(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 November 2021 (25.11.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/236983 A3

(51) International Patent Classification:
G01N 33/58 (2006.01) *G01N 33/68* (2006.01)
C07K 14/00 (2006.01)

(21) International Application Number:
PCT/US2021/033493

(22) International Filing Date:
20 May 2021 (20.05.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/027,913 20 May 2020 (20.05.2020) US
63/059,919 31 July 2020 (31.07.2020) US

(71) Applicant: **QUANTUM-SI INCORPORATED** [US/US]; 530 Old Whitfield Street, Guilford, CT 06437 (US).

(72) Inventors: **REED, Brian**; 94 Hull Road, Madison, CT 06443 (US). **PANDEY, Manjula**; 460 White Birch Drive, Guilford, CT 06437 (US).

(74) Agent: **PRITZKER, Randy, J.** et al.; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210-2206 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) **Title:** METHODS AND COMPOSITIONS FOR PROTEIN SEQUENCING

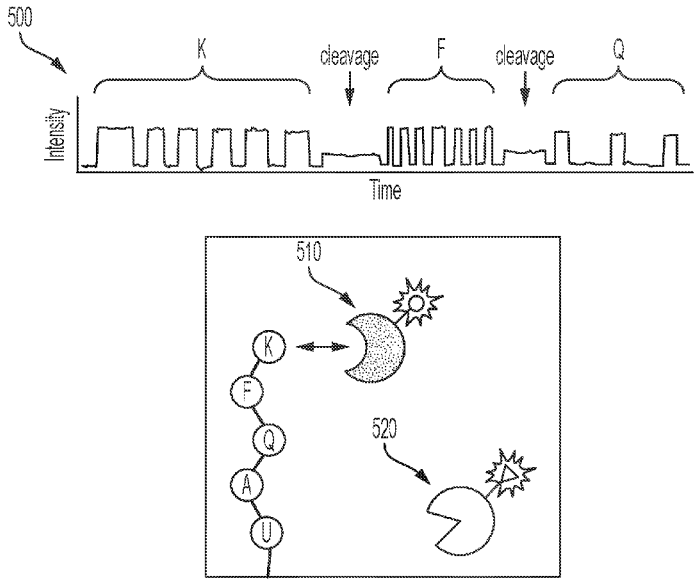


FIG. 5

(57) **Abstract:** Aspects of the application provide methods of identifying and sequencing proteins, polypeptides, and amino acids, and compositions useful for the same. In some aspects, the application provides amino acid recognition molecules, such as amino acid binding proteins and fusion polypeptides thereof. In some aspects, the application provides amino acid recognition molecules comprising a shielding element that enhances photo stability in polypeptide sequencing reactions.

WO 2021/236983 A3

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *with sequence listing part of description (Rule 5.2(a))*

(88) Date of publication of the international search report:
30 December 2021 (30.12.2021)

METHODS AND COMPOSITIONS FOR PROTEIN SEQUENCING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 63/059,919, filed July 31, 2020, and U.S. Provisional Patent Application No. 63/027,913, filed May 20, 2020, each of which is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] Proteomics has emerged as an important and necessary complement to genomics and transcriptomics in the study of biological systems. The proteomic analysis of an individual organism can provide insights into cellular processes and response patterns, which lead to improved diagnostic and therapeutic strategies. The complexity surrounding protein structure, composition, and modification present challenges in determining large-scale protein sequencing information for a biological sample.

SUMMARY

[0003] In some aspects, the application provides methods and compositions for determining amino acid sequence information from polypeptides (e.g., for sequencing one or more polypeptides). In some embodiments, amino acid sequence information can be determined for single polypeptide molecules. In some embodiments, the relative position of two or more amino acids in a polypeptide is determined, for example for a single polypeptide molecule. In some embodiments, one or more amino acids of a polypeptide are labeled (e.g., directly or indirectly) and the relative positions of the labeled amino acids in the polypeptide is determined. In some embodiments, amino acid sequence information can be determined by detecting an interaction of a polypeptide with one or more amino acid recognition molecules (e.g., one or more amino acid binding proteins).

[0004] In some aspects, the application provides an amino acid binding protein which can be used in a method for determining amino acid sequence information from polypeptides. In some aspects, the application provides a recombinant amino acid binding protein having an amino acid sequence that is at least 80% identical to a sequence selected from Table 1 or Table 2 and comprising one or more labels. In some embodiments, the one or more labels comprise a luminescent label or a conductivity label. In some embodiments, the one or more labels comprise a tag sequence. In some embodiments, the tag sequence comprises one or more of a

purification tag, a cleavage site, and a biotinylation sequence (e.g., at least one biotin ligase recognition sequence). In some embodiments, the biotinylation sequence comprises two biotin ligase recognition sequences oriented in tandem. In some embodiments, the one or more labels comprise a biotin moiety having at least one biotin molecule (e.g., a bis-biotin moiety). In some embodiments, the label comprises at least one biotin ligase recognition sequence having the at least one biotin molecule attached thereto. In some embodiments, the one or more labels comprise one or more polyol moieties (e.g., polyethylene glycol). In some embodiments, the recombinant amino acid binding protein comprises one or more unnatural amino acids having the one or more labels attached thereto. In some aspects, the application provides a composition comprising a recombinant amino acid binding protein described herein.

[0005] In some aspects, the application provides a polypeptide sequencing reaction composition comprising two or more amino acid recognition molecules, where at least one of the two or more amino acid recognition molecules is a recombinant amino acid binding protein described herein. In some embodiments, the two or more amino acid recognition molecules comprise different types of amino acid recognition molecules. For example, in some embodiments, an amino acid recognition molecule of one type interacts with a polypeptide of interest in a manner that is different (e.g., detectably different) from other types of amino acid recognition molecules in a polypeptide sequencing reaction composition. In some embodiments, the polypeptide sequencing reaction composition comprises at least one type of cleaving reagent. In some aspects, the application provides a method of polypeptide sequencing comprising contacting a polypeptide with a polypeptide sequencing reaction composition described herein. In some embodiments, the method further comprises detecting a series of interactions of the polypeptide with at least one amino acid recognition molecule while the polypeptide is being degraded, thereby sequencing the polypeptide.

[0006] In some aspects, the application provides a polypeptide sequencing reaction mixture comprising an amino acid binding protein and a peptidase. In some embodiments, the molar ratio of the labeled amino acid binding protein to the peptidase is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1. In some embodiments, the amino acid binding protein comprises one or more labels. In some embodiments, the amino acid binding protein is a ClpS protein. In some embodiments, the amino acid binding protein is a protein having an amino acid sequence that is at least 80%, 80-90%, 90-95%, or at least 95% identical to a sequence selected from Table 1 or Table 2. In some embodiments, the peptidase is an exopeptidase. In some embodiments, the peptidase is an enzyme having an amino acid sequence that is at least 80%, 80-90%, 90-95%, or at least 95% identical to a sequence selected from Table 4 or Table 5. In some embodiments, the reaction mixture comprises more than one amino acid

binding protein and/or more than one peptidase. In some embodiments, the reaction mixture comprises a polypeptide molecule immobilized to a surface.

[0007] In some aspects, the application provides a polypeptide sequencing reaction mixture comprising a single polypeptide molecule, at least one peptidase molecule, and at least three amino acid recognition molecules. In some embodiments, the reaction mixture comprises at least 1 and up to 10 peptidase molecules (e.g., at least 1 and up to 5 peptidase molecules, at least 1 and up to 3 peptidase molecules). In some embodiments, the reaction mixture comprises two or more peptidase molecules, where each peptidase molecule is of a different type. For example, in some embodiments, a peptidase molecule of one type has a cleavage preference that is different from other types of peptidase molecules in a reaction mixture. In some embodiments, the reaction mixture comprises at least 3 and up to 30 amino acid recognition molecules (e.g., up to 20, up to 10, or up to 5 amino acid recognition molecules). In some embodiments, the at least three amino acid recognition molecules comprise different types of amino acid recognition molecules. For example, in some embodiments, an amino acid recognition molecule of one type interacts with a polypeptide of interest in a manner that is different (e.g., detectably different) from other types of amino acid recognition molecules in a reaction mixture.

[0008] In some aspects, the application provides a substrate comprising an array of sample wells, wherein at least one sample well of the array comprises a polypeptide sequencing reaction mixture described herein. In some embodiments, the at least one sample well comprises a bottom surface. In some embodiments, the single polypeptide molecule is immobilized to the bottom surface.

[0009] In some aspects, the application provides an amino acid recognition molecule comprising a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end, wherein the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids. In some embodiments, the first and second amino acid binding proteins are the same. In some embodiments, the first and second amino acid binding proteins are different.

[0010] In some aspects, the application provides an amino acid recognition molecule comprising a polypeptide of Formula (**I**):

$$(Z^1–X^1)_n–Z^2$$

(**I**),

wherein: Z^1 and Z^2 are independently amino acid binding proteins; X^1 is a linker comprising at least two amino acids, where the amino acid binding proteins are joined end-to-end by the linker; and n is an integer from 1 to 5, inclusive. In some embodiments, Z^1 and Z^2 comprise amino acid binding proteins of the same type. In some embodiments, Z^1 and Z^2 comprise different types of

amino acid binding proteins. In some embodiments, Z^1 and Z^2 are independently optionally associated with a label component comprising at least one detectable label. In some embodiments, the polypeptide further comprises a tag sequence.

[0011] In some aspects, the application provides methods of polypeptide sequencing. In some embodiments, a method of polypeptide sequencing comprises contacting a single polypeptide molecule in a reaction mixture with a composition comprising a binding means and a cleaving means. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 association events between the binding means and a terminal amino acid on the polypeptide prior to removal of the terminal amino acid from the polypeptide by the cleaving means. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 and up to 1,000 association events prior to the removal of the terminal amino acid. In some embodiments, the terminal amino acid was exposed at the polypeptide terminus in a cleavage event prior to the at least 10 association events. In some embodiments, the at least 10 association events occur after the cleavage event.

[0012] In some embodiments, the binding means and the cleaving means are configured to achieve a time interval of at least 1 minute between cleavage events (e.g., between about 1 minute and about 20 minutes, between about 5 minutes and about 15 minutes, or between about 1 minute and about 10 minutes). In some embodiments, the binding means comprise one or more amino acid recognition molecules, and the cleaving means comprise one or more peptidase molecules. In some embodiments, the molar ratio of an amino acid recognition molecule to a peptidase molecule is configured to achieve the at least 10 association events prior to the removal of the terminal amino acid. In some embodiments, the molar ratio of the amino acid recognition molecule to the peptidase molecule is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1. In some embodiments, the molar ratio of the amino acid recognition molecule to the peptidase molecule is between about 1:100 and about 1:1 or between about 1:1 and about 10:1.

[0013] In some aspects, the application provides a substrate comprising an array of sample wells, where at least one sample well of the array comprises a single polypeptide molecule, a cleaving means, and a binding means. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 association events between the binding means and a terminal amino acid on the polypeptide prior to removal of the terminal amino acid from the polypeptide by the cleaving means. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 and up to 1,000 association events prior to the removal of the terminal amino acid. In some embodiments, the terminal amino acid was exposed

at the polypeptide terminus in a cleavage event prior to the at least 10 association events. In some embodiments, the at least 10 association events occur after the cleavage event.

[0014] In some aspects, the application provides amino acid recognition molecules comprising a shielding element, e.g., for enhanced photostability in polypeptide sequencing reactions. In some aspects, the application provides an amino acid recognition molecule comprising a polypeptide having an amino acid binding protein and a labeled protein joined end-to-end. In some embodiments, the amino acid binding protein and the labeled protein are separated by a linker comprising at least two amino acids (e.g., at least two and up to 100 amino acids, between about 5 and about 50 amino acids). In some embodiments, the labeled protein has a molecular weight of at least 10 kDa (e.g., between about 10 kDa and about 150 kDa, between about 15 kDa and about 100 kDa). In some embodiments, the labeled protein comprises at least 50 amino acids (e.g., between about 50 and about 1,000 amino acids, between about 100 and about 750 amino acids). In some embodiments, the labeled protein comprises a luminescent label. In some embodiments, the luminescent label comprises at least one fluorophore dye molecule. In some embodiments, the amino acid binding protein is a Gid protein, a UBR-box protein or UBR-box domain-containing fragment thereof, a p62 protein or ZZ domain-containing fragment thereof, or a ClpS protein. In some embodiments, the amino acid binding protein has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

[0015] In some aspects, the application provides an amino acid recognition molecule of Formula (**II**):

$$A\text{–}(Y)_n\text{–}D$$

(**II**),

wherein: A is an amino acid binding component comprising at least one amino acid recognition molecule; each instance of Y is a polymer that forms a covalent or non-covalent linkage group; n is an integer from 1 to 10, inclusive; and D is a label component comprising at least one detectable label. In some embodiments, A comprises at least one amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2. In some embodiments, the amino acid recognition molecule comprises a polypeptide having A and Y^1 joined end-to-end, wherein A and Y^1 are separated by a linker comprising at least two amino acids. In some embodiments, Y^1 is a protein having a molecular weight of at least 10 kDa (e.g., between about 10 kDa and about 150 kDa). In some embodiments, Y^1 is a protein comprising at least 50 amino acids (e.g., between about 50 and about 1,000 amino acids).

[0016] In some embodiments, D is less than 200 Å in diameter. In some embodiments, $-(Y)_n-$ is at least 2 nm in length (e.g., at least 5 nm, at least 10 nm, at least 20 nm, at least 30 nm, at least 50 nm, or more, in length). In some embodiments, $-(Y)_n-$ is between about 2 nm and about 200 nm in length (e.g., between about 2 nm and about 100 nm, between about 5 nm and about 50 nm, or between about 10 nm and about 100 nm in length). In some embodiments, each instance of Y is independently a biomolecule or a dendritic polymer (e.g., a polyol, a dendrimer). In some embodiments, A comprises a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end (e.g., a fusion polypeptide). In some embodiments, the application provides a composition comprising the amino acid recognition molecule of Formula (**II**). In some embodiments, the amino acid recognition molecule is soluble in the composition.

[0017] In some aspects, the application provides an amino acid recognition molecule of Formula (**III**):

$$A–Y^1–D$$

(**III**),

wherein: A is an amino acid binding component comprising at least one amino acid recognition molecule; Y^1 is a nucleic acid or a polypeptide; D is a label component comprising at least one detectable label. In some embodiments, A comprises at least one amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2. In some embodiments, when Y^1 is a nucleic acid, the nucleic acid forms a covalent or non-covalent linkage group. In some embodiments, provided that when Y^1 is a polypeptide, the polypeptide forms a non-covalent linkage group characterized by a dissociation constant (K_D) of less than 50×10^{-9} M. In some embodiments, the K_D is less than 1×10^{-9} M, less than 1×10^{-10} M, less than 1×10^{-11} M, or less than 1×10^{-12} M.

[0018] In some aspects, the application provides an amino acid recognition molecule comprising: a nucleic acid; at least one amino acid recognition molecule attached to a first attachment site on the nucleic acid; and at least one detectable label attached to a second attachment site on the nucleic acid, where the nucleic acid forms a covalent or non-covalent linkage group between the at least one amino acid recognition molecule and the at least one detectable label. In some embodiments, the nucleic acid comprises a first oligonucleotide strand. In some embodiments, the nucleic acid further comprises a second oligonucleotide strand hybridized with the first oligonucleotide strand. In some embodiments, the at least one amino acid recognition molecule comprises a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end (e.g., a fusion polypeptide).

In some embodiments, the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids.

[0019] In some aspects, the application provides an amino acid recognition molecule comprising: a multivalent protein comprising at least two ligand-binding sites; at least one amino acid recognition molecule attached to the protein through a first ligand moiety bound to a first ligand-binding site on the protein; and at least one detectable label attached to the protein through a second ligand moiety bound to a second ligand-binding site on the protein. In some embodiments, the multivalent protein is an avidin protein. In some embodiments, the at least one amino acid recognition molecule comprises a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end (e.g., a fusion polypeptide). In some embodiments, the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids.

[0020] In some embodiments, a shielded amino acid recognition molecule may be used in polypeptide sequencing methods in accordance with the application, or any method known in the art. Accordingly, in some aspects, the application provides methods of polypeptide sequencing (e.g., in an Edman-type degradation reaction, in a dynamic sequencing reaction, or other method known in the art) comprising contacting a polypeptide molecule with one or more shielded amino acid recognition molecules of the application. For example, in some embodiments, the methods comprise contacting a polypeptide molecule with at least one amino acid recognition molecule that comprises a shield or shielding element in accordance with the application, and detecting association of the at least one amino acid recognition molecule with the polypeptide molecule.

[0021] In some aspects, the application provides methods comprising obtaining data during a degradation process of a polypeptide. In some embodiments, the methods further comprise analyzing the data to determine portions of the data corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process. In some embodiments, the methods further comprise outputting an amino acid sequence representative of the polypeptide. In some embodiments, the data is indicative of amino acid identity at the terminus of the polypeptide during the degradation process. In some embodiments, the data is indicative of a signal produced by one or more amino acid recognition molecules binding to different types of terminal amino acids at the terminus during the degradation process. In some embodiments, the data is indicative of a luminescent signal generated during the degradation process. In some embodiments, the data is indicative of an electrical signal generated during the degradation process.

[0022] In some embodiments, analyzing the data further comprises detecting a series of cleavage events and determining the portions of the data between successive cleavage events. In some embodiments, analyzing the data further comprises determining a type of amino acid for each of the individual portions. In some embodiments, each of the individual portions comprises a pulse pattern (e.g., a characteristic pattern), and analyzing the data further comprises determining a type of amino acid for one or more of the portions based on its respective pulse pattern. In some embodiments, determining the type of amino acid further comprises identifying an amount of time within a portion when the data is above a threshold value and comparing the amount of time to a duration of time for the portion. In some embodiments, determining the type of amino acid further comprises identifying at least one pulse duration for each of the one or more portions. In some embodiments, the pulse pattern comprises a mean pulse duration of between about 1 millisecond and about 10 seconds. In some embodiments, determining the type of amino acid further comprises identifying at least one interpulse duration for each of the one or more portions. In some embodiments, the amino acid sequence includes a series of amino acids corresponding to the portions.

[0023] In some aspects, the application provides methods of polypeptide sequencing comprising contacting a single polypeptide molecule with one or more amino acid recognition molecules (e.g., one or more terminal amino acid recognition molecules). In some embodiments, the methods further comprise detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with successive amino acids exposed at a terminus of the single polypeptide molecule while it is being degraded, thereby obtaining sequence information about the single polypeptide molecule. In some embodiments, the amino acid sequence of most or all of the single polypeptide molecule is determined. In some embodiments, the series of signal pulses is a series of real-time signal pulses.

[0024] In some embodiments, association of the one or more amino acid recognition molecules with each type of amino acid exposed at the terminus produces a characteristic pattern in the series of signal pulses that is different from other types of amino acids exposed at the terminus. In some embodiments, signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds. In some embodiments, a signal pulse of the characteristic pattern corresponds to an individual association event between an amino acid recognition molecule and an amino acid exposed at the terminus. In some embodiments, the characteristic pattern corresponds to a series of reversible amino acid recognition molecule binding interactions with the amino acid exposed at the terminus of the single polypeptide molecule. In some embodiments, the characteristic pattern is indicative of the amino acid

exposed at the terminus of the single polypeptide molecule and an amino acid at a contiguous position (e.g., amino acids of the same type or different types).

[0025] In some embodiments, the single polypeptide molecule is degraded by a cleaving reagent that removes one or more amino acids from the terminus of the single polypeptide molecule. In some embodiments, the methods further comprise detecting a signal indicative of association of the cleaving reagent with the terminus. In some embodiments, the cleaving reagent comprises a detectable label (e.g., a luminescent label, a conductivity label). In some embodiments, the single polypeptide molecule is immobilized to a surface. In some embodiments, the single polypeptide molecule is immobilized to the surface through a terminal end distal to the terminus to which the one or more amino acid recognition molecules associate. In some embodiments, the single polypeptide molecule is immobilized to the surface through a linker (e.g., a solubilizing linker comprising a biomolecule).

[0026] In some aspects, the application provides methods of sequencing a polypeptide comprising contacting a single polypeptide molecule in a reaction mixture with a composition comprising one or more amino acid recognition molecules (e.g., one or more terminal amino acid recognition molecules) and a cleaving reagent. In some embodiments, the methods further comprise detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with a terminus of the single polypeptide molecule in the presence of the cleaving reagent. In some embodiments, the series of signal pulses is indicative of a series of amino acids exposed at the terminus over time as a result of terminal amino acid cleavage by the cleaving reagent.

[0027] In some aspects, the application provides methods of sequencing a polypeptide comprising (a) identifying a first amino acid at a terminus of a single polypeptide molecule, (b) removing the first amino acid to expose a second amino acid at the terminus of the single polypeptide molecule, and (c) identifying the second amino acid at the terminus of the single polypeptide molecule. In some embodiments, (a)-(c) are performed in a single reaction mixture. In some embodiments, (a)-(c) occur sequentially. In some embodiments, (c) occurs before (a) and (b). In some embodiments, the single reaction mixture comprises one or more amino acid recognition molecules (e.g., one or more terminal amino acid recognition molecules). In some embodiments, the single reaction mixture comprises a cleaving reagent. In some embodiments, the first amino acid is removed by the cleaving reagent. In some embodiments, the methods further comprise repeating the steps of removing and identifying one or more amino acids at the terminus of the single polypeptide molecule, thereby determining a sequence (e.g., a partial sequence or a complete sequence) of the single polypeptide molecule.

[0028] In some aspects, the application provides methods of identifying an amino acid of a polypeptide comprising contacting a single polypeptide molecule with one or more amino acid recognition molecules that bind to the single polypeptide molecule. In some embodiments, the methods further comprise detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with the single polypeptide molecule under polypeptide degradation conditions. In some embodiments, the methods further comprise identifying a first type of amino acid in the single polypeptide molecule based on a first characteristic pattern in the series of signal pulses. In some embodiments, signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds.

[0029] In some aspects, the application provides methods of identifying a terminal amino acid (e.g., the N-terminal or the C-terminal amino acid) of a polypeptide. In some embodiments, the methods comprise contacting a polypeptide with one or more labeled recognition molecules that selectively bind one or more types of terminal amino acids at a terminus of the polypeptide. In some embodiments, the methods further comprise identifying a terminal amino acid at the terminus of the polypeptide by detecting an interaction of the polypeptide with the one or more labeled recognition molecules.

[0030] In yet other aspects, the application provides methods of polypeptide sequencing by Edman-type degradation reactions. In some embodiments, Edman-type degradation reactions may be performed by contacting a polypeptide with different reaction mixtures for purposes of either detection or cleavage (e.g., as compared to a dynamic sequencing reaction, which can involve detection and cleavage using a single reaction mixture).

[0031] Accordingly, in some aspects, the application provides methods of determining an amino acid sequence of a polypeptide comprising (i) contacting a polypeptide with one or more labeled recognition molecules that selectively bind one or more types of terminal amino acids at a terminus of the polypeptide. In some embodiments, the methods further comprise (ii) identifying a terminal amino acid (e.g., the N-terminal or the C-terminal amino acid) at the terminus of the polypeptide by detecting an interaction of the polypeptide with the one or more labeled recognition molecules. In some embodiments, the methods further comprise (iii) removing the terminal amino acid. In some embodiments, the methods further comprise (iv) repeating (i)-(iii) one or more times at the terminus of the polypeptide to determine an amino acid sequence of the polypeptide.

[0032] In some embodiments, the methods further comprise, after (i) and before (ii), removing any of the one or more labeled recognition molecules that do not selectively bind the terminal amino acid. In some embodiments, the methods further comprise, after (ii) and before (iii),

removing any of the one or more labeled recognition molecules that selectively bind the terminal amino acid.

[0033] In some embodiments, removing a terminal amino acid (e.g., (iii)) comprises modifying the terminal amino acid by contacting the terminal amino acid with an isothiocyanate (e.g., phenyl isothiocyanate), and contacting the modified terminal amino acid with a protease that specifically binds and removes the modified terminal amino acid. In some embodiments cleaving a terminal amino acid (e.g., (iii)) comprises modifying the terminal amino acid by contacting the terminal amino acid with an isothiocyanate, and subjecting the modified terminal amino acid to acidic or basic conditions sufficient to remove the modified terminal amino acid.

[0034] In some embodiments, identifying a terminal amino acid comprises identifying the terminal amino acid as being one type of the one or more types of terminal amino acids to which the one or more labeled recognition molecules bind. In some embodiments, identifying a terminal amino acid comprises identifying the terminal amino acid as being a type other than the one or more types of terminal amino acids to which the one or more labeled recognition molecules bind.

[0035] In some aspects, the application provides methods of identifying a protein of interest in a mixed sample. In some embodiments, the methods comprise cleaving a mixed protein sample to produce a plurality of polypeptide fragments. In some embodiments, the methods further comprise determining an amino acid sequence of at least one polypeptide fragment of the plurality in a method in accordance with the methods of the application. In some embodiments, the methods further comprise identifying a protein of interest in the mixed sample if the amino acid sequence is uniquely identifiable to the protein of interest.

[0036] In some embodiments, methods of identifying a protein of interest in a mixed sample comprise cleaving a mixed protein sample to produce a plurality of polypeptide fragments. In some embodiments, the methods further comprise labeling one or more types of amino acids in the plurality of polypeptide fragments with one or more different luminescent labels. In some embodiments, the methods further comprise measuring luminescence over time for at least one labeled polypeptide of the plurality. In some embodiments, the methods further comprise determining an amino acid sequence of the at least one labeled polypeptide based on the luminescence detected. In some embodiments, the methods further comprise identifying a protein of interest in the mixed sample if the amino acid sequence is uniquely identifiable to the protein of interest.

[0037] Accordingly, in some embodiments, a polypeptide molecule or protein of interest to be analyzed in accordance with the application can be of a mixed or purified sample. In some embodiments, the polypeptide molecule or protein of interest is obtained from a biological

sample (e.g., blood, tissue, saliva, urine, or other biological source). In some embodiments, the polypeptide molecule or protein of interest is obtained from a patient sample (e.g., a human sample).

[0038] In some aspects, the application provides systems comprising at least one hardware processor, and at least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by the at least one hardware processor, cause the at least one hardware processor to perform a method in accordance with the application. In some aspects, the application provides at least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by at least one hardware processor, cause the at least one hardware processor to perform a method in accordance with the application.

[0039] The details of certain embodiments of the invention are set forth in the Detailed Description of Certain Embodiments, as described below. Other features, objects, and advantages of the invention will be apparent from the Examples, Figures, and Claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0040] The accompanying drawings, which constitute a part of this specification, illustrate several embodiments of the invention and together with the description, serve to explain the principles of the invention.

[0041] FIGs. 1A-1B show an example of polypeptide sequencing by detection (FIG. 1A) and analysis (FIG. 1B) of single molecule binding interactions.

[0042] FIG. 2 depicts example configurations of labeled recognition molecules, including labeled enzymes and labeled aptamers which selectively bind one or more types of terminal amino acids.

[0043] FIGs. 3A-3E show non-limiting examples of amino acid recognition molecules labeled through a shielding element. FIG. 3A illustrates single-molecule peptide sequencing with a recognition molecule labeled through a conventional covalent linkage. FIG. 3B illustrates single-molecule peptide sequencing with a recognition molecule comprising a shielding element. FIGs. 3C-3E illustrate various examples of shielding elements in accordance with the application.

[0044] FIG. 4 generically depicts a degradation-based process of polypeptide sequencing using labeled recognition molecules.

[0045] FIGs. 5-7 show examples of polypeptide sequencing in real-time by evaluating binding interactions of terminal and/or internal amino acids with labeled recognition molecules and a labeled cleaving reagent. FIG. 5 shows an example of real-time sequencing by detecting a series

of pulses in a signal output. FIG. 6 schematically depicts a temperature-dependent sequencing process. FIG. 7 shows an example of polypeptide sequencing in real-time by evaluating binding interactions of terminal and internal amino acids with labeled recognition molecules and a labeled non-specific exopeptidase.

[0046] FIGs. 8-10 show various examples of preparing samples and sample well surfaces for analysis of polypeptides and proteins in accordance with the application. FIG. 8 generically depicts an example process of preparing terminally modified polypeptides from a protein sample. FIG. 9 generically depicts an example process of conjugating a solubilizing linker to a polypeptide. FIG. 10 shows an example schematic of a sample well having modified surfaces which may be used to promote single molecule immobilization to a bottom surface.

[0047] FIG. 11 is a diagram of an illustrative sequence data processing pipeline for analyzing data obtained during a polypeptide degradation process, in accordance with some embodiments of the technology described herein.

[0048] FIG. 12 is a flow chart of an illustrative process for determining an amino acid sequence of a polypeptide molecule, in accordance with some embodiments of the technology described herein.

[0049] FIG. 13 is a flow chart of an illustrative process for determining an amino acid sequence representative of a polypeptide, in accordance with some embodiments of the technology described herein.

[0050] FIG. 14 is a block diagram of an illustrative computer system that may be used in implementing some embodiments of the technology described herein.

[0051] FIGs. 15A-15C show experimental data for select peptide-linker conjugates prepared and evaluated for enhanced solubility provided by different solubilizing linkers. FIG. 15A shows example structures of peptide-linker conjugates that were synthesized and evaluated. FIG. 15B shows results from LCMS which demonstrate peptide cleavage at the N-terminus. FIG. 15C shows results from a loading experiment.

[0052] FIG. 16 shows a summary of amino acid cleavage activities for select exopeptidases based on experimental results.

[0053] FIGs. 17A-17C show experimental data for a dye/peptide conjugate assay for detecting and cleaving terminal amino acids. FIG. 17A shows example schemes and structures used for performing a dye/peptide conjugate assay. FIG. 17B shows imaging results for peptide-linker conjugate loading into sample wells in an on-chip assay. FIG. 17C shows example signal traces which detected peptide-conjugate loading and terminal amino acid cleavage.

[0054] FIGs. 18A-18F show experimental data for a FRET dye/peptide conjugate assay for detecting and cleaving terminal amino acids. FIG. 18A shows example schemes and structures

used for performing a FRET dye/peptide conjugate assay. FIG. 18B shows FRET imaging results for different time points. FIG. 18C shows cutting efficiency at the different time points. FIG. 18D shows cutting displayed at each of the different time points. FIG. 18E shows additional FRET imaging results for different time points with a proline iminopeptidase from *Yersinia pestis* (yPIP). FIG. 18F shows FRET imaging results for different time points with an aminopeptidase from *Vibrio proteolyticus* (VPr).

[0055] FIGs. 19A-19M show experimental data for terminal amino acid discrimination by a labeled recognition molecule. FIG. 19A shows a crystal structure of a ClpS2 protein that was labeled for these experiments. FIG. 19B shows single molecule intensity traces which illustrate N-terminal amino acid discrimination by the labeled ClpS2 protein. FIG. 19C is a plot showing mean pulse duration for different terminal amino acids. FIG. 19D is a plot showing mean interpulse duration for different terminal amino acids. FIG. 19E shows plots further illustrating discriminant pulse durations among the different terminal amino acids. FIGs. 19F, 19G, and 19H show example results from dwell time analysis demonstrating leucine recognition by a ClpS protein from *Thermosynochoccus elongatus* (teClpS). FIG. 19I shows example results from dwell time analysis demonstrating differentiable recognition of phenylalanine, leucine, tryptophan, and tyrosine by *A. tumefaciens* ClpS1. FIG. 19J shows example results from dwell time analysis demonstrating leucine recognition by *S. elongatus* ClpS2. FIGs. 19K-19L show example results from dwell time analysis demonstrating proline recognition by GID4. FIG. 19M shows exemplary binding curves for atClpS2-V1 with peptides having different N-terminal amino acids.

[0056] FIGs. 20A-20D show example results from polypeptide sequencing reactions conducted in real-time using a labeled ClpS2 recognition protein and an aminopeptidase cleaving reagent in the same reaction mixture. FIG. 20A shows signal trace data for a first sequencing reaction. FIG. 20B shows pulse duration statistics for the signal trace data shown in FIG. 20A. FIG. 20C shows signal trace data for a second sequencing reaction. FIG. 20D shows pulse duration statistics for the signal trace data shown in FIG. 20C.

[0057] FIGs. 21A-21F show experimental data for terminal amino acid identification and cleavage by a labeled exopeptidase. FIG. 21A shows a crystal structure of a proline iminopeptidase (yPIP) that was site-specifically labeled for these experiments. FIG. 21B shows the degree of labeling for the purified protein product. FIG. 21C is an image of SDS page confirming site-specific labeling of yPIP. FIG. 21D is an overexposed image of the SDS page gel confirming site-specific labeling. FIG. 21E is an image of a Coomassie stained gel confirming purity of labeled protein product. FIG. 21F is an HPLC trace demonstrating cleavage activity of the labeled exopeptidase.

[0058] FIGs. 22A-22F show data from experiments evaluating recognition of amino acids containing specific post-translational modifications. FIG. 22A shows representative traces which demonstrated phospho-tyrosine recognition by an SH2 domain-containing protein; FIG. 22B shows pulse duration data corresponding to the traces of FIG. 22A; and FIG. 22C shows statistics determined for the traces. FIGs. 22D-22F show representative traces from negative control experiments.

[0059] FIG. 23 is a plot showing median pulse duration from experiments evaluating the effects of penultimate amino acids on pulse duration.

[0060] FIGs. 24A-24C show data from experiments evaluating simultaneous amino acid recognition by differentially labeled recognition molecules. FIG. 24A shows a representative trace. FIG. 24B is a plot comparing pulse duration data obtained during these experiments for each recognition molecule. FIG. 24C shows pulse duration statistics for these experiments.

[0061] FIGs. 25A-25C show data from experiments evaluating the photostability of peptides during single-molecule recognition. FIG. 25A shows a representative trace from recognition using atClpS2-V1 labeled with a dye ~2 nm from the amino acid binding site. FIG. 25B shows a visualization of the structure of the ClpS2 protein used in these experiments. FIG. 25C shows a representative trace from recognition using ClpS2 labeled with a dye >10 nm from the amino acid binding site through a DNA/protein linker.

[0062] FIGs. 26A-26D show representative traces from polypeptide sequencing reactions conducted in real-time on a complementary metal–oxide–semiconductor (CMOS) chip using a ClpS2 recognition protein labeled through a DNA/streptavidin linker in the presence of an aminopeptidase cleaving reagent.

[0063] FIG. 27 shows representative traces from polypeptide sequencing reactions conducted in real-time using atClpS2-V1 recognition protein labeled through a DNA/streptavidin linker in the presence of *Pyrococcus horikoshii* TET aminopeptidase cleaving reagent.

[0064] FIGs. 28A-28J show representative trace data from polypeptide sequencing reactions conducted in real-time using multiple types of exopeptidases with differential cleavage specificities. FIG. 28A shows a representative trace from a reaction performed with hTET exopeptidase, with expanded pulse pattern regions shown in FIG. 28B. FIG. 28C shows a representative trace from a reaction performed with both hTET and yPIP exopeptidases, with expanded pulse pattern regions shown in FIG. 28D, and additional representative traces shown in FIG. 28E. FIG. 28F shows a representative trace from a further reaction performed with both hTET and yPIP exopeptidases, with expanded pulse pattern regions shown in FIG. 28G, and additional representative traces shown in FIG. 28H. FIG. 28I shows a representative trace from a

reaction performed with both PfuTET and yPIP exopeptidases, with expanded pulse pattern regions shown in FIG. 28J.

[0065] FIGs. 29A-29G show data from experiments evaluating a newly identified ClpS homolog. FIG. 29A shows SDS-PAGE gel imaging of purified ClpS proteins. FIG. 29B shows results from biolayer interferometry screening of ClpS homologs. FIG. 29C shows select results from the screening. FIG. 29D shows response curves for a ClpS protein (PS372) with LA, IA, and VA peptides. FIG. 29E shows polarization response for PS372 and four other homologs, along with no-protein control. FIG. 29F shows biolayer interferometry response curves for PS372 with IA, IR, IQ, VA, and VR peptides. FIG. 29G shows pulse width histograms and representative traces for PS372 with IR peptide (top panels) and LF peptide (bottom panels).

[0066] FIGs. 30A-30E show data from experiments evaluating terminal amino acid discrimination by a newly identified ClpS homolog. FIG. 30A shows *in vivo* biotinylation of PS372 by SDS-PAGE. FIG. 30B shows a purification profile for PS372 conjugated to SV-Dye. FIG. 30C shows SDS-PAGE after purification of SV-Dye conjugated PS372. FIG. 30D shows representative traces showing transition from I to L (a) and L to I (b). FIG. 30E shows example data from a real-time dynamic peptide sequencing assay with dye-labeled PS327.

[0067] FIGs. 31A-31F show data for the engineering of a methionine-binding ClpS protein. FIG. 31A shows results from selections performed via fluorescence-activated cell sorting (FACS). FIG. 31B shows an example response curve for methionine-binding ClpS proteins with a peptide having N-terminal LA. FIG. 31C shows an example response curve for methionine-binding ClpS proteins with a peptide having N-terminal MA. FIG. 31D shows an example response curve for methionine-binding ClpS proteins with a peptide having N-terminal MR. FIG. 31E shows an example response curve for methionine-binding ClpS proteins with a peptide having N-terminal FA. FIG. 31F shows an example response curve for methionine-binding ClpS proteins with a peptide having N-terminal MQ.

[0068] FIGs. 32A-32I show data from experiments evaluating UBR-box domain homologs. FIGs. 32A-32B show example binding curves for UBR-box homologs PS535 (FIG. 32A) and PS522 (FIG. 32B) binding with 14 polypeptides containing N-terminal R followed by different amino acids in the penultimate position. FIG. 32C is a heatmap showing results measured for 24 UBR-box homologs binding N-terminal R peptides. FIG. 32D is a heatmap showing results measured for an expanded set of UBR-box homologs binding with polypeptides containing R, K, or H at the N-terminal position. FIG. 32E is a heatmap showing results measured for an expanded set of UBR-box homologs binding with 14 polypeptides containing N-terminal R followed by different amino acids in the penultimate position. FIG. 32F shows results from single point fluorescence polarization assays. FIG. 32G shows analysis of polarization results

for binding affinity determination. FIG. 32H shows a representative trace for PS621 in a recognition assay. FIG. 32I shows example sequencing traces from a 3-binder dynamic sequencing reaction.

[0069] FIGs. 33A-33F show data from experiments evaluating PS372-homologous proteins. FIGs. 33A-33E show example binding curves for PS372 (FIG. 33A) and homologs PS545 (FIG. 33B), PS551 (FIG. 33C), PS557 (FIG. 33D), and PS558 (FIG. 33E) binding with 4 polypeptides containing different N-terminal amino acids (I, V, L, F). FIG. 33F is a heatmap showing results measured for 34 PS372 homologs.

