

## DESCRIPTION

### TITLE OF INVENTION: [POLYPEPTIDE COMPOSITION ANALYSIS METHOD BASED ON COPPER ION MODIFIED MSPA NANOPORES]

**[0001]** The present application claims priority from Chinese invention patent application CN2021110067648 “A method for detecting various amino acids and a method for analyzing polypeptide composition based on nanopores” filed on August 30, 2021, which is incorporated by reference in its entirety.

### TECHNICAL FIELD

**[0002]** The present invention relates to the field of nanopore detection, and in particular to a polypeptide composition analysis method based on copper ion modified MspA nanopores.

### BACKGROUND

**[0003]** Nanopore sensing is a single-molecule sensing technology, with detection principles similar to those of a Coulter counter. This technology has the advantages of real-time and direct monitoring at the single-molecule level, usually without the need for labeling or modification of the analyte. These advantages make nanopores an emerging technology for biosensing and biodetection.

**[0004]** Due to the low signal-to-noise ratio and unpredictable conformational changes during the translocation of amino acids through nanopores, direct discrimination and detection of amino acids remain a challenge. Existing technologies cannot directly detect and distinguish all 20 amino acids effectively, often requiring chemical modifications of nanopores or derivatization of amino acids.

### SUMMARY

**[0005]** In view of this, the present invention provides a polypeptide composition analysis method based on M2MspA-N91H nanopore, comprising the following steps:

S1 Mixing an exopeptidase with a polypeptide to be tested uniformly and reacting to obtain a hydrolysis product;

S2 Adding the hydrolysis product to a nanopore system, wherein the nanopore system comprises an M2MspA-N91H nanopore, an insulating membrane, a first medium, and a second medium, wherein the M2MspA-N91H nanopore is embedded in the insulating membrane, the M2MspA-N91H nanopore provides a channel that connects the first medium and the second medium, and the hydrolysis product is added to the first medium;

S3 Applying a driving force between the first medium and the second medium, a free amino acid in the hydrolysis product passes through the M2MspA-N91H nanopore and generates an electrical signal;

S4 Data-analyzing the electrical signal to obtain information on amino acid type of the free amino acid.

**[0006]** In one embodiment, the free amino acid includes a natural amino acid and a non-natural amino acid.

**[0007]** In one embodiment, the natural amino acid includes one or more of glycine, alanine, valine, leucine, isoleucine, phenylalanine, tryptophan, tyrosine, aspartic acid, histidine, asparagine, glutamic acid, lysine, glutamine, methionine, arginine, serine, threonine, cysteine, and proline.

**[0008]** In one embodiment, the non-natural amino acid includes one or more of citrulline, L-norvaline, and ibotenic acid.

**[0009]** In one embodiment, the exopeptidase includes aminopeptidase and/or carboxypeptidase.

**[0010]** In one embodiment, before adding the hydrolysis product to the nanopore system, the M2MspA-N91H nanopore is combined with a copper ion.

**[0011]** In one embodiment, pH value of the first medium and the second medium includes 6.5 - 7.5.

**[0012]** In one embodiment, final concentration of solution of the polypeptide to be tested is greater than 10  $\mu\text{M}$ .

**[0013]** In one embodiment, step of the data-analyzing in the S4 further includes:

- i) Establishing a standard signal database;
- ii) Setting a signal filtering condition;
- iii) Excluding part of the electrical signal to obtain a retained electrical signal;
- iv) Determining the amino acid type corresponding to the retained electrical signal based on blockage rate of a standard signal in the standard signal database;
- v) Obtaining information on the amino acid type of the free amino acid.

**[0014]** In one embodiment, the signal filtering condition is to exclude an electrical signal with an SD value greater than an SD threshold, wherein the SD value of the electrical signal is calculated according to the following formula:

$$SD = \sqrt{\frac{\sum(x_i - x)^2}{n-1}}$$

Wherein,  $x_i$  is current value of each sampling point of the signal,  $x$  is mean current value of the signal, and  $n$  is number of sampling point of the signal.

**[0015]** In one embodiment, the method further comprises S5 Comparing the information on the amino acid type of the free amino acid with expected sequence information of the polypeptide to be tested, to determine whether there is a difference between actual sequence and the expected sequence of the polypeptide to be tested; or

S5 Comparing the information on the amino acid type of the free amino acid combined with information on the exopeptidase with polypeptide sequence in database, to predict actual sequence of the polypeptide to be tested.

**[0016] Beneficial effects**

**[0017]** Nanopores in the prior art typically could only detect amino acids with specific functional groups. However, experiments in the present invention demonstrate that the M2MspA-N91H nanopore (derived from porin A of *Mycobacterium smegmatis*) exhibits universality in detecting amino acids. It can robustly and sensitively detect and distinguish the 20 common amino acids and non-natural amino acids at the single-molecule level. Additionally, there is no need to introduce enzymes and/or modify amino acids (e.g., by adding fluorescent labels) into the M2MspA-N91H nanopore system. Both the 20 common amino acids and the non-natural amino acids can block the M2MspA-N91H nanopore and cause corresponding current blockage signals. Direct detection and discrimination of various amino acids can be achieved by analyzing different current signals.

**[0018]** Based on this, combined with the hydrolysis characteristics of exopeptidases (such as carboxypeptidase), the present invention provides a polypeptide composition analysis method based on the M2MspA-N91H nanopore. By determining the amino acid type corresponding to the detected free amino acid signals based on the blockage rate of standard signals established in the standard signal database, polypeptide composition analysis of the polypeptide to be tested can be achieved to a certain degree. The polypeptide composition analysis method provided by the present invention is applicable to various application scenarios such as judgment of expression of tumor neoantigens and synthesis quality of peptide drugs, and prediction of unknown polypeptide sequences.

