

METHODS FOR SEQUENCING BIOPOLYMERS

GOVERNMENT SUPPORT

[0001] This invention was made with government support under Grant No. R21 HG010522 awarded by the National Institutes of Health. The government has certain rights in the invention.

RELATED APPLICATIONS

[0002] This application claims priority to and the benefit of U.S. Provisional Patent Application No. 62/983,417 filed February 28, 2020, which is incorporated herein by reference in its entirety for all purposes.

INCORPORATION-BY-REFERENCE OF MATERIAL SUBMITTED ELECTRONICALLY

[0003] Incorporated by reference in its entirety herein is a computer-readable nucleotide/amino acid sequence listing submitted concurrently herewith and identified as follows: One 825 Byte ASCII (Text) file named “2021-02-24_38882-601_SQL_ST25.txt,” created on February 24, 2021.

FIELD

[0004] The present disclosure provides devices, systems, and methods related to sequencing a biopolymer. In particular, the present disclosure provides methods of obtaining a bioelectronic signature based on current fluctuations that correspond to the activity of an enzyme-of-interest. As described herein, certain aspects of the bioelectronic signature can be used to determine the sequence of a biopolymer.

BACKGROUND

[0005] As proteins perform their various functions, movements are generated that underlie these functions. The ability to develop devices, systems, and methods that measure the electrical characteristics corresponding to the fluctuations generated by an active protein can be a basis for label-free detection and analysis of protein function. For example, monitoring the functional fluctuations of an active enzyme may provide a rapid and simple method of screening candidate drug molecules that affect the enzyme's function. In other cases, the ability to monitor the fluctuations of proteins that process biopolymers (e.g., carbohydrates, polypeptides, nucleic acids, and the like) may reveal new information about their

conformational changes and how those changes are linked to function. Additionally, diagnostic and analytical devices can be developed to take advantage of the electrical characteristics produced by active proteins, providing new ways to leverage biomechanical properties for practical use.

SUMMARY

[0006] Embodiments of the present disclosure include methods for sequencing a polynucleotide using a bioelectronic device. In accordance with these embodiments, the method includes introducing a template polynucleotide to the bioelectronic device; introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide, each dNTP present in the solution at a pre-defined concentration; and obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated into the template polynucleotide. In some embodiments, at least one characteristic of the bioelectronic signature identifies each of the complementary dNTPs incorporated into the template polynucleotide. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to at least a first electrode and a second electrode. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to both a first electrode and a second electrode.

[0007] In some embodiments, the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state. In some embodiments, the duration of the open period is distinct for each dNTP monomer such that it identifies whether a particular dNTP monomer has been incorporated into the template polynucleotide.

[0008] In some embodiments, the solution comprises four dNTP monomers. In some embodiments, a first dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of a second dNTP. In some embodiments, the second dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the first dNTP and a third dNTP. In some embodiments, the third dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the second dNTP and a fourth dNTP. In some embodiments, the fourth dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the third dNTP. In some embodiments, the sequence of the polynucleotide template can be accurately determined from

the duration of each open period. In some embodiments, the sequence of the polynucleotide can be accurately determined from the duration of each open period and/or one or more characteristics of the closed period.

[0009] In some embodiments, the duration of open periods for each dNTP are determined based on a distribution of a plurality of open duration periods. In some embodiments, the first dNTP is present at a saturating concentration. In some embodiments, extent of overlap is 1% or less.

[0010] In some embodiments, the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state. In some embodiments, at least one characteristic of the closed period varies based on the previously incorporated nucleotide. In some embodiments, at least one characteristic of the closed period is identified using a method comprising machine learning. In some embodiments, the machine learning method comprises Hidden-Markov Modeling or Bayesian non-parametric analysis.

[0011] In some embodiments, a combination of at least one characteristic of the closed period and at least one characteristic of the open period is used to identify each of the complementary dNTPs incorporated into the template polynucleotide.

[0012] In some embodiments, the polynucleotide template is DNA. In some embodiments, the polynucleotide template is RNA. In some embodiments, the dNTP monomers comprise adenine (dATP), cytosine (dCTP), guanine (dGTP), thymine (dTTP), and/or uridine (dUTP), including any derivatives or variants thereof.

[0013] In some embodiments, the exonuclease activity of the polymerase is disabled. In some embodiments, the polymerase is functionally coupled to the first and second electrodes using a linker comprising thio-streptavidin. In some embodiments, linker is attached to a region of the polymerase that is inactive.

[0014] In some embodiments, the method comprises applying a voltage bias between the first and second electrodes that is 100mV or less.

[0015] Embodiments of the present disclosure also include a method of calibrating a bioelectronic device. In accordance with these embodiments, the method includes introducing a template polynucleotide to the bioelectronic device; introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide, each dNTP present in the solution at a saturating concentration; obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated

into the template polynucleotide, wherein the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state; and measuring or determining the intrinsic distribution of the open periods for each dNTP.