[0070] FIGs. 34A-34D show data from experiments evaluating an engineered multivalent amino acid binder (PS610) produced as a single polypeptide having tandem copies of atClpS2-V1. FIG. 34A shows representative trace data for peptide-on-chip recognition assays. FIG. 34B is a plot showing mean pulse rate as a function of binder concentration. FIG. 34C shows example data from a real-time dynamic peptide sequencing assay with dye-labeled PS610 and PS327. FIG. 34D shows representative trace data for binder-on-chip assays.

[0071] FIGs. 35A-35D show data from experiments evaluating tandem ClpS2-V1 constructs containing different linkers. FIG. 35A shows example binding curves for the monovalent binder atClpS2-V1. FIGs. 35B-35D show example binding curves for tandem constructs having two copies of atClpS2-V1 separated by Linker 1 (FIG. 35B), Linker 2 (FIG. 35C), or Linker 3 (FIG. 35D).

[0072] FIGs. 36A-36H show data from experiments evaluating engineered multivalent amino acid binders produced as a single polypeptide having tandem copies of the same or different ClpS proteins. FIG. 36A shows example binding curves for the monovalent binder atClpS2-V1 (left plot) and the monovalent binder PS372 (right plot). FIG. 36B shows example binding curves for a multivalent polypeptide having tandem copies of atClpS2-V1 and PS372. FIG. 36C shows example binding curves for the monovalent binder PS372. FIG. 36D shows example binding curves for a multivalent polypeptide having two tandem copies of PS372. FIG. 36E shows example binding curves for the monovalent binder PS557. FIGs. 36F-36H show example binding curves for tandem constructs having two copies of PS557 separated by Linker 1 (FIG. 36F), Linker 2 (FIG. 36G), or Linker 3 (FIG. 36H).

[0073] FIGs. 37A-37B show data from stopped-flow rapid kinetic analysis for k_{on} rate constant and k_{off} rate determination for binders and fusion proteins derived by C-terminal addition of protein shields. FIG. 37A shows a schematic illustrating assay design (top panel) and plots showing experimental results and analysis (middle and bottom panels) for determining association rate constant (k_{on}). FIG. 37B shows a schematic illustrating assay design (top panel)

and plots showing experimental results and analysis (bottom panel) for measuring dissociation rates (k_{off}).

DETAILED DESCRIPTION

[0074] Aspects of the application relate to methods of protein sequencing and identification, methods of polypeptide sequencing and identification, methods of amino acid identification, and compositions for performing such methods.

[0075] In some aspects, the application relates to the discovery of polypeptide sequencing techniques which may be implemented using existing analytic instruments with few or no device modifications. For example, previous polypeptide sequencing strategies have involved iterative cycling of different reagent mixtures through a reaction vessel containing a polypeptide being analyzed. Such strategies may require modification of an existing analytic instrument, such as a nucleic acid sequencing instrument, which may not be equipped with a flow cell or similar apparatus capable of reagent cycling. The inventors have recognized and appreciated that certain polypeptide sequencing techniques of the application do not require iterative reagent cycling, thereby permitting the use of existing instruments without significant modifications which might increase instrument size. Accordingly, in some aspects, the application provides methods of polypeptide sequencing that permit the use of smaller sequencing instruments. In some aspects, the application relates to the discovery of polypeptide sequencing techniques that allow both genomic and proteomic analyses to be performed using the same sequencing instrument.

[0076] The inventors have further recognized and appreciated that differential binding interactions can provide an additional or alternative approach to conventional labeling strategies in polypeptide sequencing. Conventional polypeptide sequencing can involve labeling each type of amino acid with a uniquely identifiable label. This process can be laborious and prone to error, as there are at least twenty different types of naturally occurring amino acids in addition to numerous post-translational variations thereof. In some aspects, the application relates to the discovery of techniques involving the use of amino acid recognition molecules which differentially associate with different types of amino acids to produce detectable characteristic signatures indicative of an amino acid sequence of a polypeptide. Accordingly, aspects of the application provide techniques that do not require polypeptide labeling and/or harsh chemical reagents used in certain conventional polypeptide sequencing approaches, thereby increasing throughput and/or accuracy of sequence information obtained from a sample.

[0077] In some aspects, the application relates to the discovery that a polypeptide sequencing reaction can be monitored in real-time using only a single reaction mixture (e.g., without requiring iterative reagent cycling through a reaction vessel). As detailed above, conventional

polypeptide sequencing reactions can involve exposing a polypeptide to different reagent mixtures to cycle between steps of amino acid detection and amino acid cleavage. Accordingly, in some aspects, the application relates to an advancement in next generation sequencing that allows for the analysis of polypeptides by amino acid detection throughout an ongoing degradation reaction in real-time. Approaches for such polypeptide analysis by dynamic sequencing are described below.

[0078] As described herein, in some aspects, the application provides methods of sequencing a polypeptide by obtaining data during a polypeptide degradation process, and analyzing the data to determine portions of the data corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process. In some embodiments, the portions of the data comprise a series of signal pulses indicative of association of one or more amino acid recognition molecules with successive amino acids exposed at the terminus of the polypeptide (e.g., during a degradation). In some embodiments, the series of signal pulses corresponds to a series of reversible single molecule binding interactions at the terminus of the polypeptide during the degradation process.

[0079] A non-limiting example of polypeptide sequencing by detecting single molecule binding interactions during a polypeptide degradation process is schematically illustrated in FIG. 1A. An example signal trace (I) is shown with a series of panels (II) that depict different association events at times corresponding to changes in the signal. As shown, an association event between an amino acid recognition molecule (stippled shape) and an amino acid at the terminus of a polypeptide (shown as beads-on-a-string) produces a change in magnitude of the signal that persists for a duration of time.

[0080] Panels (A) and (B) depict different association events between an amino acid recognition molecule and a first amino acid exposed at the terminus of the polypeptide (e.g., a first terminal amino acid). Each association event produces a change in the signal trace (I) characterized by a change in magnitude of the signal that persists for the duration of the association event. Accordingly, the time duration between the association events of panels (A) and (B) may correspond to a duration of time within which the polypeptide is not detectably associated with an amino acid recognition molecule.

[0081] Panels (C) and (D) depict different association events between an amino acid recognition molecule and a second amino acid exposed at the terminus of the polypeptide (e.g., a second terminal amino acid). As described herein, an amino acid that is "exposed" at the terminus of a polypeptide is an amino acid that is still attached to the polypeptide and that becomes the terminal amino acid upon removal of the prior terminal amino acid during degradation (e.g., either alone or along with one or more additional amino acids). Accordingly, the first and

second amino acids of the series of panels (II) provide an illustrative example of successive amino acids exposed at the terminus of the polypeptide, where the second amino acid became the terminal amino acid upon removal of the first amino acid.

[0082] As generically depicted, the association events of panels (C) and (D) produce changes in the signal trace (I) characterized by changes in magnitude that persist for time durations that are relatively shorter than that of panels (A) and (B), and the time duration between the association events of panels (C) and (D) is relatively shorter than that of panels (A) and (B). As described herein, in some embodiments, either one or both of these distinctive changes in signal may be used to determine characteristic patterns in the signal trace (I) which can discriminate between different types of amino acids. In some embodiments, a transition from one characteristic pattern to another is indicative of amino acid cleavage. As used herein, in some embodiments, amino acid cleavage refers to the removal of at least one amino acid from a terminus of a polypeptide (e.g., the removal of at least one terminal amino acid from the polypeptide). In some embodiments, amino acid cleavage is determined by inference based on a time duration between characteristic patterns. In some embodiments, amino acid cleavage is determined by detecting a change in signal produced by association of a labeled cleaving reagent with an amino acid at the terminus of the polypeptide. As amino acids are sequentially cleaved from the terminus of the polypeptide during degradation, a series of changes in magnitude, or a series of signal pulses, is detected. In some embodiments, signal pulse data can be analyzed as illustrated in FIG. 1B.

[0083] In some embodiments, signal data can be analyzed to extract signal pulse information by applying threshold levels to one or more parameters of the signal data. For example, panel (III) depicts a threshold magnitude level ("M_L") applied to the signal data of the example signal trace (I). In some embodiments, M_L is a minimum difference between a signal detected at a point in time and a baseline determined for a given set of data. In some embodiments, a signal pulse ("sp") is assigned to each portion of the data that is indicative of a change in magnitude exceeding M_L and persisting for a duration of time. In some embodiments, a threshold time duration may be applied to a portion of the data that satisfies M_L to determine whether a signal pulse is assigned to that portion. For example, experimental artifacts may give rise to a change in magnitude exceeding M_L that does not persist for a duration of time sufficient to assign a signal pulse with a desired confidence (e.g., transient association events which could be non-discriminatory for amino acid type, non-specific detection events such as diffusion into an observation region or reagent sticking within an observation region). Accordingly, in some embodiments, a signal pulse is extracted from signal data based on a threshold magnitude level and a threshold time duration.

[0084] Extracted signal pulse information is shown in panel (III) with the example signal trace (I) superimposed for illustrative purposes. In some embodiments, a peak in magnitude of a signal pulse is determined by averaging the magnitude detected over a duration of time that persists above M_L. It should be appreciated that, in some embodiments, a "signal pulse" as used herein can refer to a change in signal data that persists for a duration of time above a baseline (e.g., raw signal data, as illustrated by the example signal trace (I)), or to signal pulse information extracted therefrom (e.g., processed signal data, as illustrated in panel (IV)).

[0085] Panel (IV) shows the signal pulse information extracted from the example signal trace (I). In some embodiments, signal pulse information can be analyzed to identify different types of amino acids in a sequence based on different characteristic patterns in a series of signal pulses. For example, as shown in panel (IV), the signal pulse information is indicative of a first type of amino acid based on a first characteristic pattern ("CP_1") and a second type of amino acid based on a second characteristic pattern ("CP_2"). By way of example, the two signal pulses detected at earlier time points provide information indicative of the first amino acid at the terminus of the polypeptide based on CP_1, and the two signal pulses detected at later time points provide information indicative of the second amino acid at the terminus of the polypeptide based on CP_2.

[0086] Also as shown in panel (IV), each signal pulse comprises a pulse duration ("*pd*") corresponding to an association event between the amino acid recognition molecule and the amino acid of the characteristic pattern. In some embodiments, the pulse duration is characteristic of a dissociation rate of binding. Also as shown, each signal pulse of a characteristic pattern is separated from another signal pulse of the characteristic pattern by an interpulse duration ("*ipd*"). In some embodiments, the interpulse duration is characteristic of an association rate of binding. In some embodiments, a change in magnitude ("ΔM") can be determined for a signal pulse based on a difference between baseline and the peak of a signal pulse. In some embodiments, a characteristic pattern is determined based on pulse duration. In some embodiments, a characteristic pattern is determined based on pulse duration and interpulse duration. In some embodiments, a characteristic pattern is determined based on any one or more of pulse duration, interpulse duration, and change in magnitude.

[0087] Accordingly, as illustrated by FIGs. 1A-1B, in some embodiments, polypeptide sequencing is performed by detecting a series of signal pulses indicative of association of one or more amino acid recognition molecules with successive amino acids exposed at the terminus of a polypeptide in an ongoing degradation reaction. The series of signal pulses can be analyzed to determine characteristic patterns in the series of signal pulses, and the time course of characteristic patterns can be used to determine an amino acid sequence of the polypeptide.

[0088] In some embodiments, the series of signal pulses comprises a series of changes in magnitude of an optical signal over time. In some embodiments, the series of changes in the optical signal comprises a series of changes in luminescence produced during association events. In some embodiments, luminescence is produced by a detectable label associated with one or more reagents of a sequencing reaction. For example, in some embodiments, each of the one or more amino acid recognition molecules comprises a luminescent label. In some embodiments, a cleaving reagent comprises a luminescent label. Examples of luminescent labels and their use in accordance with the application are provided elsewhere herein.

[0089] In some embodiments, the series of signal pulses comprises a series of changes in magnitude of an electrical signal over time. In some embodiments, the series of changes in the electrical signal comprises a series of changes in conductance produced during association events. In some embodiments, conductivity is produced by a detectable label associated with one or more reagents of a sequencing reaction. For example, in some embodiments, each of the one or more amino acid recognition molecules comprises a conductivity label. Examples of conductivity labels and their use in accordance with the application are provided elsewhere herein. Methods for identifying single molecules using conductivity labels have been described (see, e.g., U.S. Patent Publication No. 2017/0037462).

[0090] In some embodiments, the series of changes in conductance comprises a series of changes in conductance through a nanopore. For example, methods of evaluating receptor-ligand interactions using nanopores have been described (see, e.g., Thakur, A.K. & Movileanu, L. (2019) *Nature Biotechnology* 37(1)). The inventors have recognized and appreciated that such nanopores may be used to monitor polypeptide sequencing reactions in accordance with the application. Accordingly, in some embodiments, the application provides methods of polypeptide sequencing comprising contacting a single polypeptide molecule with one or more amino acid recognition molecules, where the single polypeptide molecule is immobilized to a nanopore. In some embodiments, the methods further comprise detecting a series of changes in conductance through the nanopore indicative of association of the one or more terminal amino acid recognition molecules with successive amino acids exposed at a terminus of the single polypeptide while the single polypeptide is being degraded, thereby sequencing the single polypeptide molecule.

[0091] In some aspects, the application provides methods of sequencing and/or identifying an individual protein in a complex mixture of proteins by identifying one or more types of amino acids of a polypeptide from the mixture. In some embodiments, one or more amino acids (e.g., terminal amino acids and/or internal amino acids) of the polypeptide are labeled (e.g., directly or indirectly, for example using a binding agent such as an amino acid recognition molecule) and

the relative positions of the labeled amino acids in the polypeptide are determined. In some embodiments, the relative positions of amino acids in a polypeptide are determined using a series of amino acid labeling and cleavage steps. However, in some embodiments, the relative position of labeled amino acids in a polypeptide can be determined without removing amino acids from the polypeptide but by translocating a labeled polypeptide through a pore (e.g., a protein channel) and detecting a signal (e.g., a FRET signal) from the labeled amino acid(s) during translocation through the pore in order to determine the relative position of the labeled amino acids in the polypeptide molecule.

[0092] In some embodiments, the identity of a terminal amino acid (e.g., an N-terminal or a C-terminal amino acid) is assessed after which the terminal amino acid is removed and the identity of the next amino acid at the terminus is assessed, and this process is repeated until a plurality of successive amino acids in the polypeptide are assessed. In some embodiments, assessing the identity of an amino acid comprises determining the type of amino acid that is present. In some embodiments, determining the type of amino acid comprises determining the actual amino acid identity, for example by determining which of the naturally-occurring 20 amino acids is the terminal amino acid is (e.g., using a binding agent that is specific for an individual terminal amino acid). In some embodiments, the type of amino acid is selected from alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, selenocysteine, serine, threonine, tryptophan, tyrosine, and valine.

[0093] However, in some embodiments assessing the identity of a terminal amino acid type can comprise determining a subset of potential amino acids that can be present at the terminus of the polypeptide. In some embodiments, this can be accomplished by determining that an amino acid is not one or more specific amino acids (and therefore could be any of the other amino acids). In some embodiments, this can be accomplished by determining which of a specified subset of amino acids (e.g., based on size, charge, hydrophobicity, post-translational modification, binding properties) could be at the terminus of the polypeptide (e.g., using a binding agent that binds to a specified subset of two or more terminal amino acids).

[0094] In some embodiments, assessing the identity of a terminal amino acid type comprises determining that an amino acid comprises a post-translational modification. Non-limiting examples of post-translational modifications include acetylation, ADP-ribosylation, caspase cleavage, citrullination, formylation, N-linked glycosylation, O-linked glycosylation, hydroxylation, methylation, myristoylation, neddylation, nitration, oxidation, palmitoylation, phosphorylation, prenylation, S-nitrosylation, sulfation, sumoylation, and ubiquitination.

[0095] In some embodiments, assessing the identity of a terminal amino acid type comprises determining that an amino acid comprises a side chain characterized by one or more biochemical properties. For example, an amino acid may comprise a nonpolar aliphatic side chain, a positively charged side chain, a negatively charged side chain, a nonpolar aromatic side chain, or a polar uncharged side chain. Non-limiting examples of an amino acid comprising a nonpolar aliphatic side chain include alanine, glycine, valine, leucine, methionine, and isoleucine. Non-limiting examples of an amino acid comprising a positively charged side chain includes lysine, arginine, and histidine. Non-limiting examples of an amino acid comprising a negatively charged side chain include aspartate and glutamate. Non-limiting examples of an amino acid comprising a nonpolar, aromatic side chain include phenylalanine, tyrosine, and tryptophan. Non-limiting examples of an amino acid comprising a polar uncharged side chain include serine, threonine, cysteine, proline, asparagine, and glutamine.

[0096] In some embodiments, a protein or polypeptide can be digested into a plurality of smaller polypeptides and sequence information can be obtained from one or more of these smaller polypeptides (e.g., using a method that involves sequentially assessing a terminal amino acid of a polypeptide and removing that amino acid to expose the next amino acid at the terminus).

[0097] In some embodiments, a polypeptide is sequenced from its amino (N) terminus. In some embodiments, a polypeptide is sequenced from its carboxy (C) terminus. In some embodiments, a first terminus (e.g., N or C terminus) of a polypeptide is immobilized and the other terminus (e.g., the C or N terminus) is sequenced as described herein.

[0098] As used herein, sequencing a polypeptide refers to determining sequence information for a polypeptide. In some embodiments, this can involve determining the identity of each sequential amino acid for a portion (or all) of the polypeptide. However, in some embodiments, this can involve assessing the identity of a subset of amino acids within the polypeptide (e.g., and determining the relative position of one or more amino acid types without determining the identity of each amino acid in the polypeptide). However, in some embodiments, amino acid content information can be obtained from a polypeptide without directly determining the relative position of different types of amino acids in the polypeptide. The amino acid content alone may be used to infer the identity of the polypeptide that is present (e.g., by comparing the amino acid content to a database of polypeptide information and determining which polypeptide(s) have the same amino acid content).

[0099] In some embodiments, sequence information for a plurality of polypeptide products obtained from a longer polypeptide or protein (e.g., via enzymatic and/or chemical cleavage) can be analyzed to reconstruct or infer the sequence of the longer polypeptide or protein.

[0100] Accordingly, in some embodiments, the one or more types of amino acids are identified by detecting luminescence of one or more labeled recognition molecules that selectively bind the one or more types of amino acids. In some embodiments, the one or more types of amino acids are identified by detecting luminescence of a labeled polypeptide.

[0101] The inventors have further recognized and appreciated that the polypeptide sequencing techniques described herein may involve generating novel polypeptide sequencing data, particularly in contrast with conventional polypeptide sequencing techniques. Thus, conventional techniques for analyzing polypeptide sequencing data may not be sufficient when applied to the data generated using the polypeptide sequencing techniques described herein.

[0102] For example, conventional polypeptide sequencing techniques that involve iterative reagent cycling may generate data associated with individual amino acids of a polypeptide being sequenced. In such instances, analyzing the data generated may simply involve determining which amino acid is being detected at a particular time because the data being detected corresponds to only one amino acid. In contrast, the polypeptide sequencing techniques described herein may generate data during a polypeptide degradation process while multiple amino acids of the polypeptide molecule are being detected, resulting in data where it may be difficult to discern between sections of the data corresponding to different amino acids of the polypeptide. Accordingly, the inventors have developed new computational techniques for analyzing such data generated by the polypeptide sequencing techniques described herein that involve determining sections of the data that correspond to individual amino acids, such as by segmenting the data into portions that correspond to respective amino acid association events. Those sections may be then further analyzed to identify the amino acid being detected during those individual sections.

[0103] As another example, conventional sequencing techniques that involve using uniquely identifiable labels for each type of amino acid may involve simply analyzing which label is being detected at a particular time without taking into consideration any dynamics in how individual amino acids interact with other molecules. In contrast, the polypeptide sequencing techniques described herein generate data indicating how amino acids interact with recognition molecules. As discussed above, the data may include a series of characteristic patterns corresponding to association events between amino acids and their respective recognition molecules. Accordingly, the inventors have developed new computational techniques for analyzing the characteristic patterns to determine a type of amino acid corresponding to that portion of the data, allowing for an amino acid sequence of a polypeptide to be determined by analyzing a series of different characteristic patterns.

[0104] In some aspects, the polypeptide sequencing techniques described herein generate data indicating how a polypeptide interacts with a binding means while the polypeptide is being degraded by a cleaving means. As discussed above, the data can include a series of characteristic patterns corresponding to association events at a terminus of a polypeptide in between cleavage events at the terminus. In some embodiments, methods of sequencing described herein comprise contacting a single polypeptide molecule with a binding means and a cleaving means, where the binding means and the cleaving means are configured to achieve at least 10 association events prior to a cleavage event. In some embodiments, the means are configured to achieve the at least 10 association events between two cleavage events.

[0105] As described herein, in some embodiments, a plurality of single-molecule sequencing reactions are performed in parallel in an array of sample wells. In some embodiments, an array comprises between about 10,000 and about 1,000,000 sample wells. The volume of a sample well may be between about 10^{-21} liters and about 10^{-15} liters, in some implementations. Because the sample well has a small volume, detection of single-molecule events may be possible as only about one polypeptide may be within a sample well at any given time. Statistically, some sample wells may not contain a single-molecule sequencing reaction and some may contain more than one single polypeptide molecule. However, an appreciable number of sample wells may each contain a single-molecule reaction (e.g., at least 30% in some embodiments), so that single-molecule analysis can be carried out in parallel for a large number of sample wells. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 association events prior to a cleavage event in at least 10% (e.g., 10-50%, more than 50%, 25-75%, at least 80%, or more) of the sample wells in which a single-molecule reaction is occurring. In some embodiments, the binding means and the cleaving means are configured to achieve at least 10 association events prior to a cleavage event for at least 50% (e.g., more than 50%, 50-75%, at least 80%, or more) of the amino acids of a polypeptide in a single-molecule reaction.

Amino Acid Recognition Molecules

[0106] In some embodiments, methods provided herein comprise contacting a polypeptide with an amino acid recognition molecule, which may or may not comprise a label, that selectively binds at least one type of terminal amino acid. As used herein, in some embodiments, a terminal amino acid may refer to an amino-terminal amino acid of a polypeptide or a carboxy-terminal amino acid of a polypeptide. In some embodiments, a labeled recognition molecule selectively binds one type of terminal amino acid over other types of terminal amino acids. In some embodiments, a labeled recognition molecule selectively binds one type of terminal amino acid over an internal amino acid of the same type. In yet other embodiments, a labeled recognition

molecule selectively binds one type of amino acid at any position of a polypeptide, e.g., the same type of amino acid as a terminal amino acid and an internal amino acid.

[0107] As used herein, in some embodiments, a type of amino acid refers to one of the twenty naturally occurring amino acids or a subset of types thereof. In some embodiments, a type of amino acid refers to a modified variant of one of the twenty naturally occurring amino acids or a subset of unmodified and/or modified variants thereof. Examples of modified amino acid variants include, without limitation, post-translationally-modified variants (e.g., acetylation, ADP-ribosylation, caspase cleavage, citrullination, formylation, N-linked glycosylation, O-linked glycosylation, hydroxylation, methylation, myristoylation, neddylation, nitration, oxidation, palmitoylation, phosphorylation, prenylation, S-nitrosylation, sulfation, sumoylation, and ubiquitination), chemically modified variants, unnatural amino acids, and proteinogenic amino acids such as selenocysteine and pyrrolysine. In some embodiments, a subset of types of amino acids includes more than one and fewer than twenty amino acids having one or more similar biochemical properties. For example, in some embodiments, a type of amino acid refers to one type selected from amino acids with charged side chains (e.g., positively and/or negatively charged side chains), amino acids with polar side chains (e.g., polar uncharged side chains), amino acids with nonpolar side chains (e.g., nonpolar aliphatic and/or aromatic side chains), and amino acids with hydrophobic side chains.

[0108] In some embodiments, methods provided herein comprise contacting a polypeptide with one or more labeled recognition molecules that selectively bind one or more types of terminal amino acids. As an illustrative and non-limiting example, where four labeled recognition molecules are used in a method of the application, any one recognition molecule selectively binds one type of terminal amino acid that is different from another type of amino acid to which any of the other three selectively binds (e.g., a first recognition molecule binds a first type, a second recognition molecule binds a second type, a third recognition molecule binds a third type, and a fourth recognition molecule binds a fourth type of terminal amino acid). For the purposes of this discussion, one or more labeled recognition molecules in the context of a method described herein may be alternatively referred to as a set of labeled recognition molecules.

[0109] In some embodiments, a set of labeled recognition molecules comprises at least one and up to six labeled recognition molecules. For example, in some embodiments, a set of labeled recognition molecules comprises one, two, three, four, five, or six labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises ten or fewer labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises eight or fewer labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises six or fewer labeled recognition molecules. In some embodiments, a set of

labeled recognition molecules comprises four or fewer labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises three or fewer labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises two or fewer labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises four labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises at least two and up to twenty (e.g., at least two and up to ten, at least two and up to eight, at least four and up to twenty, at least four and up to ten) labeled recognition molecules. In some embodiments, a set of labeled recognition molecules comprises more than twenty (e.g., 20 to 25, 20 to 30) recognition molecules. It should be appreciated, however, that any number of recognition molecules may be used in accordance with a method of the application to accommodate a desired use.

[0110] In accordance with the application, in some embodiments, one or more types of amino acids are identified by detecting luminescence of a labeled recognition molecule. In some embodiments, a labeled recognition molecule comprises a recognition molecule that selectively binds one type of amino acid and a luminescent label having a luminescence that is associated with the recognition molecule. In this way, the luminescence (e.g., luminescence lifetime, luminescence intensity, and other luminescence properties described elsewhere herein) may be associated with the selective binding of the recognition molecule to identify an amino acid of a polypeptide. In some embodiments, a plurality of types of labeled recognition molecules may be used in a method according to the application, wherein each type comprises a luminescent label having a luminescence that is uniquely identifiable from among the plurality. Suitable luminescent labels may include luminescent molecules, such as fluorophore dyes, and are described elsewhere herein.

[0111] In some embodiments, one or more types of amino acids are identified by detecting one or more electrical characteristics of a labeled recognition molecule. In some embodiments, a labeled recognition molecule comprises a recognition molecule that selectively binds one type of amino acid and a conductivity label that is associated with the recognition molecule. In this way, the one or more electrical characteristics (e.g., charge, current oscillation color, and other electrical characteristics) may be associated with the selective binding of the recognition molecule to identify an amino acid of a polypeptide. In some embodiments, a plurality of types of labeled recognition molecules may be used in a method according to the application, wherein each type comprises a conductivity label that produces a change in an electrical signal (e.g., a change in conductance, such as a change in amplitude of conductivity and conductivity transitions of a characteristic pattern) that is uniquely identifiable from among the plurality. In some embodiments, the plurality of types of labeled recognition molecules each comprises a

conductivity label having a different number of charged groups (e.g., a different number of negatively and/or positively charged groups). Accordingly, in some embodiments, a conductivity label is a charge label. Examples of charge labels include dendrimers, nanoparticles, nucleic acids and other polymers having multiple charged groups. In some embodiments, a conductivity label is uniquely identifiable by its net charge (e.g., a net positive charge or a net negative charge), by its charge density, and/or by its number of charged groups.

[0112] In some embodiments, an amino acid recognition molecule may be engineered by one skilled in the art using conventionally known techniques. In some embodiments, desirable properties may include an ability to bind selectively and with high affinity to one type of amino acid only when it is located at a terminus (e.g., an N-terminus or a C-terminus) of a polypeptide. In yet other embodiments, desirable properties may include an ability to bind selectively and with high affinity to one type of amino acid when it is located at a terminus (e.g., an N-terminus or a C-terminus) of a polypeptide and when it is located at an internal position of the polypeptide. In some embodiments, desirable properties include an ability to bind selectively and with low affinity (e.g., with a K_D of about 50 nM or higher, for example, between about 50 nM and about 50 μM, between about 100 nM and about 10 μM, between about 500 nM and about 50 μM) to more than one type of amino acid. For example, in some aspects, the application provides methods of sequencing by detecting reversible binding interactions during a polypeptide degradation process. Advantageously, such methods may be performed using a recognition molecule that reversibly binds with low affinity to more than one type of amino acid (e.g., a subset of amino acid types).

[0113] As used herein, in some embodiments, the terms "selective" and "specific" (and variations thereof, e.g., selectively, specifically, selectivity, specificity) refer to a preferential binding interaction. For example, in some embodiments, an amino acid recognition molecule that selectively binds one type of amino acid preferentially binds the one type over another type of amino acid. A selective binding interaction will discriminate between one type of amino acid (e.g., one type of terminal amino acid) and other types of amino acids (e.g., other types of terminal amino acids), typically more than about 10- to 100-fold or more (e.g., more than about 1,000- or 10,000-fold). Accordingly, it should be appreciated that a selective binding interaction can refer to any binding interaction that is uniquely identifiable to one type of amino acid over other types of amino acids. For example, in some aspects, the application provides methods of polypeptide sequencing by obtaining data indicative of association of one or more amino acid recognition molecules with a polypeptide molecule. In some embodiments, the data comprises a series of signal pulses corresponding to a series of reversible amino acid recognition molecule binding interactions with an amino acid of the polypeptide molecule, and the data may be used to

determine the identity of the amino acid. As such, in some embodiments, a "selective" or "specific" binding interaction refers to a detected binding interaction that discriminates between one type of amino acid and other types of amino acids.

[0114] In some embodiments, an amino acid recognition molecule binds one type of amino acid with a dissociation constant (K_D) of less than about 10^{-6} M (e.g., less than about 10^{-7} M, less than about 10^{-8} M, less than about 10^{-9} M, less than about 10^{-10} M, less than about 10^{-11} M, less than about 10^{-12} M, to as low as 10^{-16} M) without significantly binding to other types of amino acids. In some embodiments, an amino acid recognition molecule binds one type of amino acid (e.g., one type of terminal amino acid) with a K_D of less than about 100 nM, less than about 50 nM, less than about 25 nM, less than about 10 nM, or less than about 1 nM. In some embodiments, an amino acid recognition molecule binds one type of amino acid with a K_D of between about 50 nM and about 50 μM (e.g., between about 50 nM and about 500 nM, between about 50 nM and about 5 μM, between about 500 nM and about 50 μM, between about 5 μM and about 50 μM, or between about 10 μM and about 50 μM). In some embodiments, an amino acid recognition molecule binds one type of amino acid with a K_D of about 50 nM.

[0115] In some embodiments, an amino acid recognition molecule binds two or more types of amino acids with a K_D of less than about 10^{-6} M (e.g., less than about 10^{-7} M, less than about 10^{-8} M, less than about 10^{-9} M, less than about 10^{-10} M, less than about 10^{-11} M, less than about 10^{-12} M, to as low as 10^{-16} M). In some embodiments, an amino acid recognition molecule binds two or more types of amino acids with a K_D of less than about 100 nM, less than about 50 nM, less than about 25 nM, less than about 10 nM, or less than about 1 nM. In some embodiments, an amino acid recognition molecule binds two or more types of amino acids with a K_D of between about 50 nM and about 50 μM (e.g., between about 50 nM and about 500 nM, between about 50 nM and about 5 μM, between about 500 nM and about 50 μM, between about 5 μM and about 50 μM, or between about 10 μM and about 50 μM). In some embodiments, an amino acid recognition molecule binds two or more types of amino acids with a K_D of about 50 nM.

[0116] In some embodiments, an amino acid recognition molecule binds at least one type of amino acid with a dissociation rate (k_{off}) of at least 0.1 s^{-1}. In some embodiments, the dissociation rate is between about 0.1 s^{-1} and about 1,000 s^{-1} (e.g., between about 0.5 s^{-1} and about 500 s^{-1}, between about 0.1 s^{-1} and about 100 s^{-1}, between about 1 s^{-1} and about 100 s^{-1}, or between about 0.5 s^{-1} and about 50 s^{-1}). In some embodiments, the dissociation rate is between about 0.5 s^{-1} and about 20 s^{-1}. In some embodiments, the dissociation rate is between about 2 s^{-1} and about 20 s^{-1}. In some embodiments, the dissociation rate is between about 0.5 s^{-1} and about 2 s^{-1}.

[0117] In some embodiments, the value for K_D or k_{off} can be a known literature value, or the value can be determined empirically. For example, the value for K_D or k_{off} can be measured in a single-molecule assay or an ensemble assay (see, e.g., Example 4 and FIG. 19M). In some embodiments, the value for k_{off} can be determined empirically based on signal pulse information obtained in a single-molecule assay as described elsewhere herein. For example, the value for k_{off} can be approximated by the reciprocal of the mean pulse duration. In some embodiments, an amino acid recognition molecule binds two or more types of amino acids with a different K_D or k_{off} for each of the two or more types. In some embodiments, a first K_D or k_{off} for a first type of amino acid differs from a second K_D or k_{off} for a second type of amino acid by at least 10% (e.g., at least 25%, at least 50%, at least 100%, or more). In some embodiments, the first and second values for K_D or k_{off} differ by about 10-25%, 25-50%, 50-75%, 75-100%, or more than 100%, for example by about 2-fold, 3-fold, 4-fold, 5-fold, or more.

[0118] In accordance with the methods and compositions provided herein, FIG. 2 shows various example configurations and uses of labeled recognition molecules. In some embodiments, a labeled recognition molecule **200** comprises a luminescent label **210** (e.g., a label) and a recognition molecule (shown as stippled shapes) that selectively binds one or more types of terminal amino acids of a polypeptide **220**. In some embodiments, a recognition molecule is selective for one type of amino acid or a subset (e.g., fewer than the twenty common types of amino acids) of types of amino acids at a terminal position or at both terminal and internal positions.

[0119] As described herein, an amino acid recognition molecule may be any biomolecule capable of selectively or specifically binding one molecule over another molecule (e.g., one type of amino acid over another type of amino acid). In some embodiments, a recognition molecule is not a peptidase or does not have peptidase activity. For example, in some embodiments, methods of polypeptide sequencing of the application involve contacting a polypeptide molecule with one or more recognition molecules and a cleaving reagent. In such embodiments, the one or more recognition molecules do not have peptidase activity, and removal of one or more amino acids from the polypeptide molecule (e.g., amino acid removal from a terminus of the polypeptide molecule) is performed by the cleaving reagent.

[0120] Recognition molecules include, for example, proteins and nucleic acids, which may be synthetic or recombinant. In some embodiments, a recognition molecule may be an antibody or an antigen-binding portion of an antibody, an SH2 domain-containing protein or fragment thereof, or an enzymatic biomolecule, such as a peptidase, an aminotransferase, a ribozyme, an aptazyme, or a tRNA synthetase, including aminoacyl-tRNA synthetases and related molecules described in U.S. Patent Application No. 15/255,433, filed September 2, 2016, titled

"MOLECULES AND METHODS FOR ITERATIVE POLYPEPTIDE ANALYSIS AND PROCESSING."

[0121] In some aspects, the application relates to the discovery and development of amino acid recognition molecules for use in accordance with methods described herein or known in the art. In some embodiments, the application provides amino acid binding proteins (e.g., ClpS proteins) having binding properties that were previously not known to exist among other homologous members of a protein family. In some embodiments, the application provides engineered amino acid binding proteins. For example, in some embodiments, the application provides fusion constructs comprising a single polypeptide having tandem copies of two or more amino acid binding proteins.

[0122] The inventors have recognized and appreciated that fusion constructs of the application allow for an effective increase in recognition molecule concentration without increasing label background noise (e.g., background fluorescence). The inventors have further recognized and appreciated that fusion constructs of the application provide increased accuracy in sequencing reactions and/or decrease the amount of time required to perform a sequencing reaction. Additionally, by providing fusion constructs having tandem copies of two or more different types of amino acid binding proteins, fewer reagents are required in reactions, which provides a more efficient and inexpensive approach for sequencing.

[0123] In some embodiments, a recognition molecule of the application is a degradation pathway protein. Examples of degradation pathway proteins suitable for use as recognition molecules include, without limitation, N-end rule pathway proteins, such as Arg/N-end rule pathway proteins, Ac/N-end rule pathway proteins, and Pro/N-end rule pathway proteins. In some embodiments, a recognition molecule is an N-end rule pathway protein selected from a Gid protein (e.g., Gid4 or Gid10 protein), a UBR-box protein (e.g., UBR1, UBR2) or UBR-box domain-containing protein fragment thereof, a p62 protein or ZZ domain-containing fragment thereof, and a ClpS protein (e.g., ClpS1, ClpS2). Accordingly, in some embodiments, labeled recognition molecule **200** comprises a degradation pathway protein. In some embodiments, labeled recognition molecule **200** comprises a ClpS protein.

[0124] In some embodiments, a recognition molecule of the application is a ClpS protein, such as *Agrobacterium tumifaciens* ClpS1, *Agrobacterium tumifaciens* ClpS2, *Synechococcus elongatus* ClpS1, *Synechococcus elongatus* ClpS2, *Thermosynechococcus elongatus* ClpS, *Escherichia coli* ClpS, or *Plasmodium falciparum* ClpS. In some embodiments, the recognition molecule is an L/F transferase, such as *Escherichia coli* leucyl/phenylalanyl-tRNA-protein transferase. In some embodiments, the recognition molecule is a D/E leucyltransferase, such as *Vibrio vulnificus* Aspartate/glutamate leucyltransferase Bpt. In some embodiments, the

recognition molecule is a UBR protein or UBR-box domain, such as the UBR protein or UBR-box domain of human UBR1 and UBR2 or *Saccharomyces cerevisiae* UBR1. In some embodiments, the recognition molecule is a p62 protein, such as *H. sapiens* p62 protein or *Rattus norvegicus* p62 protein, or truncation variants thereof that minimally include a ZZ domain. In some embodiments, the recognition molecule is a Gid4 protein, such as *H. sapiens* GID4 or *Saccharomyces cerevisiae* GID4. In some embodiments, the recognition molecule is a Gid10 protein, such as *Saccharomyces cerevisiae* GID10. In some embodiments, the recognition molecule is an N-meristoyltransferase, such as *Leishmania major* N-meristoyltransferase or *H. sapiens* N-meristoyltransferase NMT1. In some embodiments, the recognition molecule is a BIR2 protein, such as *Drosophila melanogaster* BIR2. In some embodiments, the recognition molecule is a tyrosine kinase or SH2 domain of a tyrosine kinase, such as *H. sapiens* Fyn SH2 domain, *H. sapiens* Src tyrosine kinase SH2 domain, or variants thereof, such as *H. sapiens* Fyn SH2 domain triple mutant superbinder. In some embodiments, the recognition molecule is an antibody or antibody fragment, such as a single-chain antibody variable fragment (scFv) against phosphotyrosine or another post-translationally modified amino acid variant described herein.