**DESCRIPTION OF DRAWINGS**

**[0019]** To make the embodiments of the present invention or the technical solutions in the prior art clearer, the drawings required to be used in the description

of the embodiments or the prior art will be briefly introduced below. It is obvious that the drawings described below are some embodiments of the present invention, and that other drawings can be obtained from these drawings for those of ordinary skill in the art without making inventive effort.

**[0020]** **[Fig. 1]:** illustrates the structure and electrophysiological characteristics of the M2MspA-N91H nanopore.

**[0021]** **[Fig. 2]:** illustrates the single-channel current signals of the M2MspA-N91H nanopore when different concentrations of copper ions are added (1 M KCl, pH 6).

**[0022]** **[Fig. 3]:** illustrates the current signals of amino acid detection by the M2MspA-N91H nanopore at a concentration of 3 M KCl.

**[0023]** **[Fig. 4]:** illustrates the scatter plot of amino acid volume versus current signal blockage rate (letters denote amino acid abbreviations).

**[0024]** **[Fig. 5]:** illustrates the plot of amino acid current signal blockage rate versus blockage time of amino acid (letters denote amino acid abbreviations).

**[0025]** **[Fig. 6]:** illustrates the experimental plots for extracting individual amino acid signals using Clampfit (A illustrates the single-channel statistical results by Clampfit; B illustrates the Amp S.D. values of the signals; C illustrates the distinguished single-step signals in the scatter plot based on SD values (from top to bottom: statistics of all signals for 2 minutes in the blank group, asp experiment group, asp+glu experiment group from the same experiment, respectively); D illustrates the source of background signals at -5 pA; E illustrates the source of background signals at -12 pA).

**[0026]** **[Fig. 7]:** illustrates the plots of current signals obtained after adding carboxypeptidase hydrolysis products to the nanopore system (A illustrates the plots of current signals obtained after adding hydrolysis products of 10 mM EF peptide

hydrolyzed by 10 U carboxypeptidase under the conditions of 0.5 M KCl, 10 mM MOPS, pH 7.5; B illustrates the scatter plot of blockage rate and translocation time of all current signals counted from experiments of A; C shows the detection of hydrolysis products of three peptides (P1, P2, P3) by the nanopore).

**[0027] [Fig. 8]:** illustrates the experimental plots for detecting non-natural amino acids using the M2MspA-N91H nanopore (A illustrates the statistical results of three non-natural amino acids; B illustrates the scatter plot of current blockage signals within 1 minute in the blank group (Blank) and after adding ibotenic acid from the same experiment).

**[0028] [Fig. 9]:** illustrates the current traces after adding copper ions to the M2MspA-N91H nanopore under three pH conditions.

**[0029] [Fig. 10]:** illustrates the schematic diagram of nanopore detection of polypeptide hydrolysis products in the present invention.

**[0030] [Fig. 11]:** illustrates the flowchart of establishing a standard signal database in the embodiments of the present invention.

**[0031] [Fig. 12]:** illustrates the flowchart of data analysis in the embodiments of the present invention.

## **DETAILED DESCRIPTION**

**[0032]** To make the objective, the technical solutions and advantages of the embodiments of the present invention clearer, the technical solutions of the embodiments of the present invention will be clearly and completely described below in combination with drawings. It is obvious that the described embodiments are some of the embodiments of the present invention, not all of the embodiments. Based on the embodiments of the present invention, all other embodiments obtained by those of ordinary skill in the art without making inventive effort shall belong to the protection scope of the present invention.

**[0033]** It should be noted that the term “include”, “comprise” or any variant thereof is intended to encompass nonexclusive inclusion so that a process, method, article or device including a series of elements includes not only those elements but also other elements which have been not listed definitely or an element(s) inherent to the process, method, article or device. Moreover, the expression “comprising a(n) . . .” in which an element is defined will not preclude presence of an additional identical element(s) in a process, method, article or device comprising the defined element(s) unless further defined.

**[0034]** As used herein, the term “about”, typically means  $\pm 5\%$  of the stated value, more typically  $\pm 4\%$  of the stated value, more typically  $\pm 3\%$  of the stated value, more typically,  $\pm 2\%$  of the stated value, even more typically  $\pm 1\%$  of the stated value, and even more typically  $\pm 0.5\%$  of the stated value.

**[0035]** Throughout this application, embodiments of this invention may be presented with reference to a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as “from 1 to 6” should be considered to have specifically disclosed subranges such as “from 1 to 3”, “from 1 to 4”, “from 1 to 5”, “from 2 to 4”, “from 2 to 6”, “from 3 to 6”, etc.; as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

**[0036]** Nanopore system

**[0037]** The "nanopore system" includes a nanometer-scale pore (referred to as "nanopore"), an insulating membrane, a first medium, and a second medium. In one embodiment of the present invention, the nanometer-scale pore is M2MspA-N91H nanopore. The nanometer-scale pore allows analytes to translocate from one

side of the insulating membrane to the other. Additionally, those skilled in the art may modify the aforementioned M2MspA-N91H nanopore according to actual conditions (such as mutations, truncations, fusions, chemical modifications, etc.) to obtain corresponding variants, and the modification methods are well-known in the art.

