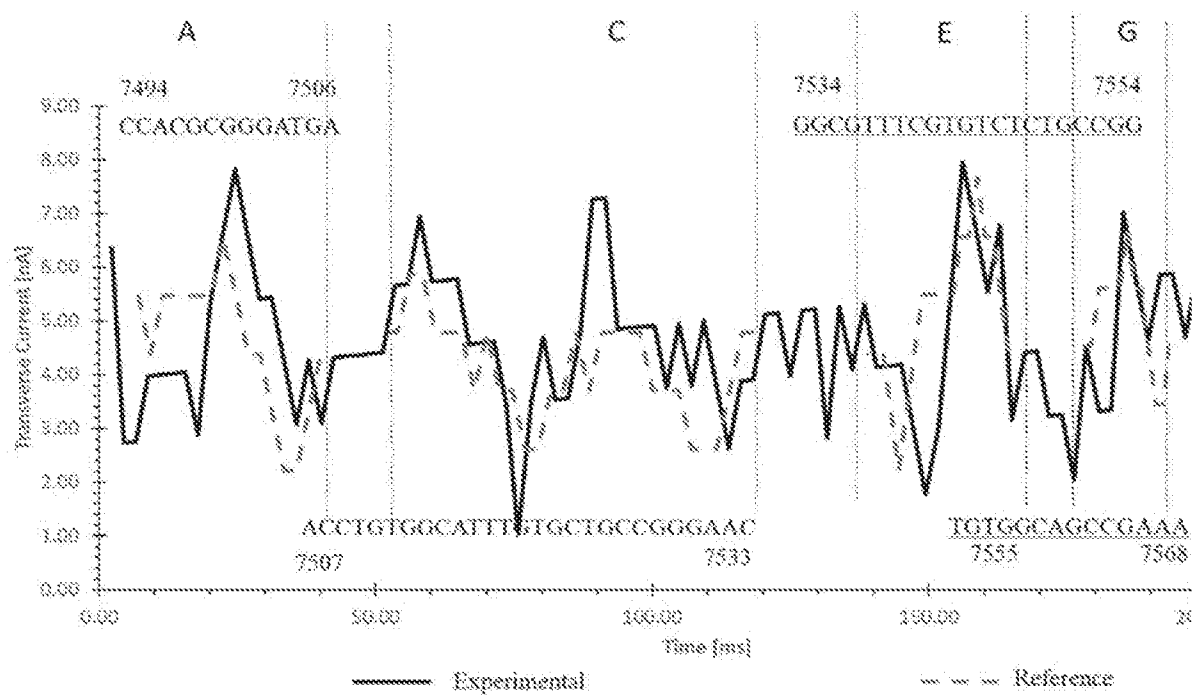


FIG. 22

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/23283

A. CLASSIFICATION OF SUBJECT MATTER

IPC - B81B 1/00, B82Y 15/00, C12Q 1/6869, C12Q 1/6825, G01N 27/327 (2020.01)

CPC - C12Q 1/6869, B81C 1/00071, G01N 27/44791, G01N 33/48721, Y10S 977/924, G01N 27/447, C12Q 1/6825, G01N 33/5438, G01N 27/3278

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/0132745 A1 (UNIVERSITY OF WASHINGTON THROUGH ITS CENTER FOR COMMERCIALIZATION) 14 May 2015 (14.05.2015) Abstract; Claim 2; Claim 5; Claim 6; Claim 15; Claim 16; Claim 19; para [0014]; para [0044]; para [0048]; para [0055]; para [0064-0067]	1-2, 4-5, 9
Y		3, 6-8
Y	US 2017/0152134 A9 (TUNG et al.) 1 June 2017 (01.06.2017) Abstract; para [0017]; para [0181-0182]; para [0187]	3, 6-7
Y	BALAN et al., Improving Signal-to-Noise Performance for DNA Translocation in Solid-State Nanopores at MHz Bandwidths. Nano Lett. 21 November 2014, Vol.14, p 7215-7220. Abstract	8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

14 June 2020

Date of mailing of the international search report

01 JUL 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/23283

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☐ forming part of the international application as filed:
 - ☐ in the form of an Annex C/ST.25 text file.
 - ☐ on paper or in the form of an image file.
 - b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
 - c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
 - ☐ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).
 - ☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).
2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

The electronic sequence listing submitted on 22 April 2020 in response to the PCT/ISA/225 issued on 31 March 2020 is acknowledged; however, it does not comply with the standard provided for in Annex C of the Administrative Instructions. Therefore, the international search was not carried out on the basis of that non-compliant electronic sequence listing.