[0125] Table 1 and Table 2 provide a list of example sequences of amino acid recognition molecules. Also shown are the amino acid binding preferences of each molecule with respect to amino acid identity at a terminal position of a polypeptide unless otherwise specified in Table 1 and Table 2. It should be appreciated that these sequences and other examples described herein are meant to be non-limiting, and recognition molecules in accordance with the application can include any homologs, variants thereof, or fragments thereof minimally containing domains or subdomains responsible for peptide recognition.

Table 1. Non-limiting examples of ClpS amino acid recognition proteins.

Name	Binding Pref.*	SEQ ID NO:	Sequence
PS368	F, Y, W, L	1	MASAPSTTLDKSTQVVKKTYPNYKVIVLNDDLNTFDHVA NCLIKYIPDMTTDRAWELTNQVHYQGQAIVWTGPQEQAE LYHQQLRREGLTMAPLEAA
PS369	F, Y, W, L	2	MTSTLRARPARDTDLQHRPYPHYRIIVLDDDVNTFQHVV NCLVTFLPGMTRDQAWAMAQQVDGEGSAVVWTGPQEQAE LYHVQLGNHGLTMAPLEPV
PS370	F, L	3	MFNSLGTVLDPKKSKAKYPEARVIVLDDNFNTFQHVANC LLAIIPRMCEQRAWDLTIKVDKAGSAEVWRGNLEQAELY HEQLFSKGLTMAPIEKT
PS371	F, Y, W, L	4	MATETIERPRTRDPGSGLGGHWLVIVLNDDHNTFDHVAK TLARVIPGVTVDDGYRFADQIHQRGQAIVWRGPKEPAEH YWEQLQDAGLSMAPLERH

PS372	L, I, V	5	MAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEM LQKIFGFPPEKGFQIAEEVDRTGRVILLTTSKEHAELKQ DQVHSYGPDPYLGRPCSGSMTCVIEPAV
PS373		6	MNRIKQEAVRTENLLICSESIRRTPGTMSNEESMFDEVV AVAVAEPETQHDERRGTKPKRQPPYHVILWDDTDHSFDY VIMMMKRLFRMPIEKGFQVAKEVDSSGRAICMTTTLELA ELKRDQIHAFGKDELLPRCKGSMSATIEPAEG
PS374	F, Y, W, L	7	MRWEDPLAAEPVTPGVAPVVEEETDAAVETPWRVILYDD DIHTFEEVILQLMKATGCTPEQGERHAWTVHTRGKDCVY QGDFFDCFRVQGVLREIQLVTEIEG
PS375	F, L	8	MEAEPETKVLASIPGVGTSEPFRVVLFNDEEHSFDEVIF QIIKAVRCSRAKAEALTMEVHNSGRSIVYTGPIEQCIRV SAVLEEIELRTEIQS
PS376	F, W, L	9	MPTNDLDLLEKQDVKIERPKMYQVVMYNDDFTPFDFVVA VLMQFFNKGMDEATAIMMQVHMQGKGICGVFPKDIAETK ATEVMKWAKVEQHPLRLQVEAQA
PS377	W	10	MADISKSRPEIGGPKGPQFGDSDRGGGVAVITKPVTKKK FKRKSQTEYEPYWHVLLHHDNVHTFEYATGAIVKVVRTV SRKTAHRITMQAHVSGVATVTTTWKAQAEEYCKGLQMHG LTSSIAPDSSFTH
PS378	F, Y, W, L	11	MXPQEVEEVSFLESKEHEIVLYNDDVNTFDHVIECLVKI CNHNYLQAEQCAYIVHHSGKCGVKTGSLEELIPKCNALL EEGLSAEVI
PS379		12	MSTQEEVLEEVKTTTQKENEIVLYNDDYNTFDHVIETLI YACEHTPVQAEQCAILVHYKGKCTVKTGSFDELKPRCSK LLEEGLSAEIV
PS380	F, W	13	MGDIYGESNPEEVSCIDSLSEEGNELILFNDNIHTFEYV IDCLVAICSLSYEQASNCAYIVDRKGLCTVKHGSYDELL IMYHALVEKDLKVEIR
PS381		14	MVAFSKKWKKDELDKSTGKQKMLILHNDSVNSFDYVIKT LCEVCDHDTIQAEQCAFLTHFKGQCEIAVGEVADLVPLK NKLLNKNLIVSIH
PS382	F, Y, W, L	15	MSDSPVIKEIKKDNIKEADEHEKKEREKETSAWKVILYN DDIHNFTYVTDVIVKVVGQISKAKAHTITVEAHSTGQAL ILSTWKSKAEKYCQELQQNGLTVSIIHESQLKDKQKK
PS388	F, Y, L	16	MVTTLSADVYGMATAPTVAPERSNQVVRKTYPNYKVIVL NDDFNTFQHVAECLMKYIPGMSSDRAWDLTNQVHYEGQA IVWVGPQEPAELYHQQLRRAGLTMAPLEAA
PS389	F, Y, L	17	MLNSAAFKAASASPVIAPERSGQVTQKPYPTYKVIVLND DFNTFQHVHDCLVKYIPGMTSDRAWQLTHQVHNDGQAIV WVGPQEQAELYHQQLSRAGLTMAPIEAA
PS390	F, Y, L	18	MLSIAAVTEAPSKGVQTADPKTVRKPYPNYKVIVLNDDF NTFQHVSSCLLKYIPGMSEARAWELTNQVHFEGLAVVWV GPQEQAELYYAQLKNAGLTMAPPEPA
PS391	F, Y, W, L	19	MGQTVEKPRVEGPGTGLGGSWRVIVRNDDHNTFDHVART LARFIPGVSLERGHEIAKVIHTTGRAVVYTGHKEAAEHY WQQLKGAGLTMAPLEQG
PS392	F, Y, W	20	MSVEIIEKRSTVRKLAPRYRVLLHNDDFNPMEYVVQTLM ATVPSLTQPQAVNVMMEAHTNGMGLVIVCALEHAEFYAE TLNNHGLGSSIEPDD
PS393	F, Y, W, L	21	MSDEDGEDGDENAVGIATRTRTRTKKPTPYRVLLLNDDY TPMEFVVLVLQRFFRMSIEDATRVMLQVHQKGVGVCGVF TYEVAETKVSQVIDFARQNQHPLQCTLEKA
PS394	F, Y, W, L	22	MAERRDTGDDEGTGLGIATKTRSKTKKPTPYRVLMLNDD YTPMEFVVLCLQRFFRMNMEEATRVMLHVHQKGVGVCGV FSYEVAETKVGQVIDFARANQHPLQCTLEKA

PS395	F, Y, W, L	23	MTVSQSKTQGAPAAQSATELEYEGLWRVVVLNDPVNLMS YVVLVFKKVFGFDETTARKHMLEVHEQGRSVVWSGMREK AEAYAFTLQQWHLTTVLEQDEVR
PS396	F, W	24	MSDNDVALKPKIKSKPKLERPKLYKVILVNDDFTPREFV IAVLKMVFRMSEETGYRVMLTAHRLGTSVVVVCARDIAE TKAKEAVDFGKEAGFPLMFTTEPEE
PS397	F, Y, W	25	MSDNEVAPKRKTRVKPKLERPRLYKVILVNDDYTPRDFV VMVLKAIFRMSEEAGYRVMMTAHKLGTSVVVVCARDIAE TKAKEATDLGKEAGFPLMFTTEPEE
PS398	F, W	26	MPLKAQNRSIVGRRDEWPPPTTQSSSETKSESKRVSDTG ADTKRKTKTVPKVEKPRLYKVILVNDDYTPREFVLVVLK AVFRMSEDQGYKVMITAHQKGSCVVAVYTRDIAETKAKE AVDLAKEIGFPLMFRTEPEE
PS404	F, Y, W	27	MPVSVTAPQTKTKTKPKVERPKLYKVILVNDDFTPREFV VRVLKAEFRMSEDQAAKVMMTAHQRGVCVVAVFTRDVAE TKATRATDAGRAKGYPLLFTTEPEE
PS405	F, Y, W	28	MVSIGAATVACAEGRPIFSGYFDWLAAMPETVTVPRTRL RPKTERPKLHKVILVNDDYTPREFVVTVLKGEFHMSEDQ AQRVMITAHRRGVCVVAVFTKDVAETKATRASDAGRAKG YPLQFTTEPEE
PS406	F, Y, W	29	MPDATTTPRTKTLTRTARPPLHKVILVNDDFTPREFVVR LLKAEFRTTGDEAQRIMITAHMKGSCVVAVFTREIAESK ATRATETARAEGFPLLFTTEPEE
PS407	F, Y, W, L	30	MPSNKRQMCLSDIKNSFNESGIVDWHISPRLANEPSEEG DSDLAVQTVPPELKRPPLYAVVLLNDDYTPMEFVIEILQ QYFAMNLDQATQVMLTVHYEGKGVAGVYPRDIAETKANQ VNNYARSQGHPLLCQIEPKD
PS408	F, Y, W, L	31	MTDPPSKGREDVDLATRTKPKTQRPPLYKVLLLNDDFTP MEFVVHILERLFGMTHAQAIEIMLTVHRKGVAVVGVFSH EIAETKVAQVMELARRQQHPLQCTMEKE
PS409	F, Y, W, L	32	MPARLTDIEGEPNTDPVEDVLLADPELKKPQMYAVVMYN DDYTPMEFVVDVLQNHFKHTLDSAISIMLAIHQQGKGIA GIYPKDIAETKAQTVNRKARQAGYPLLSQIEPQG
PS410	F, W, L	33	MGDDDQSSREGEGDVAFQTADPELKRPSLYRVVLLNDDY TPMEFVVHILEQFFAMNREKATQVMLAVHTQGKGVCGVY TKDIAETKAALVNDYSRENQHPLLCEVEELDDESR
PS411	F, Y, W, L	34	MTRPDAPEYDDDLAVEPAEPELARPPLYKVVLHNDDFTP MEFVVEVLQEFFNMDSEQAVQVMLAVHTQGKATCGIFTR DIAETKSYQVNEYARECEHPLMCDIEAAD
PS412	F, Y, W, L	35	MATKREGSTLLEPTAAKVKPPPLYKVLLLNDDYTPMEFV VLVLKKFFGIDQERATQIMLKVHTEGVGVCGVYPRDIAH TKVEQVVDFARQHQHPLQCTMEES
PS413	F, Y, W, L	36	MMKQCGSYFLIKAVQDFKPLSKHRSDTDVITETKIQVKQ PKLYTVIMYNDNYTTMDFVVYVLVEIFQHSIDKAYQLMM QIHESGQAAVALLPYDLAEMKVDEVTALAEQESYPLLTT IEPA
PS414	F, Y, W, L	37	MQAAGNEPPDPQNPGDVGNGGDGGNQDGSNTGVVVKTRT RTRKPAMYKVLMLNDDYTPMEFVVHVLERFFQKNREEAT RIMLHVHRRGVGVCGVYTYEVAETKVTQVMDLARQNQHP LQCTIEKE
PS415	F, Y, W	38	MALPETRTKIKPDVNIKEPPNYRVIYLNDDKTSMEFVIG SLMQHFSYPQQQAVEKTEEVHEHGSSTVAVLPYEMAEQK GIEVTLDARAEGFPLQVKIEPAER
PS416	F, Y, W, L	39	MTSQTDTLVKPNIQPPSLFKVIYINDSVTTMEFVVESLM SVFNHSADEATRLTQLVHEEGAAVVAILPYELAEQKGME VTLLARNNGFPLAIRLEPAV

PS417	F, Y, W	40	MSNLDTDVLIDEKVKVVTTEPEKYRVILLNDDVTPMDFV INILVSIFKHSTDTAKDLTLKIHKEGSAIVGVYTYEIAE QKGIEATNESRQHGFPLQVKIERENTL
PS418	F, Y, W, L	41	MSDHNIDHDTSVAVHLDVVVREPPMYRVVLLNDDFTPME FVVELLMHFFRKTAEQATQIMLNIHHEGVGVCGTYPREI AETKVAQVHQHARTNGHPLKCRMEPS
PS419	F, Y, W, L	42	MEKEQSLCKEKTHVELSEPKHYKVVFHNDDFTTMDFVVK VLQLVFFKSQLQAEDLTMKIHLEGSATAGIYSYDIAQSK AQKTTQMAREEGFPLRLTVEPEDN
PS420	F, Y, W, L	43	MSDYSNQISQAGSGVAEDASITLPPERKVVFYNDDFTTM EFVVDVLVSIFNKSHSEAEELMQTVHQEGSSVVGVYTYD IAVSRTNLTIQAARKNGFPLRVEVE
PS421	F, Y, W, L	44	MTTPNKRPEFEPEIGLEDEVGEPRKYKVLLHNDDYTTMD FVVQVLIEVFRKSETEATHIMLTIHEKGVGTCGIYPAEV AETKINEVHTRARREGFPLRASMEEV
PS422	F, Y, W, L	45	MTQIKPQTIPDTDVISQTQSDWQMPDLYAVIMHNDDYTT MDFVVFLLNAVFDKPIEQAYQLMMQIHQTGRAVVAILPY EIAEMKVDEATSLAEQEQFPLFISIEQA
PS423	F, Y, W	46	MAPTPAGAAVLDKQQQRRHKHASRYRVLLHNDPVNTMEY VVESLRQVVPQLSEQDAIAVMVEAHNTGVGLVIVCDIEP AEFYCEQLKTKGLTSSIEPED
PS424	F, Y, W	47	MSVETIEKRSTTRKLAPQYRVLLHNDDYNSMEYVVQVLM TSVPSITQPQAVNIMMEAHNSGLALVITCAQEHAEFYCE TLKGHGLSSTIEPD
PS425	F, Y, W, L	48	MTHYFSNILRDQESPKINPKELEQIDVLEEKEHQIILYN DDVNTFEHVIDCLVKICEHNYLQAEQCAYIVHHSGKCSV KTGSLDELVPKCNALLEEGLSAEVV
PS426	F, Y, W, L	49	MSIIEKTQENVAILEKVSINHEIILYNDDVNTFDHVIET LIRVCNHEELQAEQCAILVHYTGKCAVKTGSFDELQPLC LALLDAGLSAEIT
PS427	F, W	50	MSTKEKVKERVREKEAISFNNEIIVYNDDVNTFDHVIET LIRVCNHTPEQAEQCSLIVHYNGKCTVKTGSMDKLKPQC TQLLEAGLSAEIV
PS428	F	51	MSTKEKVKERVREKEAVGFNNEIIVYNDDVNTFDHVIDT LMRVCSHTPEQAEQCSLIVHYNGKCTVKTGPMDKLKPQC TQLLEAGLSAEIV
PS429		52	MSVQEEVLEEVKTKERVNKQNQIIVFNDDVNTFDHVIDM LIATCDHDPIQAEQCTMLIHYKGKCEVKTGDYDDLKPRC SKLLDAGISAEIQ
PS430	F, Y, W, L	53	MQPFEETYTDVLDEVVDTDVHNLVVFNDDVNTFDHVIET LIDVCKHTPEQAEQCTLLIHYKGKCSVKNGSWEELVPMR NEICRRGISAEVLK
PS431		54	MIISSVKSSPSTETLSRTELQLGGVWRVVVLNDPVNLMS YVMMIFKKIFGFNETVARRHMLEVHEKGRSVVWSGLREK AEAYVFTLQQWHLTAVLESDETH
PS432	F, W	55	MIGVEARTSSAPELAIETEIRLAGLWHVIVINDPVNLMS YVVMVLRKIFGFDDTKARKHMLEVHENGRSIVWSGEREP AEAYANTLHQWHLSAVLERDETD
PS433	F, Y, W, L	56	MMSSLKECSIQALPSLDEKTKTEEDLSVPWKVIVLNDPV NLMSYVVMVFRKVFGYNENKATKHMMEVHQLGKSVLWTG QREEAECYAYQLQRWRLQTILEKDD
PS434	F, Y, W, L	57	MSRLPWKQEAKFAATVIIDFPDATLEAPTIEKKEATEQQ IEMPWNVVVHNDPVNLMSYVTMVFQRVFGYPRERAEKHM LEVHHSGRSILWSGLRERAELYVQQLHGYLLLATIEKTV
PS435	F, Y, W, L	58	MTLSVALGPDTQESTQTGTAVSTDTLTAPDIPWNLVIWN DPVNLMSYVSYVFQSYFGYSETKANKLMMEVHKKGRSIV

			AHGSKEQVEQHAVAMHGYGLWATVEKATGGNSGGGKSGG PGKGKGKRG
Planctomycetales bacterium (PS545)	I, L, V	59	MSEPMTLPAIPQPRLKERTQRQPPYNVILLNDDDHSYEY VIAMLQVLFGYPREKGYQMAKEVDSTGRVILLTTTREHA ELKQEQIHAFGPDPLMARCQGSMTAVIEPAV
Planctomycetia bacterium (PS546)	I, L, V	60	MSDTITLPGRPEVERDERTRRQPPYNVILHNDDDHTFEY VIVMLNQLFGYPPEKGYEMAKEVHLNGRVIVLTTSKEHA ELKRDQIHAFGPDPFSSKDCKGSMSASIEPAY
Gemmataceae bacterium (PS547)	I, L, V	61	MGFPTDFRQSIEISTPLGSQQPRFSNASSEPALADPVLV INPRIQPRYHVILLNDDDHTYRYVIEMMLIVFGHPPEKG FLIAKEVDKAGRAICLTTSLEHAEFKQEQVHAYGADPYF GPKCKGSMTAVLEPAE
Gemmataceae bacterium (PS548)	I, L, V	62	MSDTITLPEEKTDVRTKRQPPYHVILLNDDDHTYQYVIY MLQTLFGHPPETGFKMAQEVDKTGRVIVDTTSLERAELK RDQIHAFGPDPYIERCKGSMSAMIEPSE
Planctomycetes bacterium (PS549)	I, L, V	63	MSESITTLPKKSRRLKEEEEQKTKRQPPYNVILLNDDDH TFEYVIFMLQKLFGHPPERGMQMAKEVHTTGRVIVMTTA LELAELKRDQIHAFGPDPLIDRCKGSMSATIEPAPI
Planctomycetes bacterium (PS550)	I, L, V	64	MPTFTEPEVVNDTRILPPYHVILLNDDDHTYEYVIHMLQ TLFGHPQERGFQLAVEVDKKGKAIVFTTSKEHAEFKRDQ IHAFGADPLSSKNCKGSMSAVIEPSF
Rubrobacter indicoceani (PS551)	I, L, V	65	MPSAAPAKPKTKRQSRTQGMPPYNVVLLDDDDHTYGYVI EMLNKVFGHPPEKGFELATEVDKNGRVIVMTTNLEVAEL KRDEVHAFGPDPLMPRSKGSMSAVVERAG
Fimbriiglobus ruber (PS552)	I, L, V	66	MSKTSTLPEVESESAQKLKYQPPYHVILLNDDDHSYVYV ITMLKELFGHPEQKGYQLADAVDKQGRAIVFTTTREHAE LKQEQIHAYGPDPTIPRCKGAMTAVIEPAE
Planctomycetes bacterium (PS553)	I, L, V	67	MPASASAVTEPPVSLPEAAAPRPKDRPKRQPRYHVILWN DDDHTYQYVVAMLRQLFGHPPEKGFTLAKQVDKDGRVVV LTTTKEHAELKRDQIHAFGADRLLARSKGSMSASIEPEA STG
Planctomycetia bacterium (PS554)	I, L, V	68	MSDSASATVEVQADPPADATARSQPTPARSTGSKPKRQP RYHVVLWNDDDHTYEYVIAMLRRLFGIEPEKGFRIAEEV DQSGRAVVLTTTREHAELKRDQIHAFGADRLLARSKGSM SASIEPEA
Planctomycetes bacterium RBG_16_64_12 (PS555)	I, L, V	69	MADSAQTGVAEPIQETLRRRKLRDDRRPKRQPPYHVILW NDNDHTYAYVVVMLMQLFGYPAEKGYQLASEVDTQGRAV VLTTTKEHAELKRDQIHAYGKDGLIEKCKGSMWATIEPA PGE
Blastopirellula marina (PS556)	I, L, V	70	MGDSNTSVAEPGEVTVVTTKPAPKKAKPKRQPKYHVVLW NDDDHTYEYVILMMHELFGHPVEKGFQIAKTVDADGRAI CLTTTKEHAELKRDQIHAYGKDELIARCRGSMSSTIEPE C
Planctomycetia bacterium (PS557)	I, L, V	71	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVL WNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEVDTQGRV IVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEA EE
Planctomycetia bacterium (PS558)	I, L, V	72	MTATTADPDRTTAEKTTKKARRSGQPKRQPRYHVILWND NDHTYQYVVAMLQQLFGHPATTGLKLATEVDRTGRAVIL TTTREHAELKRDQIHAFGADRLLARSKGAMSASIEPEAE
Planctomycetaceae bacterium (PS559)	I, L, V	73	MNQAAISPNPDIKPNPSTHKKRASQRQPRYHVILWNDND HTYHYVVTMLQKLFGHPPRTGIKMATEVDKKGKVIVLTT SREHAELKRDQIHAFGADKLIRRSKGAMAASIEPES
Planctomycetes bacterium (PS560)	I, L, V	74	MTETITTPAERTQTQAEPRSDRAWLWNVVLLDDDEHTYE YVIRMLHTLFGMPVERAFRLAEEVDARGRAVVLTTHKEH AELKRDQVHAFGKDALIASCAGSMSAVLEPAECGSDDED

Roseimaritima sp. JC651 (PS561)	I, L, V	75	MAELQTAVVEPTTRPEQDEKQSQSRPKRQPRYNVILWDD PDHSYDYVIMMLKELFGHPRQRGHQMAEEVDTTGRVICL TTTMEHAELKRDQIHAYGSDEGITRCKGSMSASIEPVPE
Rubripirellula amarantea (PS562)	I, L, V	76	MSDQQSMVAEPEVVVHTQDEKKLEKQNKRKKQPRYNVVL WDDTDHSYDYVVLMMKQLFHHPIETGFQIAKQVDKGGKA ICLTTTMEHAELKRDQIHAFGKDDLIARCTGSMSATIEP VPE
Acidobacteria bacterium (PS563)	I, L, V	77	MSSRSATAYPEVEDDTSDQLQPLYHVILLNDEDHTYDYV IEMLQKIFGFPESKAFSHAVEVDTKGTTILLTCDLEQAE RKRDLIHSYGPDWRLPRSLGSMAAVVEPAAG
Planctomycetes bacterium Poly21 (PS564)	I, L, V	78	MFEEVVSVAVAEPKTKKQSRTKPKRQPPYHVILWDDTDH TFDYVIKMMGELFRMPREKGYQLAKEVDTSGRAICMTTT LELAELKRDQIHAFGRDDASAHCKGSMSATIEPAEG
Aquisphaera sp. JC650 (PS565)	I, L, V	79	MSEFDHEHSGDTSVADPIVTTKTAPKPQKHAENETETRR QPPYNVIILNDEEHTFDYVIELLCKVFRHSLATAQELTW RIHLTGRAVVLTTHKELAELKRDQVLAYGPDPRMSVSKG PLDCFIEPAPGG
Planctomycetaceae bacterium (PS566)	I, L, V	80	MSSPSSLDDVQVSTSRAKPANETRTRKQPPYAVIVENDD HHTFLYVIEALMKVCGHAPEKGFVLAQQIHTQGKAMVWS GTLELAELKRDQLRGFGPDNYAPRPVTFPLGVTIEPLP
Planctomycetaceae bacterium (PS567)	I, L, V	81	MADYEDAGEDALEDDFDHGTVTVAPQKPEPKKQSENKRQ ANRQPRYNVLLWDSEDHTFEYVEKMLRELFGHIKKQCQI IAEQVDQEGRAVVLTTTLEHAELKRDQIHAYGKDQLEGS KGSMWSTIEAVD
Dehalococcoidia bacterium (PS568)	I, L, V	82	MTTPSLPTRETEVEERTEVEPERLYHLVLLDDDQHSYQY VIEMLASIFGYGSEKAWTLARIVDTEGRAILETASHAQC ERHQSQIHAYGADSRIPTSVGSMSAVIEEAGTPPQT
Planctomycetes bacterium (PS569)	I, L, V	83	MYSKNQIKIYCSEDDKGQTATPLLEKKPKFAPLYHVILW DDNTHTYEYVIKLLMSLFRMTFEKAYQHTLEVDKKGRTI CITTHLEKAELKQEQISNFGPDILMQNSKGPMSATIEPA N
Leptospira congkakensis (PS570)	I, L, V	84	MTGAGASQPSILEETEVRPRLSDGPWKVVLWDDDFHTYE YVIEMLMDVCQMPWEKAFQHAVEVDTRKKTIVFSGELEH AEFVHERILNYGPDPRMGSSKGSMTATLEQ
Leptospira meyeri (PS571)	I, L, V	85	MTSSGASQPSILEETERKPRLSDGPWKVVLWDDDFHTYE YVIEMLMDVCQMPWEKAFQHAVEVDTRKKTIVFFGELEH AEFVHERILNYGPDPRMGTSKGSMTATLEK
Blastopirellula marina (PS572)	I, L, V	86	MSSEELSLQTRPKRQPPFGVILHNDDLNSFDYVIDSIRK VFHYELEKCFQLTLEAHETGRSLLWTGTLEGAELKQELL LSCGPDPIMLDKGGLPLKVTLEELPQ
Leptospira fluminis (PS573)	I, L, V	87	MSQTPVIEETTVKDPVKTGGPWKVVLWDDDEHTYDYVIE MLMEVCVMTMEQAFHHAVEVDTQKKTVVYSGEFEHAEHI QELILEYGPDPRMAVSKGSMSATLEKS
Gemmata obscuriglobus (PS574)	I, L, V	88	MANATPTPDVVPEEETETRTRRQPPYAVVLHNDDTNTMD FVVTVLRKVFGYTVEKCVELMLEAHTQGKVAVWIGALEV AELKADQIKSFGPDPHVTKNGHPLGVTVEPAA
Leptospira kmetyi (PS575)	I, L, V	89	MASTQTPDLNEITEESTKSTGGPWRVVLWDDNEHTYEYV IEMLMEICTMTVEKAFLHAVQVDQEKRTVVFSGEFEHAE HVQERILTYGADPRMSNSKGSMSATLEK
Leptospira interrogans (PS576)	I, L, V	90	MASTQTPDLNEITEESTKSTGGPWRVVLWDDNEHTYEYV IEMLVEICMMTVEKAFLHAVQVDKEKRTVDFSGELEHAE HVQERILNYGADPRMSNSKGSMSATLER
Tuwongella immobilis (PS577)	I, L, V	91	MSASSSQPGTTTKPDLDIQPRLLPPFHVILENDEFHSME FVIDTLRKVLGVSIERAYQLMMTAHESGQAIIWTGPKEV AELKYEQVIGFHEKRSDGRDLGPLGCRIEPAV

Planctomycetes bacterium (PS578)	I, L, V	92	MSGTVVESKPRNSTQLAPRWKVIVHDDPVTTFDFVLGVL RRVFAKPPGEARRITREAHDTGSALVDVLALEQAEFRRD QAHSLARAEGFPLTLTLEPAD
Agrobacterium tumifaciens ClpS1 (atClpS1)	F, W, Y, L	93	MIAEPICMQGEGDGEDGGTNRGTSVITRVKPKTKRPNLY RVLLLNDDYTPMEFVIHILERFFQKDREAATRIMLHVHQ HGVGECGVFTYEVAETKVSQVMDFARQHQHPLQCVMEKK
Agrobacterium tumifaciens ClpS2 (atClpS2)	F, W, Y	94	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPREFV TVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAE TKAKEATDLGKEAGFPLMFTTEPEE
atClpS2 thermostable variant	F, W, Y	95	MSDSPVDLKPKPKVKPKLERPKLYKVILLNDDYTPMEFV VEVLKRVFNMSEEQARRVMMTAHKKGKAVVGVCPRDIAE TKAKQATDLAREAGFPLMFTTEPEE
PS489	M	96	MSDSPVDLKPKPKVKPKLERLKLYKVILLNDDYTTAFFV VKVLKRVFNMSEEQARRVMMTAHKKGKAVVGVCPRDIAE TKAKQATDLAREAGFPLMFTTEPEE
PS490	M	97	MSDSPVDLKPKPKVKPKLERLKLYKVILLNDDYTTMRFV VLVLKRVFNMSEEQARRVMMTAHKKGKAVVGVCPRDIAE TKAKQATDLAREAGFPLMFTTEPEE
PS218	F, W, Y, L	98	MIAEPICMQGEGDGEDGGTNRGTSVITRVKPKTKRPNLY RVLLLNDDYTPFQFVIHILERFFQKDREAAWRITLHVHQ HGVGECGVFTYEVAETKVSQVMDFARQHQHPLQCVMEKK
atClpS2-V1	F, W, Y	99	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFV TVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAE TKAKEATDLGKEAGFPLMFTTEPEE
atClpS2 C72S	F, W, Y	100	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPREFV TVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVSERDIAE TKAKEATDLGKEAGFPLMFTTEPEE
atClpS2-V1 + C72S	F, W, Y	101	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFV TVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVSERDIAE TKAKEATDLGKEAGFPLMFTTEPEE
atClpS2 thermostable variant + C72S	F, W, Y	102	MSDSPVDLKPKPKVKPKLERPKLYKVILLNDDYTPMEFV VEVLKRVFNMSEEQARRVMMTAHKKGKAVVGVSPRDIAE TKAKQATDLAREAGFPLMFTTEPEE
atClpS1 C7S	F, W, Y, L	103	MIAEPISMQGEGDGEDGGTNRGTSVITRVKPKTKRPNLY RVLLLNDDYTPMEFVIHILERFFQKDREAATRIMLHVHQ HGVGECGVFTYEVAETKVSQVMDFARQHQHPLQCVMEKK
atClpS1 C7S, C84S, C112S	F, W, Y, L	104	MIAEPISMQGEGDGEDGGTNRGTSVITRVKPKTKRPNLY RVLLLNDDYTPMEFVIHILERFFQKDREAATRIMLHVHQ HGVGESGVFTYEVAETKVSQVMDFARQHQHPLQSVMEKK
Synechococcus elongatus ClpS1	F, W, Y	105	MAVETIQKPETTTKRKIAPRYRVLLHNDDFNPMEYVVMV LMQTVPSLTQPQAVDIMMEAHTNGTGLVITCDIEPAEFY CEQLKSHGLSSSIEPDD
Synechococcus elongatus ClpS2	F, W, Y, L	106	MSPQPDESVLSILGVPRPCVKKRSRNDAFVLTVLTCSLQ AIAAPATAPGTTTTRVRQPYPHFRVIVLDDDVNTFQHVA ECLLKYIPGMTGDRAWDLTNQVHYEGAATVWSGPQEQAE LYHEQLRREGLTMAPLEAA
Thermosynechococcus elongatus ClpS	F, W, Y, L	107	MPQERQQVTRKHYPNYKVIVLNDDFNTFQHVAACLMKYI PNMTSDRAWELTNQVHYEGQAIVWVGPQEQAELYHEQLL RAGLTMAPLEPE

Escherichia coli ClpS	F, W, Y, L	108	MGKTNDWLDFDQLAEEKVRDALKPPSMYKVILVNDDYTP MEFVIDVLQKFFSYDVERATQLMLAVHYQGKAICGVFTA EVAETKVAMVNKYARENEHPLLCTLEKA
Escherichia coli ClpS M40A	F, W, Y, L	109	MGKTNDWLDFDQLAEEKVRDALKPPSMYKVILVNDDYTP AEFVIDVLQKFFSYDVERATQLMLAVHYQGKAICGVFTA EVAETKVAMVNKYARENEHPLLCTLEKA
Plasmodium falciparum ClpS	F, W, Y, L	110	MFKDLKPFFLCIILLLLLIYKCTHSYNIKNKNCPLNFMN SCVRINNVNKNTNISFPKELQKRPSLVYSQKNFNLEKIK KLRNVIKEIKKDNIKEADEHEKKEREKETSAWKVILYND DIHNFTYVTDVIVKVVGQISKAKAHTITVEAHSTGQALI LSTWKSKAEKYCQELQQNGLTVSIIHESQLKDKQKK

*Binding preferences are inferred from published scientific literature and/or further demonstrated by the inventors in single-molecule and/or ensemble experiments, as described herein.

** Binding to phosphotyrosine may occur at a peptide terminus or at an internal position.

Table 2. Non-limiting examples of amino acid recognition proteins.