**[0038]** In one embodiment of the present invention, the nanometer-scale pore is embedded in the insulating membrane, where the insulating membrane (also understood as the composite of the nanometer-scale pore and the insulating membrane) separates the first medium from the second medium, and the channel of the nanometer-scale pore provides a channel that connects the first medium with the second medium; after applying a driving force between the first medium and the second medium, the analyte in the first medium interacts with the M2MspA-N91H nanopore to generate a current (i.e., an electrical signal). In the present invention, the "first medium" refers to the medium in which the analyte is located when added to the nanopore system; the "second medium" refers to the other side of the insulating membrane from the first medium. In the present invention, the driving force refers to the force that drives the interaction of the analyte with the nanopore, such as through potential difference, electro-osmotic flow, or concentration gradient.

**[0039]** The first medium and the second medium may be the same or different, and the first medium and the second medium may include a conductive liquid. The conductive liquid is an alkali metal halide aqueous solution, specifically including sodium chloride (NaCl), lithium chloride (LiCl), cesium chloride (CsCl), potassium chloride (KCl), and sodium bromide (NaBr). In one embodiment of the present invention, the concentrations of the conductive liquid in the first medium and the second medium are different, in other words, there is a difference in the concentrations of conductive liquid between the first medium and the second medium, resulting in a difference in osmotic pressure on both sides of the insulating membrane. The first medium and/or the second medium may also include a buffer, such as HEPES, MOPS. The concentration range of the first medium and/or the second medium may be 0.3 M - 3 M. The pH range of the first medium and/or the second medium may be 6 – 9 (preferably 6.5 - 7.5).

**[0040]** The insulating membrane refers to a membrane capable of carrying a nanopore and blocking the passage of ion currents not through the nanopore. The insulating membrane may include a lipid membrane and/or a polymer membrane. Exemplary lipid membranes include DPHPC, DOPC, E. coli lipid, and exemplary polymer membranes include triblock copolymer polymer membranes.

**[0041]** In a specific embodiment of the present invention, the nanopore system comprises two electrolyte chambers separated by the insulating membrane to form a -trans chamber and a -cis chamber, with the pore of the M2MspA-N91H channel embedded in the insulating membrane, and only the M2MspA-N91H channel present on the insulating membrane connects the two electrolyte chambers. When a potential is applied to the two electrolyte chambers, electrolyte ions in the solution of the electrolyte chambers move by electrophoresis and pass through the the M2MspA-N91H channel.

**[0042]** Analyte

**[0043]** The analyte is a charged substance. An analyte is charged if it has a net charge. The analyte may be negatively charged or positively charged. An analyte is negatively charged if it has a net negative charge. An analyte is positively charged if it has a net positive charge. Preferably, suitable analytes are amino acids.

**[0044]** In one embodiment of the present invention, the amino acid may be a natural amino acid. Natural amino acids are naturally occurring amino acids and post-translationally modified variants thereof, such as the 20 common amino acids (glycine, alanine, valine, leucine, isoleucine, methionine, proline, tryptophan, serine, threonine, cysteine, phenylalanine, asparagine, glutamine, tyrosine, aspartic acid, glutamic acid, lysine, arginine, and histidine), etc.

**[0045]** In another embodiment of the present invention, the analyte may be a non-natural amino acid. "Non-natural amino acid" refers to amino acids that are not protein amino acids or post-translationally modified variants thereof (i.e., not the 20 common amino acids or pyrrolysine or selenocysteine or post-translationally

modified variants thereof). In a specific embodiment of the present invention, the non-natural amino acids are citrulline, L-norvaline, and ibotenic acid.

**[0046]** Interaction between the nanopore and analyte

**[0047]** The analyte may be contacted with the nanopore on either side of the insulating membrane. The analyte may be contacted with the either side of the insulating membrane that allows the analyte to pass through the channel of the nanopore to the other side of the insulating membrane. In this case, the analyte interacts with the nanopore as it passes across the insulating membrane via the channel of the pore. Alternatively, the analyte may be contacted with the side of the insulating membrane that allows the analyte to interact with the nanopore, dissociate from the nanopore and remain on the same side of the insulating membrane. The analyte can interact with the nanopore in any manner and at any position. Further, the analyte may collide with the nanopore that allows the analyte to interact with the nanopore, dissociate from the nanopore and remain on the same side of the insulating membrane.

**[0048]** During the interaction between the analyte and the nanopore, the analyte affects the current flowing through the nanopore in a manner specific for that analyte. In other words, the current flowing through the nanopore is distinctive for a particular analyte. Control experiments may be carried out to determine the effect of a particular analyte has on the current flowing through the nanopore, and then to identify a particular analyte in the sample or determine whether a particular analyte is present in the sample. More specifically, the current pattern obtained by detecting an analyte can be compared with a known current pattern obtained under the same condition using a known analyte to identify the presence or concentration of the analyte.

**[0049]** The nanopore system of the present invention may also include one or more measurement devices for measuring the current passing through the nanopore, such as patch clamp amplifiers or data acquisition devices.

**[0050]** A polypeptide composition analysis method based on M2MspA-N91H nanopore

**[0051]** The present invention also provides a polypeptide composition analysis method based on M2MspA-N91H nanopore, comprising the following steps:

S1 Mixing an exopeptidase with a polypeptide to be tested uniformly and reacting to obtain a hydrolysis product;

S2 Adding the hydrolysis product to a nanopore system, wherein the nanopore system comprises an M2MspA-N91H nanopore, an insulating membrane, a first medium, and a second medium, wherein the M2MspA-N91H nanopore is embedded in the insulating membrane, the M2MspA-N91H nanopore provides a channel that connects the first medium and the second medium, and the hydrolysis product is added to the first medium;

S3 Applying a driving force between the first medium and the second medium, a free amino acid in the hydrolysis product passes through the M2MspA-N91H nanopore and generates an electrical signal;

S4 Data-analyzing the electrical signal to obtain information on amino acid type of the free amino acid.