[0016] In some embodiments, the bioelectronic device is calibrated based on the distribution of open periods. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to at least a first electrode and a second electrode. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to both a first electrode and a second electrode.

[0017] In some embodiments, the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state, and the bioelectronic device is calibrated based on at least one characteristic of the closed period.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] FIG. 1: Representative graph showing a bioelectronic signature of enzyme activity based on current fluctuations, according to one embodiment of the present disclosure.

[0019] FIG. 2: Representative graph of an expanded portion of the bioelectronic signature of FIG. 1, which includes the portion of time during which a new monomer (e.g., dNTP) is incorporated into a biopolymer as the enzyme transitions from an open state (201) to a closed state (202), according to one embodiment of the present disclosure.

[0020] FIG. 3: Representative graph of distribution times for which an enzyme-of-interest (e.g., polymerase) resides in an open state or conformation, awaiting arrival of a monomer (e.g., dNTP), according to one embodiment of the present disclosure.

[0021] FIG. 4: Representative graph of distribution times for which an enzyme-of-interest (e.g., polymerase) resides in an open state or conformation, awaiting arrival of a monomer (e.g., dNTP), with each monomer present in a solution at pre-defined concentrations, according to one embodiment of the present disclosure.

[0022] FIG. 5: Representative graph of an expanded portion of the bioelectronic signature of FIG. 1, which includes the portion of time during which a new monomer (e.g., dNTP) is incorporated into a biopolymer as the enzyme transitions from an open state (201) to a closed state (202); certain characteristics of the bioelectronic signature can be extracted (e.g., current fluctuations in the closed state) using Hidden Markov Modeling and/or Bayesian Nonparametric Modeling, which can form a basis for determining the sequence of a biopolymer.

DETAILED DESCRIPTION

[0023] Section headings as used in this section and the entire disclosure herein are merely for organizational purposes and are not intended to be limiting.

1. Definitions

[0024] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art. In case of conflict, the present document, including definitions, will control. Preferred methods and materials are described below, although methods and materials similar or equivalent to those described herein can be used in practice or testing of the present disclosure. All publications, patent applications, patents and other references mentioned herein are incorporated by reference in their entirety. The materials, methods, and examples disclosed herein are illustrative only and not intended to be limiting.

[0025] As noted herein, the disclosed embodiments have been presented for illustrative purposes only and are not limiting. Other embodiments are possible and are covered by the disclosure, which will be apparent from the teachings contained herein. Thus, the breadth and scope of the disclosure should not be limited by any of the above-described embodiments but should be defined only in accordance with claims supported by the present disclosure and their equivalents. Moreover, embodiments of the subject disclosure may include methods, compositions, systems and apparatuses/devices which may further include any and all elements from any other disclosed methods, compositions, systems, and devices, including any and all elements corresponding to detecting protein activity. In other words, elements from one or another disclosed embodiments may be interchangeable with elements from other disclosed embodiments. Moreover, some further embodiments may be realized by combining one and/or another feature disclosed herein with methods, compositions, systems and devices, and one or more features thereof, disclosed in materials incorporated by reference. In addition, one or more features/elements of disclosed embodiments may be removed and still result in patentable subject matter (and thus, resulting in yet more embodiments of the subject disclosure). Furthermore, some embodiments correspond to methods, compositions, systems, and devices which specifically lack one and/or another element, structure, and/or steps (as applicable), as compared to teachings of the prior art, and therefore represent patentable subject matter and

are distinguishable therefrom (i.e. claims directed to such embodiments may contain negative limitations to note the lack of one or more features prior art teachings).

[0026] All definitions, as defined and used herein, should be understood to control over dictionary definitions, definitions in documents incorporated by reference, and/or ordinary meanings of the defined terms.

[0027] The indefinite articles “a” and “an,” as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean “at least one.”

[0028] The phrase “and/or,” as used herein in the specification and in the claims, should be understood to mean “either or both” of the elements so conjoined, i.e., elements that are conjunctively present in some cases and disjunctively present in other cases. Multiple elements listed with “and/or” should be construed in the same fashion, i.e., “one or more” of the elements so conjoined. Other elements may optionally be present other than the elements specifically identified by the “and/or” clause, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, a reference to “A and/or B”, when used in conjunction with open-ended language such as “comprising” can refer, in one embodiment, to A only (optionally including elements other than B); in another embodiment, to B only (optionally including elements other than A); in yet another embodiment, to both A and B (optionally including other elements); etc.

[0029] As used herein in the specification and in the claims, “or” should be understood to have the same meaning as “and/or” as defined above. For example, when separating items in a list, “or” or “and/or” shall be interpreted as being inclusive, i.e., the inclusion of at least one, but also including more than one, of a number or list of elements, and, optionally, additional unlisted items. Only terms clearly indicated to the contrary, such as “only one of” or “exactly one of” or, when used in the claims, “consisting of,” will refer to the inclusion of exactly one element of a number or list of elements. In general, the term “or” as used herein shall only be interpreted as indicating exclusive alternatives (i.e. “one or the other but not both”) when preceded by terms of exclusivity, such as “either,” “one of” “only one of” or “exactly one of.” “Consisting essentially of,” when used in the claims, shall have its ordinary meaning as used in the field of patent law.