Name	Binding Pref.*	SEQ ID NO:	Sequence
Escherichia coli leucyl/phenylalanyl-tRNA-protein transferase	K, R	111	MRLVQLSRHSIAFPSPEGALREPNGLLALGGDLSPARLL MAYQRGIFPWFSPGDPILWWSPDPRAVLWPESLHISRSM KRFHKRSPYRVTMNYAFGQVIEGCASDREEGTWITRGVV EAYHRLHELGHAHSIEVWREDELVGGMYGVAQGTLFCGE SMFSRMENASKTALLVFCEEFIGHGGKLIDCQVLNDHTA SLGACEIPRRDYLNYLNQMRLGRLPNNFWVPRCLFSPQE LE
Vibrio vulnificus Aspartate/glutamate leucyltransferase Bpt	D, E	112	MSSDIHQIKIGLTDNHPCSYLPERKERVAVALEADMHTA DNYEVLLANGFRRSGNTIYKPHCDSCHSCQPIRISVPDI ELSRSQKRLLAKARSLSWSMKRNMDENWFDLYSRYIVAR HRNGTMYPPKKDDFAHFSRNQWLTTQFLHIYEGQRLIAV AVTDIMDHCASAFYTFFEPEHELSLGTLAVLFQLEFCQE EKKQWLYLGYQIDECPAMNYKVRFHRHQKLVNQRWQ
H. sapiens GID4	P	113	MSGSKFRGHQKSKGNSYDVEVVLQHVDTGNSYLCGYLKI KGLTEEYPTLTTFFEGEIISKKHPFLTRKWDADEDVDRK HWGKFLAFYQYAKSFNSDDFDYEELKNGDYVFMRWKEQF LVPDHTIKDISGASFAGFYYICFQKSAASIEGYYYHRSS EWYQSLNLTHV
Saccharomyces cerevisiae GID4	P	114	MINNPKVDSVAEKPKAVTSKQSEQAASPEPTPAPPVSRN QYPITFNLTSTAPFHLHDRHRYLQEQDLYKCASRDSLSS LQQLAHTPNGSTRKKYIVEDQSPYSSENPVIVTSSYNHT VCTNYLRPRMQFTGYQISGYKRYQVTVNLKTVDLPKKDC TSLSPHLSGFLSIRGLTNQHPEISTYFEAYAVNHKELGF LSSSWKDEPVLNEFKATDQTDLEHWINFPSFRQLFLMSQ KNGLNSTDDNGTTNAAKKLPPQQLPTTPSADAGNISRIF SQEKQFDNYLNERFIFMKWKEKFLVPDALLMEGVDGASY DGFYYIVHDQVTGNIQGFYYHQDAEKFQQLELVPSLKNK VESSDCSFEFA
Single-chain antibody variable	phospho-Y	115	MMEVQLQQSGPELVKPGASVMISCRTSAYTFTENTVHWV KQSHGESLEWIGGINPYYGGSIFSPKFKGKATLTVDKSS

fragment (scFv) against phosphotyrosine**			STAYMELRSLTSEDSAVYYCARRAGAYYFDYWGQGTTLT VSSGGGSGGGSGGGSENVLTQSPAIMSASPGEKVTMTCR ASSSVSSSYLHWYRQKSGASPKLWIYSTSNLASGVPARF SGSGSGTSYSLTISSVEAEDAATYYCQQYSGYRTFGGGT KLEIKR
H. sapiens Fyn SH2 domain**	phospho-Y	116	MGAMDSIQAEEWYFGKLGRKDAERQLLSFGNPRGTFLIR ESETTKGAYSLSIRDWDDMKGDHVKHYKIRKLDNGGYYI TTRAQFETLQQLVQHYSERAAGLSSRLVVPSHK
H. sapiens Fyn SH2 domain triple mutant superbinder**	phospho-Y	117	MGAMDSIQAEEWYFGKLGRKDAERQLLSFGNPRGTFLIR ESETVKGAYALSIRDWDDMKGDHVKHYLIRKLDNGGYYI TTRAQFETLQQLVQHYSERAAGLSSRLVVPSHK
H. sapiens Src tyrosine kinase SH2 domain**	phospho-Y	118	MGAMDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVR ESETTKGAYSLSVSDFDNAKGLNVKHYKIRKLDSGGFYI TSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTSK
H. sapiens Src tyrosine kinase SH2 domain triple mutant**	phospho-Y	119	MGAMDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVR ESEVTKGAYALSVSDFDNAKGLNVKHYLIRKLDSGGFYI TSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTSK
H. sapiens p62 fragment 1-310	K, R, H, W, F, Y	120	MASLTVKAYLLGKEDAAREIRRFSFCCSPEPEAEAEAAA GPGPCERLLSRVAALFPALRPGGFQAHYRDEDGDLVAFS SDEELTMAMSYVKDDIFRIYIKEKKECRRDHRPPCAQEA PRNMVHPNVICDGCNGPVVGTRYKCSVCPDYDLCSVCEG KGLHRGHTKLAFPSPFGHLSEGFSHSRWLRKVKHGHFGW PGWEMGPPGNWSPRPPRAGEARPGPTAESASGPSEDPSV NFLKNVGESVAAALSPLGIEVDIDVEHGGKRSRLTPVSP ESSSTEEKSSSQPSSCCSDPSKPGGNVEGATQSLAEQ
H. sapiens p62 fragment 1-180	K, R, H, W, F, Y	121	MASLTVKAYLLGKEDAAREIRRFSFCCSPEPEAEAEAAA GPGPCERLLSRVAALFPALRPGGFQAHYRDEDGDLVAFS SDEELTMAMSYVKDDIFRIYIKEKKECRRDHRPPCAQEA PRNMVHPNVICDGCNGPVVGTRYKCSVCPDYDLCSVCEG KGLHRGHTKLAFPSPFGHLSEGFSHSRWLRKVKHGHFGW PGWEMGPPGNWSPRPPRAGEARPGPTAESASGPSEDPSV NFLKNVGESVAAALSPLGIEVDIDVEHGGKRSRLTPVSP ESSSTEEKSSSQPSSCCSDPSKPGGNVEGATQSLAEQ
H. sapiens p62 fragment 126-180	K, R, H, W, F, Y	122	MASLTVKAYLLGKEDAAREIRRFSFCCSPEPEAEAEAAA GPGPCERLLSRVAALFPALRPGGFQAHYRDEDGDLVAFS SDEELTMAMSYVKDDIFRIYIKEKKECRRDHRPPCAQEA PRNMVHPNVICDGCNGPVVGTRYKCSVCPDYDLCSVCEG KGLHRGHTKLAFPSPFGHLSEGFSHSRWLRKVKHGHFGW PGWEMGPPGNWSPRPPRAGEARPGPTAESASGPSEDPSV NFLKNVGESVAAALSPLGIEVDIDVEHGGKRSRLTPVSP ESSSTEEKSSSQPSSCCSDPSKPGGNVEGATQSLAEQ
H. sapiens p62 protein	K, R, H, W, F, Y	123	MASLTVKAYLLGKEDAAREIRRFSFCCSPEPEAEAEAAA GPGPCERLLSRVAALFPALRPGGFQAHYRDEDGDLVAFS SDEELTMAMSYVKDDIFRIYIKEKKECRRDHRPPCAQEA PRNMVHPNVICDGCNGPVVGTRYKCSVCPDYDLCSVCEG KGLHRGHTKLAFPSPFGHLSEGFSHSRWLRKVKHGHFGW PGWEMGPPGNWSPRPPRAGEARPGPTAESASGPSEDPSV NFLKNVGESVAAALSPLGIEVDIDVEHGGKRSRLTPVSP

			ESSSTEEKSSSQPSSCCSDPSKPGGNVEGATQSLAEQMR KIALESEGRPEEQMESDNCSGGDDDWTHLSSKEVDPSTG ELQSLQMPESEGPSSLDPSQEGPTGLKEAALYPHLPPEA DPRLIESLSQMLSMGFSDEGGWLTRLLQTKNYDIGAALD TIQYSKHPPPL
Rattus norvegicus p62 protein	K, R, H, W, F, Y	124	MASLTVKAYLLGKEEAAREIRRFSFCFSPEPEAEAAAGP GPCERLLSRVAVLFPALRPGGFQAHYRDEDGDLVAFSSD EELTMAMSYVKDDIFRIYIKEKKECRREHRPPCAQEARS MVHPNVICDGCNGPVVGTRYKCSVCPDYDLCSVCEGKGL HREHSKLIFPNPFGHLSDSFSHSRWLRKLKHGHFGWPGW EMGPPGNWSPRPPRAGDGRPCPTAESASAPSEDPNVNFL KNVGESVAAALSPLGIEVDIDVEHGGKRSRLTPTSAESS STGTEDKSGTQPSSCSSEVSKPDGAGEGPAQSLTEQMKK IALESVGQPEELMESDNCSGGDDDWTHLSSKEVDPSTGE LQSLQMPESEGPSSLDPSQEGPTGLKEAALYPHLPPEAD PRLIESLSQMLSMGFSDEGGWLTRLLQTKNYDIGAALDT IQYSKHPPPL
Saccharomyces cerevisiae GID10	P, M, V	125	MTSLNIMGRKFILERAKRNDNIEEIYTSAYVSLPSSTDT RLPHFKAKEEDCDVYEEGTNLVGKNAKYTYRSLGRHLDF LRPGLRFGGSQSSKYTYYTVEVKIDTVNLPLYKDSRSLD PHVTGTFTIKNLTPVLDKVVTLFEGYVINYNQFPLCSLH WPAEETLDPYMAQRESDCSHWKRFGHFGSDNWSLTERNF GQYNHESAEFMNQRYIYLKWKERFLLDDEEQENQMLDDN HHLEGASFEGFYYVCLDQLTGSVEGYYYHPACELFQKLE LVPTNCDALNTYSSGFEIA
Leishmania major N-meristoyltransferase	G	126	MSRNPSNSDAAHAFWSTQPVPQTEDETEKIVFAGPMDEP KTVADIPEEPYPIASTFEWWTPNMEAADDIHAIYELLRD NYVEDDDSMFRFNYSEEFLQWALCPPNYIPDWHVAVRRK ADKKLLAFIAGVPVTLRMGTPKYMKVKAQEKGEGEEAAK YDEPRHICEINFLCVHKQLREKRLAPILIKEATRRVNRT NVWQAVYTAGVLLPTPYASGQYFHRSLNPEKLVEIRFSG IPAQYQKFQNPMAMLKRNYQLPSAPKNSGLREMKPSDVP QVRRILMNYLDSFDVGPVFSDAEISHYLLPRDGVVFTYV VENDKKVTDFFSFYRIPSTVIGNSNYNLLNAAYVHYYAA TSIPLHQLILDLLIVAHSRGFDVCNMVEILDNRSFVEQL KFGAGDGHLRYYFYNWAYPKIKPSQVALVML
H. sapiens N-meristoyltransferase NMT1	G	127	MADESETAVKPPAPPLPQMMEGNGNGHEHCSDCENEEDN SYNRGGLSPANDTGAKKKKKKQKKKKEKGSETDSAQDQP VKMNSLPAERIQEIQKAIELFSVGQGPAKTMEEASKRSY QFWDTQPVPKLGEVVNTHGPVEPDKDNIRQEPYTLPQGF TWDALDLGDRGVLKELYTLLNENYVEDDDNMFRFDYSPE FLLWALRPPGWLPQWHCGVRVVSSRKLVGFISAIPANIH IYDTEKKMVEINFLCVHKKLRSKRVAPVLIREITRRVHL EGIFQAVYTAGVVLPKPVGTCRYWHRSLNPRKLIEVKFS HLSRNMTMQRTMKLYRLPETPKTAGLRPMETKDIPVVHQ LLTRYLKQFHLTPVMSQEEVEHWFYPQENIIDTFVVENA NGEVTDFLSFYTLPSTIMNHPTHKSLKAAYSFYNVHTQT PLLDLMSDALVLAKMKGFDVFNALDLMENKTFLEKLKFG IGDGNLQYYLYNWKCPSMGAEKVGLVLQ

Drosophila melanogaster BIR2	A	128	MGDVQPETCRPSAASGNYFPQYPEYAIETARLRTFEAWP RNLKQKPHQLAEAGFFYTGVGDRVRCFSCGGGLMDWNDN DEPWEQHALWLSQCRFVKLMKGQLYIDTVAAKPVLAEEK EESTSIGGDT
Amanita thiersii Skay4041 UBR-box domain (PS501)	K, R, H	129	MICGQIIGKGESCFRCRDCGLDESCVMCSQCFHATDHIN HNVSFFVSQQPGGCCDCGDEEAWKKPMNCPYHPP
Helobdella robusta UBR-box domain (PS502)	K, R, H	130	MVCLKVFKLGEPTYSCRSVTCGMDPTCVLCVDCFQNSSH KLHKYKMSTSGGGGYCDCGDLEAWKADPLCDLHKL
Hydra vulgaris UBR-box domain (PS503)	K, R, H	131	MFCGRLFKVGDPTYTCKDCAADPTCVFCHDCFHQSVHTK HKYKLFASQGRGGYCDCGDKEAWTNDPACNKHKE
Galleria mellonella UBR-box domain (PS504)	K, R, H	132	MLCGKVFKQGEPAYSCRECGMDNTCVLCVECFKVSPHRH HKYKMGQSGGGGCCDCGDTEAWKRDPFCERHAK
Brachionus plicatilis UBR-box domain (PS505)	K, R, H	133	MVCGRVFKSGEPSYFCRECGTDPTCVLCSICFRHSKHRY HKYVMMTSGGGGYCDCGDPEAWKSDPCCELHMP
Capitella teleta UBR-box domain (PS506)	K, R, H	134	MLCGKVFKMGELTYSCRDCGTDPTCVLCMDCFQHSAHKK HRYKMAASGGGGYCDCGDREAWKAEPFCDVHKR
Sparassis crispa UBR-box domain (PS507)	K, R, H	135	MPCGHIFKKGESCFRCKDCALDDSCVLCSKCFEATDHAN HNVSFFIAQQSGGCCDCGDIEAWLVPIDCPFHPV
Anabarilius graham UBR-box domain (PS508)	K, R, H	136	MLCGRVFKEGETVYSCRDCAIDPTCVLCIECFQKSVHKS HRYKMHASAGGGFCDCGDLEAWKTGPCCSQHDP
Lottia gigantean UBR-box domain (PS509)	K, R, H	137	MICGHGFKTGEPTYSCRDCATDPTCVLCISCFQKSPHRE HRYKMSASGGGGYCDCGDPEAWKIEPFCEQHKP
Camponotus floridanus UBR-box domain (PS510)	K, R, H	138	MICGRMFKMGEPTYSCRQCGMDSTCVLCVDCFKQSAHRN HKYKMGTSSGGGCCDCGDTEAWKNEPFCKIHLA
Habropoda laboriosa UBR-box domain (PS511)	K, R, H	139	MICGKVFKMGEATYSCKECGVDPTCVLCADCFKQSAHRH HKYRMGTSSGGGFCDCGDIEAWKKEPFCNTHLA
Mastacembelus armatus UBR-box domain (PS512)	K, R, H	140	MLCGRVFKEGETVYSCRDCAIDPTCVLCMDCFQDSVHKS HRYKMHASAGGGFCDCGDVEAWKIGPYCSKHDP
Pyrenophora seminiperda CCB06 UBR-box domain (PS513)	K, R, H	141	MPCGHIFKNGEATYRCKTCTADDTCVLCARCFDASDHEG HQVFVSVSPGNSGCCDCGDDEAWVRPVHCNIHSA
Tribolium castaneum UBR-box domain (PS514)	K, R, H	142	MVCGRVFKLGEPTYSCRECGMDNTCVLCVNCFKNSEHRF HKYKMGTSQGGGCCDCGDVEAWKKAPFCDVHIA

Wasmannia auropunctata UBR-box domain (PS515)	K, R, H	143	MICGKMFKIGEPTYSCRECGMDSTCVLCVDCFKQSAHRN HKYKMGTSSGGGCCDCGDTEAWKKEPFCKTHVV
Crassostrea gigas UBR-box domain (PS516)	K, R, H	144	MLCGKVFKTGEPTYSCRDCANDPTCVLCIDCFQNGAHKN HRYKMNTSGGGGYCDCGDQEAWTSHPFCNLHSP
Harpegnathos saltator UBR-box domain (PS517)	K, R, H	145	MMCGRVFKMGEPTYSCRECGVDSTCVLCVGCFQQSAHRD HKYKMGTSGGGGCCDCGDTEAWKRDPFCEIHMV
Nilaparvata lugens UBR-box domain (PS518)	K, R, H	146	MVCGRVFKMGEPSYHCRECGMDATCVLCVDCFKKSSHRN HKYKMGTSIGGGCCDCGDVEAWKTEPYCEVHIA
Manduca sexta UBR-box domain (PS519)	K, R, H	147	MLCGRVFKQGEPAYSCRECGMDNTCVLCVECFKVSAHRH HKYKMGQSGGGGCCDCGDTEAWKRDPFCELHAA
Monopterus albus UBR-box domain (PS520)	K, R, H	148	MLCGRVFKEGETVYSCRDCAIDPTCVLCMDCFQDSVHKS HRYKMHASSGGGFCDCGDVEAWKIGPCCSKHDP
Lingula anatine UBR-box domain (PS521)	K, R, H	149	MLCGRVFRSGEPTYSCRDCAVDPTCVLCIDCFNNGAHRK HKYRMSTSSGGGYCDCGDKEAWKTDPLCEIHRK
Vombatus ursinus UBR-box domain (PS522)	K, R, H	150	MLCGKVFKSGETTYSCRDCAIDPTCVLCMNCFQSSVHKN HRYKMHTSTGGGFCDCGDTEAWKTGPFCTIHEP
Saccharomycetaceae sp. Ashbya aceri UBR-box domain (PS523)	K, R, H	151	MAKSHRHTGRNCGRAFQPGEPLYRCQECAYDDTCVLCIS CFNPDDHVNHHVSTHICNELHDGICDCGDAEAWNVPLHC KAEED
Drosophila ficusphila UBR-box domain (PS524)	K, R, H	152	MVCGKVFKNGEPTYSCRECGVDPTCVLCVNCFKRSAHRF HKYKMSTSGGGGCCDCGDDEAWKKDHYCQLHLA
Mus musculus UBR-box domain (PS525)	K, R, H	153	MLCGKVFKSGETTYSCRDCAIDPTCVLCMDCFQSSVHKN HRYKMHTSTGGGFCDCGDTEAWKTGPFCVDHEP
Maylandia zebra UBR-box domain (PS526)	K, R, H	154	MLCGRVFKEGETVYSCRDCAIDPTCVLCMDCFQDSVHKS HRYKMHASSGGGFCDCGDVEAWKIGPYCSKHDP
Mizuhopecten yessoensis UBR-box domain (PS527)	K, R, H	155	MLCGKVFKYGEPTYSCRDCANDPTCVLCIDCFQKSAHKK HRYKMSTSGGGGYCDCGDSEAWKTAPFCSNHKA
Kluyveromyces lactis UBR-box domain (PS528)	K, R, H	156	MHSKFNHAGRICGAKFRVGEPIYRCKECSFDDTCVLCVN CFNPKDHVGHHVYTSICTEFNNGICDCGDKEAWNHELNC KGAED
Chelonia mydas UBR-box domain (PS529)	K, R, H	157	MLCGKVFKGGETTYSCRDCAIDPTCVLCMDCFQNSIHKN HRYKMHTSTGGGFCDCGDTEAWKTGPLCANHEP
Acropora millepora UBR-box domain (PS530)	K, R, H	158	MLCGKVFKVGEPTYSCRDCGYDNTCVLCINCFQKSIHKN HHYKMNTSGGGGVCDCGDVEAWKEGEACEIHQQ
Musca domestica UBR-box domain	K, R, H	159	MVCGKVFKIGEPTYSCRECGMDQTCVLCVNCFKQSAHRY HKYKMSTSGGGGCCDCGDEEAWKKDHYCEEHLR

(PS531)			
Schizosaccharomyces cryophilus OY26 UBR-box domain (PS532)	K, R, H	160	MSCGRIFKKGEVFYRCKTCSVDSNSALCVKCFRATDHHG HETSFTISAGSGGCCDCGNSAAWIRDMPCKIHNR
Contarinia nasturtii UBR-box domain (PS533)	K, R, H	161	MVCGRVFKMNEPFYSCRECGMDPTCVLCVNCFKQSAHRH HKYKMGTSAGGGCCDCGDNEAWKQDHYCDEHTK
Schizosaccharomyces pombe UBR-box domain (PS534)	K, R, H	162	MKCGHIFRKGEVFYRCKTCSVDSNSALCVKCFRATSHKD HETSFTVSAGSGGCCDCGNAAAWIGDVSCKIHSH
Mus musculus UBR-box domain (PS535)	K, R, H	163	MLCGRVFKVGEPTYSCRDCAVDPTCVLCMECFLGSIHRD HRYRMTTSGGGGFCDCGDTEAWKEGPYCQKHKL
Aphis gossypii UBR-box domain (PS536)	K, R, H	164	MVCGRVFKMGEPTYNCRECGMDSTCVLCVDCFKRSPHKN HKYKMGTSYGGGCCDCGDVEAWKHDPYCQTHKL
Aedes aegypti UBR-box domain (PS537)	K, R, H	165	MVCGRVFKIGEPTYSCRECSMDPTCVLCSSCFKKSSHRL HKYKMSTSGGGGCCDCGDHEAWKRDPSCEEHAV
Saccharomyces cerevisiae UBR-box domain (PS538)	K, R, H	166	MGDVHKHTGRNCGRKFKIGEPLYRCHECGCDDTCVLCIH CFNPKDHVNHHVCTDICTEFTSGICDCGDEEAWNSPLHC KAEEQ
Saccharomyces cerevisiae UBR1 D3S variant (PS25)	K, R, H	167	MGSVHKHTGRNCGRKFKIGEPLYRCHECGCDDTCVLCIH CFNPKDHVNHHVCTDICTEFTSGICDCGDEEAWNSPLHC KAEEQ
Kazachstania africana CBS 2517 UBR-box domain (PS539)	K, R, H	168	MQTSFTHKGRNCGRKFKVGEPLYRCHECGFDDTCVLCIH CFNPADHENHHIYTDICNDFTSGICDCGDTEAWNGDLHC KAEEI
Clathrospora elynae UBR-box domain (PS540)	K, R, H	169	MPCGHIFKNGEATYRCKTCTADDTCVLCARCFDASDHEG HQVFVSVSPGNSGCCDCGDDEAWVRPVHCNMHSA
Aspergillus neoniger CBS 115656 UBR-box domain (PS541)	K, R, H	170	MRCGHIFRAGEATYRCITCAADDTCVLCSRCFDASDHTG HQYQISLSSGNCGCCDCGDEEAWRLPLFCAIHTD
Trichuris suis UBR-box domain (PS542)	K, R, H	171	MRCNHVFANGEATYSCRGCAADPTCVLCASCFELSAHKE HKYMITTSSGTGYCDCGDPEAWKADPFCQQHQP
Trichinella spiralis UBR-box domain (PS543)	K, R, H	172	MKCNRQLICGEPTYCCLDCACDQTCIFCHACFQSSEHKN HRYSMSTSEGSGTCDCGDKEAWKSNYYCLNHKP
Homo sapiens UBR1 (PS544)	K, R, H	173	MGPLGSLCGRVFKSGETTYSCRDCAIDPTCVLCMDCFQD SVHKNHRYKMHTSTGGGFCDCGDTEAWKTGPFCVNHEP
Homo sapiens UBR2	K, R, H	174	MGPLGSLCGRVFKVGEPTYSCRDCAVDPTCVLCMECFLG SIHRDHRYRMTTSGGGGFCDCGDTEAWKEGPYCQKHE

Kluyveromyces marxianus UBR2 (PS615)	K, R, H	175	MVNEHRGSQCSKQCHGTETVYYCFDCTKNPLYEICEECF DETQHMGHRYTSRVVTRPEGKVCHCGDISGYNNPEKAFQ CKI
Kluyveromyces lactis UBR2 (PS616)	K, R, H	176	MHNDHRGSQCSKQCHGTETVYYCFDCTKNPLYEICEDCF DESQHIGHRYTSRVVTRPEGKVCHCGDISSYNDPKKAFQ CRI
Eremothecium sinecaudum UBR2 (PS617)	K, R, H	177	MPKEHRGTSCNKHCQPTETVYYCFDCTKNPLYEICEECF DADKHLGHRWTSKVVSRPEGKICHCGDPSGLTDPENGYE CKN
Zygosaccharomyces bailii UBR2 (PS618)	K, R, H	178	MNASHKGAMCSKQCYPTETVFYCFTCTTNPLYEICESCF DEEKHRGHLYTAKVVVRPEGRVCHCGDPFVFKEPRFAFL CKN
Vanderwaltozyma polyspora UBR2 (PS619)	K, R, H	179	MENLHIGSCCNRQCYPTQTVYYCLTCTINPLYEICELCF DEDKHVGHTYISKSVIRPEGKVCHCGNPNVFKKPEFAFN CKN
Saccharomyces cerevisiae UBR2 (PS620)	K, R, H	180	MGNMHIGTACTRLCFPSETIYYCFTCSTNPLYEICELCF DKEKHVNHSYVAKVVMRPEGRICHCGDPFAFNDPSDAFK CKN
Kluyveromyces marxianus UBR1 (PS621)	K, R, H	181	MHSKFSHAGRICGAKFKVGEPIYRCKECSFDDTCVLCVN CFNPKDHTGHHVYTTICTEFNNGICDCGDKEAWNHTLFC KAEEG
Kluyveromyces dobzhanskii UBR1 (PS622)	K, R, H	182	MHSRFNHAGRICASKFKVGEPIYRCKECSFDDTCVICVN CFNPKDHVGHHVYTSICSEFNNGICDCGDTEAWNHDMHC KADEN
Kazachstania naganishii UBR1 (PS623)	K, R, H	183	MSKQFRHKGRNCGRKFRLGEPLYRCQECGYDDTCVLCIN CFNPKDHEGHHIYTDICNDFTSGICDCGDEEAWLSPLHC KAEED
Eremothecium sinecaudum UBR1 (PS624)	K, R, H	184	MPKNHNHKGRNCGRSFQPGEPLYRCQECAYDDTCVLCIR CFNPLDHVNHHVSTHICSEFNDGICDCGDVEAWNVELNC KAEED
Saccharomyces eubayanus UBR1 (PS625)	K, R, H	185	MGDVHKHTGRNCGRKFKIGEPLYRCHECGCDDTCVLCIH CFNPKDHINHHVCTDICSEFTSGICDCGDEEAWNSSLHC KAEEQ
Zygosaccharomyces parabailii UBR1 (PS626)	K, R, H	186	MYHVYKHSGRNCGRKFKVGEPIYRCHECGYDETCVLCIH CFNPKDHDSHHVYIDICSEFSTGICDCGDTEAFVNPLHC KAEED
Zygosaccharomyces mellis UBR1 (PS627)	K, R, H	187	MPKYHQHSGRYCGRKFKVGEPIYRCHECGFDETCVICIH CFNAKDHETHHVSVSICSEYSTGICDCGDTEAFVNPLHC RAEEV
Candida albicans UBR1 (PS628)	K, R, H	188	MSHRAYHKNSPCGRIFRKGEPIHRCLTCGFDDTCALCSH CFQPEYHEGHKVHIGICQRENGGVCDCGDPEAWTQELFC PYAVD
Pichia pastoris UBR1 (PS629)	K, R, H	189	MCPNYKHHGRPCARQFKQGEPIYRCYECGFDETCVMCMH CFNREQHRDHEVSISIASSSNDGICDCGDPQAWNIELHC QSELD

*Binding preferences are inferred from published scientific literature and/or further demonstrated by the inventors in single-molecule and/or ensemble experiments, as described herein.

** Binding to phosphotyrosine may occur at a peptide terminus or at an internal position.

[0126] Accordingly, in some embodiments, the application provides an amino acid recognition molecule having an amino acid sequence selected from Table 1 or Table 2 (or having an amino acid sequence that has at least 50%, at least 60%, at least 70%, at least 80%, 80-90%, 90-95%, 95-99%, or higher, amino acid sequence identity to an amino acid sequence selected from Table 1 or Table 2). In some embodiments, an amino acid recognition molecule has 25-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-95%, or 95-99%, or higher, amino acid sequence identity to an amino acid recognition molecule listed in Table 1 or Table 2. In some embodiments, an amino acid recognition molecule is a modified amino acid recognition molecule and includes one or more amino acid deletions, additions, or mutations relative to a sequence set forth in Table 1 or Table 2. In some embodiments, a modified amino acid recognition molecule includes a deletion, addition, or mutation of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more amino acids (which may or may not be consecutive amino acids) relative to a sequence set forth in Table 1 or Table 2.

[0127] In some embodiments, an amino acid recognition molecule comprises a single polypeptide having tandem copies of two or more amino acid binding proteins (e.g., two or more binders). As used herein, in some embodiments, a tandem arrangement or orientation of elements in a molecule refers to an end-to-end joining of each element to the next element in a linear fashion such that the elements are fused in series. For example, in some embodiments, a polypeptide having tandem copies of two binders refers to a fusion polypeptide in which the C-terminus of one binder is fused to the N-terminus of the other binder. Similarly, a polypeptide having tandem copies of two or more binders refers to a fusion polypeptide in which the C-terminus of a first binder is fused to the N-terminus of a second binder, the C-terminus of the second binder is fused to the N-terminus of a third binder, and so forth. Such fusion polypeptides can comprise multiple copies of the same binder or multiple copies of different binders. In some embodiments, a fusion polypeptide of the application has at least two and up to ten binders (e.g., at least 2 binders and up to eight, six, five, four, or three binders). In some embodiments, a fusion polypeptide of the application has five or fewer binders (e.g., two, three, four, or five binders). Accordingly, in some embodiments, labeled recognition molecule **200** comprises a fusion polypeptide of the application.

[0128] In some embodiments, a fusion polypeptide is provided by expression of a single coding sequence containing segments encoding monomeric binder subunits separated by segments encoding flexible linkers, where expression of the single coding sequence produces a single full-length polypeptide having two or more independent binding sites. In some embodiments, one or more of the monomeric subunits (e.g., binders) are ClpS proteins. In some embodiments, ClpS

subunits may be identical or non-identical. Where non-identical, ClpS subunits may be distinct variants of the same parent ClpS protein, or they may be derived from different parent ClpS proteins. In some embodiments, a fusion polypeptide comprises one or more ClpS monomers and one or more non-ClpS monomers. In some embodiments, the monomeric subunits comprise non-ClpS monomers. In some embodiments, the monomeric subunits comprise one or more degradation pathway proteins. For example, in some embodiments, the monomeric subunits comprise one or more of a Gid protein, a UBR-box protein or UBR-box domain-containing protein fragment thereof, a p62 protein or ZZ domain-containing fragment thereof, and a ClpS protein (e.g., ClpS1, ClpS2).

[0129] In some embodiments, at least one binder of a fusion polypeptide has an amino acid sequence selected from Table 1 or Table 2 (or having an amino acid sequence that has at least 50%, at least 60%, at least 70%, at least 80%, 80-90%, 90-95%, 95-99%, or higher, amino acid sequence identity to an amino acid sequence selected from Table 1 or Table 2). In some embodiments, each binder of a fusion polypeptide has an amino acid sequence that is at least 80% (e.g., 80-90%, 90-95%, 95-99%, or higher) identical to an amino acid sequence selected from Table 1 or Table 2 (or having an amino acid sequence that has at least 50%, at least 60%, at least 70%, at least 80%, 80-90%, 90-95%, 95-99%, or higher, amino acid sequence identity to an amino acid sequence selected from Table 1 or Table 2). In some embodiments, a binder of a fusion polypeptide is modified and includes one or more amino acid deletions, additions, or mutations relative to a sequence set forth in Table 1 or Table 2. In some embodiments, a binder of a fusion polypeptide includes a deletion, addition, or mutation of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more amino acids (which may or may not be consecutive amino acids) relative to a sequence set forth in Table 1 or Table 2.

[0130] In some embodiments, binders of a fusion polypeptide recognize the same set of one or more amino acids. In some embodiments, binders of a fusion polypeptide recognize a distinct set of one or more amino acids. In some embodiments, binders of a fusion polypeptide recognize an overlapping set of amino acids. In some embodiments, where the binders of a fusion polypeptide recognize the same amino acid, they may recognize the amino acid with the same characteristic pulsing pattern or with different characteristic pulsing patterns.

[0131] In some embodiments, binders of a fusion polypeptide are joined end-to-end, either by a covalent bond or a linker that covalently joins the C-terminus of one binder to the N-terminus of another binder. In the context of fusion polypeptides of the application, a linker refers to one or more amino acids within a fusion polypeptide that joins two binders and that does not form part of the polypeptide sequence corresponding to either of the two binders. In some embodiments, a

linker comprises at least two amino acids (e.g., at least 2, 3, 4, 5, 6, 8, 10, 15, 25, 50, 100, or more, amino acids). In some embodiments, a linker comprises up to 5, up to 10, up to 15, up to 25, up to 50, or up to 100, amino acids. In some embodiments a linker comprises between about 2 and about 200 amino acids (e.g., between about 2 and about 100, between about 5 and about 50, between about 2 and about 20, between about 5 and about 20, or between about 2 and about 30, amino acids).

[0132] In some aspects, the application provides a nucleic acid encoding a single polypeptide having tandem copies of two or more amino acid binding proteins. In some embodiments, the nucleic acid is an expression construct encoding a fusion polypeptide of the application. In some embodiments, an expression construct encodes a fusion polypeptide having at least two and up to ten binders (e.g., at least 2 binders and up to eight, six, five, four, or three binders). In some embodiments, an expression construct encodes a fusion polypeptide having five or fewer binders (e.g., two, three, four, or five binders).

[0133] In some embodiments, an amino acid recognition molecule comprises one or more labels. In some embodiments, the one or more labels comprise a luminescent label or a conductivity label as described elsewhere herein. In some embodiments, the one or more labels comprise one or more polyol moieties (e.g., one or more moieties selected from dextran, polyvinylpyrrolidone, polyethylene glycol, polypropylene glycol, polyoxyethylene glycol, and polyvinyl alcohol). For example, in some embodiments, an amino acid recognition molecule is PEGylated. In some embodiments, polyol modification (e.g., PEGylation) can limit the extent of non-specific sticking to a substrate (e.g., sequencing chip) surface. In some embodiments, polyol modification can limit the extent of aggregation or interaction between an amino acid recognition molecule with other recognition molecules, with a cleaving reagent, or with other species present in a sequencing reaction mixture. PEGylation can be performed by incubating a recognition molecule (e.g., an amino acid binding protein, such as a ClpS protein) with mPEG4-NHS ester, which labels primary amines such as surface-exposed lysine side chains. Other types of PEG and other methods of polyol modification are known in the art.

[0134] In some embodiments, the one or more labels comprise a tag sequence. For example, in some embodiments, an amino acid recognition molecule comprises a tag sequence that provides one or more functions other than amino acid binding. In some embodiments, a tag sequence comprises at least one biotin ligase recognition sequence that permits biotinylation of the recognition molecule (e.g., incorporation of one or more biotin molecules, including biotin and bis-biotin moieties). In some embodiments, the tag sequence comprises two biotin ligase recognition sequences oriented in tandem. In some embodiments, a biotin ligase recognition sequence refers to an amino acid sequence that is recognized by a biotin ligase, which catalyzes a

covalent linkage between the sequence and a biotin molecule. Each biotin ligase recognition sequence of a tag sequence can be covalently linked to a biotin moiety, such that a tag sequence having multiple biotin ligase recognition sequences can be covalently linked to multiple biotin molecules. A region of a tag sequence having one or more biotin ligase recognition sequences can be generally referred to as a biotinylation tag or a biotinylation sequence. In some embodiments, a bis-biotin or bis-biotin moiety can refer to two biotins bound to two biotin ligase recognition sequences oriented in tandem.

[0135] Additional examples of functional sequences in a tag sequence include purification tags, cleavage sites, and other moieties useful for purification and/or modification of recognition molecules. Table 3 provides a list of non-limiting sequences of tag sequences, any one or more of which may be used in combination with any one of the amino acid recognition molecules of the application (e.g., in combination with a sequence set forth in Table 1 or Table 2). It should be appreciated that the tag sequences shown in Table 3 are meant to be non-limiting, and recognition molecules in accordance with the application can include any one or more of the tag sequences (e.g., His-tags and/or biotinylation tags) at the N- or C-terminus of a recognition molecule polypeptide or at an internal position, split between the N- and C-terminus, or otherwise rearranged as practiced in the art.

Table 3. Non-limiting examples of tag sequences.

Name	SEQ ID NO:	Sequence
Biotinylation tag	190	GGGSGGGSGGGSGLNDFFEAQKIEWHE
Bis-biotinylation tag	191	GGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
Bis-biotinylation tag	192	GSGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
His/biotinylation tag	193	GHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHE
His/bis-biotinylation tag	194	GHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
His/bis-biotinylation tag	195	GGSHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
His/bis-biotinylation tag	196	GSHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
Bis-biotinylation/His tag	197	GGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHEGHHHHHH

[0136] In some embodiments, a recognition molecule of the application is an amino acid binding protein which can be used with other types of amino acid binding molecules, such as a peptidase and/or a nucleic acid aptamer, in a method sequencing. A peptidase, also referred to as a protease or proteinase, is an enzyme that catalyzes the hydrolysis of a peptide bond. Peptidases digest polypeptides into shorter fragments and may be generally classified into endopeptidases

and exopeptidases, which cleave a polypeptide chain internally and terminally, respectively. In some embodiments, labeled recognition molecule **200** comprises a peptidase that has been modified to inactivate exopeptidase or endopeptidase activity. In this way, labeled recognition molecule **200** selectively binds without also cleaving the amino acid from a polypeptide. In yet other embodiments, a peptidase that has not been modified to inactivate exopeptidase or endopeptidase activity may be used with an amino acid binding protein of the application. For example, in some embodiments, a labeled recognition molecule comprises a labeled exopeptidase **202**.

[0137] In accordance with certain embodiments of the application, protein sequencing methods may comprise iterative detection and cleavage at a terminal end of a polypeptide. In some embodiments, labeled exopeptidase **202** may be used as a single reagent that performs both steps of detection and cleavage of an amino acid. As generically depicted, in some embodiments, labeled exopeptidase **202** has aminopeptidase or carboxypeptidase activity such that it selectively binds and cleaves an N-terminal or C-terminal amino acid, respectively, from a polypeptide. It should be appreciated that, in certain embodiments, labeled exopeptidase **202** may be catalytically inactivated by one skilled in the art such that labeled exopeptidase **202** retains selective binding properties for use as a non-cleaving labeled recognition molecule **200**, as described herein.

[0138] An exopeptidase generally requires a polypeptide substrate to comprise at least one of a free amino group at its amino-terminus or a free carboxyl group at its carboxy-terminus. In some embodiments, an exopeptidase in accordance with the application hydrolyses a bond at or near a terminus of a polypeptide. In some embodiments, an exopeptidase hydrolyses a bond not more than three residues from a polypeptide terminus. For example, in some embodiments, a single hydrolysis reaction catalyzed by an exopeptidase cleaves a single amino acid, a dipeptide, or a tripeptide from a polypeptide terminal end.

[0139] In some embodiments, an exopeptidase in accordance with the application is an aminopeptidase or a carboxypeptidase, which cleaves a single amino acid from an amino- or a carboxy-terminus, respectively. In some embodiments, an exopeptidase in accordance with the application is a dipeptidyl-peptidase or a peptidyl-dipeptidase, which cleave a dipeptide from an amino- or a carboxy-terminus, respectively. In yet other embodiments, an exopeptidase in accordance with the application is a tripeptidyl-peptidase, which cleaves a tripeptide from an amino-terminus. Peptidase classification and activities of each class or subclass thereof is well known and described in the literature (see, e.g., Gurupriya, V. S. & Roy, S. C. *Proteases and Protease Inhibitors in Male Reproduction. Proteases in Physiology and Pathology* 195–216 (2017); and Brix, K. & Stöcker, W. *Proteases: Structure and Function*. Chapter 1). In some

embodiments, a peptidase in accordance with the application removes more than three amino acids from a polypeptide terminus. Accordingly, in some embodiments, the peptidase is an endopeptidase, e.g., that cleaves preferentially at particular positions (e.g., before or after a particular amino acid). In some embodiments, the size of a polypeptide cleavage product of endopeptidase activity will will depend on the distribution of cleavage sites (e.g., amino acids) within the polypeptide being analyzed.

[0140] An exopeptidase in accordance with the application may be selected or engineered based on the directionality of a sequencing reaction. For example, in embodiments of sequencing from an amino-terminus to a carboxy-terminus of a polypeptide, an exopeptidase comprises aminopeptidase activity. Conversely, in embodiments of sequencing from a carboxy-terminus to an amino-terminus of a polypeptide, an exopeptidase comprises carboxypeptidase activity. Examples of carboxypeptidases that recognize specific carboxy-terminal amino acids, which may be used as labeled exopeptidases or inactivated to be used as non-cleaving labeled recognition molecules described herein, have been described in the literature (see, e.g., Garcia-Guerrero, M.C., et al. (2018) *PNAS* 115(17)).

[0141] Suitable peptidases for use as cleaving reagents and/or recognition molecules include aminopeptidases that selectively bind one or more types of amino acids. In some embodiments, an aminopeptidase recognition molecule is modified to inactivate aminopeptidase activity. In some embodiments, an aminopeptidase cleaving reagent is non-specific such that it cleaves most or all types of amino acids from a terminal end of a polypeptide. In some embodiments, an aminopeptidase cleaving reagent is more efficient at cleaving one or more types of amino acids from a terminal end of a polypeptide as compared to other types of amino acids at the terminal end of the polypeptide. For example, an aminopeptidase in accordance with the application specifically cleaves alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, selenocysteine, serine, threonine, tryptophan, tyrosine, and/or valine. In some embodiments, an aminopeptidase is a proline aminopeptidase. In some embodiments, an aminopeptidase is a proline iminopeptidase. In some embodiments, an aminopeptidase is a glutamate/aspartate-specific aminopeptidase. In some embodiments, an aminopeptidase is a methionine-specific aminopeptidase. In some embodiments, an aminopeptidase is an aminopeptidase set forth in Table 4. In some embodiments, an aminopeptidase cleaving reagent cleaves a peptide substrate as set forth in Table 4.

[0142] In some embodiments, an aminopeptidase is a non-specific aminopeptidase. In some embodiments, a non-specific aminopeptidase is a zinc metalloprotease. In some embodiments, a

non-specific aminopeptidase is an aminopeptidase set forth in Table 5. In some embodiments, a non-specific aminopeptidase cleaves a peptide substrate as set forth in Table 5.

[0143] Accordingly, in some embodiments, the application provides an aminopeptidase (e.g., an aminopeptidase recognition molecule, an aminopeptidase cleaving reagent) having an amino acid sequence selected from Table 4 or Table 5 (or having an amino acid sequence that has at least 50%, at least 60%, at least 70%, at least 80%, 80-90%, 90-95%, 95-99%, or higher, amino acid sequence identity to an amino acid sequence selected from Table 4 or Table 5). In some embodiments, an aminopeptidase has 25-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-95%, or 95-99%, or higher, amino acid sequence identity to an aminopeptidase listed in Table 4 or Table 5. In some embodiments, an aminopeptidase is a modified aminopeptidase and includes one or more amino acid mutations relative to a sequence set forth in Table 4 or Table 5.

Table 4. Non-limiting examples of aminopeptidases.