**[0052]** "Polypeptide (peptide)" refers to a peptide or protein containing two or more amino acids connected by peptide bonds. Polypeptides may contain natural, modified, or synthetic amino acids. Polypeptides may also be naturally modified (e.g., through post-translational processing) or chemically modified (e.g., amidation, acylation, cross-linking, etc.). In the present invention, the sequence information of the polypeptide may be known or unknown.

**[0053]** "Exopeptidase" refers to a hydrolytic enzyme that removes terminal amino acids from peptides by cleaving peptide bonds. Terminal amino acids are amino acids within approximately 10 amino acids from the N-terminus or C-terminus of the peptide. Therefore, "exopeptidases" are divided into aminopeptidases, which

hydrolyze amino acids one by one from the N-terminus of the peptide, and carboxypeptidases, which hydrolyze amino acids one by one from the C-terminus of the peptide. Exemplary aminopeptidases include cysteine aminopeptidase, prolyl aminopeptidase, and arginyl aminopeptidase. Exemplary carboxypeptidases include carboxypeptidase A, B, C, and Y. In the present invention, the exopeptidase may be carboxypeptidase and/or aminopeptidase.

**[0054]** Under suitable conditions, the exopeptidase can completely or partially hydrolyze the polypeptide to be tested. For example, as shown in Fig. 10, carboxypeptidase A1 can cleave amino acids at the C-terminus except for lysine (K), arginine (R), and proline (P) (i.e., the stop point of hydrolysis is R, K, P). In other words, if the polypeptide to be tested does not contain these three amino acids (R, K, P), it will be completely hydrolyzed by carboxypeptidase A, and the information on the amino acid type of the free amino acid obtained from the hydrolysis product will represent the amino acid composition of the polypeptide to be tested. If the polypeptide to be tested contains one or more of these three amino acids (R, K, P), it will be partially hydrolyzed by carboxypeptidase A (hydrolysis stops at the stop point), and the information on the amino acid type of the free amino acid obtained from the hydrolysis product will represent the amino acid composition of the polypeptide to be tested before the stop point.

**[0055]** When the sequence of the polypeptide to be tested is known, the information on the amino acid type of the free amino acid obtained from the hydrolysis product can be compared to an expected sequence of the polypeptide to be tested, to determine if there is a difference between the actual sequence of the polypeptide to be tested and the expected sequence of the polypeptide to be tested, (e.g., absence of certain amino acid(s) that should be present in the hydrolysis product, presence of certain amino acid(s) that should not be present in the hydrolysis product). This method is applicable to application scenarios such as judgement of the expression of tumor neoantigens and the quality control of synthesized peptide drugs.

**[0056]** When the sequence of the polypeptide to be tested is unknown, the information on the amino acid type of the free amino acid obtained from the hydrolysis product can be compared to a database by machine learning or other methods in combination with information on the exopeptidase used (e.g., carboxypeptidase or aminopeptidase, whether there is a "stop point of hydrolysis"), thereby predicting the sequence of the polypeptide to be tested or narrowing down the possible range of the polypeptide to be tested to some extent (e.g., starting from the N-terminus or C-terminus, the composition of amino acids (including the determined types of free amino acids and the possible types corresponding to the "stop point of hydrolysis").

**[0057] Embodiment 1**

**[0058]** Methods and materials

**[0059]** Amino acid detection:

**[0060]** Single-channel experiments were conducted using a classical vertical lipid bilayer setup (Warner Instruments), consisting of two chambers (cis and trans) connected by a 150  $\mu\text{m}$  diameter pore. Two Ag-AgCl electrodes were used to apply transmembrane voltage and measure transmembrane ion currents. The cis chamber was grounded. Buffer for amino acid detection is as follows:

- 1) Buffer: 0.3 M - 3 M salt solutions such as KCl and NaCl (preferably 1 M KCl aqueous solution), 10 mM MOPS;
- 2) Copper ion concentration range: 1 nM to 200  $\mu\text{M}$ ;
- 3) pH range: 6 - 9 (preferably 6.5 - 7.5).

**[0061]** As the salt concentration or pH of the buffer is further reduced, the efficiency and discrimination of amino acid detection may decrease. All experiments were conducted at temperatures of 15°C - 60°C (preferably 22.0°C  $\pm$  0.5°C). A planar lipid bilayer membrane was formed by coating the pore connecting the two chambers with a membrane of 1,2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC) (Avanti Polar Lipids) dissolved in decane (40 mg/ml). Protein solution was

then added to the cis chamber, and after protein corporation into the lipid membrane to form a single channel, copper chloride was added to the trans chamber, and amino acids were added to the cis chamber. The solution was gently mixed, and the current signal was recorded.

**[0062]**      Data processing:

**[0063]**      Single-channel current recordings were amplified using an Axopatch 200B amplifier (Molecular Devices) and filtered at 2 kHz using the built-in four-pole low-pass Bessel filter. Data were digitized at a sampling rate of 100 kHz using a Digidata 1550B converter (Molecular Devices). The sampled current signal files were extracted to obtain amino acid signal data. Individual amino acid signals were identified and defined based on parameters such as signal blockage rate, blockage time, standard deviation (SD), etc.

**[0064]**      Background noise processing:

**[0065]**      For the distribution of current values of individual amino acid signals, the SD value was calculated to determine the deviation of the current values at each sampling point from the overall mean, thereby distinguishing individual amino acid signals from background noise and overlapping signals of multiple amino acids.