[0030] As used herein in the specification and in the claims, the phrase “at least one,” in reference to a list of one or more elements, should be understood to mean at least one element selected from any one or more of the elements in the list of elements, but not necessarily

including at least one of each and every element specifically listed within the list of elements and not excluding any combinations of elements in the list of elements. This definition also allows that elements may optionally be present other than the elements specifically identified within the list of elements to which the phrase “at least one” refers, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, “at least one of A and B” (or, equivalently, “at least one of A or B,” or, equivalently “at least one of A and/or B”) can refer, in one embodiment, to at least one, optionally including more than one, A, with no B present (and optionally including elements other than B); in another embodiment, to at least one, optionally including more than one, B, with no A present (and optionally including elements other than A); in yet another embodiment, to at least one, optionally including more than one, A, and at least one, optionally including more than one, B (and optionally including other elements); etc.

[0031] In the claims, as well as in the specification above, all transitional phrases such as “comprising,” “including,” “carrying,” “having,” “containing,” “involving,” “holding,” “composed of,” and the like are to be understood to be open-ended, i.e., to mean including but not limited to. Only the transitional phrases “consisting of” and “consisting essentially of” shall be closed or semi-closed transitional phrases, respectively, as set forth in the United States Patent Office Manual of Patent Examining Procedures, Section 2111.03.

2. Bioelectronic Devices and Systems

[0032] Embodiments of the present disclosure include devices, systems, and methods related to sequencing a biopolymer. In particular, the present disclosure provides methods of obtaining a bioelectronic signature based on current fluctuations that corresponds to the activity of an enzyme-of-interest. As described further herein, certain aspects of the bioelectronic signature of the enzyme-of-interest can be used to determine the sequence of a biopolymer.

[0033] In accordance with these embodiments, the enzyme-of-interest can be a polymerase, and various aspects of a bioelectronic signature of a polymerase as it adds nucleotide monomers to a template polynucleotide strand can be used to determine the sequence of that template polynucleotide. For example, the bioelectronic signature of polymerase activity can be based on current fluctuations as each complementary nucleotide monomer is incorporated into the template polynucleotide; and the signature can be obtained using a bioelectronic device comprising a polymerase functionally coupled to at least a first electrode and a second electrode. In some embodiments, the bioelectronic device comprises a polymerase functionally

coupled to both a first electrode and a second electrode. The term “nucleotide” generally refers to a base-sugar-phosphate combination and includes ribonucleoside triphosphates ATP, UTP, CTG, GTP and deoxyribonucleoside triphosphates such as dATP, dCTP, dTTP, dUTP, dGTP, dTTP, or derivatives thereof.

[0034] In some embodiments, the conductance of a polymerase molecule doubles when the molecule makes a transition from the open state (poised and ready to accept an incoming dNTP monomer) and the closed state (incorporating the incoming dNTP and translating the new double helix within the polymerase). A representative illustration of a typical electrical signal obtained during this process is shown in FIG. 1. In the inactive state, the current through the polymerase is at a low baseline level (101). Once dNTPs are added, the current jumps to the new higher conductance (102) associated with the closed state (103). After each new nucleotide is incorporated, the current dips down (104), which indicates a transition to the next open state. In the trace shown in FIG. 1, the downward sweep in current is limited by the response time of the electronics. That is, slower openings do come down all the way to the background level of current (101). In this particular example dataset, the polymerase has not captured a new template (105), and the current drops back to the baseline level (101).

[0035] The bioelectronic signatures of polymerase activity contain information both in the transient open states and in the closed regions in between. For example, FIG. 2 includes an expanded portion of the bioelectronic signature of FIG. 1, which includes the portion of time during which a new monomer (e.g., dNTP) is incorporated into a biopolymer as the enzyme transitions from an open state (201) to a closed state (202). The two transient openings (201) in FIG. 2 demarcate the time for the reaction of cleaving the triphosphate, incorporating the new nucleotide, and translating the DNA, shown as τ_c (the closed interval, or 202). The width of the open state is shown between arrows as τ_o . The current (203) in between the first opening (204) and the subsequent reopening contains features or characteristics (203) that reflect the incorporation of the nucleotide captured in the first opening (204).

[0036] As described further herein, the various features or characteristics of the bioelectronic signature of an active polymerase can be used to determine the sequence of a polynucleotide template. Additionally, as would be recognized by one of skill in the art based on the present disclosure, the methods of obtaining a bioelectronic signature and extracting various characteristics described herein can be used to determine the sequence of any biopolymer and any corresponding enzyme-of-interest, including but not limited to

polymerase, a nuclease, a proteasome, a glycopeptidase, a glycosidase, a kinase and an endonuclease.