Name	SEQ ID NO:	Sequence
L. pneumophila M1 Aminopeptidase (Glu/Asp Specific)	198	MMVKQGVFMKTDQSKVKKLSDYKSLDYFVIHVDLQIDLSKKPVESK ARLTVVPNLNVDSHSNDLVLDGENMTLVSLQMNDNLLKENEYELTK DSLIIKNIPQNTPFTIEMTSLLGENTDLFGLYETEGVALVKAESEG LRRVFYLPDRPDNLATYKTTIIANQEDYPVLLSNGVLIEKKELPLG LHSVTWLDDVPKPSYLFALVAGNLQRSVTYYQTKSGRELPIEFYVP PSATSKCDFAKEVLKEAMAWDERTFNLECALRQHMVAGVDKYASGA SEPTGLNLFNTENLFASPETKTDLGILRVLEVVAHEFFHYWSGDRV TIRDWFNLPLKEGLTTFRAAMFREELFGTDLIRLLDGKNLDERAPR QSAYTAVRSLYTAAAYEKSADIFRMMMLFIGKEPFIEAVAKFFKDN DGGAVTLEDFIESISNSSGKDLRSFLSWFTESGIPELIVTDELNPD TKQYFLKIKTVNGRNRPIPILMGLLDSSGAEIVADKLLIVDQEEIE FQFENIQTRPIPSLLRSFSAPVHMKYEYSYQDLLLLMQFDTNLYNR CEAAKQLISALINDFCIGKKIELSPQFFAVYKALLSDNSLNEWMLA ELITLPSLEELIENQDKPDFEKLNEGRQLIQNALANELKTDFYNLL FRIQISGDDDKQKLKGFDLKQAGLRRLKSVCFSYLLNVDFEKTKEK LILQFEDALGKNMTETALALSMLCEINCEEADVALEDYYHYWKNDP GAVNNWFSIQALAHSPDVIERVKKLMRHGDFDLSNPNKVYALLGSF IKNPFGFHSVTGEGYQLVADAIFDLDKINPTLAANLTEKFTYWDKY DVNRQAMMISTLKIIYSNATSSDVRTMAKKGLDKVKEDLPLPIHLT FHGGSTMQDRTAQLIADGNKENAYQLH
E. coli methionine aminopeptidase (Met specific)	199	MGTAISIKTPEDIEKMRVAGRLAAEVLEMIEPYVKPGVSTGELDRI CNDYIVNEQHAVSACLGYHGYPKSVCISINEVVCHGIPDDAKLLKD GDIVNIDVTVIKDGFHGDTSKMFIVGKPTIMGERLCRITQESLYLA LRMVKPGINLREIGAAIQKFVEAEGFSVVREYCGHGIGRGFHEEPQ VLHYDSRETNVVLKPGMTFTIEPMVNAGKKEIRTMKDGWTVKTKDR SLSAQYEHTIVVTDNGCEILTLRKDDTIPAIISHD
M. smegmatis Proline iminopeptidase (Pro specific)	200	MGTLEANTNGPGSMLSRMPVSSRTVPFGDHETWVQVTTPENAQPHA LPLIVLHGGPGMAHNYVANIAALADETGRTVIHYDQVGCGNSTHLP DAPADFWTPQLFVDEFHAVCTALGIERYHVLGQSWGGMLGAEIAVR QPSGLVSLAICNSPASMRLWSEAAGDLRAQLPAETRAALDRHEAAG TITHPDYLQAAAEFYRRHVCRVVPTPQDFADSVAQMEAEPTVYHTM NGPNEFHVVGTLGDWSVIDRLPDVTAPVLVIAGEHDEATPKTWQPF VDHIPDVRSHVFPGTSHCTHLEKPEEFRAVVAQFLHQHDLAADARV

Y. pestis Proline iminopeptidase (Pro Specific)	201	MTQQEYQNRRQALLAKMAPGSAAIIFAAPEATRSADSEYPYRQNSD FSYLTGFNEPEAVLILVKSDETHNHSVLFNRIRDLTAEIWFGRRLG QEAAPTKLAVDRALPFDEINEQLYLLLNRLDVIYHAQGQYAYADNI VFAALEKLRHGFRKNLRAPATLTDWRPWLHEMRLFKSAEEIAVLRR AGEISALAHTRAMEKCRPGMFEYQLEGEILHEFTRHGARYPAYNTI VGGGENGCILHYTENECELRDGDLVLIDAGCEYRGYAGDITRTFPV NGKFTPAQRAVYDIVLAAINKSLTLFRPGTSIREVTEEVVRIMVVG LVELGILKGDIEQLIAEQAHRPFFMHGLSHWLGMDVHDVGDYGSSD RGRILEPGMVLTVEPGLYIAPDADVPPQYRGIGIRIEDDIVITATG NENLTASVVKDPDDIEALMALNHAGENLYFQE
P. furiosus methionine aminopeptidase	202	MDTEKLMKAGEIAKKVREKAIKLARPGMLLLELAESIEKMIMELGG KPAFPVNLSINEIAAHYTPYKGDTTVLKEGDYLKIDVGVHIDGFIA DTAVTVRVGMEEDELMEAAKEALNAAISVARAGVEIKELGKAIENE IRKRGFKPIVNLSGHKIERYKLHAGISIPNIYRPHDNYVLKEGDVF AIEPFATIGAGQVIEVPPTLIYMYVRDVPVRVAQARFLLAKIKREY GTLPFAYRWLQNDMPEGQLKLALKTLEKAGAIYGYPVLKEIRNGIV AQFEHTIIVEKDSVIVTQDMINKSTLE
Aeromonas sobria Proline aminopeptidase	203	HMSSPLHYVLDGIHCEPHFFTVPLDHQQPDDEETITLFGRTLCRKD RLDDELPWLLYLQGGPGFGAPRPSANGGWIKRALQEFRVLLLDQRG TGHSTPIHAELLAHLNPRQQADYLSHFRADSIVRDAELIREQLSPD HPWSLLGQSFGGFCSLTYLSLFPDSLHEVYLTGGVAPIGRSADEVY RATYQRVADKNRAFFARFPHAQAIANRLATHLQRHDVRLPNGQRLT VEQLQQQGLDLGASGAFEELYYLLEDAFIGEKLNPAFLYQVQAMQP FNTNPVFAILHELIYCEGAASHWAAERVRGEFPALAWAQGKDFAFT GEMIFPWMFEQFRELIPLKEAAHLLAEKADWGPLYDPVQLARNKVP VACAVYAEDMYVEFDYSRETLKGLSNSRAWITNEYEHNGLRVDGEQ ILDRLIRLNRDCLE
Pyrococcus furiosus Proline Aminopeptidase (X-/-Pro)	204	MKERLEKLVKFMDENSIDRVFIAKPVNVYYFSGTSPLGGGYIIVDG DEATLYVPELEYEMAKEESKLPVVKFKKFDEIYEILKNTETLGIEG TLSYSMVENFKEKSNVKEFKKIDDVIKDLRIIKTKEEIEIIEKACE IADKAVMAAIEEITEGKREREVAAKVEYLMKMNGAEKPAFDTIIAS GHRSALPHGVASDKRIERGDLVVIDLGALYNHYNSDITRTIVVGSP NEKQREIYEIVLEAQKRAVEAAKPGMTAKELDSIAREIIKEYGYGD YFIHSLGHGVGLEIHEWPRISQYDETVLKEGMVITIEPGIYIPKLG GVRIEDTVLITENGAKRLTKTERELL
Elizabethkingia meningoseptica Proline aminopeptidase	205	MIPITTPVGNFKVWTKRFGTNPKIKVLLLHGGPAMTHEYMECFETF FQREGFEFYEYDQLGSYYSDQPTDEKLWNIDRFVDEVEQVRKAIHA DKENFYVLGNSWGGILAMEYALKYQQNLKGLIVANMMASAPEYVKY AEVLSKQMKPEVLAEVRAIEAKKDYANPRYTELLFPNYYAQHICRL KEWPDALNRSLKHVNSTVYTLMQGPSELGMSSDARLAKWDIKNRLH EIATPTLMIGARYDTMDPKAMEEQSKLVQKGRYLYCPNGSHLAMWD DQKVFMDGVIKFIKDVDTKSFN
N. gonorrhoeae Proline Iminopeptidase	206	MYEIKQPFHSGYLQVSEIHQIYWEESGNPDGVPVIFLHGGPGAGAS PECRGFFNPDVFRIVIIDQRGCGRSHPYACAEDNTTWDLVADIEKV REMLGIGKWLVFGGSWGSTLSLAYAQTHPERVKGLVLRGIFLCRPS ETAWLNEAGGVSRIYPEQWQKFVAPIAENRRNRLIEAYHGLLFHQD EEVCLSAAKAWADWESYLIRFEPEGVDEDAYASLAIARLENHYFVN GGWLQGDKAILNNIGKIRHIPTVIVQGRYDLCTPMQSAWELSKAFP EAELRVVQAGHCAFDPPLADALVQAVEDILPRLL

Table 5. Non-limiting examples of non-specific aminopeptidases.

Name	SEQ ID NO:	Sequence
E. coli Aminopeptidase N*	207	MTQQPQAKYRHDYRAPDYQITDIDLTFDLDAQKTVVTAVSQAVRHG ASDAPLRLNGEDLKLVSVHINDEPWTAWKEEEGALVISNLPERFTL KIINEISPAANTALEGLYQSGDALCTQCEAEGFRHITYYLDRPDVL

(Zinc Metalloprotease)		ARFTTKIIADKIKYPFLLSNGNRVAQGELENGRHWVQWQDPFPKPC YLFALVAGDFDVLRDTFTTRSGREVALELYVDRGNLDRAPWAMTSL KNSMKWDEERFGLEYDLDIYMIVAVDFFNMGAMENKGLNIFNSKYV LARTDTATDKDYLDIERVIGHEYFHNWTGNRVTCRDWFQLSLKEGL TVFRDQEFSSDLGSRAVNRINNVRTMRGLQFAEDASPMAHPIRPDM VIEMNNFYTLTVYEKGAEVIRMIHTLLGEENFQKGMQLYFERHDGS AATCDDFVQAMEDASNVDLSHFRRWYSQSGTPIVTVKDDYNPETEQ YTLTISQRTPATPDQAEKQPLHIPFAIELYDNEGKVIPLQKGGHPV NSVLNVTQAEQTFVFDNVYFQPVPALLCEFSAPVKLEYKWSDQQLT FLMRHARNDFSRWDAAQSLLATYIKLNVARHQQGQPLSLPVHVADA FRAVLLDEKIDPALAAEILTLPSVNEMAELFDIIDPIAIAEVREAL TRTLATELADELLAIYNANYQSEYRVEHEDIAKRTLRNACLRFLAF GETHLADVLVSKQFHEANNMTDALAALSAAVAAQLPCRDALMQEYD DKWHQNGLVMDKWFILQATSPAANVLETVRGLLQHRSFTMSNPNRI RSLIGAFAGSNPAAFHAEDGSGYLFLVEMLTDLNSRNPQVASRLIE PLIRLKRYDAKRQEKMRAALEQLKGLENLSGDLYEKITKALA
P. falciparum M1 aminopeptidase**	208	PKIHYRKDYKPSGFIINQVTLNINIHDQETIVRSVLDMDISKHNVG EDLVFDGVGLKINEISINNKKLVEGEEYTYDNEFLTIFSKFVPKSK FAFSSEVIIHPETNYALTGLYKSKNIIVSQCEATGFRRITFFIDRP DMMAKYDVTVTADKEKYPVLLSNGDKVNEFEIPGGRHGARFNDPPL KPCYLFAVVAGDLKHLSATYITKYTKKKVELYVFSEEKYVSKLQWA LECLKKSMAFDEDYFGLEYDLSRLNLVAVSDFNVGAMENKGLNIFN ANSLLASKKNSIDFSYARILTVVGHEYFHQYTGNRVTLRDWFQLTL KEGLTVHRENLFSEEMTKTVTTRLSHVDLLRSVQFLEDSSPLSHPI RPESYVSMENFYTTTVYDKGSEVMRMYLTILGEEYYKKGFDIYIKK NDGNTATCEDFNYAMEQAYKMKKADNSANLNQYLLWFSQSGTPHVS FKYNYDAEKKQYSIHVNQYTKPDENQKEKKPLFIPISVGLINPENG KEMISQTTLELTKESDTFVFNNIAVKPIPSLFRGFSAPVYIEDQLT DEERILLLKYDSDAFVRYNSCTNIYMKQILMNYNEFLKAKNEKLES FQLTPVNAQFIDAIKYLLEDPHADAGFKSYIVSLPQDRYIINFVSN LDTDVLADTKEYIYKQIGDKLNDVYYKMFKSLEAKADDLTYFNDES HVDFDQMNMRTLRNTLLSLLSKAQYPNILNEIIEHSKSPYPSNWLT SLSVSAYFDKYFELYDKTYKLSKDDELLLQEWLKTVSRSDRKDIYE ILKKLENEVLKDSKNPNDIRAVYLPFTNNLRRFHDISGKGYKLIAE VITKTDKFNPMVATQLCEPFKLWNKLDTKRQELMLNEMNTMLQEPQ ISNNLKEYLLRLTNK
Puromycin-sensitive aminopeptidase (“NPEPPS”)	209	MWLAAAAPSLARRLLFLGPPPPPLLLLVFSRSSRRRLHSLGLAAMP EKRPFERLPADVSPINYSLCLKPDLLDFTFEGKLEAAAQVRQATNQ IVMNCADIDIITASYAPEGDEEIHATGFNYQNEDEKVTLSFPSTLQ TGTGTLKIDFVGELNDKMKGFYRSKYTTPSGEVRYAAVTQFEATDA RRAFPCWDEPAIKATFDISLVVPKDRVALSNMNVIDRKPYPDDENL VEVKFARTPVMSTYLVAFVVGEYDFVETRSKDGVCVRVYTPVGKAE QGKFALEVAAKTLPFYKDYFNVPYPLPKIDLIAIADFAAGAMENWG LVTYRETALLIDPKNSCSSSRQWVALVVGHELAHQWFGNLVTMEWW THLWLNEGFASWIEYLCVDHCFPEYDIWTQFVSADYTRAQELDALD NSHPIEVSVGHPSEVDEIFDAISYSKGASVIRMLHDYIGDKDFKKG MNMYLTKFQQKNAATEDLWESLENASGKPIAAVMNTWTKQMGFPLI YVEAEQVEDDRLLRLSQKKFCAGGSYVGEDCPQWMVPITISTSEDP NQAKLKILMDKPEMNVVLKNVKPDQWVKLNLGTVGFYRTQYSSAML ESLLPGIRDLSLPPVDRLGLQNDLFSLARAGIISTVEVLKVMEAFV NEPNYTVWSDLSCNLGILSTLLSHTDFYEEIQEFVKDVFSPIGERL GWDPKPGEGHLDALLRGLVLGKLGKAGHKATLEEARRRFKDHVEGK QILSADLRSPVYLTVLKHGDGTTLDIMLKLHKQADMQEEKNRIERV LGATLLPDLIQKVLTFALSEEVRPQDTVSVIGGVAGGSKHGRKAAW KFIKDNWEELYNRYQGGFLISRLIKLSVEGFAVDKMAGEVKAFFES HPAPSAERTIQQCCENILLNAAWLKRDAESIHQYLLQRKASPPTV
NPEPPS E366V	210	MWLAAAAPSLARRLLFLGPPPPPLLLLVFSRSSRRRLHSLGLAAMP EKRPFERLPADVSPINYSLCLKPDLLDFTFEGKLEAAAQVRQATNQ

		IVMNCADIDIITASYAPEGDEEIHATGFNYQNEDEKVTLSFPSTLQ TGTGTLKIDFVGELNDKMKGFYRSKYTTPSGEVRYAAVTQFEATDA RRAFPCWDEPAIKATFDISLVVPKDRVALSNMNVIDRKPYPDDENL VEVKFARTPVMSTYLVAFVVGEYDFVETRSKDGVCVRVYTPVGKAE QGKFALEVAAKTLPFYKDYFNVPYPLPKIDLIAIADFAAGAMENWG LVTYRETALLIDPKNSCSSSRQWVALVVGHVLAHQWFGNLVTMEWW THLWLNEGFASWIEYLCVDHCFPEYDIWTQFVSADYTRAQELDALD NSHPIEVSVGHPSEVDEIFDAISYSKGASVIRMLHDYIGDKDFKKG MNMYLTKFQQKNAATEDLWESLENASGKPIAAVMNTWTKQMGFPLI YVEAEQVEDDRLLRLSQKKFCAGGSYVGEDCPQWMVPITISTSEDP NQAKLKILMDKPEMNVVLKNVKPDQWVKLNLGTVGFYRTQYSSAML ESLLPGIRDLSLPPVDRLGLQNDLFSLARAGIISTVEVLKVMEAFV NEPNYTVWSDLSCNLGILSTLLSHTDFYEEIQEFVKDVFSPIGERL GWDPKPGEGHLDALLRGLVLGKLGKAGHKATLEEARRRFKDHVEGK QILSADLRSPVYLTVLKHGDGTTLDIMLKLHKQADMQEEKNRIERV LGATLLPDLIQKVLTFALSEEVRPQDTVSVIGGVAGGSKHGRKAAW KFIKDNWEELYNRYQGGFLISRLIKLSVEGFAVDKMAGEVKAFFES HPAPSAERTIQQCCENILLNAAWLKRDAESIHQYLLQRKASPPTV
Francisella tularensis Aminopeptidase N	211	MIYEFVMTDPKIKYLKDYKPSNYLIDETHLIFELDESKTRVTANLY IVANRENRENNTLVLDGVELKLLSIKLNNKHLSPAEFAVNENQLII NNVPEKFVLQTVVEINPSANTSLEGLYKSGDVFSTQCEATGFRKIT YYLDRPDVMAAFTVKIIADKKKYPIILSNGDKIDSGDISDNQHFAV WKDPFKKPCYLFALVAGDLASIKDTYITKSQRKVSLEIYAFKQDID KCHYAMQAVKDSMKWDEDRFGLEYDLDTFMIVAVPDFNAGAMENKG LNIFNTKYIMASNKTATDKDFELVQSVVGHEYFHNWTGDRVTCRDW FQLSLKEGLTVFRDQEFTSDLNSRDVKRIDDVRIIRSAQFAEDASP MSHPIRPESYIEMNNFYTVTVYNKGAEIIRMIHTLLGEEGFQKGMK LYFERHDGQAVTCDDFVNAMADANNRDFSLFKRWYAQSGTPNIKVS ENYDASSQTYSLTLEQTTLPTADQKEKQALHIPVKMGLINPEGKNI AEQVIELKEQKQTYTFENIAAKPVASLFRDFSAPVKVEHKRSEKDL LHIVKYDNNAFNRWDSLQQIATNIILNNADLNDEFLNAFKSILHDK DLDKALISNALLIPIESTIAEAMRVIMVDDIVLSRKNVVNQLADKL KDDWLAVYQQCNDNKPYSLSAEQIAKRKLKGVCLSYLMNASDQKVG TDLAQQLFDNADNMTDQQTAFTELLKSNDKQVRDNAINEFYNRWRH EDLVVNKWLLSQAQISHESALDIVKGLVNHPAYNPKNPNKVYSLIG GFGANFLQYHCKDGLGYAFMADTVLALDKFNHQVAARMARNLMSWK RYDSDRQAMMKNALEKIKASNPSKNVFEIVSKSLES
Pyrococcus horikoshii TET Aminopeptidase	212	MEVRNMVDYELLKKVVEAPGVSGYEFLGIRDVVIEEIKDYVDEVKV DKLGNVIAHKKGEGPKVMIAAHMDQIGLMVTHIEKNGFLRVAPIGG VDPKTLIAQRFKVWIDKGKFIYGVGASVPPHIQKPEDRKKAPDWDQ IFIDIGAESKEEAEDMGVKIGTVITWDGRLERLGKHRFVSIAFDDR IAVYTILEVAKQLKDAKADVYFVATVQEEVGLRGARTSAFGIEPDY GFAIDVTIAADIPGTPEHKQVTHLGKGTAIKIMDRSVICHPTIVRW LEELAKKHEIPYQLEILLGGGTDAGAIHLTKAGVPTGALSVPARYI HSNTEVVDERDVDATVELMTKALENIHELKI
T. aquaticus Aminopeptidase T	213	MDAFTENLNKLAELAIRVGLNLEEGQEIVATAPIEAVDFVRLLAEK AYENGASLFTVLYGDNLIARKRLALVPEAHLDRAPAWLYEGMAKAF HEGAARLAVSGNDPKALEGLPPERVGRAQQAQSRAYRPTLSAITEF VTNWTIVPFAHPGWAKAVFPGLPEEEAVQRLWQAIFQATRVDQEDP VAAWEAHNRVLHAKVAFLNEKRFHALHFQGPGTDLTVGLAEGHLWQ GGATPTKKGRLCNPNLPTEEVFTAPHRERVEGVVRASRPLALSGQL VEGLWARFEGGVAVEVGAEKGEEVLKKLLDTDEGARRLGEVALVPA DNPIAKTGLVFFDTLFDENAASHIAFGQAYAENLEGRPSGEEFRRR GGNESMVHVDWMIGSEEVDVDGLLEDGTRVPLMRRGRWVI
Bacillus stearothermophilus Peptidase M28	214	MAKLDETLTMLKALTDAKGVPGNEREARDVMKTYIAPYADEVTTDG LGSLIAKKEGKSGGPKVMIAGHLDEVGFMVTQIDDKGFIRFQTLGG WWSQVMLAQRVTIVTKKGDITGVIGSKPPHILPSEARKKPVEIKDM FIDIGATSREEAMEWGVRPGDMIVPYFEFTVLNNEKMLLAKAWDNR

		IGCAVAIDVLKQLKGVDHPNTVYGVGTVQEEVGLRGARTAAQFIQP DIAFAVDVGIAGDTPGVSEKEAMGKLGAGPHIVLYDATMVSHRGLR EFVIEVAEELNIPHHFDAMPGVGTDAGAIHLTGIGVPSLTIAIPTR YIHSHAAILHRDDYENTVKLLVEVIKRLDADKVKQLTFDE
Vibrio cholera Aminopeptidase	215	MEDKVWISMGADAVGSLNPALSESLLPHSFASGSQVWIGEVAIDEL AELSHTMHEQHNRCGGYMVHTSAQGAMAALMMPESIANFTIPAPSQ QDLVNAWLPQVSADQITNTIRALSSFNNRFYTTTSGAQASDWLANE WRSLISSLPGSRIEQIKHSGYNQKSVVLTIQGSEKPDEWVIVGGHL DSTLGSHTNEQSIAPGADDDASGIASLSEIIRVLRDNNFRPKRSVA LMAYAAEEVGLRGSQDLANQYKAQGKKVVSVLQLDMTNYRGSAEDI VFITDYTDSNLTQFLTTLIDEYLPELTYGYDRCGYACSDHASWHKA GFSAAMPFESKFKDYNPKIHTSQDTLANSDPTGNHAVKFTKLGLAY VIEMANAGSSQVPDDSVLQDGTAKINLSGARGTQKRFTFELSQSKP LTIQTYGGSGDVDLYVKYGSAPSKSNWDCRPYQNGNRETCSFNNAQ PGIYHVMLDGYTNYNDVALKASTQ
Photobacterium halotolerans Aminopeptidase	216	MEDKVWISIGSDASQTVKSVMQSNARSLLPESLASNGPVWVGQVDY SQLAELSHHMHEDHQRCGGYMVHSSPESAIAASNMPQSLVAFSIPE ISQQDTVNAWLPQVNSQAITGTITSLTSFINRFYTTTSGAQASDWL ANEWRSLSASLPNASVRQVSHFGYNQKSVVLTITGSEKPDEWIVLG GHLDSTIGSHTNEQSVAPGADDDASGIASVTEIIRVLSENNFQPKR SIAFMAYAAEEVGLRGSQDLANQYKAEGKQVISALQLDMTNYKGSV EDIVFITDYTDSNLTTFLSQLVDEYLPSLTYGFDTCGYACSDHASW HKAGFSAAMPFEAKFNDYNPMIHTPNDTLQNSDPTASHAVKFTKLG LAYAIEMASTTGGTPPPTGNVLKDGVPVNGLSGATGSQVHYSFELP AQKNLQISTAGGSGDVDLYVSFGSEATKQNWDCRPYRNGNNEVCTF AGATPGTYSIMLDGYRQFSGVTLKASTQ
Yersinia pestis AminopeptidaseN	217	MTQQPQAKYRHDYRAPDYTITDIDLDFALDAQKTTVTAVSKVKRQG TDVTPLILNGEDLTLISVSVDGQAWPHYRQQDNTLVIEQLPADFTL TIVNDIHPATNSALEGLYLSGEALCTQCEAEGFRHITYYLDRPDVL ARFTTRIVADKSRYPYLLSNGNRVGQGELDDGRHWVKWEDPFPKPS YLFALVAGDFDVLQDKFITRSGREVALEIFVDRGNLDRADWAMTSL KNSMKWDETRFGLEYDLDIYMIVAVDFFNMGAMENKGLNVFNSKYV LAKAETATDKDYLNIEAVIGHEYFHNWTGNRVTCRDWFQLSLKEGL TVFRDQEFSSDLGSRSVNRIENVRVMRAAQFAEDASPMAHAIRPDK VIEMNNFYTLTVYEKGSEVIRMMHTLLGEQQFQAGMRLYFERHDGS AATCDDFVQAMEDVSNVDLSLFRRWYSQSGTPLLTVHDDYDVEKQQ YHLFVSQKTLPTADQPEKLPLHIPLDIELYDSKGNVIPLQHNGLPV HHVLNVTEAEQTFTFDNVAQKPIPSLLREFSAPVKLDYPYSDQQLT FLMQHARNEFSRWDAAQSLLATYIKLNVAKYQQQQPLSLPAHVADA FRAILLDEHLDPALAAQILTLPSENEMAELFTTIDPQAISTVHEAI TRCLAQELSDELLAVYVANMTPVYRIEHGDIAKRALRNTCLNYLAF GDEEFANKLVSLQYHQADNMTDSLAALAAAVAAQLPCRDELLAAFD VRWNHDGLVMDKWFALQATSPAANVLVQVRTLLKHPAFSLSNPNRT RSLIGSFASGNPAAFHAADGSGYQFLVEILSDLNTRNPQVAARLIE PLIRLKRYDAGRQALMRKALEQLKTLDNLSGDLYEKITKALAA
Vibrio anguillarum Aminopeptidase	218	MEEKVWISIGGDATQTALRSGAQSLLPENLINQTSVWVGQVPVSEL ATLSHEMHENHQRCGGYMVHPSAQSAMSVSAMPLNLNAFSAPEITQ QTTVNAWLPSVSAQQITSTITTLTQFKNRFYTTSTGAQASNWIADH WRSLSASLPASKVEQITHSGYNQKSVMLTITGSEKPDEWVVIGGHL DSTLGSRTNESSIAPGADDDASGIAGVTEIIRLLSEQNFRPKRSIA FMAYAAEEVGLRGSQDLANRFKAEGKKVMSVMQLDMTNYQGSREDI VFITDYTDSNFTQYLTQLLDEYLPSLTYGFDTCGYACSDHASWHAV GYPAAMPFESKFNDYNPNIHSPQDTLQNSDPTGFHAVKFTKLGLAY VVEMGNASTPPTPSNQLKNGVPVNGLSASRNSKTWYQFELQEAGNL SIVLSGGSGDADLYVKYQTDADLQQYDCRPYRSGNNETCQFSNAQP GRYSILLHGYNNYSNASLVANAQ

Salinivibrio spYCSC6 Aminopeptidase	219	MEDKKVWISIGADAQQTALSSGAQPLLAQSVAHNGQAWIGEVSESE LAALSHEMHENHHRCGGYIVHSSAQSAMAASNMPLSRASFIAPAIS QQALVTPWISQIDSALIVNTIDRLTDFPNRFYTTTSGAQASDWIKQ RWQSLSAGLAGASVTQISHSGYNQASVMLTIEGSESPDEWVVVGGH LDSTIGSRTNEQSIAPGADDDASGIAAVTEVIRVLAQNNFQPKRSI AFVAYAAEEVGLRGSQDVANQFKQAGKDVRGVLQLDMTNYQGSAED IVFITDYTDNQLTQYLTQLLDEYLPTLNYGFDTCGYACSDHASWHQ VGYPAAMPFEAKFNDYNPNIHTPQDTLANSDSEGAHAAKFTKLGLA YTVELANADSSPNPGNELKLGEPINGLSGARGNEKYFNYRLDQSGE LVIRTYGGSGDVDLYVKANGDVSTGNWDCRPYRSGNDEVCRFDNAT PGNYAVMLRGYRTYDNVSLIVE
Vibrio proteolyticus Aminopeptidase I	220	MPPITQQATVTAWLPQVDASQITGTISSLESFTNRFYTTTSGAQAS DWIASEWQALSASLPNASVKQVSHSGYNQKSVVMTITGSEAPDEWI VIGGHLDSTIGSHTNEQSVAPGADDDASGIAAVTEVIRVLSENNFQ PKRSIAFMAYAAEEVGLRGSQDLANQYKSEGKNVVSALQLDMTNYK GSAQDVVFITDYTDSNFTQYLTQLMDEYLPSLTYGFDTCGYACSDH ASWHNAGYPAAMPFESKFNDYNPRIHTTQDTLANSDPTGSHAKKFT QLGLAYAIEMGSATGDTPTPGNQLE
Vibrio proteolyticus Aminopeptidase I (A55F)	221	MPPITQQATVTAWLPQVDASQITGTISSLESFTNRFYTTTSGAQAS DWIASEWQFLSASLPNASVKQVSHSGYNQKSVVMTITGSEAPDEWI VIGGHLDSTIGSHTNEQSVAPGADDDASGIAAVTEVIRVLSENNFQ PKRSIAFMAYAAEEVGLRGSQDLANQYKSEGKNVVSALQLDMTNYK GSAQDVVFITDYTDSNFTQYLTQLMDEYLPSLTYGFDTCGYACSDH ASWHNAGYPAAMPFESKFNDYNPRIHTTQDTLANSDPTGSHAKKFT QLGLAYAIEMGSATGDTPTPGNQLE
P. furiosus Aminopeptidase I	222	MVDWELMKKIIESPGVSGYEHLGIRDLVVDILKDVADEVKIDKLGN VIAHFKGSAPKVMVAAHMDKIGLMVNHIDKDGYLRVVPIGGVLPET LIAQKIRFFTEKGERYGVVGVLPPHLRREAKDQGGKIDWDSIIVDV GASSREEAEEMGFRIGTIGEFAPNFTRLSEHRFATPYLDDRICLYA MIEAARQLGEHEADIYIVASVQEEIGLRGARVASFAIDPEVGIAMD VTFAKQPNDKGKIVPELGKGPVMDVGPNINPKLRQFADEVAKKYEI PLQVEPSPRPTGTDANVMQINREGVATAVLSIPIRYMHSQVELADA RDVDNTIKLAKALLEELKPMDFTPLE

*Cleavage efficiency (from most to least): arginine > lysine > hydrophobic residues (including alanine, leucine, methionine, and phenylalanine) > proline (see, e.g., Matthews Biochemistry 47, 2008, 5303-5311).
**Cleavage efficiency (from most to least): leucine > alanine > arginine > phenylalanine > proline; does not cleave after glutamate and aspartate.

[0144] For the purposes of comparing two or more amino acid sequences, the percentage of "sequence identity" between a first amino acid sequence and a second amino acid sequence (also referred to herein as "amino acid identity") may be calculated by dividing [the number of amino acid residues in the first amino acid sequence that are identical to the amino acid residues at the corresponding positions in the second amino acid sequence] by [the total number of amino acid residues in the first amino acid sequence] and multiplying by [100], in which each deletion, insertion, substitution or addition of an amino acid residue in the second amino acid sequence compared to the first amino acid sequence is considered as a difference at a single amino acid residue (position). Alternatively, the degree of sequence identity between two amino acid

sequences may be calculated using a known computer algorithm (e.g., by the local homology algorithm of Smith and Waterman (1970) Adv. Appl. Math. 2:482c, by the homology alignment algorithm of Needleman and Wunsch, J. Mol. Biol. (1970) 48:443, by the search for similarity method of Pearson and Lipman. Proc. Natl. Acad. Sci. USA (1998) 85:2444, or by computerized implementations of algorithms available as Blast, Clustal Omega, or other sequence alignment algorithms) and, for example, using standard settings. Usually, for the purpose of determining the percentage of "sequence identity" between two amino acid sequences in accordance with the calculation method outlined hereinabove, the amino acid sequence with the greatest number of amino acid residues will be taken as the "first" amino acid sequence, and the other amino acid sequence will be taken as the "second" amino acid sequence.

[0145] Additionally, or alternatively, two or more sequences may be assessed for the identity between the sequences. The terms "identical" or percent "identity" in the context of two or more nucleic acids or amino acid sequences, refer to two or more sequences or subsequences that are the same. Two sequences are "substantially identical" if two sequences have a specified percentage of amino acid residues or nucleotides that are the same (e.g., at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.6%, 99.7%, 99.8%, or 99.9% identical) over a specified region or over the entire sequence, when compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the above sequence comparison algorithms or by manual alignment and visual inspection. Optionally, the identity exists over a region that is at least about 25, 50, 75, or 100 amino acids in length, or over a region that is 100 to 150, 150 to 200, 100 to 200, or 200 or more, amino acids in length.

[0146] Additionally, or alternatively, two or more sequences may be assessed for the alignment between the sequences. The terms "alignment" or percent "alignment" in the context of two or more nucleic acids or amino acid sequences, refer to two or more sequences or subsequences that are the same. Two sequences are "substantially aligned" if two sequences have a specified percentage of amino acid residues or nucleotides that are the same (e.g., at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.6%, 99.7%, 99.8% or 99.9% identical) over a specified region or over the entire sequence, when compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the above sequence comparison algorithms or by manual alignment and visual inspection. Optionally, the alignment exists over a region that is at least about 25, 50, 75, or 100 amino acids in length, or over a region that is 100 to 150, 150 to 200, 100 to 200, or 200 or more amino acids in length.

[0147] In addition to protein molecules, nucleic acid molecules possess a variety of advantageous properties for use as amino acid recognition molecules in accordance with the application.

[0148] Nucleic acid aptamers are nucleic acid molecules that have been engineered to bind desired targets with high affinity and selectivity. Accordingly, nucleic acid aptamers may be engineered to selectively bind a desired type of amino acid using selection and/or enrichment techniques known in the art. Thus, in some embodiments, a recognition molecule comprises a nucleic acid aptamer (e.g., a DNA aptamer, an RNA aptamer). As shown in FIG. 2, in some embodiments, a labeled recognition molecule is a labeled aptamer **204** that selectively binds one type of terminal amino acid. For example, in some embodiments, labeled aptamer **204** selectively binds one type of amino acid (e.g., a single type of amino acid or a subset of types of amino acids) at a terminus of a polypeptide, as described herein. Although not shown, it should be appreciated that labeled aptamer **204** may be engineered to selectively bind one type of amino acid at any position of a polypeptide (e.g., at a terminal position or at terminal and internal positions of a polypeptide) in accordance with a method of the application.

[0149] In some embodiments, a labeled recognition molecule comprises a label having binding-induced luminescence. For example, in some embodiments, a labeled aptamer **206** comprises a donor label **212** and an acceptor label **214** and functions as illustrated in panels (I) and (II) of FIG. 2. As depicted in panel (I), labeled aptamer **206** as a free molecule adopts a conformation in which donor label **212** and acceptor label **214** are separated by a distance that limits detectable FRET between the labels (e.g., about 10 nm or more). As depicted in panel (II), labeled aptamer **206** as a selectively bound molecule adopts a conformation in which donor label **212** and acceptor label **214** are within a distance that promotes detectable FRET between the labels (e.g., about 10 nm or less). In yet other embodiments, labeled aptamer **206** comprises a quenching moiety and functions analogously to a molecular beacon, wherein luminescence of labeled aptamer **206** is internally quenched as a free molecule and restored as a selectively bound molecule (see, e.g., Hamaguchi, et al. (2001) *Analytical Biochemistry* 294, 126-131). Without wishing to be bound by theory, it is thought that these and other types of mechanisms for binding-induced luminescence may advantageously reduce or eliminate background luminescence to increase overall sensitivity and accuracy of the methods described herein.

Shielded Recognition Molecules

[0150] In accordance with embodiments described herein, single-molecule polypeptide sequencing methods can be carried out by illuminating a surface-immobilized polypeptide with excitation light, and detecting luminescence produced by a label attached to an amino acid recognition molecule. In some cases, radiative and/or non-radiative decay produced by the label can result in photodamage to the polypeptide. For example, FIG. 3A illustrates an example

sequencing reaction in which a recognition molecule is shown associated with a polypeptide immobilized to a surface.

[0151] In the presence of excitation illumination, the label can produce fluorescence through radiative decay which results in a detectable association event. However, in some cases, the label produces non-radiative decay which can result in the formation of reactive oxygen species **300**. The reactive oxygen species **300** can eventually damage the immobilized peptide, such that the reaction ends before obtaining complete sequence information for the polypeptide. This photodamage can occur, for example, at the exposed polypeptide terminus (top open arrow), at an internal position (middle open arrow), or at the surface linker attaching the polypeptide to the surface (bottom open arrow).

[0152] The inventors have found that photodamage can be mitigated and recognition times extended by incorporation of a shielding element into an amino acid recognition molecule. FIG. 3B illustrates an example sequencing reaction using a shielded recognition molecule that includes a shield **302**. Shield **302** forms a covalent or non-covalent linkage group that provides increased distance between the label and polypeptide, such that damaging effects from reactive oxygen species **300** can be reduced due to free radical decay over the label-polypeptide separation distance. Shield **302** can also provide a steric barrier that shields the polypeptide from the label by absorbing damage from reactive oxygen species **300** and radiative and/or non-radiative decay.

[0153] Without wishing to be bound by theory, it is thought that a shield, positioned between a recognition component and a label component, can absorb, deflect, or otherwise block radiative and/or non-radiative decay emitted by the label component. In some embodiments, the shield prevents or limits the extent to which one or more labels (e.g., luminescent labels) interact with one or more amino acid recognition molecules. In some embodiments, the shield prevents or limits the extent to which one or more labels interact with one or more molecules associated with an amino acid recognition molecule (e.g., a polypeptide associated with the recognition molecule). Accordingly, in some embodiments, the term shield can generally refer to a protective or shielding effect that is provided by some portion of a linkage group formed between a recognition component and a label component.

[0154] In some embodiments, a shield is attached to one or more amino acid recognition molecules (e.g., a recognition component) and to one or more labels (e.g., a label component). In some embodiments, the recognition and label components are attached at non-adjacent sites on the shield. For example, one or more amino acid recognition molecules can be attached to a first side of the shield, and one or more labels can be attached to a second side of the shield,

where the first and second sides of the shield are distant from each other. In some embodiments, the attachment sites are on approximately opposite sides of the shield.