**[0066]**      Polypeptide composition analysis method:

**[0067]**      The buffer conditions for the experiment are as described above. The experimental temperature may be set at 15°C - 60°C (preferably 25°C). Outside the nanopore system, exopeptidases (such as carboxypeptidases) are mixed with polypeptides at certain concentration ratios, for example, 10 U of carboxypeptidase, 0.5 M KCl, 10 mM MOPS, pH 7.5, and 10 mM polypeptide (Fig. 7A). After incubating at 25°C for 20 minutes, the mixture is added to the nanopore sample chamber for detection. The composition of the polypeptide is determined based on the types and quantities of amino acid signals observed.

**[0068]        Embodiment 2: Detection of 20 natural amino acids using  
M2MspA-N91H Nanopore**

**[0069]**        Structure and characteristics of copper ion modified nanopore used:

**[0070]**        As shown in Fig. 1, the M2MspA-N91H nanopore has an octameric structure, with the eight amino acid residues at the narrowest part of the channel facing the central axis of the M2MspA-N91H nanopore. Adding  $\text{Cu}^{2+}$  to the trans chamber, and observing the current traces.  $\text{Cu}^{2+}$  can significantly reduce the current fluctuations (Fig. 1). The reduction of the current level is related to the concentration of  $\text{Cu}^{2+}$  added (Fig. 2). The copper ion concentration may range from 1 nM to 200  $\mu\text{M}$ .

**[0071]**        Detection of 20 natural amino acids:

**[0072]**        In this embodiment, 20 natural amino acids were detected using the M2MspA-N91H nanopore. Wherein, the amino acid sequence of the MspA monomer (MspA consists of eight identical monomers forming the channel) is:

GLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGA  
DEFEGTLELGYQIGFPWSLGVGINFSYTTPNILINHGNIAPPFGLNSVITPNLFPGVS  
ISARLGNGPGIQEVATFSVRVSGAKGGVAVSNAHGTVTGAAGGVLLRPFARLIAS  
TGD SVTTYGEPWNMN (SEQ ID NO:1)

**[0073]**        The amino acid sequence of M2MspA-N91H is:

GLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGA  
DEFEGTLELGYQIGFPWSLGVGINFSYTTPNILINHGNIAPPFGLNSVITPNLFPGVS  
ISARLGNGPGIQEVATFSVRVSGAKGGVAVSNAHGTVTGAAGGVLLRPFARLIAS  
TGD SVTTYGEPWNMGLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREW  
FHSGRAKYIVAGPGADEFEGTLELGYQIGFPWSLGVGINFSYTTPNILINHGNIAPP  
FGLNSVITPNLFPGVSISARLGNGPGIQEVATFSVRVSGAKGGVAVSNAHGTVTGA  
AGGVLLRPFARLIAS TGD SVTTYGEPWNMGLDNELSLVDGQDRTLTVQQWDTFL  
NGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEGTLELGYQIGFPWSLGVGINFS  
YTTPNILINHGNIAPPFGLNSVITPNLFPGVSISARLGNGPGIQEVATFSVRVSGAK

GGVAVSNAHGTVTGAAGGVLLRPFARLIASGDSVTTYGEPWNMNGLDNELSLVD  
 GQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEGTLELG  
 YQIGFPWSLGVGINFSYTTNPILINHGNIAPPFGLNSVITPNLFPGVVISARLGNGPG  
 IQEVATFSVRVSGAKGGVAVSNAHGTVTGAAGGVLLRPFARLIASGDSVTTYGEP  
 WNMNGLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIV  
 AGPGADEFEGTLELGYQIGFPWSLGVGINFSYTTNPILINHGNIAPPFGLNSVITPN  
 LFPGVVISARLGNGPGIQEVATFSVRVSGAKGGVAVSNAHGTVTGAAGGVLLRPFAR  
 LIASGDSVTTYGEPWNMNGLDNELSLVDGQDRTLTVQQWDTFLNGVFPLDRNR  
 LTREWFHSGRAKYIVAGPGADEFEGTLELGYQIGFPWSLGVGINFSYTTNPILINHG  
 NITAPPFGLNSVITPNLFPGVVISARLGNGPGIQEVATFSVRVSGAKGGVAVSNAHG  
 TVTGAAGGVLLRPFARLIASGDSVTTYGEPWNMNGLDNELSLVDGQDRTLTVQQ  
 WDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEGTLELGYQIGFPWSLG  
 VGINFSYTTNPILINHGNIAPPFGLNSVITPNLFPGVVISARLGNGPGIQEVATFSVR  
 VSGAKGGVAVSNAHGTVTGAAGGVLLRPFARLIASGDSVTTYGEPWNMNGLDNE  
 LSLVDGQDRTLTVQQWDTFLNGVFPLDRNRLTREWFHSGRAKYIVAGPGADEFEG  
 TLELGYQIGFPWSLGVGINFSYTTNPILINHGNIAPPFGLNSVITPNLFPGVVISARL  
 GNGPGIQEVATFSVRVSGAKGGVAVSNAHGTVTGAAGGVLLRPFARLIASGDSVT  
 TYGEPWNMN (SEQ ID NO:2)

**[0074]** When the target analyte passes through the Cu<sup>2+</sup>-chelated M2MspA-N91H, specific current signals are generated, as shown in Fig. 3. The current traces after the addition of copper ions at three pH conditions are shown in Fig. 9. The results indicate that copper ion binding is unstable at pH 7.9, while stable at pH 7.5 and 6.5. Considering that higher pH of solution favours more amino acids carrying negative charges, thereby enhancing detection efficiency, in this embodiment, MOPS buffer salt is added to the solution and the pH of the solution is stabilized around 7.5 to achieve optimal detection results.