[0037] As further provided in FIG. 3, the selection of a complementary dNTP occurs during the open state (201), and the distribution of values of τ_0 is sensitive to the composition of the dNTP solution (e.g., concentration). FIG. 3 shows measured probabilities of a given value of τ_0 for (301 - dots) a homopolymer template consisting of 10 A bases in a 1 mM solution of the complementary dNTP, i.e., dTTP. The fitted distribution curve (302) is a sharp Gaussian with a peak at 0.16 ms and a width of 0.3 ms.

[0038] In contrast, when the polymerase has to search for the complementary dNTP, the distribution of open times is much broader, as shown by the squares (303). In this example, the template comprised a five-fold repeat of the sequence ATC in a mM solution of all four dNTPs. Now the distribution is fitted by an exponential (304), with open times that are as long as 3 ms (though less than 1% of all values exceed about 2 ms, as marked by the box 305).

[0039] These times reflect the intrinsic response of the polymerase, as can be seen by considering the time between incoming dNTPs at these mM concentrations.

[0040] The flux, I , of molecules into a sphere of radius R_p (by which is meant the radius of the polymerase, ~ 3 nm) for a molecule diffusion constant D and a concentration $[C]$ particles/m³ is

$$[0041] \quad I = 4\pi R_p D [C]$$

[0042] The diffusion constant of the dNTPs is given by the Einstein-Smoluchowski relation and Stoke's law as

$$[0043] \quad D = \frac{kT}{6\pi\eta R_N}$$

[0044] Here, R_N is the radius of a dNTP and η is the viscosity of water (10^{-3} Pa.s at 300K).

[0045] Accordingly

$$[0046] \quad I = \frac{2}{3} \frac{kT}{\eta} \left(\frac{R_p}{R_N} \right) [C]$$

[0047] where I is the number of particles entering the polymerase per second for a concentration in the bulk of $[C]$ is in particles/m³. This can be expressed in terms of the Molarity (M) times $1000 \times N_A$ (N_A is Avogadro's Number) giving

$$[0048] \quad I = \frac{2}{3} \frac{kT}{\eta} \left(\frac{R_p}{R_N} \right) [M] \times 6 \times 10^{26}$$

[0049] At 300K, $kT = 4.14 \times 10^{-21}$ J, and η for water is 10^{-3} Pa.s, so

$$\{0050\} \quad I \approx 3 \left(\frac{R_p}{R_N} \right) [M] \times 6 \times 10^8 = 1.8 \left(\frac{R_p}{R_N} \right) [M] \times 10^9 \text{ per second} \quad (1)$$

\{0051\} With $R_p/R_n \sim 3$ this give a flux of about 5×10^6 at 1 mM dNTP (or a time between arrivals of about 0.2 μ s). Thus, dNTPs are arriving at the polymerase at the rate of more than 100 per fastest opening event. T times shown in FIG. 3, therefore, represent the intrinsic response time of the polymerase. In the case where the correct dNTP is always present (301), this is consistently about 0.16 ms. When the polymerase has to stay open to find the correct complementary nucleotide among the four possible dNTPs, the distribution of times is much broader (303). Nearly all search events are over by about 3 ms at mM concentration of dNTPs.

\{0052\} Equation 1 can be used to predict the arrival rate of dNTPs at a reduced concentration. Importantly, this arrival rate will be distributed according to a Poisson distribution:

$$\{0053\} \quad \exp \left(-\frac{(t-\mu)^2}{2\mu} \right) \quad (2)$$

\{0054\} which has the special property that the mean interval between arrivals, μ is also the variance of the distribution.

\{0055\} FIG. 4 shows the intrinsic distribution of open times for the polymerase at mM dNTP's as the exponential close to the origin (401). The adjacent distribution (402) is calculated using equation 2 with $\mu = 17$ ms. Equation 1 gives

$$\{0056\} \quad [M] = \frac{I}{5.4} \text{ nM}$$

\{0057\} As shown in this exemplary data, a 17 ms interval corresponds to $I=58.8$ or an 11 nM solution to give the distribution of arrival times shown by (402). In this case, the interval between arrivals is much longer than the intrinsic response of the polymerase, so this dominates the open state lifetime. Thus, if a first dNTP is present at mM concentration, and a second dNTP at 11 nM, the overlap of the open time distributions will be given by the overlap between curves 401 and 402 (these curves are all normalized so that the total probability =1). This overlap (405) is about 0.001 or about one tenth of a percent. A further dilution of a third dNTP to 3.8 nM would result in $\mu = 50$ ms, giving the curve labeled 403. A dilution of a fourth dNTP to 1.9 nM gives $\mu=100$ ms, with the resulting distribution plotted as the curve labeled as 404. For the dNTP represented by the curve 403, the overlap with curve 402 is 0.004 (406) and with curve 404 also 0.004 (407). Thus, the nucleotide being incorporated following an opening event can be identified by the duration of the open event to better than 1%.