[0155] The distance between the site at which a shield is attached to a recognition molecule and the site at which the shield is attached to a label can be a linear measurement through space or a non-linear measurement across the surface of the shield. The distance between the recognition molecule and label attachment sites on a shield can be measured by modeling the three-dimensional structure of the shield. In some embodiments, this distance can be at least 2 nm, at least 4 nm, at least 6 nm, at least 8 nm, at least 10 nm, at least 12 nm, at least 15 nm, at least 20 nm, at least 30 nm, at least 40 nm, or more. Alternatively, the relative positions of the recognition molecule and label on a shield can be described by treating the structure of the shield as a quadratic surface (e.g., ellipsoid, elliptic cylinder). In some embodiments, the recognition molecule and label attachment sites are separated by a distance that is at least one eighth of the distance around an ellipsoidal shape representing the shield. In some embodiments, the recognition molecule and label are separated by a distance that is at least one quarter of the distance around an ellipsoidal shape representing the shield. In some embodiments, the recognition molecule and label are separated by a distance that is at least one third of the distance around an ellipsoidal shape representing the shield. In some embodiments, the recognition molecule and label are separated by a distance that is one half of the distance around an ellipsoidal shape representing the shield.

[0156] The size of a shield should be such that a label is unable or unlikely to directly contact the polypeptide when the amino acid recognition molecule is associated with the polypeptide. The size of a shield should also be such that an attached label is detectable when the amino acid recognition molecule is associated with the polypeptide. For example, the size should be such that an attached luminescent label is within an illumination volume to be excited.

[0157] It should be appreciated that there are a variety of parameters by which a practitioner could evaluate shielding effects. Generally, the effects of a shielding element can be evaluated by conducting a comparative assessment between a composition having the shielding element and a composition lacking the shielding element. For example, a shielding element can increase recognition time of an amino acid recognition molecule. In some embodiments, recognition time refers to the length of time in which association events between the recognition molecule and a polypeptide are observable in a polypeptide sequencing reaction as described herein. In some embodiments, recognition time is increased by about 10-25%, 25-50%, 50-75%, 75-100%, or more than 100%, for example by about 2-fold, 3-fold, 4-fold, 5-fold, or more, relative to a polypeptide sequencing reaction performed under the same conditions, with the exception that the amino acid recognition molecule lacks the shielding element but is otherwise similar or

identical. In some embodiments, a shielding element can increase sequencing accuracy and/or sequence read length (e.g., by at least 5%, at least 10%, at least 15%, at least 25% or more, relative to a sequencing reaction performed under comparative conditions as described above).

[0158] Accordingly, in some aspects, the application provides shielded recognition molecules comprising at least one amino acid recognition molecule, at least one detectable label, and a shielding element (e.g., a "shield") that forms a covalent or non-covalent linkage group between the recognition molecule and label. In some embodiments, a shielding element is at least 2 nm, at least 5 nm, at least 10 nm, at least 12 nm, at least 15 nm, at least 20 nm, or more, in length (e.g., in an aqueous solution). In some embodiments, a shielding element is between about 2 nm and about 100 nm in length (e.g., between about 2 nm and about 50 nm, between about 10 nm and about 50 nm, between about 20 nm and about 100 nm).

[0159] In some embodiments, a shield (e.g., shielding element) forms a covalent or non-covalent linkage group between one or more amino acid recognition molecules (e.g., a recognition component) and one or more labels (e.g., a label component). As used herein, in some embodiments, covalent and non-covalent linkages or linkage groups refer to the nature of the attachments of the recognition and label components to the shield.

[0160] In some embodiments, a covalent linkage, or a covalent linkage group, refers to a shield that is attached to each of the recognition and label components through a covalent bond or a series of contiguous covalent bonds. Covalent attachment one or both components can be achieved by covalent conjugation methods known in the art. For example, in some embodiments, click chemistry techniques (e.g., copper-catalyzed, strain-promoted, copper-free click chemistry, etc.) can be used to attach one or both components to the shield. Such methods generally involve conjugating one reactive moiety to another reactive moiety to form one or more covalent bonds between the reactive moieties. Accordingly, in some embodiments, a first reactive moiety of a shield can be contacted with a second reactive moiety of a recognition or label component to form a covalent attachment. Examples of reactive moieties include, without limitation, reactive amines, azides, alkynes, nitrones, alkenes (e.g., cycloalkenes), tetrazines, tetrazoles, and other reactive moieties suitable for click reactions and similar coupling techniques.

[0161] In some embodiments, a non-covalent linkage, or a non-covalent linkage group, refers to a shield that is attached to one or both of the recognition and label components through one or more non-covalent coupling means, including but not limited to receptor-ligand interactions and oligonucleotide strand hybridization. Examples of receptor-ligand interactions are provided herein and include, without limitation, protein-protein complexes, protein-ligand complexes, protein-aptamer complexes, and aptamer-nucleic acid complexes. Various configurations and

strategies for oligonucleotide strand hybridization are described herein and are known in the art (see, e.g., U.S. Patent Publication No. 2019/0024168).

[0162] In some embodiments, shield **302** comprises a polymer, such as a biomolecule or a dendritic polymer. FIG. 3C depicts examples of polymer shields and configurations of shielded recognition molecules of the application. A first shielded construct **304** shows an example of a protein shield **330**. In some embodiments, protein shield **330** forms a covalent linkage group between a recognition molecule and a label. For example, in some embodiments, protein shield **330** is attached to each of the recognition molecule and label through one or more covalent bonds, e.g., by covalent attachment through a side-chain of a natural or unnatural amino acid of protein shield **330**. In some embodiments, an amino acid recognition molecule comprises a single polypeptide having at least one amino acid binding protein and protein shield **330** joined end-to-end.

[0163] Accordingly, in some aspects, the application provides a shielded recognition molecule comprising a fusion polypeptide having an amino acid binding protein and a protein shield joined end-to-end (e.g., in a C-terminal to N-terminal fashion). In some embodiments, the binder and protein shield are joined end-to-end, either by a covalent bond or a linker that covalently joins the C-terminus of one protein to the N-terminus of the other protein. In some embodiments, a linker in the context of a fusion polypeptide refers to one or more amino acids within the fusion polypeptide that joins the binder and protein shield and that does not form part of the polypeptide sequence corresponding to either the binder or protein shield. In some embodiments, a linker comprises at least two amino acids (e.g., at least 2, 3, 4, 5, 6, 8, 10, 15, 25, 50, 100, or more, amino acids). In some embodiments, a linker comprises up to 5, up to 10, up to 15, up to 25, up to 50, or up to 100, amino acids. In some embodiments a linker comprises between about 2 and about 200 amino acids (e.g., between about 2 and about 100, between about 5 and about 50, between about 2 and about 20, between about 5 and about 20, or between about 2 and about 30, amino acids).

[0164] In some embodiments, a protein shield of a fusion polypeptide is a protein having a molecular weight of at least 10 kDa. For example, in some embodiments, a protein shield is a protein having a molecular weight of at least 10 kDa and up to 500 kDa (e.g., between about 10 kDa and about 250 kDa, between about 10 kDa and about 150 kDa, between about 10 kDa and about 100 kDa, between about 20 kDa and about 80 kDa, between about 15 kDa and about 100 kDa, or between about 15 kDa and about 50 kDa). In some embodiments, a protein shield of a fusion polypeptide is a protein comprising at least 25 amino acids. For example, in some embodiments, a protein shield is a protein comprising at least 25 and up to 1,000 amino acids (e.g., between about 100 and about 1,000 amino acids, between about 100 and about 750 amino

acids, between about 500 and about 1,000 amino acids, between about 250 and about 750 amino acids, between about 50 and about 500 amino acids, between about 100 and about 400 amino acids, or between about 50 and about 250 amino acids).

[0165] In some embodiments, a protein shield is a polypeptide comprising one or more tag proteins. In some embodiments, a protein shield is a polypeptide comprising at least two tag proteins. In some embodiments, the at least two tag proteins are the same (e.g., the polypeptide comprises at least two copies of a tag protein sequence). In some embodiments, the at least two tag proteins are different (e.g., the polypeptide comprises at least two different tag protein sequences). Examples of tag proteins include, without limitation, *Fasciola hepatica* 8-kDa antigen (Fh8), Maltose-binding protein (MBP), N-utilization substance (NusA), Thioredoxin (Trx), Small ubiquitin-like modifier (SUMO), Glutathione-S-transferase (GST), Solubility-enhancer peptide sequences (SET), IgG domain B1 of Protein G (GB1), IgG repeat domain ZZ of Protein A (ZZ), Mutated dehalogenase (HaloTag), Solubility eNhancing Ubiquitous Tag (SNUT), Seventeen kilodalton protein (Skp), Phage T7 protein kinase (T7PK), *E. coli* secreted protein A (EspA), Monomeric bacteriophage T7 0.3 protein (Orc protein; Mocr), *E. coli* trypsin inhibitor (Ecotin), Calcium-binding protein (CaBP), Stress-responsive arsenate reductase (ArsC), N-terminal fragment of translation initiation factor IF2 (IF2-domain I), Stress-responsive proteins (e.g., RpoA, SlyD, Tsf, RpoS, PotD, Crr), and *E. coli* acidic proteins (e.g., msyB, yjgD, rpoD). *See,* e.g., Costa, S., et al. "Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system." *Front Microbiol.* 2014 Feb 19; 5:63, the relevant content of which is incorporated herein by reference.

[0166] As described herein, a shielding element of the application can advantageously absorb, deflect, or otherwise block radiative and/or non-radiative decay emitted by a label component of an amino acid recognition molecule. Thus, it should be appreciated that a suitable protein shield of a fusion polypeptide can be readily selected by those skilled in the art. For example, the inventors have demonstrated the use of a variety of types of protein shields in the context of a fusion polypeptide, including polypeptides having an amino acid binding protein fused to an enzyme (e.g., DNA polymerase, glutathione S-transferase), a transport protein (e.g., maltose-binding protein), a fluorescent protein (e.g., GFP), and a commercially available tag protein (e.g., SNAP-tag®). The inventors have further demonstrated the use of fusion polypeptides having multiple copies of a protein shield oriented in tandem.

[0167] Accordingly, in some embodiments, the application provides a fusion polypeptide having one or more tandemly-oriented amino acid binding proteins fused to one or more tandemly-oriented protein shields. In some embodiments, where a fusion polypeptide comprises two or more tandemly-oriented binders and/or two or more tandemly-oriented shields, a terminal end of

one of the two or more binders is joined end-to-end with a terminal end of one of the two or more shields. Fusion polypeptides having tandem copies of two or more binders are described elsewhere herein, and in some embodiments, such fusions can further comprise a protein shield joined end-to-end with one of the two or more binders.

[0168] In some embodiments, protein shield **330** forms a non-covalent linkage group between a recognition molecule and a label. For example, in some embodiments, protein shield **330** is a monomeric or multimeric protein comprising one or more ligand-binding sites. In some embodiments, a non-covalent linkage group is formed through one or more ligand moieties bound to the one or more ligand-binding sites. Additional examples of non-covalent linkages formed by protein shields are described elsewhere herein.

[0169] A second shielded construct **306** shows an example of a double-stranded nucleic acid shield comprising a first oligonucleotide strand **332** hybridized with a second oligonucleotide strand **334**. As shown, in some embodiments, the double-stranded nucleic acid shield can comprise a recognition molecule attached to first oligonucleotide strand **332**, and a label attached to second oligonucleotide strand **334**. In this way, the double-stranded nucleic acid shield forms a non-covalent linkage group between the recognition molecule and the label through oligonucleotide strand hybridization. In some embodiments, a recognition molecule and a label can be attached to the same oligonucleotide strand, which can provide a single-stranded nucleic acid shield or a double-stranded nucleic acid shield through hybridization with another oligonucleotide strand. In some embodiments, strand hybridization can provide increased rigidity within a linkage group to further enhance separation between the recognition molecule and the label.

[0170] Where shielding element **302** comprises a nucleic acid, the separation distance between a label and a recognition molecule can be measured by the distance between attachment sites on the nucleic acid (e.g., direct attachment or indirect attachment, such as through one or more additional shield polymers). In some embodiments, the distance between attachment sites on a nucleic acid can be measured by the number of nucleotides within the nucleic acid that occur between the label and the recognition molecule. It should be understood that the number of nucleotides can refer to either the number of nucleotide bases in a single-stranded nucleic acid or the number of nucleotide base pairs in a double-stranded nucleic acid.

[0171] Accordingly, in some embodiments, the attachment site of a recognition molecule and the attachment site of a label can be separated by between 5 and 200 nucleotides (e.g., between 5 and 150 nucleotides, between 5 and 100 nucleotides, between 5 and 50 nucleotides, between 10 and 100 nucleotides). It should be appreciated that any position in a nucleic acid can serve as an attachment site for a recognition molecule, a label, or one or more additional polymer shields. In

some embodiments, an attachment site can be at or approximately at the 5′ or 3′ end, or at an internal position along a strand of the nucleic acid.

[0172] The non-limiting configuration of second shielded construct **306** illustrates an example of a shield that forms a non-covalent linkage through strand hybridization. A further example of non-covalent linkage is illustrated by a third shielded construct **308** comprising an oligonucleotide shield **336**. In some embodiments, oligonucleotide shield **336** is a nucleic acid aptamer that binds a recognition molecule to form a non-covalent linkage. In some embodiments, the recognition molecule is a nucleic acid aptamer, and oligonucleotide shield **336** comprises an oligonucleotide strand that hybridizes with the aptamer to form a non-covalent linkage.

[0173] A fourth shielded construct **310** shows an example of a dendritic polymer shield **338**. As used herein, in some embodiments, a dendritic polymer refers generally to a polyol or a dendrimer. Polyols and dendrimers have been described in the art, and may include branched dendritic structures optimized for a particular configuration. In some embodiments, dendritic polymer shield **338** comprises polyethylene glycol, tetraethylene glycol, poly(amidoamine), poly(propyleneimine), poly(propyleneamine), carbosilane, poly(L-lysine), or a combination of one or more thereof.

[0174] A dendrimer, or dendron, is a repetitively branched molecule that is typically symmetric around the core and that may adopt a spherical three-dimensional morphology. See, e.g., Astruc et al. (2010) Chem. Rev. 110:1857. Incorporation of such structures into a shield of the application can provide for a protective effect through the steric inhibition of contacts between a label and one or more biomolecules associated therewith (e.g., a recognition molecule and/or a polypeptide associated with the recognition molecule). Refinement of the chemical and physical properties of the dendrimer through variation in primary structure of the molecule, including potential functionalization of the dendrimer surface, allows the shielding effects to be adjusted as desired. Dendrimers may be synthesized by a variety of techniques using a wide range of materials and branching reactions, as is known in the art. Such synthetic variation allows the properties of the dendrimer to be customized as necessary. Examples of polyol and dendrimer compounds which can be used in accordance with shields of the application include, without limitation, compounds described in U.S. Patent Publication No. 20180346507.

[0175] FIG. 3D depicts further example configurations of shielded recognition molecules of the application. A protein-nucleic acid construct **312** shows an example of a shield comprising more than one polymer in the form of a protein and a double-stranded nucleic acid. In some embodiments, the protein portion of the shield is attached to the nucleic acid portion of the shield through a covalent linkage. In some embodiments, the attachment is through a non-covalent

linkage. For example, in some embodiments, the protein portion of the shield is a monovalent or multivalent protein that forms at least one non-covalent linkage through a ligand moiety attached to a ligand-binding site of the monovalent or multivalent protein. In some embodiments, the protein portion of the shield comprises an avidin protein.

[0176] In some embodiments, a shielded recognition molecule of the application is an avidin-nucleic acid construct **314**. In some embodiments, avidin-nucleic acid construct **314** includes a shield comprising an avidin protein **340** and a double-stranded nucleic acid. As described herein, avidin protein **340** may be used to form a non-covalent linkage between one or more amino acid recognition molecules and one or more labels, either directly or indirectly, such as through one or more additional shield polymers described herein.

[0177] Avidin proteins are biotin-binding proteins, generally having a biotin binding site at each of four subunits of the avidin protein. Avidin proteins include, for example, avidin, streptavidin, traptavidin, tamavidin, bradavidin, xenavidin, and homologs and variants thereof. In some cases, the monomeric, dimeric, or tetrameric form of the avidin protein can be used. In some embodiments, the avidin protein of an avidin protein complex is streptavidin in a tetrameric form (e.g., a homotetramer). In some embodiments, the biotin binding sites of an avidin protein provide attachment sites for one or more amino acid recognition molecules, one or more labels, and/or one or more additional shield polymers described herein.

[0178] An illustrative diagram of an avidin protein complex is shown in the inset panel of FIG. 3D. As shown in the inset panel, avidin protein **340** can include a binding site **342** at each of four subunits of the protein which can be bound to a biotin moiety (shown as white circles). The multivalency of avidin protein **340** can allow for various linkage configurations, which are generally shown for illustrative purposes. For example, in some embodiments, a biotin linkage moiety **344** can be used to provide a single point of attachment to avidin protein **340**. In some embodiments, a bis-biotin linkage moiety **346** can be used to provide two points of attachment to avidin protein **340**. As illustrated by avidin-nucleic acid construct **314**, an avidin protein complex may be formed by two bis-biotin linkage moieties, which form a trans-configuration to provide an increased separation distance between a recognition molecule and a label.

[0179] Various further examples of avidin protein shield configurations are shown. A first avidin construct **316** shows an example of an avidin shield attached to a recognition molecule through a bis-biotin linkage moiety and to two labels through separate biotin linkage moieties. A second avidin construct **318** shows an example of an avidin shield attached to two recognition molecules through separate biotin linkage moieties and to a label through a bis-biotin linkage moiety. A third avidin construct **320** shows an example of an avidin shield attached to two recognition molecules through separate biotin linkage moieties and to a labeled nucleic acid

through a biotin linkage moiety of each strand of the nucleic acid. A fourth avidin construct **322** shows an example of an avidin shield attached to a recognition molecule and to a labeled nucleic acid through separate bis-biotin linkage moieties. As shown, the label is further shielded from the recognition molecule by a dendritic polymer between the label and nucleic acid. A fifth avidin construct **324** shows an example of an internal label **326** attached to two avidin-shielded recognition molecules. As shown, each recognition molecule is attached to a different avidin protein through a bis-biotin linkage moiety, and internal label **326** is attached to both avidin proteins through separate bis-biotin linkage moieties.

[0180] It should be appreciated that the example configurations of shielded recognition molecules shown in FIGs. 3A-3D are provided for illustrative purposes. The inventors have conceived of various other shield configurations using one or more different polymers that form a covalent or non-covalent linkage between recognition and label components of a shielded recognition molecule. By way of example, FIG. 3E illustrates the modularity of shield configuration in accordance with the application.

[0181] As shown at the top of FIG. 3E, a shielded recognition molecule generally comprises a recognition component **350**, a shielding element **352**, and a label component **354**. For ease of illustration, recognition component **350** is depicted as one amino acid recognition molecule, and label component **354** is depicted as one label.

[0182] It should be appreciated that shielded recognition molecules of the application can comprise shielding element **352** attached to one or more amino acid recognition molecules and one or more labels. Where recognition component **350** comprises more than one recognition molecule, each recognition molecule can be attached to shielding element **352** at one or more attachment sites on shielding element **352**. In some embodiments, recognition component **350** comprises a single polypeptide fusion construct having tandem copies of two or more amino acid binding proteins, as described elsewhere herein. Where label component **354** comprises more than one label, each label can be attached to shielding element **352** at one or more attachment sites on shielding element **352**. While label component **354** is generically shown as having a single attachment point, it is not limited in this respect. For example, in some embodiments, an internal label having more than one attachment point can be used to join more than one recognition component **350** and/or shielding element **352**, as illustrated by avidin construct **324**.

[0183] In some embodiments, shielding element **352** comprises a protein **360**. In some embodiments, protein **360** is a monovalent or multivalent protein. In some embodiments, protein **360** is a monomeric or multimeric protein, such as a protein homodimer, protein heterodimer, protein oligomer, or other proteinaceous molecule. In some embodiments, shielding element **352** comprises a protein complex formed by a protein non-covalently bound to at least one other

molecule. For example, in some embodiments, shielding element **352** comprises a protein-protein complex **362**. In some embodiments, protein-protein complex **362** comprises one proteinaceous molecule specifically bound to another proteinaceous molecule. In some embodiments, protein-protein complex **362** comprises an antibody or antibody fragment (e.g., scFv) bound to an antigen. In some embodiments, protein-protein complex **362** comprises a receptor bound to a protein ligand. Additional examples of protein-protein complexes include, without limitation, trypsin-aprotinin, barnase-barstar, and colicin E9-Im9 immunity protein.

[0184] In some embodiments, shielding element **352** comprises a protein-ligand complex **364**. In some embodiments, protein-ligand complex **364** comprises a monovalent protein and a non-proteinaceous ligand moiety. For example, in some embodiments, protein-ligand complex **364** comprises an enzyme bound to a small-molecule inhibitor moiety. In some embodiments, protein-ligand complex **364** comprises a receptor bound to a non-proteinaceous ligand moiety.

[0185] In some embodiments, shielding element **352** comprises a multivalent protein complex formed by a multivalent protein non-covalently bound to one or more ligand moieties. In some embodiments, shielding element **352** comprises an avidin protein complex formed by an avidin protein non-covalently bound to one or more biotin linkage moieties. Constructs **366**, **368**, **370**, and **372** provide illustrative examples of avidin protein complexes, any one or more of which may be incorporated into shielding element **352**.

[0186] In some embodiments, shielding element **352** comprises a two-way avidin complex **366** comprising an avidin protein bound to two bis-biotin linkage moieties. In some embodiments, shielding element **352** comprises a three-way avidin complex **368** comprising an avidin protein bound to two biotin linkage moieties and a bis-biotin linkage moiety. In some embodiments, shielding element **352** comprises a four-way avidin complex **370** comprising an avidin protein bound to four biotin linkage moieties.

[0187] In some embodiments, shielding element **352** comprises an avidin protein comprising one or two non-functional binding sites engineered into the avidin protein. For example, in some embodiments, shielding element **352** comprises a divalent avidin complex **372** comprising an avidin protein bound to a biotin linkage moiety at each of two subunits, where the avidin protein comprises a non-functional ligand-binding site **348** at each of two other subunits. As shown, in some embodiments, divalent avidin complex **372** comprises a trans-divalent avidin protein, although a cis-divalent avidin protein may be used depending on a desired implementation. In some embodiments, the avidin protein is a trivalent avidin protein. In some embodiments, the trivalent avidin protein comprises non-functional ligand-binding site **348** at one subunit and is bound to three biotin linkage moieties, or one biotin linkage moiety and one bis-biotin linkage moiety, at the other subunits.

[0188] In some embodiments, shielding element **352** comprises a dendritic polymer **374**. In some embodiments, dendritic polymer **374** is a polyol or a dendrimer, as described elsewhere herein. In some embodiments, dendritic polymer **374** is a branched polyol or a branched dendrimer. In some embodiments, dendritic polymer **374** comprises a monosaccharide-TEG, a disaccharide, an N-acetyl monosaccharide, a TEMPO-TEG, a trolox-TEG, or a glycerol dendrimer. Examples of polyols useful in accordance with shielded recognition molecules of the application include polyether polyols and polyester polyols, e.g., polyethylene glycol, polypropylene glycol, and similar such polymers well known in the art. In some embodiments, dendritic polymer **374** comprises a compound of the following formula: $-(CH_2CH_2O)_n-$, where n is an integer from 1 to 500, inclusive. In some embodiments, dendritic polymer **374** comprises a compound of the following formula: $-(CH_2CH_2O)_n-$, wherein n is an integer from 1 to 100, inclusive.

[0189] In some embodiments, shielding element **352** comprises a nucleic acid. In some embodiments, the nucleic acid is single-stranded. In some embodiments, label component **354** is attached directly or indirectly to one end of the single-stranded nucleic acid (e.g., the 5′ end or the 3′ end) and recognition component **350** is attached directly or indirectly to the other end of the single-stranded nucleic acid (e.g., the 3′ end or the 5′ end). For example, the single-stranded nucleic acid can comprise a label attached to the 5′ end of the nucleic acid and an amino acid recognition molecule attached to the 3′ end of the nucleic acid.

[0190] In some embodiments, shielding element **352** comprises a double-stranded nucleic acid **376**. As shown, in some embodiments, double-stranded nucleic acid **376** can form a non-covalent linkage between recognition component **350** and label component **354** through strand hybridization. However, in some embodiments, double-stranded nucleic acid **376** can form a covalent linkage between recognition component **350** and label component **354** through attachment to the same oligonucleotide strand. In some embodiments, label component **354** is attached directly or indirectly to one end of the double-stranded nucleic acid and recognition component **350** is attached directly or indirectly to the other end of the double-stranded nucleic acid. For example, the double-stranded nucleic acid can comprise a label attached to the 5′ end of one strand and an amino acid recognition molecule attached to the 5′ end of the other strand.

[0191] In some embodiments, shielding element **352** comprises a nucleic acid that forms one or more structural motifs which can be useful for increasing steric bulk of the shield. Examples of nucleic acid structural motifs include, without limitation, stem-loops, three-way junctions (e.g., formed by two or more stem-loop motifs), four-way junctions (e.g., Holliday junctions), and bulge loops.

[0192] In some embodiments, shielding element **352** comprises a nucleic acid that forms a stem-loop **378**. A stem-loop, or hairpin loop, is an unpaired loop of nucleotides on an oligonucleotide strand that is formed when the oligonucleotide strand folds and forms base pairs with another section of the same strand. In some embodiments, the unpaired loop of stem-loop **378** comprises three to ten nucleotides. Accordingly, stem-loop **378** can be formed by two regions of an oligonucleotide strand having inverted complementary sequences that hybridize to form a stem, where the two regions are separated by the three to ten nucleotides that form the unpaired loop. In some embodiments, the stem of stem-loop **378** can be designed to have one or more G/C nucleotides, which can provide added stability with the addition hydrogen bonding interaction that forms compared to A/T nucleotides. In some embodiments, the stem of stem-loop **378** comprises G/C nucleotides immediately adjacent to an unpaired loop sequence. In some embodiments, the stem of stem-loop **378** comprises G/C nucleotides within the first 2, 3, 4, or 5 nucleotides adjacent to an unpaired loop sequence. In some embodiments, an unpaired loop of stem-loop **378** comprises one or more attachment sites. In some embodiments, an attachment site occurs at an abasic site in the unpaired loop. In some embodiments, an attachment site occurs at a base of the unpaired loop.

[0193] In some embodiments, stem-loop **378** is formed by a double-stranded nucleic acid. As described herein, in some embodiments, the double-stranded nucleic acid can form a non-covalent linkage group through strand hybridization of first and second oligonucleotide strands. However, in some embodiments, shielding element **352** comprises a single-stranded nucleic acid that forms a stem-loop motif, e.g., to provide a covalent linkage group. In some embodiments, shielding element **352** comprises a nucleic acid that forms two or more stem-loop motifs. For example, in some embodiments, the nucleic acid comprises two stem-loop motifs. In some embodiments, a stem of one stem-loop motif is adjacent to the stem of the other such that the motifs together form a three-way junction. In some embodiments, shielding element **352** comprises a nucleic acid that forms a four-way junction **378**. In some embodiments, four-way junction **378** is formed through hybridization of two or more oligonucleotide strands (e.g., 2, 3, or 4 oligonucleotide strands).

[0194] In some embodiments, shielding element **352** comprises one or more polymers selected from **360**, **362**, **364**, **366**, **368**, **370**, **372**, **374**, **376**, **378**, and **380** of FIG. 3E. It should be appreciated that the linkage moieties and attachment sites shown on each of **360**, **362**, **364**, **366**, **368**, **370**, **372**, **374**, **376**, **378**, and **380** are shown for illustrative purposes and are not intended to depict a preferred linkage or attachment site configuration.

[0195] In some aspects, the application provides an amino acid recognition molecule of Formula (**II**):

$$A–(Y)_n–D$$

(II),

wherein: A is an amino acid binding component comprising at least one amino acid recognition molecule; each instance of Y is a polymer that forms a covalent or non-covalent linkage group; n is an integer from 1 to 10, inclusive; and D is a label component comprising at least one detectable label. In some embodiments, the application provides a composition comprising a soluble amino acid recognition molecule of Formula (**II**).

[0196] In some embodiments, A comprises a plurality of amino acid recognition molecules. In some embodiments, each amino acid recognition molecule of the plurality is attached to a different attachment site on Y. In some embodiments, at least two amino acid recognition molecules of the plurality are attached to a single attachment site on Y. In some embodiments, the amino acid recognition molecule is a recognition protein or a nucleic acid aptamer, e.g., as described elsewhere herein.

[0197] In some embodiments, the detectable label is a luminescent label or a conductivity label. In some embodiments, the luminescent label comprises at least one fluorophore dye molecule. In some embodiments, D comprises 20 or fewer fluorophore dye molecules. In some embodiments, the ratio of the number of fluorophore dye molecules to the number of amino acid recognition molecules is between 1:1 and 20:1. In some embodiments, the luminescent label comprises at least one FRET pair comprising a donor label and an acceptor label. In some embodiments, the ratio of the donor label to the acceptor label is 1:1, 2:1, 3:1, 4:1, or 5:1. In some embodiments, the ratio of the acceptor label to the donor label is 1:1, 2:1, 3:1, 4:1, or 5:1.

[0198] In some embodiments, D is less than 200 Å in diameter. In some embodiments, $–(Y)_n–$ is at least 2 nm in length. In some embodiments, $–(Y)_n–$ is at least 5 nm in length. In some embodiments, $–(Y)_n–$ is at least 10 nm in length. In some embodiments, each instance of Y is independently a biomolecule, a polyol, or a dendrimer. In some embodiments, the biomolecule is a nucleic acid, a polypeptide, or a polysaccharide.

[0199] In some embodiments, the amino acid recognition molecule is of one of the following formulae:

$$A–Y^1–(Y)_m–D \text{ or } A–(Y)_m–Y^1–D,$$

wherein: Y^1 is a nucleic acid or a polypeptide; and m is an integer from 0 to 10, inclusive.

[0200] In some embodiments, the nucleic acid comprises a first oligonucleotide strand. In some embodiments, the nucleic acid comprises a second oligonucleotide strand hybridized with the first oligonucleotide strand. In some embodiments, the nucleic acid forms a covalent linkage through the first oligonucleotide strand. In some embodiments, the nucleic acid forms a non-covalent linkage through the hybridized first and second oligonucleotide strands.

[0201] In some embodiments, the polypeptide is a monovalent or multivalent protein. In some embodiments, the monovalent or multivalent protein forms at least one non-covalent linkage through a ligand moiety attached to a ligand-binding site of the monovalent or multivalent protein. In some embodiments, A, Y, or D comprises the ligand moiety.

[0202] In some embodiments, the amino acid recognition molecule is of one of the following formulae:

$$A–(Y)_m–Y^2–D \text{ or } A–Y^2–(Y)_m–D,$$

wherein: Y^2 is a polyol or dendrimer; and m is an integer from 0 to 10, inclusive. In some embodiments, the polyol or dendrimer comprises polyethylene glycol, tetraethylene glycol, poly(amidoamine), poly(propyleneimine), poly(propyleneamine), carbosilane, poly(L-lysine), or a combination of one or more thereof.

[0203] In some aspects, the application provides an amino acid recognition molecule of Formula (**III**):

$$A–Y^1–D$$

(**III**),

wherein: A is an amino acid binding component comprising at least one amino acid recognition molecule; Y^1 is a nucleic acid or a polypeptide; D is a label component comprising at least one detectable label. In some embodiments, when Y^1 is a nucleic acid, the nucleic acid forms a covalent or non-covalent linkage group. In some embodiments, when Y^1 is a polypeptide, the polypeptide forms a non-covalent linkage group characterized by a dissociation constant (K_D) of less than 50×10^{-9} M.

[0204] In some embodiments, Y^1 is a nucleic acid comprising a first oligonucleotide strand. In some embodiments, the nucleic acid comprises a second oligonucleotide strand hybridized with the first oligonucleotide strand. In some embodiments, A is attached to the first oligonucleotide strand, and wherein D is attached to the second oligonucleotide strand. In some embodiments, A is attached to a first attachment site on the first oligonucleotide strand, and wherein D is attached to a second attachment site on the first oligonucleotide strand. In some embodiments, each oligonucleotide strand of the nucleic acid comprises fewer than 150, fewer than 100, or fewer than 50 nucleotides.

[0205] In some embodiments, Y^1 is a monovalent or multivalent protein. In some embodiments, the monovalent or multivalent protein forms at least one non-covalent linkage through a ligand moiety attached to a ligand-binding site of the monovalent or multivalent protein. In some embodiments, at least one of A and D comprises the ligand moiety. In some embodiments, the polypeptide is an avidin protein (e.g., avidin, streptavidin, traptavidin, tamavidin, bradavidin,

xenavidin, or a homolog or variant thereof). In some embodiments, the ligand moiety is a biotin moiety.

[0206] In some embodiments, the amino acid recognition molecule is of one of the following formulae:

$$A–Y^1–(Y)_n–D \text{ or } A–(Y)_n–Y^1–D,$$

wherein: each instance of Y is a polymer that forms a covalent or non-covalent linkage group; and n is an integer from 1 to 10, inclusive. In some embodiments, each instance of Y is independently a biomolecule, a polyol, or a dendrimer.

[0207] In other aspects, the application provides an amino acid recognition molecule comprising: a nucleic acid; at least one amino acid recognition molecule attached to a first attachment site on the nucleic acid; and at least one detectable label attached to a second attachment site on the nucleic acid. In some embodiments, the nucleic acid forms a covalent or non-covalent linkage group between the at least one amino acid recognition molecule and the at least one detectable label.

[0208] In some embodiments, the nucleic acid is a double-stranded nucleic acid comprising a first oligonucleotide strand hybridized with a second oligonucleotide strand. In some embodiments, the first attachment site is on the first oligonucleotide strand, and wherein the second attachment site is on the second oligonucleotide strand. In some embodiments, the at least one amino acid recognition molecule is attached to the first attachment site through a protein that forms a covalent or non-covalent linkage group between the at least one amino acid recognition molecule and the nucleic acid. In some embodiments, the at least one detectable label is attached to the second attachment site through a protein that forms a covalent or non-covalent linkage group between the at least one detectable label and the nucleic acid. In some embodiments, the first and second attachment sites are separated by between 5 and 100 nucleotide bases or nucleotide base pairs on the nucleic acid.

[0209] In yet other aspects, the application provides an amino acid recognition molecule comprising: a multivalent protein comprising at least two ligand-binding sites; at least one amino acid recognition molecule attached to the protein through a first ligand moiety bound to a first ligand-binding site on the protein; and at least one detectable label attached to the protein through a second ligand moiety bound to a second ligand-binding site on the protein.

[0210] In some embodiments, the multivalent protein is an avidin protein comprising four ligand-binding sites. In some embodiments, the ligand-binding sites are biotin binding sites, and wherein the ligand moieties are biotin moieties. In some embodiments, at least one of the biotin moieties is a bis-biotin moiety, and wherein the bis-biotin moiety is bound to two biotin binding sites on the avidin protein. In some embodiments, the at least one amino acid recognition

molecule is attached to the protein through a nucleic acid comprising the first ligand moiety. In some embodiments, the at least one detectable label is attached to the protein through a nucleic acid comprising the second ligand moiety.

[0211] As described elsewhere herein, shielded recognition molecules of the application may be used in a polypeptide sequencing method in accordance with the application, or any method known in the art. For example, in some embodiments, a shielded recognition molecule provided herein may be used in an Edman-type degradation reaction provided herein, or conventionally known in the art, which can involve iterative cycling of multiple reaction mixtures in a polypeptide sequencing reaction. In some embodiments, a shielded recognition molecule provided herein may be used in a dynamic sequencing reaction of the application, which involves amino acid recognition and degradation in a single reaction mixture.

Polypeptide Sequencing

[0212] In addition to methods of identifying a terminal amino acid of a polypeptide, the application provides methods of sequencing polypeptides using labeled recognition molecules. In some embodiments, methods of sequencing may involve subjecting a polypeptide terminus to repeated cycles of terminal amino acid detection and terminal amino acid cleavage. For example, in some embodiments, the application provides a method of determining an amino acid sequence of a polypeptide comprising contacting a polypeptide with one or more labeled recognition molecules described herein and subjecting the polypeptide to Edman degradation.

[0213] Conventional Edman degradation involves repeated cycles of modifying and cleaving the terminal amino acid of a polypeptide, wherein each successively cleaved amino acid is identified to determine an amino acid sequence of the polypeptide. As an illustrative example of a conventional Edman degradation, the N-terminal amino acid of a polypeptide is modified using phenyl isothiocyanate (PITC) to form a PITC-derivatized N-terminal amino acid. The PITC-derivatized N-terminal amino acid is then cleaved using acidic conditions, basic conditions, and/or elevated temperatures. It has also been shown that the step of cleaving the PITC-derivatized N-terminal amino acid may be accomplished enzymatically using a modified cysteine protease from the protozoa *Trypanosoma cruzi*, which involves relatively milder cleavage conditions at a neutral or near-neutral pH. Non-limiting examples of useful enzymes are described in U.S. Patent Application No. 15/255,433, filed September 2, 2016, titled "MOLECULES AND METHODS FOR ITERATIVE POLYPEPTIDE ANALYSIS AND PROCESSING."

[0214] An example of sequencing by Edman degradation using labeled recognition molecules in accordance with the application is depicted in FIG. 4. In some embodiments, sequencing by

Edman degradation comprises providing a polypeptide **420** that is immobilized to a surface **430** of a solid support (e.g., immobilized to a bottom or sidewall surface of a sample well) through a linker **424**. In some embodiments, as described herein, polypeptide **420** is immobilized at one terminus (e.g., an amino-terminal amino acid or a carboxy-terminal amino acid) such that the other terminus is free for detecting and cleaving of a terminal amino acid. Accordingly, in some embodiments, the reagents used in Edman degradation methods described herein preferentially interact with terminal amino acids at the non-immobilized (e.g., free) terminus of polypeptide **420**. In this way, polypeptide **420** remains immobilized over repeated cycles of detecting and cleaving. To this end, in some embodiments, linker **424** may be designed according to a desired set of conditions used for detecting and cleaving, e.g., to limit detachment of polypeptide **420** from surface **430** under chemical cleavage conditions. Suitable linker compositions and techniques for immobilizing a polypeptide to a surface are described in detail elsewhere herein.

[0215] In accordance with the application, in some embodiments, a method of sequencing by Edman degradation comprises a step (1) of contacting polypeptide **420** with one or more labeled recognition molecules that selectively bind one or more types of terminal amino acids. As shown, in some embodiments, a labeled recognition molecule **400** interacts with polypeptide **420** by selectively binding the terminal amino acid. In some embodiments, step (1) further comprises removing any of the one or more labeled recognition molecules that do not selectively bind the terminal amino acid (e.g., the free terminal amino acid) of polypeptide **420**.