**[0075]** From the scatter plot of amino acid volume and current signal blockage rate (Fig. 4), it can be observed that the blockage rate caused by amino acids is positively correlated with their hydrophobic volume.

**[0076]** From the plot of current signal blockage rate versus blockage time of amino acid (Fig. 5), the blockage rates caused by the 20 natural amino acids are as follows: Glycine (G):  $0.1207 \pm 0.0006$ ; Alanine (A):  $0.1471 \pm 0.0025$ ; Valine (V):  $0.1904 \pm 0.0012$ ; Leucine (L):  $0.1995 \pm 0.0003$ ; Isoleucine (I):  $0.2072 \pm 0.0003$ ; Phenylalanine (F):  $0.2201 \pm 0.0007$ ; Tryptophan (W):  $0.2274 \pm 0.0002$ ; Tyrosine (Y):  $0.2127 \pm 0.0040$ ; Aspartic acid (D):  $0.2140 \pm 0.0028$ ; Histidine (H):  $0.2465 \pm 0.0009$ ; Asparagine (N):  $0.1651 \pm 0.0002$ ; Glutamic acid (E):  $0.2452 \pm 0.0016$ ; Lysine (K):  $0.1687 \pm 0.0031$ ; Glutamine (Q):  $0.1882 \pm 0.0009$ ; Methionine (M):  $0.1977 \pm 0.0005$ ; Arginine (R):  $0.1714 \pm 0.0022$ ; Serine (S):  $0.1313 \pm 0.0002$ ; Threonine (T):  $0.1610 \pm 0.0004$ ; Cysteine (C):  $0.1862 \pm 0.0023$ ; Proline (P):  $0.2183 \pm 0.0010$ . Based on the blockage rate of amino acids and their standard deviation, amino acids that do not overlap include G, S, A, T, N, R, V, Q, M, L, I, D, F, W, E, totaling 15 amino acids that can be distinguished from each other.

**[0077]** Background noise processing:

**[0078]** The method described above for directly detecting the 20 amino acids may generate background noise that interferes with subsequent analysis. Based on this, the experimenters of the present invention calculate the standard deviation (SD) of the current value distribution for individual amino acid signals using the following formula. This calculation allows for assessing the deviation of the current values at each sampling point from the overall mean value, thereby distinguishing individual amino acid signals from background noise and overlapping signals from multiple amino acids.

**[0079]** 
$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

**[0080]** Where  $x_i$  represents the current value at each sampling point of the signal;  $\bar{x}$  represents the mean current value of the signal;  $n$  represents the number of sampling points of the signal.

**[0081]** As shown in Fig. 6C, the scatter plots represent the current blockade signals within 2 minutes for the blank group, the group with added aspartic acid (Asp), and the group with added mixture of aspartic acid and glutamic acid (Asp+Glu) from the same experiment. An SD threshold can be set to filter the signals. For example, the SD threshold can be set to 2, wherein signals with an SD value less than 2 (shown in blue) are considered valid signals, while signals with an SD value greater than 2 are considered background noise or overlapping signals generated by multiple amino acids passing through simultaneously, and can thus be filtered out. After filtering, only the blue signals are considered, making it possible to clearly observe the newly generated signal peaks after the addition of aspartic acid and glutamic acid. By using this background noise processing method, amino acid signals can be identified and counted more accurately.

**[0082]** To demonstrate the processing procedure more effectively, the experimenters of the present invention use Clampfit software to extract individual amino acid signals. Initially, using the fitted SD values to filter out background signals and double-step signals, and retaining individual amino acid signals. As shown in Fig. 6A, in the Clampfit single-channel statistics results (with only level 0 and level 1 set), the "Amp S.D." value represents the fitting effect of the signal. The SD values corresponding to the specific signals are shown in Fig. 6B. Among the six signals shown, it can be observed that the SD values for double-step signals are generally greater than 4. A single-step signal refers to a current distribution curve of the electrical signal with a single peak, and the current does not have a distribution between 0 and 60% of the peak value. In contrast, a double-step signal refers to a current distribution curve of the electrical signal with two peaks.

**[0083]** Setting the SD threshold within the range of 0 - 2 may filter out most multiple-step signals. However, background noise interference still exists near -5 pA and -12 pA. The signals near -5 pA (as shown in Fig. 6D) have excessive deviation from the baseline at level 0, and can be filtered out later. The signals near -12 pA (Fig. 6E) are normal background signals with short blockage times. Background signals only interfere with the signals of nearby glycine (Gly), serine (Ser), and alanine (Ala), and these three amino acid signals can be distinguished.

**[0084]** Based on this, a standard signal database can be established using known detection data, and the specific process is shown in Fig. 11. First, detecting a standard free amino acid using an M2MspA-N91H nanopore to obtain electrical signal data. Second, setting a filtering condition (for example, setting the SD threshold of the electrical signal to 2). Third, excluding part of the electrical signal to obtain a retained electrical signal based on the filtering condition. Fourth, determining if the retained electrical signal is a single-step signal. If the retained electrical signal is the single-step signal, the single-step signal is included in a standard signal database, and the single-step electrical signal is considered as a standard signals for the corresponding amino acid. If the retained electrical signal is not the single-step signal (such as a double-step signal), the filtering condition is reset (for example, raising the SD threshold), and the retained electrical signal is obtained again to determine if it is the single-step signal. Finally, establishing a standard signal database using the single-step electrical signal.