[0058] In some embodiments, a first dNTP can be present in a solution at a concentration ranging from about 1 mM to about 10 mM, from about 1 mM to about 8 mM, from about 1 mM to about 6 mM, from about 1 mM to about 5 mM, from about 1 mM to about 4 mM, from about 1 mM to about 3 mM, or from about 1 mM to about 2 mM. In some embodiments, the first dNTP can be present in a solution at a concentration ranging from about 2 mM to about 10 mM, from about 4 mM to about 10 mM, from about 5 mM to about 10 mM, from about 6 mM to about 10 mM, from about 7 mM to about 10 mM, from about 8 mM to about 10 mM, or from about 8 mM to about 10 mM. In some embodiments, the first dNTP can be present in the solution at a concentration of about 1 mM, 2 mM, 3 mM, 4 mM, 5 mM, 6 mM, 7 mM, 8 mM, 9 mM, or 10 mM. Accordingly, in some embodiments, the second dNTP can be present in a solution at a concentration ranging from about 5 nM to about 15 nM, from about 10 nM to about 15 nM, from about 12 nM to about 15 nM, from about 5 nM to about 12 nM, from about 5 nM to about 10 nM, from about 5 nM to about 8 nM, from about 7 nM to about 12 nM, or from about 8 nM to about 10 nM. In some embodiments, the second dNTP can be present in the solution at a concentration of about 5 nM, 6 nM, 7 nM, 8 nM, 9 nM, 10 nM, 11 nM, 12 nM, 13 nM, 14 nM, or 15 nM. Accordingly, in some embodiments, the third dNTP can be present in a solution at a concentration ranging from about 1 nM to about 10 nM, from about 1 nM to about 8 nM, from about 1 nM to about 6 nM, from about 1 nM to about 5 nM, from about 2 nM to about 10 nM, from about 3 nM to about 10 nM, from about 2 nM to about 8 nM, or from about 2 nM to about 6 nM. In some embodiments, the third dNTP can be present in the solution at a concentration of about 1 nM, 2 nM, 3 nM, 4 nM, 5 nM, 6 nM, 7 nM, 8 nM, 9 nM, or 10 nM. Accordingly, in some embodiments, the fourth dNTP can be present in a solution at a concentration ranging from about 0.1 nM to about 5 nM, from about 0.1 nM to about 2.5 nM, from about 0.1 nM to about 1 nM, from about 0.5 nM to about 5 nM, from about 0.5 nM to about 2.5 nM, from about 1 nM to about 5 nM, from about 1 nM to about 4 nM, or from about 1 nM to about 2.5 nM. In some embodiments, the fourth dNTP can be present in the solution at a concentration of about 1.5 nM, 1.6 nM, 1.7 nM, 1.8 nM, 1.9 nM, 2 nM, 2.1 nM, 2.2 nM, 2.3 nM, 2.4 nM or 2.5 nM. As would be recognized by one of ordinary skill in the art based on the present disclosure, these concentrations of dNTPs are not meant to be limiting, and can be adjusted based on various aspects of the methods described herein (e.g., template sequence and structure).

[0059] As shown in Tables 1 and 2 (below), the distribution of open and closed times varies with sequence and template structure. These data show the measured distributions of times for the following sequences: (1) AAAAAAAAAA (SEQ ID NO: 1) - single stranded oligomer (A10), dTTP only in the polymerization buffer; (2) ATCATCATCATCATC (SEQ ID NO: 2) - single stranded oligomer (ATC5), all 4 dNTPs present; and (3) catctactacgcttagcttgctatcatctatgcttagcatga (SEQ ID NO: 3) - circular template, all 4 dNTPs present.

[0060] Table 1: Open state times for three template sequences.

Sequence	Open state 1	Fraction state 1	Open State 2	Fraction state 2
A10	0.26 ms	1.00		
ATC5	0.22 ms	0.76	1.47 ms	0.24
42 nt circle	0.12 ms	0.85	0.78 ms	0.15

[0061] Table 2: Closed state times for three template sequences.

Sequence	Closed state 1	Fraction state 1	Closed State 2	Fraction state 2
A10	0.37 ms	1.00		
ATC5	0.35 ms	0.71	2.7 ms	0.29
42 nt circle	0.12 ms	0.56	1.96 ms	0.44

[0062] The homopolymer A10 is characterized by just one open state and one closed state. The heteropolymer ATC5 is characterized by 2 open times, one as fast as that for A10 being about 3/4 of the events with 1/4 being much slower (Table 1). Likewise, the majority (about 3/4) of its closed states are of as short a duration as for the homopolymer, with the remaining 1/4 being much slower (Table 2). The circular template, with a heteropolymer sequence, also manifests two states in both its open and closed states, but both events are faster than the events in the linear polymers.

[0063] Taken together, these data show that the open times are sensitive to the nucleotide composition of the buffer. In the case of A10, where only dTTPs are present, there is just one (fast) open time. When all four nucleotides are present, there are two open states. The short open state likely corresponds to capture of the correct nucleotide at first try, while the longer open times likely correspond to capture, followed by rejection of a non-complementary nucleotide. The data also show how the closed times (corresponding to the catalytic part of the cycle) also depend on sequence. For the homopolymer A10, there is only one fast closed time. The heteropolymers both have two distinct closed times, one fast and the second almost ten times longer, which illustrates how some of the nucleotide incorporations take longer.