[0216] In some embodiments, the method further comprises identifying the terminal amino acid of polypeptide **420** by detecting labeled recognition molecule **400**. In some embodiments, detecting comprises detecting a luminescence from labeled recognition molecule **400**. As described herein, in some embodiments, the luminescence is uniquely associated with labeled recognition molecule **400**, and the luminescence is thereby associated with the type of amino acid to which labeled recognition molecule **400** selectively binds. As such, in some embodiments, the type of amino acid is identified by determining one or more luminescence properties of labeled recognition molecule **400**.

[0217] In some embodiments, a method of sequencing by Edman degradation comprises a step (2) of removing the terminal amino acid of polypeptide **420**. In some embodiments, step (2) comprises removing labeled recognition molecule **400** (e.g., any of the one or more labeled recognition molecules that selectively bind the terminal amino acid) from polypeptide **420**. In some embodiments, step (2) comprises modifying the terminal amino acid (e.g., the free terminal amino acid) of polypeptide **420** by contacting the terminal amino acid with an isothiocyanate (e.g., PITC) to form an isothiocyanate-modified terminal amino acid. In some embodiments, an

isothiocyanate-modified terminal amino acid is more susceptible to removal by a cleaving reagent (e.g., a chemical or enzymatic cleaving reagent) than an unmodified terminal amino acid.

[0218] In some embodiments, step (2) comprises removing the terminal amino acid by contacting polypeptide **420** with a protease **440** that specifically binds and cleaves the isothiocyanate-modified terminal amino acid. In some embodiments, protease **440** comprises a modified cysteine protease. In some embodiments, protease **440** comprises a modified cysteine protease, such as a cysteine protease from *Trypanosoma cruzi* (see, e.g., Borgo, et al. (2015) *Protein Science* 24:571-579). In yet other embodiments, step (2) comprises removing the terminal amino acid by subjecting polypeptide **420** to chemical (e.g., acidic, basic) conditions sufficient to cleave the isothiocyanate-modified terminal amino acid.

[0219] In some embodiments, a method of sequencing by Edman degradation comprises a step (3) of washing polypeptide **420** following terminal amino acid cleavage. In some embodiments, washing comprises removing protease **440**. In some embodiments, washing comprises restoring polypeptide **420** to neutral pH conditions (e.g., following chemical cleavage by acidic or basic conditions). In some embodiments, a method of sequencing by Edman degradation comprises repeating steps (1) through (3) for a plurality of cycles.

[0220] In some aspects, the application provides methods of polypeptide sequencing in real-time by evaluating binding interactions of terminal amino acids with labeled amino acid recognition molecules and a labeled cleaving reagent (e.g., a labeled non-specific exopeptidase). FIG. 5 shows an example of a method of sequencing in which discrete association events give rise to signal pulses of a signal output **500**. The inset panel of FIG. 5 illustrates a general scheme of real-time sequencing by this approach. As shown, a labeled recognition molecule **510** selectively associates with (e.g., binds to) and dissociates from a terminal amino acid (shown here as lysine), which gives rise to a series of pulses in signal output **500** which may be used to identify the terminal amino acid. In some embodiments, the series of pulses provide a pulsing pattern (e.g., a characteristic pattern) which may be diagnostic of the identity of the corresponding terminal amino acid.

[0221] Without wishing to be bound by theory, labeled recognition molecule **510** selectively binds according to a binding affinity (K_D) defined by an association rate, or an "on" rate, of binding (k_{on}) and a dissociation rate, or an "off" rate, of binding (k_{off}). The rate constants k_{off} and k_{on} are the critical determinants of pulse duration (e.g., the time corresponding to a detectable association event) and interpulse duration (e.g., the time between detectable association events), respectively. In some embodiments, these rates can be engineered to achieve pulse durations and pulse rates (e.g., the frequency of signal pulses) that give the best sequencing accuracy.

[0222] As shown in the inset panel, a sequencing reaction mixture further comprises a labeled non-specific exopeptidase **520** comprising a luminescent label that is different than that of labeled recognition molecule **510**. In some embodiments, labeled non-specific exopeptidase **520** is present in the mixture at a concentration that is less than that of labeled recognition molecule **510**. In some embodiments, labeled non-specific exopeptidase **520** displays broad specificity such that it cleaves most or all types of terminal amino acids. Accordingly, a dynamic sequencing approach can involve monitoring recognition molecule binding at a terminus of a polypeptide over the course of a degradation reaction catalyzed by exopeptidase cleavage activity.

[0223] As illustrated by the progress of signal output **500**, in some embodiments, terminal amino acid cleavage by labeled non-specific exopeptidase **520** gives rise to a signal pulse, and these events occur with lower frequency than the binding pulses of a labeled recognition molecule **510**. In this way, amino acids of a polypeptide may be counted and/or identified in a real-time sequencing process. As further illustrated in signal output **500**, in some embodiments, a plurality of labeled recognition molecules may be used, each with a diagnostic pulsing pattern (e.g., characteristic pattern) which may be used to identify a corresponding terminal amino acid. For example, in some embodiments, different characteristic patterns (as illustrated by each of lysine, phenylalanine, and glutamine in signal output **500**) correspond to the association of more than one labeled recognition molecule with different types of terminal amino acids. As described herein, it should be appreciated that a single recognition molecule that associates with more than one type of amino acid may be used in accordance with the application. Accordingly, in some embodiments, different characteristic patterns correspond to the association of one labeled recognition molecule with different types of terminal amino acids.

[0224] As described herein, signal pulse information may be used to identify an amino acid based on a characteristic pattern in a series of signal pulses. In some embodiments, a characteristic pattern comprises a plurality of signal pulses, each signal pulse comprising a pulse duration. In some embodiments, the plurality of signal pulses may be characterized by a summary statistic (e.g., mean, median, time decay constant) of the distribution of pulse durations in a characteristic pattern. In some embodiments, the mean pulse duration of a characteristic pattern is between about 1 millisecond and about 10 seconds (e.g., between about 1 ms and about 1 s, between about 1 ms and about 100 ms, between about 1 ms and about 10 ms, between about 10 ms and about 10 s, between about 100 ms and about 10 s, between about 1 s and about 10 s, between about 10 ms and about 100 ms, or between about 100 ms and about 500 ms). In some embodiments, the mean pulse duration is between about 50 milliseconds and about 2 seconds,

between about 50 milliseconds and about 500 milliseconds, or between about 500 milliseconds and about 2 seconds.

[0225] In some embodiments, different characteristic patterns corresponding to different types of amino acids in a single polypeptide may be distinguished from one another based on a statistically significant difference in the summary statistic. For example, in some embodiments, one characteristic pattern may be distinguishable from another characteristic pattern based on a difference in mean pulse duration of at least 10 milliseconds (e.g., between about 10 ms and about 10 s, between about 10 ms and about 1 s, between about 10 ms and about 100 ms, between about 100 ms and about 10 s, between about 1 s and about 10 s, or between about 100 ms and about 1 s). In some embodiments, the difference in mean pulse duration is at least 50 ms, at least 100 ms, at least 250 ms, at least 500 ms, or more. In some embodiments, the difference in mean pulse duration is between about 50 ms and about 1 s, between about 50 ms and about 500 ms, between about 50 ms and about 250 ms, between about 100 ms and about 500 ms, between about 250 ms and about 500 ms, or between about 500 ms and about 1 s. In some embodiments, the mean pulse duration of one characteristic pattern is different from the mean pulse duration of another characteristic pattern by about 10-25%, 25-50%, 50-75%, 75-100%, or more than 100%, for example by about 2-fold, 3-fold, 4-fold, 5-fold, or more. It should be appreciated that, in some embodiments, smaller differences in mean pulse duration between different characteristic patterns may require a greater number of pulse durations within each characteristic pattern to distinguish one from another with statistical confidence.

[0226] In some embodiments, a characteristic pattern generally refers to a plurality of association events between an amino acid of a polypeptide and a means for binding the amino acid (e.g., an amino acid recognition molecule). In some embodiments, a characteristic pattern comprises at least 10 association events (e.g., at least 25, at least 50, at least 75, at least 100, at least 250, at least 500, at least 1,000, or more, association events). In some embodiments, a characteristic pattern comprises between about 10 and about 1,000 association events (e.g., between about 10 and about 500 association events, between about 10 and about 250 association events, between about 10 and about 100 association events, or between about 50 and about 500 association events). In some embodiments, the plurality of association events is detected as a plurality of signal pulses.

[0227] In some embodiments, a characteristic pattern refers to a plurality of signal pulses which may be characterized by a summary statistic as described herein. In some embodiments, a characteristic pattern comprises at least 10 signal pulses (e.g., at least 25, at least 50, at least 75, at least 100, at least 250, at least 500, at least 1,000, or more, signal pulses). In some embodiments, a characteristic pattern comprises between about 10 and about 1,000 signal pulses

(e.g., between about 10 and about 500 signal pulses, between about 10 and about 250 signal pulses, between about 10 and about 100 signal pulses, or between about 50 and about 500 signal pulses).

[0228] In some embodiments, a characteristic pattern refers to a plurality of association events between an amino acid recognition molecule and an amino acid of a polypeptide occurring over a time interval prior to removal of the amino acid (e.g., a cleavage event). In some embodiments, a characteristic pattern refers to a plurality of association events occurring over a time interval between two cleavage events (e.g., prior to removal of the amino acid and after removal of an amino acid previously exposed at the terminus). In some embodiments, the time interval of a characteristic pattern is between about 1 minute and about 30 minutes (e.g., between about 1 minute and about 20 minutes, between about 1 minute and 10 minutes, between about 5 minutes and about 20 minutes, between about 5 minutes and about 15 minutes, or between about 5 minutes and about 10 minutes).

[0229] In some embodiments, polypeptide sequencing reaction conditions can be configured to achieve a time interval that allows for sufficient association events which provide a desired confidence level with a characteristic pattern. This can be achieved, for example, by configuring the reaction conditions based on various properties, including: reagent concentration, molar ratio of one reagent to another (e.g., ratio of amino acid recognition molecule to cleaving reagent, ratio of one recognition molecule to another, ratio of one cleaving reagent to another), number of different reagent types (e.g., the number of different types of recognition molecules and/or cleaving reagents, the number of recognition molecule types relative to the number of cleaving reagent types), cleavage activity (e.g., peptidase activity), binding properties (e.g., kinetic and/or thermodynamic binding parameters for recognition molecule binding), reagent modification (e.g., polyol and other protein modifications which can alter interaction dynamics), reaction mixture components (e.g., one or more components, such as pH, buffering agent, salt, divalent cation, surfactant, and other reaction mixture components described herein), temperature of the reaction, and various other parameters apparent to those skilled in the art, and combinations thereof. The reaction conditions can be configured based on one or more aspects described herein, including, for example, signal pulse information (e.g., pulse duration, interpulse duration, change in magnitude), labeling strategies (e.g., number and/or type of fluorophore, linkers with or without shielding element), surface modification (e.g., modification of sample well surface, including polypeptide immobilization), sample preparation (e.g., polypeptide fragment size, polypeptide modification for immobilization), and other aspects described herein.

[0230] In some embodiments, a polypeptide sequencing reaction in accordance with the application is performed under conditions in which recognition and cleavage of amino acids can

occur simultaneously in a single reaction mixture. For example, in some embodiments, a polypeptide sequencing reaction is performed in a reaction mixture having a pH at which association events and cleavage events can occur. In some embodiments, a polypeptide sequencing reaction is performed in a reaction mixture at a pH of between about 6.5 and about 9.0. In some embodiments, a polypeptide sequencing reaction is performed in a reaction mixture at a pH of between about 7.0 and about 8.5 (e.g., between about 7.0 and about 8.0, between about 7.5 and about 8.5, between about 7.5 and about 8.0, or between about 8.0 and about 8.5).

[0231] In some embodiments, a polypeptide sequencing reaction is performed in a reaction mixture comprising one or more buffering agents. In some embodiments, a reaction mixture comprises a buffering agent in a concentration of at least 10 mM (e.g., at least 20 mM and up to 250 mM, at least 50 mM, 10-250 mM, 10-100 mM, 20-100 mM, 50-100 mM, or 100-200 mM). In some embodiments, a reaction mixture comprises a buffering agent in a concentration of between about 10 mM and about 50 mM (e.g., between about 10 mM and about 25 mM, between about 25 mM and about 50 mM, or between about 20 mM and about 40 mM). Examples of buffering agents include, without limitation, HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), Tris (tris(hydroxymethyl)aminomethane), and MOPS (3-(N-morpholino)propanesulfonic acid).

[0232] In some embodiments, a polypeptide sequencing reaction is performed in a reaction mixture comprising salt in a concentration of at least 10 mM. In some embodiments, a reaction mixture comprises salt in a concentration of at least 10 mM (e.g., at least 20 mM, at least 50 mM, at least 100 mM, or more). In some embodiments, a reaction mixture comprises salt in a concentration of between about 10 mM and about 250 mM (e.g., between about 20 mM and about 200 mM, between about 50 mM and about 150 mM, between about 10 mM and about 50 mM, or between about 10 mM and about 100 mM). Examples of salts include, without limitation, sodium salts, potassium salts, and acetates, such as sodium chloride (NaCl), sodium acetate (NaOAc), and potassium acetate (KOAc).

[0233] Additional examples of components for use in a reaction mixture include divalent cations (e.g., Mg^{2+}, Co^{2+}) and surfactants (e.g., polysorbate 20). In some embodiments, a reaction mixture comprises a divalent cation in a concentration of between about 0.1 mM and about 50 mM (e.g., between about 10 mM and about 50 mM, between about 0.1 mM and about 10 mM, or between about 1 mM and about 20 mM). In some embodiments, a reaction mixture comprises a surfactant in a concentration of at least 0.01% (e.g., between about 0.01% and about 0.10%). In some embodiments, a reaction mixture comprises one or more components useful in single-molecule analysis, such as an oxygen-scavenging system (e.g., a PCA/PCD system or a Pyranose

oxidase/Catalase/glucose system) and/or one or more triplet state quenchers (e.g., trolox, COT, and NBA).

[0234] In some embodiments, a polypeptide sequencing reaction is performed at a temperature at which association events and cleavage events can occur. In some embodiments, a polypeptide sequencing reaction is performed at a temperature of at least 10 °C. In some embodiments, a polypeptide sequencing reaction is performed at a temperature of between about 10 °C and about 50 °C (e.g., 15-45 °C, 20-40 °C, at or around 25 °C, at or around 30 °C, at or around 35 °C, at or around 37 °C). In some embodiments, a polypeptide sequencing reaction is performed at or around room temperature.

[0235] As detailed above, a real-time sequencing process as illustrated by FIG. 5 can generally involve cycles of terminal amino acid recognition and terminal amino acid cleavage, where the relative occurrence of recognition and cleavage can be controlled by a concentration differential between a labeled recognition molecule **510** and a labeled non-specific exopeptidase **520**. In some embodiments, the concentration differential can be optimized such that the number of signal pulses detected during recognition of an individual amino acid provides a desired confidence interval for identification. For example, if an initial sequencing reaction provides signal data with too few signal pulses between cleavage events to permit determination of characteristic patterns with a desired confidence interval, the sequencing reaction can be repeated using a decreased concentration of non-specific exopeptidase relative to recognition molecule.

[0236] In some embodiments, polypeptide sequencing in accordance with the application may be carried out by contacting a polypeptide with a sequencing reaction mixture comprising one or more amino acid recognition molecules and/or one or more cleaving reagents (e.g., peptidases). In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of between about 10 nM and about 10 μM. In some embodiments, a sequencing reaction mixture comprises a cleaving reagent at a concentration of between about 500 nM and about 500 μM.

[0237] In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of between about 100 nM and about 10 μM, between about 250 nM and about 10 μM, between about 100 nM and about 1 μM, between about 250 nM and about 1 μM, between about 250 nM and about 750 nM, or between about 500 nM and about 1 μM. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of about 100 nM, about 250 nM, about 500 nM, about 750 nM, or about 1 μM.

[0238] In some embodiments, a sequencing reaction mixture comprises a cleaving reagent at a concentration of between about 500 nM and about 250 μM, between about 500 nM and about

100 μM, between about 1 μM and about 100 μM, between about 500 nM and about 50 μM, between about 1 μM and about 100 μM, between about 10 μM and about 200 μM, or between about 10 μM and about 100 μM. In some embodiments, a sequencing reaction mixture comprises a cleaving reagent at a concentration of about 1 μM, about 5 μM, about 10 μM, about 30 μM, about 50 μM, about 70 μM, or about about 100 μM.

[0239] In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of between about 10 nM and about 10 μM, and a cleaving reagent at a concentration of between about 500 nM and about 500 μM. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of between about 100 nM and about 1 μM, and a cleaving reagent at a concentration of between about 1 μM and about 100 μM. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of between about 250 nM and about 1 μM, and a cleaving reagent at a concentration of between about 10 μM and about 100 μM. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule at a concentration of about 500 nM, and a cleaving reagent at a concentration of between about 25 μM and about 75 μM. In some embodiments, the concentration of an amino acid recognition molecule and/or the concentration of a cleaving reagent in a reaction mixture is as described elsewhere herein.

[0240] In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule and a cleaving reagent in a molar ratio of about 500:1, about 400:1, about 300:1, about 200:1, about 100:1, about 75:1, about 50:1, about 25:1, about 10:1, about 5:1, about 2:1, or about 1:1. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule and a cleaving reagent in a molar ratio of between about 10:1 and about 200:1. In some embodiments, a sequencing reaction mixture comprises an amino acid recognition molecule and a cleaving reagent in a molar ratio of between about 50:1 and about 150:1. In some embodiments, the molar ratio of an amino acid recognition molecule to a cleaving reagent in a reaction mixture is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1 (e.g., 1:1,000, about 1:500, about 1:200, about 1:100, about 1:10, about 1:5, about 1:2, about 1:1, about 5:1, about 10:1, about 50:1, about 100:1). In some embodiments, the molar ratio of an amino acid recognition molecule to a cleaving reagent in a reaction mixture is between about 1:100 and about 1:1 or between about 1:1 and about 10:1. In some embodiments, the molar ratio of an amino acid recognition molecule to a cleaving reagent in a reaction mixture is as described elsewhere herein.

[0241] In some embodiments, a sequencing reaction mixture comprises one or more amino acid recognition molecules and one or more cleaving reagents. In some embodiments, a sequencing

reaction mixture comprises at least three amino acid recognition molecules and at least one cleaving reagent. In some embodiments, the sequencing reaction mixture comprises two or more cleaving reagents. In some embodiments, the sequencing reaction mixture comprises at least one and up to ten cleaving reagents (e.g., 1-3 cleaving reagents, 2-10 cleaving reagents, 1-5 cleaving reagents, 3-10 cleaving reagents). In some embodiments, the sequencing reaction mixture comprises at least three and up to thirty amino acid recognition molecules (e.g., between 3 and 25, between 3 and 20, between 3 and 10, between 3 and 5, between 5 and 30, between 5 and 20, between 5 and 10, or between 10 and 20, amino acid recognition molecules). In some embodiments, the one or more amino acid recognition molecules include at least one recognition molecule selected from Table 1 or Table 2. In some embodiments, the one or more cleaving reagents include at least one peptidase molecule selected from Table 4.

[0242] In some embodiments, a sequencing reaction mixture comprises more than one amino acid recognition molecule and/or more than one cleaving reagent. In some embodiments, a sequencing reaction mixture described as comprising more than one amino acid recognition molecule (or cleaving reagent) refers to the mixture as having more than one type of amino acid recognition molecule (or cleaving reagent). For example, in some embodiments, a sequencing reaction mixture comprises two or more amino acid binding proteins. In some embodiments, the two or more amino acid binding proteins refer to two or more types of amino acid binding proteins. In some embodiments, one type of amino acid binding protein has an amino acid sequence that is different from another type of amino acid binding protein in the reaction mixture. In some embodiments, one type of amino acid binding protein has a label that is different from a label of another type of amino acid binding protein in the reaction mixture. In some embodiments, one type of amino acid binding protein associates with (e.g., binds to) an amino acid that is different from an amino acid with which another type of amino acid binding protein in the reaction mixture associates. In some embodiments, one type of amino acid binding protein associates with (e.g., binds to) a subset of amino acids that is different from a subset of amino acids with which another type of amino acid binding protein in the reaction mixture associates.

[0243] While the example illustrated by FIG. 5 relates to a sequencing process using a labeled cleaving reagent, the sequencing process is not intended to be limited in this respect. As described elsewhere herein, the inventors have demonstrated single-molecule sequencing using an unlabeled cleaving reagent. In some embodiments, the approximate frequency with which a cleaving reagent removes successive terminal amino acids is known, e.g., based on a known activity and/or concentration of the enzyme being used. In some embodiments, terminal amino

acid cleavage by the reagent is inferred, e.g., based on signal detected for amino acid recognition or a lack of signal detected.

[0244] The inventors have recognized further techniques for controlling real-time sequencing reactions, which may be used in combination with, or alternatively to, the concentration differential approach as described. An example of a temperature-dependent real-time sequencing process is shown in FIG. 6. Panels (I) through (III) illustrate a sequencing reaction involving cycles of temperature-dependent terminal amino acid recognition and terminal amino acid cleavage. Each cycle of the sequencing reaction is carried out over two temperature ranges: a first temperature range ("T_1") that is optimal for recognition molecule activity over exopeptidase activity (e.g., to promote terminal amino acid recognition), and a second temperature range ("T_2") that is optimal for exopeptidase activity over recognition molecule activity (e.g., to promote terminal amino acid cleavage). The sequencing reaction progresses by alternating the reaction mixture temperature between the first temperature range T_1 (to initiate amino acid recognition) and the second temperature range T_2 (to initiate amino acid cleavage). Accordingly, progression of a temperature-dependent sequencing process is controllable by temperature, and alternating between different temperature ranges (e.g., between T_1 and T_2) may be carried through manual or automated processes. In some embodiments, recognition molecule activity (e.g., binding affinity (K_D) for an amino acid) within the first temperature range T_1 as compared to the second temperature range T_2 is increased by at least 10-fold, at least 100-fold, at least 1,000-fold, at least 10,000-fold, at least 100,000-fold, or more. In some embodiments, exopeptidase activity (e.g., rate of substrate conversion to cleavage product) within the second temperature range T_2 as compared to the first temperature range T_1 is increased by at least 2-fold, 10-fold, at least 25-fold, at least 50-fold, at least 100-fold, at least 1,000-fold, or more.

[0245] In some embodiments, the first temperature range T_1 is lower than the second temperature range T_2. In some embodiments, the first temperature range T_1 is between about 15 °C and about 40 °C (e.g., between about 25 °C and about 35 °C, between about 15 °C and about 30 °C, between about 20 °C and about 30 °C). In some embodiments, the second temperature range T_2 is between about 40 °C and about 100 °C (e.g., between about 50 °C and about 90 °C, between about 60 °C and about 90 °C, between about 70 °C and about 90 °C). In some embodiments, the first temperature range T_1 is between about 20 °C and about 40 °C (e.g., approximately 30 °C), and the second temperature range T_2 is between about 60 °C and about 100 °C (e.g., approximately 80 °C).

[0246] In some embodiments, the first temperature range T_1 is higher than the second temperature range T_2. In some embodiments, the first temperature range T_1 is between about 40 °C and about 100 °C (e.g., between about 50 °C and about 90 °C, between about 60 °C and about

90 °C, between about 70 °C and about 90 °C). In some embodiments, the second temperature range T_2 is between about 15 °C and about 40 °C (e.g., between about 25 °C and about 35 °C, between about 15 °C and about 30 °C, between about 20 °C and about 30 °C). In some embodiments, the first temperature range T_1 is between about 60 °C and about 100 °C (e.g., approximately 80 °C), and the second temperature range T_2 is between about 20 °C and about 40 °C (e.g., approximately 30 °C).

[0247] Panel (I) depicts a sequencing reaction mixture at a temperature that is within a first temperature range T_1 which is optimal for recognition molecule activity over exopeptidase activity. For illustrative purposes, a polypeptide of amino acid sequence "KFVAG..." is shown. When the reaction mixture temperature is within the first temperature range T_1, labeled recognition molecules in the mixture are activated (e.g., renatured) to initiate amino acid recognition by associating with the polypeptide terminus. Also within the first temperature range T_1, labeled exopeptidases in the mixture are inactivated (e.g., denatured) to prevent amino acid cleavage during recognition. In panel (I), a first recognition molecule is shown reversibly associating with lysine at the polypeptide terminus, while a labeled exopeptidase (e.g., Pfu aminopeptidase I (Pfu API)) is shown denatured. In some embodiments, amino acid recognition occurs for a predetermined duration of time before initiating cleavage of the amino acid. In some embodiments, amino acid recognition occurs for a duration of time required to reach a desired confidence interval for identification before initiating cleavage of the amino acid. Following amino acid recognition, the reaction proceeds by changing the temperature of the mixture to within a second temperature range T_2.

[0248] Panel (II) depicts the sequencing reaction mixture at a temperature that is within a second temperature range T_2 which is optimal for exopeptidase activity over recognition molecule activity. For illustrative purposes of this example, the second temperature range T_2 is higher than the first temperature range T_1, although it should be appreciated that reagent activity may be optimized for any desired temperature range. Accordingly, progression from panel (I) to panel (II) is carried out by raising the reaction mixture temperature using a suitable source of heat. When the reaction mixture reaches a temperature that is within the second temperature range T_2, labeled exopeptidases in the mixture are activated (e.g., renatured) to initiate terminal amino acid cleavage by exopeptidase activity. Also within the second temperature range T_2, labeled recognition molecules in the mixture are inactivated (e.g., denatured) to prevent amino acid recognition during cleavage. In panel (II), a labeled exopeptidase is shown cleaving the terminal lysine residue, while labeled recognition molecules are denatured. In some embodiments, amino acid cleavage occurs for a predetermined duration of time before initiating recognition of a successive amino acid at the polypeptide terminus. In some embodiments, amino acid cleavage

occurs for a duration of time required to detect cleavage before initiating recognition of a successive amino acid. Following amino acid cleavage, the reaction proceeds by changing the temperature of the mixture to within the first temperature range T_1.

[0249] Panel (III) depicts the beginning of the next cycle in the sequencing reaction, wherein the reaction mixture temperature has been reduced back to within the first temperature range T_1. Accordingly, in this example, progression from panel (II) to panel (III) can be carried out by removing the reaction mixture from the source of heat or otherwise cooling the reaction mixture (e.g., actively or passively) to within the first temperature range T_1. As shown, labeled recognition molecules are renatured, including a second recognition molecule that reversibly associates with phenylalanine at the polypeptide terminus, while the labeled exopeptidase is shown denatured. The sequencing reaction continues by further cycling between amino acid recognition and amino acid cleavage in a temperature-dependent fashion as illustrated by this example.

[0250] Accordingly, a dynamic sequencing approach can involve reaction cycling that is controlled at the level of protein activity or function of one or more proteins within a reaction mixture. It should be appreciated that the temperature-dependent polypeptide sequencing process depicted in FIG. 6 and described above may be illustrative of a general approach to polypeptide sequencing by controllable cycling of condition-dependent recognition and cleavage. For example, in some embodiments, the application provides a luminescence-dependent sequencing process using luminescence-activated reagents. In some embodiments, a luminescence-dependent sequencing process involves cycles of luminescence-dependent amino acid recognition and cleavage. Each cycle of the sequencing reaction may be carried out by exposing a sequencing reaction mixture to two different luminescent conditions: a first luminescent condition that is optimal for recognition molecule activity over exopeptidase activity (e.g., to promote amino acid recognition), and a second luminescent condition that is optimal for exopeptidase activity over recognition molecule activity (e.g., to promote amino acid cleavage). The sequencing reaction progresses by alternating between exposing the reaction mixture to the first luminescent condition (to initiate amino acid recognition) and exposing the reaction mixture to the second luminescent condition (to initiate amino acid cleavage). By way of example and not limitation, in some embodiments, the two different luminescent conditions comprise a first wavelength and a second wavelength.

[0251] In some aspects, the application provides methods of polypeptide sequencing in real-time by evaluating binding interactions of one or more labeled recognition molecules with terminal and internal amino acids and binding interactions of a labeled non-specific exopeptidase with terminal amino acids. FIG. 7 shows an example of a method of sequencing in which the method

described and illustrated for the approach in FIGs. 5-6 is modified by using a labeled recognition molecule **710** that selectively binds to and dissociates from one type of amino acid (shown here as lysine) at both terminal and internal positions (FIG. 7, inset panel). As described in the previous approach, the selective binding gives rise to a series of pulses in signal output **700**. In this approach, however, the series of pulses occur at a rate that is determined by the number of the type of amino acid throughout the polypeptide. Accordingly, in some embodiments, the rate of pulsing corresponding to association events would be diagnostic of the number of cognate amino acids currently present in the polypeptide.

[0252] As in the previous approach, a labeled non-specific peptidase **720** would be present at a relatively lower concentration than labeled recognition molecule **710**, e.g., to give optimal time windows in between cleavage events (FIG. 7, inset panel). Additionally, in certain embodiments, uniquely identifiable luminescent label of labeled non-specific peptidase **720** would indicate when cleavage events have occurred. As the polypeptide undergoes iterative cleavage, the rate of pulsing corresponding to binding by labeled recognition molecule **710** would drop in a step-wise manner whenever a terminal amino acid is cleaved by labeled non-specific peptidase **720**. This concept is illustrated by plot **702**, which generally depicts pulse rate as a function of time, with cleavage events in time denoted by arrows. Thus, in some embodiments, amino acids may be identified—and polypeptides thereby sequenced—in this approach based on a pulsing pattern and/or on the rate of pulsing that occurs within a pattern detected between cleavage events.

[0253] In some embodiments, terminal polypeptide sequence information (e.g., determined as described herein) can be combined with polypeptide sequence information obtained from one or more other sources. For example, terminal polypeptide sequence information could be combined with internal polypeptide sequence information. In some embodiments, internal polypeptide sequence information can be obtained using one or more amino acid recognition molecules that associate with internal amino acids, as described herein. Internal or other polypeptide sequence information can be obtained before or during a polypeptide degradation process. In some embodiments, sequence information obtained from these methods can be combined with polypeptide sequence information using other techniques, e.g., sequence information obtained using one or more internal amino acid recognition molecules.

Preparation of Samples for Sequencing

[0254] A polypeptide sample can be modified prior to sequencing. In some embodiments, the N-terminal amino acid or the C-terminal amino acid of a polypeptide is modified. FIG. 8 illustrates a non-limiting example of terminal end modification for preparing terminally modified

polypeptides from a protein sample. In step (1), protein sample **800** is fragmented to produce polypeptide fragments **802**. A polypeptide can be fragmented by cleaving (e.g., chemically) and/or digesting (e.g., enzymatically, for example using a peptidase, for example trypsin) a polypeptide of interest. Fragmentation can be performed before or after labeling. In some embodiments, fragmentation is performed after labeling of whole proteins. One or more amino acids can be labeled before or after cleavage to produce labeled polypeptides. In some embodiments, polypeptides are size selected after chemical or enzymatic fragmentation. In some embodiments, smaller polypeptides (e.g., < 2 kDa) are removed and larger polypeptides are retained for sequence analysis. Size selection can be achieved using a technique such as gel filtration, SEC, dialysis, PAGE gel extraction, microfluidic tension flow, or any other suitable technique. In step (2), the N-termini or C-termini of polypeptide fragments **802** are modified to produce terminally modified polypeptides **804**. In some embodiments, modification comprises adding an immobilization moiety. In some embodiments, modification comprises adding a coupling moiety.

[0255] Accordingly, provided herein are methods of modifying terminal ends of proteins and polypeptides with moieties that enable immobilization to a surface (e.g., a surface of a sample well on a chip used for protein analysis). In some embodiments, such methods comprise modifying a terminal end of a labeled polypeptide to be analyzed in accordance with the application. In yet other embodiments, such methods comprise modifying a terminal end of a protein or enzyme that degrades or translocates a protein or polypeptide substrate in accordance with the application.

[0256] In some embodiments, a carboxy-terminus of a protein or polypeptide is modified in a method comprising: (i) blocking free carboxylate groups of the protein or polypeptide; (ii) denaturing the protein or polypeptide (e.g., by heat and/or chemical means); (iii) blocking free thiol groups of the protein or polypeptide; (iv) digesting the protein or polypeptide to produce at least one polypeptide fragment comprising a free C-terminal carboxylate group; and (v) conjugating (e.g., chemically) a functional moiety to the free C-terminal carboxylate group. In some embodiments, the method further comprises, after (i) and before (ii), dialyzing a sample comprising the protein or polypeptide.

[0257] In some embodiments, a carboxy-terminus of a protein or polypeptide is modified in a method comprising: (i) denaturing the protein or polypeptide (e.g., by heat and/or chemical means); (ii) blocking free thiol groups of the protein or polypeptide; (iii) digesting the protein or polypeptide to produce at least one polypeptide fragment comprising a free C-terminal carboxylate group; (iv) blocking the free C-terminal carboxylate group to produce at least one polypeptide fragment comprising a blocked C-terminal carboxylate group; and (v) conjugating

(e.g., enzymatically) a functional moiety to the blocked C-terminal carboxylate group. In some embodiments, the method further comprises, after (iv) and before (v), dialyzing a sample comprising the protein or polypeptide.

[0258] In some embodiments, blocking free carboxylate groups refers to a chemical modification of these groups which alters chemical reactivity relative to an unmodified carboxylate. Suitable carboxylate blocking methods are known in the art and should modify side-chain carboxylate groups to be chemically different from a carboxy-terminal carboxylate group of a polypeptide to be functionalized. In some embodiments, blocking free carboxylate groups comprises esterification or amidation of free carboxylate groups of a polypeptide. In some embodiments, blocking free carboxylate groups comprises methyl esterification of free carboxylate groups of a polypeptide, e.g., by reacting the polypeptide with methanolic HCl. Additional examples of reagents and techniques useful for blocking free carboxylate groups include, without limitation, 4-sulfo-2,3,5,6-tetrafluorophenol (STP) and/or a carbodiimide such as N-(3-Dimethylaminopropyl)-N′-ethylcarbodiimide hydrochloride (EDAC), uronium reagents, diazomethane, alcohols and acid for Fischer esterification, the use of N-hydroxylsuccinimide (NHS) to form NHS esters (potentially as an intermediate to subsequent ester or amine formation), or reaction with carbonyldiimidazole (CDI) or the formation of mixed anhydrides, or any other method of modifying or blocking carboxylic acids, potentially through the formation of either esters or amides.

[0259] In some embodiments, blocking free thiol groups refers to a chemical modification of these groups which alters chemical reactivity relative to an unmodified thiol. In some embodiments, blocking free thiol groups comprises reducing and alkylating free thiol groups of a protein or polypeptide. In some embodiments, reduction and alkylation is carried out by contacting a polypeptide with dithiothreitol (DTT) and one or both of iodoacetamide and iodoacetic acid. Examples of additional and alternative cysteine-reducing reagents which may be used are well known and include, without limitation, 2-mercaptoethanol, Tris (2-carboxyehtyl) phosphine hydrochloride (TCEP), tributylphosphine, dithiobutylamine (DTBA), or any reagent capable of reducing a thiol group. Examples of additional and alternative cysteine-blocking (e.g., cysteine-alkylating) reagents which may be used are well known and include, without limitation, acrylamide, 4-vinylpyridine, N-Ethylmalemide (NEM), N-ε-maleimidocaproic acid (EMCA), or any reagent that modifies cysteines so as to prevent disulfide bond formation.

[0260] In some embodiments, digestion comprises enzymatic digestion. In some embodiments, digestion is carried out by contacting a protein or polypeptide with an endopeptidase (e.g., trypsin) under digestion conditions. In some embodiments, digestion comprises chemical

digestion. Examples of suitable reagents for chemical and enzymatic digestion are known in the art and include, without limitation, trypsin, chemotrypsin, Lys-C, Arg-C, Asp-N, Lys-N, BNPS-Skatole, CNBr, caspase, formic acid, glutamyl endopeptidase, hydroxylamine, iodosobenzoic acid, neutrophil elastase, pepsin, proline-endopeptidase, proteinase K, staphylococcal peptidase I, thermolysin, and thrombin.

[0261] In some embodiments, the functional moiety comprises a biotin molecule. In some embodiments, the functional moiety comprises a reactive chemical moiety, such as an alkynyl. In some embodiments, conjugating a functional moiety comprises biotinylation of carboxy-terminal carboxy-methyl ester groups by carboxypeptidase Y, as known in the art.

[0262] In some embodiments, a solubilizing moiety is added to a polypeptide. FIG. 9 illustrates a non-limiting example of a solubilizing moiety added to a terminal amino acid of a polypeptide, for example using a process of conjugating a solubilizing linker to the polypeptide.

[0263] In some embodiments, a terminally modified polypeptide **910** comprising a linker conjugating moiety **912** is conjugated to a solubilizing linker **920** comprising a polypeptide conjugating moiety **922.** In some embodiments, the solubilizing linker comprises a solubilizing polymer, such as a biomolecule (e.g., shown as stippled shape). In some embodiments, a resulting linker-conjugated polypeptide **930** comprising a linkage **932** formed between **912** and **922** further comprises a surface conjugating moiety **934.** Accordingly, in some embodiments methods and compositions provided herein are useful for modifying terminal ends of polypeptides with moieties that increase their solubility. In some embodiments, a solubilizing moiety is useful for small polypeptides that result from fragmentation (e.g., enzymatic fragmentation, for example using trypsin) and that are relatively insoluble. For example, in some embodiments, short polypeptides in a polypeptide pool can be solubilized by conjugating a polymer (e.g., a short oligo, a sugar, or other charged polymer) to the polypeptides.

[0264] In some embodiments, one or more surfaces of a sample well (e.g., sidewalls of a sample well) can be modified. A non-limiting example of passivation and/or antifouling of a sample well sidewall is shown in FIG. 10 where an example schematic of a sample well is illustrated with modified surfaces which may be used to promote single molecule immobilization to a bottom surface. In some embodiments, **1040** is SiO_2. In some embodiments, **1042** is a polypeptide conjugating moiety (e.g., TCO, tetrazine, N3, alkyne, aldehyde, NCO, NHS, thiol, alkene, DBCO, BCN, TPP, biotin, or other suitable conjugating moiety). In some embodiments, **1050** is TiO_2 or Al_2O_3. In some embodiments, **1052** is a hydrophobic C_{4-18} molecule, a polytetrafluoroethylene compound (e.g., $(CF_2)_{4-12}$), a polyol, such as a polyethylene glycol (e.g., PEG_{3-100}), polypropylene glycol, polyoxyethylene glycol, or combinations or variations thereof,

or a zwitterion, such as sulfobetaine. In some embodiments, **1060** is Si. In some embodiments, **1070** is Al. In some embodiments, **1080** is TiN.