**[0085] Embodiment 3: Detection of 3 non-natural amino acids using M2MspA-N91H nanopore**

**[0086]** Based on the above method, this embodiment further detected 3 representative non-natural amino acids, namely, citrulline, L-norvaline, and ibotenic acid. As shown in Fig. 8A, the blockage rates caused by these 3 non-natural amino acids are: citrulline:  $0.1912 \pm 0.0011$ ; L-norvaline:  $0.1806 \pm 0.0017$ ; ibotenic acid:  $0.2333 \pm 0.0029$ . According to the blockage rates of amino acids and their standard deviations, the signals of the 3 non-natural amino acids do not overlap and can be distinguished from each other.

**[0087]** As shown in Fig. 8B, the upper and lower plots represent the scatter plots of current blockade signals within 1 minute for the blank group and the group with added ibotenic acid from the same experiment. Signals with an SD value less than 2 are considered valid signals. After filtering, it can be clearly observed that new signal peaks are generated after adding ibotenic acid.

**[0088]** Peptides can serve as substrates in enzyme-catalyzed reactions. Using the above method for amino acid detection, continuous monitoring of enzyme-catalyzed reactions can be achieved: by detecting the types of free amino acids in the reaction system, it is possible to determine whether enzyme-catalyzed reactions have occurred and the progress of enzyme-catalyzed cascade reactions.

**[0089] Embodiment 4: Polypeptide composition analysis method based on M2MspA-N91H Nanopore:**

**[0090]** After the embedding of the M2MspA-N91H nanopore protein, immediately adjust the voltage to +50 mV and record for 10 minutes. Then, add 1  $\mu\text{L}$  of 20 mM copper chloride solution to the trans side of the sample chamber and record for 10 minutes. At the same time, take 8  $\mu\text{L}$  of a 2 mM polypeptide solution and place it in a PCR tube, and incubate at 37°C for 5 minutes. Add 2  $\mu\text{L}$  of carboxypeptidase A1 solution, mix uniformly, and react at 37°C for 5 minutes. After the reaction is complete, add all the reaction products to the cis side of the sample chamber, mix uniformly, and record for 2 hours under +50 mV voltage. The detection principle is as shown in Fig. 10, carboxypeptidase A1 (a type of exopeptidase) can sequentially hydrolyze amino acids from the C-terminus of the polypeptide, the hydrolyzed free amino acid products are captured by the nanopore and detected, and the polypeptide composition can be determined based on the types and quantities of amino acid signals observed.

**[0091]** Since 1 U exopeptidase can cleave one terminal amino acid of 1  $\mu\text{mol}$  of polypeptide within 1 minute, the final concentration of the polypeptide solution can be calculated using the following formula (assuming that the polypeptide can be completely hydrolyzed):

**[0092]** 
$$b = \frac{c \times d}{a - 1}$$

**[0093]** Where, a represents the polypeptide length, b represents the amount of substance ( $\mu\text{mol}$ ), d represents the exopeptidase activity (U), and c represents the reaction time (min).

**[0094]** In the nanopore system, when the concentration of free amino acids is 1  $\mu\text{M}$ , the detection effect is better; if the final concentration of free amino acids is lower than 1  $\mu\text{M}$ , it may reduce the signal of free amino acids and thus affect the detection effect. Considering that the hydrolysis products (reaction products) may be diluted at least 10 times after being added to the M2MspA-N91H nanopore system, based on the above formula, the final concentration of the polypeptide solution in the reaction system should be greater than 10 times of 1  $\mu\text{M}$  (i.e., 10  $\mu\text{M}$ ).

**[0095]** By hydrolyzing the terminal amino acids of polypeptides with exopeptidase, a hydrolysis product (reaction product) is obtained. After adding the above reaction product to the M2MspA-N91H nanopore system and applying a driving force, the free amino acid in the reaction product passes through the M2MspA-N91H nanopore and generates an electrical signal, and the corresponding electrical signal data are obtained. The specific data analysis process is shown in Fig. 12. First, obtaining electrical signal data by detecting the above reaction product with M2MspA-N91H nanopore. Second, establishing a standard signal database (as described above). Third, set a filtering condition (SD threshold value for an electrical signal). Fourth, based on the filtering condition, excluding part of the electrical signal to obtain a retained electrical signal. Fifth, based on blockage rate of a standard signal in the standard signal database, determining amino acid type corresponding to the retained electrical signal. Sixth, obtaining information on the amino acid type of the free amino acid.

**[0096]** Fig. 7 shows the analysis of the composition of four different polypeptides using the above polypeptide composition analysis method and data analysis method. Fig. 7A shows the current signal plot obtained by adding the reaction solution into the nanopore system after hydrolyzing 10 mM EF polypeptide for 20 minutes under the conditions of 0.5 M KCl, 10 mM MOPS, pH 7.5. Fig. 7B shows the scatter plot of the blockage rate and translocation time of all current signals counted from the experiment involving Fig. 7A. It can be observed that the signals with SD values between 1-2 (red) have three distribution peaks, among which the signal near the blockage rate of 0.4 corresponds to the EF polypeptide; the

two peaks on the left are closest to the blockage rates of glutamic acid (E) and phenylalanine (F), thus determined to be signals of amino acids E and F.