[0064] Referring to the data represented in FIG. 5, the ability to associate a given closed event (202) with the incorporation of a particular nucleotide allows for further identification of signal features or characteristics associated with a particular nucleotide. Additionally, the incorporation of a particular nucleotide on top of a particular previously incorporated nucleotide also allows for further identification of signal features or characteristics associated with a particular nucleotide, such that signal features in the closed interval reflect 16 such combinations (“base stacking”). While the open states represent large changes in current relative to the closed states, the changes of current in the closed state are stochastic and subject to noise. However, the underlying levels can be extracted in a model-independent manner using, for example, the infinite Hidden Markov model together with Bayesian nonparametric methods. In consequence, characteristic levels can be located in a model-free manner, as shown by the “hidden” underlying states (501) in FIG. 5.

[0065] In accordance with the embodiments described herein, the present disclosure provides methods for sequencing a polynucleotide using a bioelectronic device. In some embodiments, the method includes introducing a template polynucleotide to the bioelectronic device, and introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide. In some embodiments, each dNTP is present in the solution at a pre-defined concentration. In some embodiments, the method includes obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated into the template polynucleotide. In some embodiments, at least one characteristic of the bioelectronic signature identifies each of the complementary dNTPs incorporated into the template polynucleotide. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to at least first electrode and a second electrode. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to both a first electrode and a second electrode.

[0066] In some embodiments, the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state. In some embodiments, the duration of the open period is distinct for each dNTP monomer such that it identifies whether a particular dNTP monomer has been incorporated into the template polynucleotide. In some embodiments, the solution comprises four dNTP monomers. In some embodiments, a first dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of a second dNTP. In some embodiments, a second dNTP is

present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the first dNTP and a third dNTP. In some embodiments, the third dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the second dNTP and a fourth dNTP. In some embodiments, the fourth dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the third dNTP. In accordance with these embodiments, the sequence of the polynucleotide template can be accurately determined from the duration of each open period. In some embodiments, the sequence of the polynucleotide can be accurately determined from the duration of each open period and/or one or more characteristics of the closed period.

[0067] In some embodiments, the duration of open periods for each dNTP are determined based on a distribution of a plurality of open duration periods. In some embodiments, the first dNTP is present at a saturating concentration. In some embodiments, extent of overlap is 1% or less. In some embodiments, the extent to which the distributions minimally overlap is 1% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.9% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.8% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.7% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.6% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.5% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.4% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.3% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.2% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.1% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.075% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.050% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.025% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.010% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.005% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.001% or less. In some embodiments, the extent to which the distributions minimally overlap is 0.0001% or less.

[0068] In some embodiments, the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state. In some embodiments, at least one characteristic of the closed period varies based on the previously incorporated nucleotide. In some embodiments, at least one characteristic of the closed period is identified using a method comprising machine learning. In some embodiments, the machine learning method comprises Hidden-Markov Modeling or Bayesian non-parametric analysis. In some embodiments, a combination of at least one characteristic of the closed period and at least one characteristic of the open period is used to identify each of the complementary dNTPs incorporated into the template polynucleotide. In some embodiments, the polynucleotide template is DNA. In some embodiments, the polynucleotide template is RNA. In some embodiments, the dNTP monomers comprise adenine (dATP), cytosine (dCTP), guanine (dGTP), thymine (dTTP), and/or uridine (dUTP), including any derivatives or variants thereof.

[0069] As persons of ordinary skill in the art will readily recognize and appreciate after having benefited from the teachings of the present disclosure, the methods described herein can be used with any bioelectronic device that senses the duration of the open and closed states of an enzyme (e.g., polymerase). Exemplary devices include, but are not limited to, the bioelectronic devices and systems disclosed in U.S. Patent No. 10,422,787 and PCT Appl. No. PCT/US2019/032707, both of which are herein incorporated by reference in their entirety and for all purposes. Additionally, it will be readily recognized and appreciated by those of ordinary skill in the art based on the present disclosure that the foregoing embodiments apply equally to (and include) sequencing RNAs with the substitution of rNTPs for dNTPs and the use of an RNA polymerase.

[0070] In accordance with these embodiments, the polymerase can be functionally coupled to the first and second electrodes using a linker comprising thio-streptavidin. In some embodiments, the polymerase is biotinylated. In some embodiments, the linker is attached to a region of the polymerase that is inactive. In some embodiments, the polymerase and the first and second electrodes are biotinylated, and the linker comprises a streptavidin molecule comprising at least two biotin binding sites. In some embodiments, the exonuclease activity of the polymerase is disabled. In some embodiments, the gap has a width of about 1.0 nm to about 20.0 nm. In some embodiments, the first and second electrodes are separated by a dielectric layer. In some embodiments, the method comprises applying a voltage bias between the first and second electrodes that is 100mV or less.

[0071] Embodiments of the present disclosure also include a system for direct electrical measurement of polymerase activity. In accordance with these embodiments, the system includes any of the bioelectronic devices described herein, a means for introducing dNTPs capable of interacting with the polymerase, a means for applying a voltage bias between the first and second electrodes that is 100mV or less, and a means for monitoring fluctuations that occur as the dNTPs are incorporated into a template polynucleotide by the polymerase.

[0072] Embodiments of the present disclosure also include a method of calibrating a bioelectronic device. In accordance with these embodiments, the method includes introducing a template polynucleotide to the bioelectronic device; introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide, each dNTP present in the solution at a saturating concentration; obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated into the template polynucleotide, wherein the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state; and measuring or determining the intrinsic distribution of the open periods for each dNTP.

[0073] In some embodiments, the bioelectronic device is calibrated based on the distribution of open periods. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to at least a first electrode and a second electrode. In some embodiments, the bioelectronic device comprises a polymerase functionally coupled to both a first electrode and a second electrode. In some embodiments, the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state, and the bioelectronic device is calibrated based on at least one characteristic of the closed period.

[0074] Embodiments of the present disclosure also include methods of calling bases in electrical signals from a polymerase protein spanning a junction. These methods include measuring the intrinsic distribution of opening times for a polymerase functioning in a saturating concentration of dNTPs; repeating the measurement in a solution in which at least one dNTP is diluted such that its incorporation can be identified by corresponding increased time of the open state; characterizing signal features of both the open state and the following closed state in terms of the nucleotide being incorporated and the previously incorporated nucleotide, wherein the nucleotide is first identified using the dilution method described herein; and optimizing the dilutions of each nucleotide and the use of signals parameters so that the desired sequencing accuracy is obtained at the fastest read rate.

[0075] In some embodiments, the methods provided herein include measuring or determining the opening state of a polymerase by introducing a first solution comprising a DNA template to a device, wherein the device comprises a first and a second electrode separated by a gap, and a polymerase attached to the first and second electrodes; introducing a second solution comprising four dNTPs to the product of step the previous step under conditions that allow for incorporation of the dNTP complementary to the DNA template, wherein the dNTPs are present in the solution at saturating concentrations; and measuring the intrinsic distribution of opening times for the polymerase.

[0076] In some embodiments, the methods of the present disclosure include calibrating a sequencing device comprising a first and a second electrode separated by a gap and a polymerase attached to the first and the second electrode. In accordance with these embodiments, the method includes introducing a first solution comprising a DNA template to a device, wherein the device comprises a first and a second electrode separated by a gap, and a polymerase attached to the first and second electrodes; introducing a second solution comprising four dNTPs to the product of the previous step under conditions that allow for incorporation of the dNTP complementary to the DNA template, wherein the dNTPs are present in the solution at saturating concentrations; and measuring the intrinsic distribution of opening times for the polymerase, wherein the sequencing device is calibrated from the measured intrinsic distribution of opening times.

[0077] In some embodiments, the methods of the present disclosure include identifying a base incorporated into a strand of DNA. In accordance with these embodiments, the method includes introducing a first solution comprising a DNA template to a device, wherein the device comprises a first and a second electrode separated by a gap, and a polymerase attached to the first and second electrodes, and wherein the device has been calibrated according to the method described above; introducing a second solution comprising four dNTPs to the product of the previous step under conditions that allow for incorporation of the dNTP complementary to the DNA template, wherein the first dNTP is present in the solution at a concentration such that its distribution of arrival times minimally overlaps with the distribution of polymerase opening times in a saturated concentration of the second dNTP, the second dNTP is present in the solution at a concentration such that its distribution of arrival times minimally overlaps with the distribution of arrival times of the first dNTP, the third dNTP is present in the solution at a concentration such that its distribution of arrival time minimally overlaps with the distribution

of arrival times of the second dNTP and the fourth dNTP is present in the solution at a concentration such that its distribution of arrival times minimally overlaps with the distribution of arrival times of the third dNTP; and measuring current over time; wherein the base is identified from (or based on) the known distribution opening time of the polymerase in a given set of concentrations of nucleotides.

[0078] In some embodiments, the methods of the present disclosure include sequencing DNA. In accordance with these embodiments, the method includes introducing a first solution comprising a DNA template to a device, wherein the device comprises a first and a second electrode separated by a gap, and a polymerase attached to the first and second electrodes, and wherein the device has been calibrated according to the method described above; introducing a second solution comprising four dNTPs to the product of the previous step under conditions that allow for incorporation of the dNTP complementary to the DNA template, wherein the first dNTP is present at a saturating concentration in the solution, the second dNTP is present in the solution at a concentration such that its distribution of arrival times overlaps minimally with the distribution of arrival times of the first dNTP, the third dNTP is present in the solution at a concentration such that its distribution of arrival times overlaps minimally with the distribution of arrival times of the second dNTP and the fourth dNTP is present in the solution at a concentration such that its distribution of arrival time overlaps minimally with the distribution of arrival times of the third dNTP; and measuring current over time; wherein the DNA is sequenced from (or based on) the known distribution opening times of the polymerase in the given concentrations of first, second, third and fourth dNTPs.