**[0097]** Fig. 7C shows the experimental results of detecting the hydrolysis products of three polypeptides (P1: APRLRFYSL; P2: DRVYIHPFHL; P3: RPKVKVYPNGAEDESAEAFPLEF; wherein the underline represents that these amino acids at the C-terminus of the polypeptides can be hydrolyzed by carboxypeptidase A1 into free amino acids) using the M2MspA-N91H nanopore. The above analysis results indicate that the polypeptide composition analysis method provided by the present invention can analyze the composition of the terminal amino acids of polypeptides.

**[0098]** The above polypeptide composition analysis method provided by the present invention can be applied to various scenarios, such as the identification of known polypeptides and the prediction of unknown polypeptides:

**[0099]** Tumor neoantigens, as antigens expressed only in tumor tissues but not in normal tissues, are ideal targets for immunotherapy. Each patient has a unique spectrum of new antigens. Tumor neoantigen peptide can be screened using the above polypeptide composition analysis method. First, extracting and purifying a known tumor cell surface antigen peptide using a highly specific extraction method. The purified mixture theoretically contains only a small amount of and known antigen peptides. Second, using the above polypeptide composition analysis method to detect the purified mixture. Finally, comparing a detected amino acid type of detection result with amino acid type of the known antigen peptide to determine whether certain amino acids are present or absent, thereby determining whether tumor neoantigens may exist on the tumor cell surface antigen.

**[0100]** Currently, some peptide drugs have been developed and applied clinically, but due to limitations in synthesis processes, the product quality and purity of peptide drugs cannot meet the requirements. Through the polypeptide composition analysis method described above, it is possible to detect the purity of peptide drugs: comparing amino acid composition detection result of the peptide

drug to be tested with expected synthesized amino acid types can to some extent determine the synthesis quality of the peptide drug.

**[0101]** In addition, the peptide composition analysis method provided by the present invention can also be combined with machine learning to predict unknown peptides: hydrolyzing an unknown polypeptide to be tested by an exopeptidase and detecting a free amino acid; comparing it to a known polypeptide in database using machine learning methods, combined with types of exopeptidases, hydrolysis characteristics, and amino acid composition detection results, can to some extent predict the sequence of the unknown polypeptide and narrow down the range of the unknown polypeptide.

**[0102]** The embodiments of the present invention are described above with reference to the accompanying drawings, but the present invention is not limited to the aforementioned specific embodiments. The aforementioned embodiments are merely illustrative and not limiting. For those of ordinary skill in the art, many forms can be made under the teaching of present invention without departing from the spirit of the present invention and the scope of the claims, all of which shall fall within the protection scope of the present invention.

## CLAIMS

Claim 1. A polypeptide composition analysis method based on M2MspA-N91H nanopore, comprising the following steps:

S1 Mixing an exopeptidase with a polypeptide to be tested uniformly and reacting to obtain a hydrolysis product;

S2 Adding the hydrolysis product to a nanopore system, wherein the nanopore system comprises an M2MspA-N91H nanopore, an insulating membrane, a first medium, and a second medium, wherein the M2MspA-N91H nanopore is embedded in the insulating membrane, the M2MspA-N91H nanopore provides a channel that connects the first medium and the second medium, and the hydrolysis product is added to the first medium;

S3 Applying a driving force between the first medium and the second medium, a free amino acid in the hydrolysis product passes through the M2MspA-N91H nanopore and generates an electrical signal;

S4 Data-analyzing the electrical signal to obtain information on amino acid type of the free amino acid.

Claim 2. The method according to claim 1, wherein the free amino acid includes a natural amino acid and a non-natural amino acid.

Claim 3. The method according to claim 1, wherein the natural amino acid includes one or more of glycine, alanine, valine, leucine, isoleucine, phenylalanine, tryptophan, tyrosine, aspartic acid, histidine, asparagine, glutamic acid, lysine, glutamine, methionine, arginine, serine, threonine, cysteine, and proline; optionally, the non-natural amino acid includes one or more of citrulline, L-norvaline, and ibotenic acid.

Claim 4. The method according to claim 1, wherein the exopeptidase includes aminopeptidase and/or carboxypeptidase.

Claim 5. The method according to claim 1, wherein before adding the hydrolysis product to the nanopore system, the M2MspA-N91H nanopore is combined with a copper ion.

Claim 6. The method according to claim 1, wherein pH value of the first medium and the second medium includes 6.5 - 7.5.

Claim 7. The method according to claim 1, wherein final concentration of solution of the polypeptide to be tested is greater than 10  $\mu$ M.

Claim 8. The method according to claim 1, wherein step of the data-analyzing in the S4 further includes:

- i) Establishing a standard signal database;
- ii) Setting a signal filtering condition;
- iii) Excluding part of the electrical signal to obtain a retained electrical signal;
- iv) Determining the amino acid type corresponding to the retained electrical signal based on blockage rate of a standard signal in the standard signal database;
- v) Obtaining information on the amino acid type of the free amino acid.

Claim 9. The method according to claim 8, wherein the signal filtering condition is to exclude an electrical signal with an SD value greater than an SD threshold, wherein the SD value of the electrical signal is calculated according to the following formula:

$$SD = \sqrt{\frac{\sum(x_i - x)^2}{n-1}}$$

Wherein,  $x_i$  is current value of each sampling point of the signal,  $x$  is mean current value of the signal, and  $n$  is number of sampling point of the signal.

Claim 10. The method according to claim 1, wherein further comprising S5 Comparing the information on the amino acid type of the free amino acid with expected sequence information of the polypeptide to be tested, to determine whether there is a difference between actual sequence and the expected sequence of the polypeptide to be tested; or  
S5 Comparing the information on the amino acid type of the free amino acid combined with information on the exopeptidase with polypeptide sequence in database, to predict actual sequence of the polypeptide to be tested.