[0079] In some embodiments, the methods of the present disclosure include improving the accuracy of biopolymer sequencing systems and methods (e.g., DNA sequencing, RNA sequencing, or other biopolymer sequencing). In accordance with these embodiments, the method includes collecting recordings of current over time according to the methods described above, and collecting the portions of the current signal from the closed state in between the open state signals and sorting them in terms of the nucleotide incorporated at a given opening event, and the nucleotide incorporated in the prior event to yield a collection of a plurality of sets (e.g., 16 sets) of closed states signals, each one of which is associated with incorporation of a given pair of nucleotides in two sequential incorporation events. In some embodiments, the method includes applying one or more machine learning methods to locate signal features in the closed-state current associated with a given pair of nucleotides in two sequential

incorporation events. In some embodiments, machine-learning methods may include Hidden-Markov Modeling or Bayesian non-parametric analysis, for example.

CLAIMS

What is claimed is:

1. A method for sequencing a polynucleotide using a bioelectronic device, the method comprising:
 - (a) introducing a template polynucleotide to the bioelectronic device, wherein the bioelectronic device comprises a polymerase functionally coupled to at least one of a first electrode and a second electrode;
 - (b) introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide, each dNTP present in the solution at a pre-defined concentration; and
 - (c) obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated into the template polynucleotide;wherein at least one characteristic of the bioelectronic signature identifies each of the complementary dNTPs incorporated into to the template polynucleotide.
2. The method of claim 1, wherein the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state.
3. The method of claim 1 or claim 2, wherein duration of the open period is distinct for each dNTP monomer such that it identifies whether a particular dNTP monomer has been incorporated into the template polynucleotide.
4. The method of any of claims 1 to 3, wherein the solution comprises four dNTP monomers, and wherein a first dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of a second dNTP; wherein the second dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the first dNTP and a third dNTP; wherein the third dNTP is present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the second dNTP and a fourth dNTP; and wherein the fourth dNTP is

present in the solution at a concentration such that the duration of its open period minimally overlaps with the duration of the open period of the third dNTP.

5. The method of any of claims 1 to 4, wherein the duration of open periods for each dNTP are determined based on a distribution of a plurality of open duration periods.
6. The method of any of claims 1 to 5, wherein the first dNTP is present at a saturating concentration.
7. The method of any of claims 4 to 6, wherein extent of overlap is 1% or less.
8. The method of claim 1, wherein the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state.
9. The method of claim 8, wherein at least one characteristic of the closed period varies based on the previously incorporated nucleotide.
10. The method of claim 8 or claim 9, wherein the at least one characteristic of the closed period is identified using a method comprising machine learning.
11. The method of claim 10, wherein the machine learning method comprises Hidden-Markov Modeling or Bayesean non-parametric analysis.
12. The method of any of claims 1 to 11, wherein a combination of at least one characteristic of the closed period and at least one characteristic of the open period is used to identify each of the complementary dNTPs incorporated into the template polynucleotide.
13. The method of any of claims 1 to 12, wherein the polynucleotide template is DNA.
14. The method of any of claims 1 to 13, wherein the exonuclease activity of the polymerase is disabled.

15. The method of any of claims 1 to 14, wherein the polymerase is functionally coupled to the first and second electrodes using a linker comprising thio-streptavidin.

16. The method of claim 15, wherein linker is attached to a region of the polymerase that is inactive.

17. The method of any of claims 1 to 16, wherein the method comprises applying a voltage bias between the first and second electrodes that is 100mV or less.

18. The method of any of claims 1 to 17, wherein the dNTP monomers comprise adenine (dATP), cytosine (dCTP), guanine (dGTP), and thymine (dTTP).

19. A method of calibrating a bioelectronic device, the method comprising:

(a) introducing a template polynucleotide to the bioelectronic device, wherein the bioelectronic device comprises a polymerase functionally coupled to at least a first electrode and a second electrode;

(b) introducing a solution comprising dNTP monomers to the device comprising the template polynucleotide, each dNTP present in the solution at a saturating concentration;

(c) obtaining a bioelectronic signature of polymerase activity based on current fluctuations as each complementary dNTP monomer is incorporated into the template polynucleotide, wherein the bioelectronic signature comprises an open period corresponding to the polymerase being in an open state; and

(d) measuring or determining the intrinsic distribution of the open periods for each dNTP, wherein the bioelectronic device is calibrated based on the distribution of open periods.

20. The method of claim 19, wherein the bioelectronic signature comprises a closed period corresponding to the polymerase being in a closed state, and wherein the bioelectronic device is calibrated based on at least one characteristic of the closed period.