Luminescent Labels

[0265] As used herein, a luminescent label is a molecule that absorbs one or more photons and may subsequently emit one or more photons after one or more time durations. In some embodiments, the term is used interchangeably with "label" or "luminescent molecule" depending on context. A luminescent label in accordance with certain embodiments described herein may refer to a luminescent label of a labeled recognition molecule, a luminescent label of a labeled peptidase (e.g., a labeled exopeptidase, a labeled non-specific exopeptidase), a luminescent label of a labeled peptide, a luminescent label of a labeled cofactor, or another labeled composition described herein. In some embodiments, a luminescent label in accordance with the application refers to a labeled amino acid of a labeled polypeptide comprising one or more labeled amino acids.

[0266] In some embodiments, a luminescent label may comprise a first and second chromophore. In some embodiments, an excited state of the first chromophore is capable of relaxation via an energy transfer to the second chromophore. In some embodiments, the energy transfer is a Förster resonance energy transfer (FRET). Such a FRET pair may be useful for providing a luminescent label with properties that make the label easier to differentiate from amongst a plurality of luminescent labels in a mixture—e.g., as illustrated and described herein for labeled aptamer **206** of FIG. 2. In yet other embodiments, a FRET pair comprises a first chromophore of a first luminescent label and a second chromophore of a second luminescent label. In certain embodiments, the FRET pair may absorb excitation energy in a first spectral range and emit luminescence in a second spectral range.

[0267] In some embodiments, a luminescent label refers to a fluorophore or a dye. Typically, a luminescent label comprises an aromatic or heteroaromatic compound and can be a pyrene, anthracene, naphthalene, naphthylamine, acridine, stilbene, indole, benzindole, oxazole, carbazole, thiazole, benzothiazole, benzoxazole, phenanthridine, phenoxazine, porphyrin, quinoline, ethidium, benzamide, cyanine, carbocyanine, salicylate, anthranilate, coumarin, fluoroscein, rhodamine, xanthene, or other like compound.

[0268] In some embodiments, a luminescent label comprises a dye selected from one or more of the following: 5/6-Carboxyrhodamine 6G, 5-Carboxyrhodamine 6G, 6-Carboxyrhodamine 6G, 6-TAMRA, Abberior® STAR 440SXP, Abberior® STAR 470SXP, Abberior® STAR 488, Abberior® STAR 512, Abberior® STAR 520SXP, Abberior® STAR 580, Abberior® STAR 600, Abberior® STAR 635, Abberior® STAR 635P, Abberior® STAR RED, Alexa Fluor® 350,

Alexa Fluor® 405, Alexa Fluor® 430, Alexa Fluor® 480, Alexa Fluor® 488, Alexa Fluor® 514, Alexa Fluor® 532, Alexa Fluor® 546, Alexa Fluor® 555, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 610-X, Alexa Fluor® 633, Alexa Fluor® 647, Alexa Fluor® 660, Alexa Fluor® 680, Alexa Fluor® 700, Alexa Fluor® 750, Alexa Fluor® 790, AMCA, ATTO 390, ATTO 425, ATTO 465, ATTO 488, ATTO 495, ATTO 514, ATTO 520, ATTO 532, ATTO 542, ATTO 550, ATTO 565, ATTO 590, ATTO 610, ATTO 620, ATTO 633, ATTO 647, ATTO 647N, ATTO 655, ATTO 665, ATTO 680, ATTO 700, ATTO 725, ATTO 740, ATTO Oxa12, ATTO Rho101, ATTO Rho11, ATTO Rho12, ATTO Rho13, ATTO Rho14, ATTO Rho3B, ATTO Rho6G, ATTO Thio12, BD Horizon™ V450, BODIPY® 493/501, BODIPY® 530/550, BODIPY® 558/568, BODIPY® 564/570, BODIPY® 576/589, BODIPY® 581/591, BODIPY® 630/650, BODIPY® 650/665, BODIPY® FL, BODIPY® FL-X, BODIPY® R6G, BODIPY® TMR, BODIPY® TR, CAL Fluor® Gold 540, CAL Fluor® Green 510, CAL Fluor® Orange 560, CAL Fluor® Red 590, CAL Fluor® Red 610, CAL Fluor® Red 615, CAL Fluor® Red 635, Cascade® Blue, CF™350, CF™405M, CF™405S, CF™488A, CF™514, CF™532, CF™543, CF™546, CF™555, CF™568, CF™594, CF™620R, CF™633, CF™633-V1, CF™640R, CF™640R-V1, CF™640R-V2, CF™660C, CF™660R, CF™680, CF™680R, CF™680R-V1, CF™750, CF™770, CF™790, Chromeo™ 642, Chromis 425N, Chromis 500N, Chromis 515N, Chromis 530N, Chromis 550A, Chromis 550C, Chromis 550Z, Chromis 560N, Chromis 570N, Chromis 577N, Chromis 600N, Chromis 630N, Chromis 645A, Chromis 645C, Chromis 645Z, Chromis 678A, Chromis 678C, Chromis 678Z, Chromis 770A, Chromis 770C, Chromis 800A, Chromis 800C, Chromis 830A, Chromis 830C, Cy®3, Cy®3.5, Cy®3B, Cy®5, Cy®5.5, Cy®7, DyLight® 350, DyLight® 405, DyLight® 415-Co1, DyLight® 425Q, DyLight® 485-LS, DyLight® 488, DyLight® 504Q, DyLight® 510-LS, DyLight® 515-LS, DyLight® 521-LS, DyLight® 530-R2, DyLight® 543Q, DyLight® 550, DyLight® 554-R0, DyLight® 554-R1, DyLight® 590-R2, DyLight® 594, DyLight® 610-B1, DyLight® 615-B2, DyLight® 633, DyLight® 633-B1, DyLight® 633-B2, DyLight® 650, DyLight® 655-B1, DyLight® 655-B2, DyLight® 655-B3, DyLight® 655-B4, DyLight® 662Q, DyLight® 675-B1, DyLight® 675-B2, DyLight® 675-B3, DyLight® 675-B4, DyLight® 679-C5, DyLight® 680, DyLight® 683Q, DyLight® 690-B1, DyLight® 690-B2, DyLight® 696Q, DyLight® 700-B1, DyLight® 700-B1, DyLight® 730-B1, DyLight® 730-B2, DyLight® 730-B3, DyLight® 730-B4, DyLight® 747, DyLight® 747-B1, DyLight® 747-B2, DyLight® 747-B3, DyLight® 747-B4, DyLight® 755, DyLight® 766Q, DyLight® 775-B2, DyLight® 775-B3, DyLight® 775-B4, DyLight® 780-B1, DyLight® 780-B2, DyLight® 780-B3, DyLight® 800, DyLight® 830-B2, Dyomics-350, Dyomics-350XL, Dyomics-360XL, Dyomics-370XL, Dyomics-375XL, Dyomics-380XL, Dyomics-390XL, Dyomics-405, Dyomics-415, Dyomics-430, Dyomics-431, Dyomics-478,

Dyomics-480XL, Dyomics-481XL, Dyomics-485XL, Dyomics-490, Dyomics-495, Dyomics-505, Dyomics-510XL, Dyomics-511XL, Dyomics-520XL, Dyomics-521XL, Dyomics-530, Dyomics-547, Dyomics-547P1, Dyomics-548, Dyomics-549, Dyomics-549P1, Dyomics-550, Dyomics-554, Dyomics-555, Dyomics-556, Dyomics-560, Dyomics-590, Dyomics-591, Dyomics-594, Dyomics-601XL, Dyomics-605, Dyomics-610, Dyomics-615, Dyomics-630, Dyomics-631, Dyomics-632, Dyomics-633, Dyomics-634, Dyomics-635, Dyomics-636, Dyomics-647, Dyomics-647P1, Dyomics-648, Dyomics-648P1, Dyomics-649, Dyomics-649P1, Dyomics-650, Dyomics-651, Dyomics-652, Dyomics-654, Dyomics-675, Dyomics-676, Dyomics-677, Dyomics-678, Dyomics-679P1, Dyomics-680, Dyomics-681, Dyomics-682, Dyomics-700, Dyomics-701, Dyomics-703, Dyomics-704, Dyomics-730, Dyomics-731, Dyomics-732, Dyomics-734, Dyomics-749, Dyomics-749P1, Dyomics-750, Dyomics-751, Dyomics-752, Dyomics-754, Dyomics-776, Dyomics-777, Dyomics-778, Dyomics-780, Dyomics-781, Dyomics-782, Dyomics-800, Dyomics-831, eFluor® 450, Eosin, FITC, Fluorescein, HiLyte™ Fluor 405, HiLyte™ Fluor 488, HiLyte™ Fluor 532, HiLyte™ Fluor 555, HiLyte™ Fluor 594, HiLyte™ Fluor 647, HiLyte™ Fluor 680, HiLyte™ Fluor 750, IRDye® 680LT, IRDye® 750, IRDye® 800CW, JOE, LightCycler® 640R, LightCycler® Red 610, LightCycler® Red 640, LightCycler® Red 670, LightCycler® Red 705, Lissamine Rhodamine B, Napthofluorescein, Oregon Green® 488, Oregon Green® 514, Pacific Blue™, Pacific Green™, Pacific Orange™, PET, PF350, PF405, PF415, PF488, PF505, PF532, PF546, PF555P, PF568, PF594, PF610, PF633P, PF647P, Quasar® 570, Quasar® 670, Quasar® 705, Rhodamine 123, Rhodamine 6G, Rhodamine B, Rhodamine Green, Rhodamine Green-X, Rhodamine Red, ROX, Seta™ 375, Seta™ 470, Seta™ 555, Seta™ 632, Seta™ 633, Seta™ 650, Seta™ 660, Seta™ 670, Seta™ 680, Seta™ 700, Seta™ 750, Seta™ 780, Seta™ APC-780, Seta™ PerCP-680, Seta™ R-PE-670, Seta™ 646, SeTau 380, SeTau 425, SeTau 647, SeTau 405, Square 635, Square 650, Square 660, Square 672, Square 680, Sulforhodamine 101, TAMRA, TET, Texas Red®, TMR, TRITC, Yakima Yellow™, Zenon®, Zy3, Zy5, Zy5.5, and Zy7.

Luminescence

[0269] In some aspects, the application relates to polypeptide sequencing and/or identification based on one or more luminescence properties of a luminescent label. In some embodiments, a luminescent label is identified based on luminescence lifetime, luminescence intensity, brightness, absorption spectra, emission spectra, luminescence quantum yield, or a combination of two or more thereof. In some embodiments, a plurality of types of luminescent labels can be distinguished from each other based on different luminescence lifetimes, luminescence intensities, brightnesses, absorption spectra, emission spectra, luminescence quantum yields, or

combinations of two or more thereof. Identifying may mean assigning the exact identity and/or quantity of one type of amino acid (e.g., a single type or a subset of types) associated with a luminescent label, and may also mean assigning an amino acid location in a polypeptide relative to other types of amino acids.

[0270] In some embodiments, luminescence is detected by exposing a luminescent label to a series of separate light pulses and evaluating the timing or other properties of each photon that is emitted from the label. In some embodiments, information for a plurality of photons emitted sequentially from a label is aggregated and evaluated to identify the label and thereby identify an associated type of amino acid. In some embodiments, a luminescence lifetime of a label is determined from a plurality of photons that are emitted sequentially from the label, and the luminescence lifetime can be used to identify the label. In some embodiments, a luminescence intensity of a label is determined from a plurality of photons that are emitted sequentially from the label, and the luminescence intensity can be used to identify the label. In some embodiments, a luminescence lifetime and luminescence intensity of a label is determined from a plurality of photons that are emitted sequentially from the label, and the luminescence lifetime and luminescence intensity can be used to identify the label.

[0271] In some aspects of the application, a single polypeptide molecule is exposed to a plurality of separate light pulses and a series of emitted photons are detected and analyzed. In some embodiments, the series of emitted photons provides information about the single polypeptide molecule that is present and that does not change in the reaction sample over the time of the experiment. However, in some embodiments, the series of emitted photons provides information about a series of different molecules that are present at different times in the reaction sample (e.g., as a reaction or process progresses). By way of example and not limitation, such information may be used to sequence and/or identify a polypeptide subjected to chemical or enzymatic degradation in accordance with the application.

[0272] In certain embodiments, a luminescent label absorbs one photon and emits one photon after a time duration. In some embodiments, the luminescence lifetime of a label can be determined or estimated by measuring the time duration. In some embodiments, the luminescence lifetime of a label can be determined or estimated by measuring a plurality of time durations for multiple pulse events and emission events. In some embodiments, the luminescence lifetime of a label can be differentiated amongst the luminescence lifetimes of a plurality of types of labels by measuring the time duration. In some embodiments, the luminescence lifetime of a label can be differentiated amongst the luminescence lifetimes of a plurality of types of labels by measuring a plurality of time durations for multiple pulse events and emission events. In certain embodiments, a label is identified or differentiated amongst a

plurality of types of labels by determining or estimating the luminescence lifetime of the label. In certain embodiments, a label is identified or differentiated amongst a plurality of types of labels by differentiating the luminescence lifetime of the label amongst a plurality of the luminescence lifetimes of a plurality of types of labels.

[0273] Determination of a luminescence lifetime of a luminescent label can be performed using any suitable method (e.g., by measuring the lifetime using a suitable technique or by determining time-dependent characteristics of emission). In some embodiments, determining the luminescence lifetime of one label comprises determining the lifetime relative to another label. In some embodiments, determining the luminescence lifetime of a label comprises determining the lifetime relative to a reference. In some embodiments, determining the luminescence lifetime of a label comprises measuring the lifetime (e.g., fluorescence lifetime). In some embodiments, determining the luminescence lifetime of a label comprises determining one or more temporal characteristics that are indicative of lifetime. In some embodiments, the luminescence lifetime of a label can be determined based on a distribution of a plurality of emission events (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more emission events) occurring across one or more time-gated windows relative to an excitation pulse. For example, a luminescence lifetime of a label can be distinguished from a plurality of labels having different luminescence lifetimes based on the distribution of photon arrival times measured with respect to an excitation pulse.

[0274] It should be appreciated that a luminescence lifetime of a luminescent label is indicative of the timing of photons emitted after the label reaches an excited state and the label can be distinguished by information indicative of the timing of the photons. Some embodiments may include distinguishing a label from a plurality of labels based on the luminescence lifetime of the label by measuring times associated with photons emitted by the label. The distribution of times may provide an indication of the luminescence lifetime which may be determined from the distribution. In some embodiments, the label is distinguishable from the plurality of labels based on the distribution of times, such as by comparing the distribution of times to a reference distribution corresponding to a known label. In some embodiments, a value for the luminescence lifetime is determined from the distribution of times.

[0275] As used herein, in some embodiments, luminescence intensity refers to the number of emitted photons per unit time that are emitted by a luminescent label which is being excited by delivery of a pulsed excitation energy. In some embodiments, the luminescence intensity refers to the detected number of emitted photons per unit time that are emitted by a label which is being excited by delivery of a pulsed excitation energy, and are detected by a particular sensor or set of sensors.

[0276] As used herein, in some embodiments, brightness refers to a parameter that reports on the average emission intensity per luminescent label. Thus, in some embodiments, "emission intensity" may be used to generally refer to brightness of a composition comprising one or more labels. In some embodiments, brightness of a label is equal to the product of its quantum yield and extinction coefficient.

[0277] As used herein, in some embodiments, luminescence quantum yield refers to the fraction of excitation events at a given wavelength or within a given spectral range that lead to an emission event, and is typically less than 1. In some embodiments, the luminescence quantum yield of a luminescent label described herein is between 0 and about 0.001, between about 0.001 and about 0.01, between about 0.01 and about 0.1, between about 0.1 and about 0.5, between about 0.5 and 0.9, or between about 0.9 and 1. In some embodiments, a label is identified by determining or estimating the luminescence quantum yield.

[0278] As used herein, in some embodiments, an excitation energy is a pulse of light from a light source. In some embodiments, an excitation energy is in the visible spectrum. In some embodiments, an excitation energy is in the ultraviolet spectrum. In some embodiments, an excitation energy is in the infrared spectrum. In some embodiments, an excitation energy is at or near the absorption maximum of a luminescent label from which a plurality of emitted photons are to be detected. In certain embodiments, the excitation energy is between about 500 nm and about 700 nm (e.g., between about 500 nm and about 600 nm, between about 600 nm and about 700 nm, between about 500 nm and about 550 nm, between about 550 nm and about 600 nm, between about 600 nm and about 650 nm, or between about 650 nm and about 700 nm). In certain embodiments, an excitation energy may be monochromatic or confined to a spectral range. In some embodiments, a spectral range has a range of between about 0.1 nm and about 1 nm, between about 1 nm and about 2 nm, or between about 2 nm and about 5 nm. In some embodiments, a spectral range has a range of between about 5 nm and about 10 nm, between about 10 nm and about 50 nm, or between about 50 nm and about 100 nm.

Sequencing

[0279] Aspects of the application relate to sequencing biological polymers, such as polypeptides and proteins. As used herein, "sequencing," "sequence determination," "determining a sequence," and like terms, in reference to a polypeptide or protein includes determination of partial sequence information as well as full sequence information of the polypeptide or protein. That is, the terminology includes sequence comparisons, fingerprinting, probabalistic fingerprinting, and like levels of information about a target molecule, as well as the express identification and ordering of each amino acid of the target molecule within a region of interest.

In some embodiments, the terminology includes identifying a single amino acid of a polypeptide. In yet other embodiments, more than one amino acid of a polypeptide is identified. As used herein, in some embodiments, "identifying," "determining the identity," and like terms, in reference to an amino acid includes determination of an express identity of an amino acid as well as determination of a probability of an express identity of an amino acid. For example, in some embodiments, an amino acid is identified by determining a probability (e.g., from 0% to 100%) that the amino acid is of a specific type, or by determining a probability for each of a plurality of specific types. Accordingly, in some embodiments, the terms "amino acid sequence," "polypeptide sequence," and "protein sequence" as used herein may refer to the polypeptide or protein material itself and is not restricted to the specific sequence information (e.g., the succession of letters representing the order of amino acids from one terminus to another terminus) that biochemically characterizes a specific polypeptide or protein.

[0280] In some embodiments, sequencing of a polypeptide molecule comprises identifying at least two (e.g., at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, or more) amino acids in the polypeptide molecule. In some embodiments, the at least two amino acids are contiguous amino acids. In some embodiments, the at least two amino acids are non-contiguous amino acids.

[0281] In some embodiments, sequencing of a polypeptide molecule comprises identification of less than 100% (e.g., less than 99%, less than 95%, less than 90%, less than 85%, less than 80%, less than 75%, less than 70%, less than 65%, less than 60%, less than 55%, less than 50%, less than 45%, less than 40%, less than 35%, less than 30%, less than 25%, less than 20%, less than 15%, less than 10%, less than 5%, less than 1% or less) of all amino acids in the polypeptide molecule. For example, in some embodiments, sequencing of a polypeptide molecule comprises identification of less than 100% of one type of amino acid in the polypeptide molecule (e.g., identification of a portion of all amino acids of one type in the polypeptide molecule). In some embodiments, sequencing of a polypeptide molecule comprises identification of less than 100% of each type of amino acid in the polypeptide molecule.

[0282] In some embodiments, sequencing of a polypeptide molecule comprises identification of at least 1, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100 or more types of amino acids in the polypeptide.

[0283] In some embodiments, the application provides compositions and methods for sequencing a polypeptide by identifying a series of amino acids that are present at a terminus of a

polypeptide over time (e.g., by iterative detection and cleavage of amino acids at the terminus). In yet other embodiments, the application provides compositions and methods for sequencing a polypeptide by identifying labeled amino content of the polypeptide and comparing to a reference sequence database.

[0284] In some embodiments, the application provides compositions and methods for sequencing a polypeptide by sequencing a plurality of fragments of the polypeptide. In some embodiments, sequencing a polypeptide comprises combining sequence information for a plurality of polypeptide fragments to identify and/or determine a sequence for the polypeptide. In some embodiments, combining sequence information may be performed by computer hardware and software. The methods described herein may allow for a set of related polypeptides, such as an entire proteome of an organism, to be sequenced. In some embodiments, a plurality of single molecule sequencing reactions are performed in parallel (e.g., on a single chip) according to aspects of the present application. For example, in some embodiments, a plurality of single molecule sequencing reactions are each performed in separate sample wells on a single chip.

[0285] In some embodiments, methods provided herein may be used for the sequencing and identification of an individual protein in a sample comprising a complex mixture of proteins. In some embodiments, the application provides methods of uniquely identifying an individual protein in a complex mixture of proteins. In some embodiments, an individual protein is detected in a mixed sample by determining a partial amino acid sequence of the protein. In some embodiments, the partial amino acid sequence of the protein is within a contiguous stretch of approximately 5 to 50 amino acids.

[0286] Without wishing to be bound by any particular theory, it is believed that most human proteins can be identified using incomplete sequence information with reference to proteomic databases. For example, simple modeling of the human proteome has shown that approximately 98% of proteins can be uniquely identified by detecting just four types of amino acids within a stretch of 6 to 40 amino acids (see, e.g., Swaminathan, et al. *PLoS Comput Biol.* 2015, 11(2):e1004080; and Yao, et al. *Phys. Biol.* 2015, 12(5):055003). Therefore, a complex mixture of proteins can be degraded (e.g., chemically degraded, enzymatically degraded) into short polypeptide fragments of approximately 6 to 40 amino acids, and sequencing of this polypeptide library would reveal the identity and abundance of each of the proteins present in the original complex mixture. Compositions and methods for selective amino acid labeling and identifying polypeptides by determining partial sequence information are described in detail in U.S. Patent Application No. 15/510,962, filed September 15, 2015, titled "SINGLE MOLECULE PEPTIDE SEQUENCING," which is incorporated by reference in its entirety.

[0287] Embodiments are capable of sequencing single polypeptide molecules with high accuracy, such as an accuracy of at least about 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%. In some embodiments, the target molecule used in single molecule sequencing is a polypeptide that is immobilized to a surface of a solid support such as a bottom surface or a sidewall surface of a sample well. The sample well also can contain any other reagents needed for a sequencing reaction in accordance with the application, such as one or more suitable buffers, co-factors, labeled recognition molecules, and enzymes (e.g., catalytically active or inactive exopeptidase enzymes, which may be luminescently labeled or unlabeled).

[0288] As described above, in some embodiments, sequencing in accordance with the application comprises identifying an amino acid by determining a probability that the amino acid is of a specific type. Conventional protein identification systems require identification of each amino acid in a polypeptide to identify the polypeptide. However, it is difficult to accurately identify each amino acid in a polypeptide. For example, data collected from an interaction in which a first recognition molecule associates with a first amino acid may not be sufficiently different from data collected from an interaction in which a second recognition molecule associates with a second amino acid to differentiate between the two amino acids. In some embodiments, sequencing in accordance with the application avoids this problem by using a protein identification system that, unlike conventional protein identification systems, does not require (but does not preclude) identification of each amino acid in the protein.

[0289] Accordingly, in some embodiments, sequencing in accordance with the application may be carried out using a protein identification system that uses machine learning techniques to identify proteins. In some embodiments, the system operates by: (1) collecting data about a polypeptide of a protein using a real-time protein sequencing device; (2) using a machine learning model and the collected data to identify probabilities that certain amino acids are part of the polypeptide at respective locations; and (3) using the identified probabilities, as a "probabilistic fingerprint" to identify the protein. In some embodiments, data about the polypeptide of the protein may be obtained using reagents that selectively bind amino acids. As an example, the reagents and/or amino acids may be labeled with luminescent labels that emit light in response to application of excitation energy. In this example, a protein sequencing device may apply excitation energy to a sample of a protein (e.g., a polypeptide) during binding interactions of reagents with amino acids in the sample. In some embodiments, one or more sensors in the sequencing device (e.g., a photodetector, an electrical sensor, and/or any other suitable type of sensor) may detect binding interactions. In turn, the data collected and/or derived from the detected light emissions may be provided to the machine learning model.

Machine learning models and associated systems and methods are described in detail in U.S. Provisional Patent Appl. No. 62/860,750, filed June 12, 2019, titled "MACHINE LEARNING ENABLED PROTEIN IDENTIFICATION," which is incorporated by reference in its entirety.

[0290] Sequencing in accordance with the application, in some aspects, may involve immobilizing a polypeptide on a surface of a substrate (e.g., of a solid support, for example a chip, for example an integrated device as described herein). In some embodiments, a polypeptide may be immobilized on a surface of a sample well (e.g., on a bottom surface of a sample well) on a substrate. In some embodiments, the N-terminal amino acid of the polypeptide is immobilized (e.g., attached to the surface). In some embodiments, the C-terminal amino acid of the polypeptide is immobilized (e.g., attached to the surface). In some embodiments, one or more non-terminal amino acids are immobilized (e.g., attached to the surface). The immobilized amino acid(s) can be attached using any suitable covalent or non-covalent linkage, for example as described in this application. In some embodiments, a plurality of polypeptides are attached to a plurality of sample wells (e.g., with one polypeptide attached to a surface, for example a bottom surface, of each sample well), for example in an array of sample wells on a substrate.

[0291] Sequencing in accordance with the application, in some aspects, may be performed using a system that permits single molecule analysis. The system may include an integrated device and an instrument configured to interface with the integrated device. The integrated device may include an array of pixels, where individual pixels include a sample well and at least one photodetector. The sample wells of the integrated device may be formed on or through a surface of the integrated device and be configured to receive a sample placed on the surface of the integrated device. Collectively, the sample wells may be considered as an array of sample wells. The plurality of sample wells may have a suitable size and shape such that at least a portion of the sample wells receive a single sample (e.g., a single molecule, such as a polypeptide). In some embodiments, the number of samples within a sample well may be distributed among the sample wells of the integrated device such that some sample wells contain one sample while others contain zero, two or more samples.

[0292] Excitation light is provided to the integrated device from one or more light source external to the integrated device. Optical components of the integrated device may receive the excitation light from the light source and direct the light towards the array of sample wells of the integrated device and illuminate an illumination region within the sample well. In some embodiments, a sample well may have a configuration that allows for the sample to be retained in proximity to a surface of the sample well, which may ease delivery of excitation light to the sample and detection of emission light from the sample. A sample positioned within the

illumination region may emit emission light in response to being illuminated by the excitation light. For example, the sample may be labeled with a fluorescent marker, which emits light in response to achieving an excited state through the illumination of excitation light. Emission light emitted by a sample may then be detected by one or more photodetectors within a pixel corresponding to the sample well with the sample being analyzed. When performed across the array of sample wells, which may range in number between approximately 10,000 pixels to 1,000,000 pixels according to some embodiments, multiple samples can be analyzed in parallel.

[0293] The integrated device may include an optical system for receiving excitation light and directing the excitation light among the sample well array. The optical system may include one or more grating couplers configured to couple excitation light to the integrated device and direct the excitation light to other optical components. The optical system may include optical components that direct the excitation light from a grating coupler towards the sample well array. Such optical components may include optical splitters, optical combiners, and waveguides. In some embodiments, one or more optical splitters may couple excitation light from a grating coupler and deliver excitation light to at least one of the waveguides. According to some embodiments, the optical splitter may have a configuration that allows for delivery of excitation light to be substantially uniform across all the waveguides such that each of the waveguides receives a substantially similar amount of excitation light. Such embodiments may improve performance of the integrated device by improving the uniformity of excitation light received by sample wells of the integrated device. Examples of suitable components, e.g., for coupling excitation light to a sample well and/or directing emission light to a photodetector, to include in an integrated device are described in U.S. Patent Application No. 14/821,688, filed August 7, 2015, titled "INTEGRATED DEVICE FOR PROBING, DETECTING AND ANALYZING MOLECULES," and U.S. Patent Application No. 14/543,865, filed November 17, 2014, titled "INTEGRATED DEVICE WITH EXTERNAL LIGHT SOURCE FOR PROBING, DETECTING, AND ANALYZING MOLECULES," both of which are incorporated by reference in their entirety. Examples of suitable grating couplers and waveguides that may be implemented in the integrated device are described in U.S. Patent Application No. 15/844,403, filed December 15, 2017, titled "OPTICAL COUPLER AND WAVEGUIDE SYSTEM," which is incorporated by reference in its entirety.

[0294] Additional photonic structures may be positioned between the sample wells and the photodetectors and configured to reduce or prevent excitation light from reaching the photodetectors, which may otherwise contribute to signal noise in detecting emission light. In some embodiments, metal layers which may act as a circuitry for the integrated device, may also act as a spatial filter. Examples of suitable photonic structures may include spectral filters, a

polarization filters, and spatial filters and are described in U.S. Patent Application No. 16/042,968, filed July 23, 2018, titled "OPTICAL REJECTION PHOTONIC STRUCTURES," which is incorporated by reference in its entirety.

[0295] Components located off of the integrated device may be used to position and align an excitation source to the integrated device. Such components may include optical components including lenses, mirrors, prisms, windows, apertures, attenuators, and/or optical fibers. Additional mechanical components may be included in the instrument to allow for control of one or more alignment components. Such mechanical components may include actuators, stepper motors, and/or knobs. Examples of suitable excitation sources and alignment mechanisms are described in U.S. Patent Application No. 15/161,088, filed May 20, 2016, titled "PULSED LASER AND SYSTEM," which is incorporated by reference in its entirety. Another example of a beam-steering module is described in U.S. Patent Application No. 15/842,720, filed December, 14, 2017, titled "COMPACT BEAM SHAPING AND STEERING ASSEMBLY," which is incorporated herein by reference. Additional examples of suitable excitation sources are described in U.S. Patent Application No. 14/821,688, filed August 7, 2015, titled "INTEGRATED DEVICE FOR PROBING, DETECTING AND ANALYZING MOLECULES," which is incorporated by reference in its entirety.

[0296] The photodetector(s) positioned with individual pixels of the integrated device may be configured and positioned to detect emission light from the pixel's corresponding sample well. Examples of suitable photodetectors are described in U.S. Patent Application No. 14/821,656, filed August 7, 2015, titled "INTEGRATED DEVICE FOR TEMPORAL BINNING OF RECEIVED PHOTONS," which is incorporated by reference in its entirety. In some embodiments, a sample well and its respective photodetector(s) may be aligned along a common axis. In this manner, the photodetector(s) may overlap with the sample well within the pixel.

[0297] Characteristics of the detected emission light may provide an indication for identifying the marker associated with the emission light. Such characteristics may include any suitable type of characteristic, including an arrival time of photons detected by a photodetector, an amount of photons accumulated over time by a photodetector, and/or a distribution of photons across two or more photodetectors. In some embodiments, a photodetector may have a configuration that allows for the detection of one or more timing characteristics associated with a sample's emission light (e.g., luminescence lifetime). The photodetector may detect a distribution of photon arrival times after a pulse of excitation light propagates through the integrated device, and the distribution of arrival times may provide an indication of a timing characteristic of the sample's emission light (e.g., a proxy for luminescence lifetime). In some embodiments, the one or more photodetectors provide an indication of the probability of emission light emitted by the

marker (e.g., luminescence intensity). In some embodiments, a plurality of photodetectors may be sized and arranged to capture a spatial distribution of the emission light. Output signals from the one or more photodetectors may then be used to distinguish a marker from among a plurality of markers, where the plurality of markers may be used to identify a sample within the sample. In some embodiments, a sample may be excited by multiple excitation energies, and emission light and/or timing characteristics of the emission light emitted by the sample in response to the multiple excitation energies may distinguish a marker from a plurality of markers.

[0298] In operation, parallel analyses of samples within the sample wells are carried out by exciting some or all of the samples within the wells using excitation light and detecting signals from sample emission with the photodetectors. Emission light from a sample may be detected by a corresponding photodetector and converted to at least one electrical signal. The electrical signals may be transmitted along conducting lines in the circuitry of the integrated device, which may be connected to an instrument interfaced with the integrated device. The electrical signals may be subsequently processed and/or analyzed. Processing or analyzing of electrical signals may occur on a suitable computing device either located on or off the instrument.

[0299] The instrument may include a user interface for controlling operation of the instrument and/or the integrated device. The user interface may be configured to allow a user to input information into the instrument, such as commands and/or settings used to control the functioning of the instrument. In some embodiments, the user interface may include buttons, switches, dials, and a microphone for voice commands. The user interface may allow a user to receive feedback on the performance of the instrument and/or integrated device, such as proper alignment and/or information obtained by readout signals from the photodetectors on the integrated device. In some embodiments, the user interface may provide feedback using a speaker to provide audible feedback. In some embodiments, the user interface may include indicator lights and/or a display screen for providing visual feedback to a user.

[0300] In some embodiments, the instrument may include a computer interface configured to connect with a computing device. The computer interface may be a USB interface, a FireWire interface, or any other suitable computer interface. A computing device may be any general purpose computer, such as a laptop or desktop computer. In some embodiments, a computing device may be a server (e.g., cloud-based server) accessible over a wireless network via a suitable computer interface. The computer interface may facilitate communication of information between the instrument and the computing device. Input information for controlling and/or configuring the instrument may be provided to the computing device and transmitted to the instrument via the computer interface. Output information generated by the instrument may be received by the computing device via the computer interface. Output information may

include feedback about performance of the instrument, performance of the integrated device, and/or data generated from the readout signals of the photodetector.

[0301] In some embodiments, the instrument may include a processing device configured to analyze data received from one or more photodetectors of the integrated device and/or transmit control signals to the excitation source(s). In some embodiments, the processing device may comprise a general purpose processor, a specially-adapted processor (e.g., a central processing unit (CPU) such as one or more microprocessor or microcontroller cores, a field-programmable gate array (FPGA), an application-specific integrated circuit (ASIC), a custom integrated circuit, a digital signal processor (DSP), or a combination thereof). In some embodiments, the processing of data from one or more photodetectors may be performed by both a processing device of the instrument and an external computing device. In other embodiments, an external computing device may be omitted and processing of data from one or more photodetectors may be performed solely by a processing device of the integrated device.

[0302] According to some embodiments, the instrument that is configured to analyze samples based on luminescence emission characteristics may detect differences in luminescence lifetimes and/or intensities between different luminescent molecules, and/or differences between lifetimes and/or intensities of the same luminescent molecules in different environments. The inventors have recognized and appreciated that differences in luminescence emission lifetimes can be used to discern between the presence or absence of different luminescent molecules and/or to discern between different environments or conditions to which a luminescent molecule is subjected. In some cases, discerning luminescent molecules based on lifetime (rather than emission wavelength, for example) can simplify aspects of the system. As an example, wavelength-discriminating optics (such as wavelength filters, dedicated detectors for each wavelength, dedicated pulsed optical sources at different wavelengths, and/or diffractive optics) may be reduced in number or eliminated when discerning luminescent molecules based on lifetime. In some cases, a single pulsed optical source operating at a single characteristic wavelength may be used to excite different luminescent molecules that emit within a same wavelength region of the optical spectrum but have measurably different lifetimes. An analytic system that uses a single pulsed optical source, rather than multiple sources operating at different wavelengths, to excite and discern different luminescent molecules emitting in a same wavelength region can be less complex to operate and maintain, more compact, and may be manufactured at lower cost.

[0303] Although analytic systems based on luminescence lifetime analysis may have certain benefits, the amount of information obtained by an analytic system and/or detection accuracy may be increased by allowing for additional detection techniques. For example, some embodiments of the systems may additionally be configured to discern one or more properties of

a sample based on luminescence wavelength and/or luminescence intensity. In some implementations, luminescence intensity may be used additionally or alternatively to distinguish between different luminescent labels. For example, some luminescent labels may emit at significantly different intensities or have a significant difference in their probabilities of excitation (e.g., at least a difference of about 35%) even though their decay rates may be similar. By referencing binned signals to measured excitation light, it may be possible to distinguish different luminescent labels based on intensity levels.

[0304] According to some embodiments, different luminescence lifetimes may be distinguished with a photodetector that is configured to time-bin luminescence emission events following excitation of a luminescent label. The time binning may occur during a single charge-accumulation cycle for the photodetector. A charge-accumulation cycle is an interval between read-out events during which photo-generated carriers are accumulated in bins of the time-binning photodetector. Examples of a time-binning photodetector are described in U.S. Patent Application No. 14/821,656, filed August 7, 2015, titled "INTEGRATED DEVICE FOR TEMPORAL BINNING OF RECEIVED PHOTONS," which is incorporated herein by reference. In some embodiments, a time-binning photodetector may generate charge carriers in a photon absorption/carrier generation region and directly transfer charge carriers to a charge carrier storage bin in a charge carrier storage region. In such embodiments, the time-binning photodetector may not include a carrier travel/capture region. Such a time-binning photodetector may be referred to as a "direct binning pixel." Examples of time-binning photodetectors, including direct binning pixels, are described in U.S. Patent Application No. 15/852,571, filed December, 22, 2017, titled "INTEGRATED PHOTODETECTOR WITH DIRECT BINNING PIXEL," which is incorporated herein by reference.

[0305] In some embodiments, different numbers of fluorophores of the same type may be linked to different reagents in a sample, so that each reagent may be identified based on luminescence intensity. For example, two fluorophores may be linked to a first labeled recognition molecule and four or more fluorophores may be linked to a second labeled recognition molecule. Because of the different numbers of fluorophores, there may be different excitation and fluorophore emission probabilities associated with the different recognition molecules. For example, there may be more emission events for the second labeled recognition molecule during a signal accumulation interval, so that the apparent intensity of the bins is significantly higher than for the first labeled recognition molecule.

[0306] The inventors have recognized and appreciated that distinguishing nucleotides or any other biological or chemical samples based on fluorophore decay rates and/or fluorophore intensities may enable a simplification of the optical excitation and detection systems. For

example, optical excitation may be performed with a single-wavelength source (e.g., a source producing one characteristic wavelength rather than multiple sources or a source operating at multiple different characteristic wavelengths). Additionally, wavelength discriminating optics and filters may not be needed in the detection system. Also, a single photodetector may be used for each sample well to detect emission from different fluorophores. The phrase "characteristic wavelength" or "wavelength" is used to refer to a central or predominant wavelength within a limited bandwidth of radiation (e.g., a central or peak wavelength within a 20 nm bandwidth output by a pulsed optical source). In some cases, "characteristic wavelength" or "wavelength" may be used to refer to a peak wavelength within a total bandwidth of radiation output by a source.

Computational Techniques

[0307] Aspects of the present application relate to computational techniques for analyzing the data generated by the polypeptide sequencing techniques described herein. As discussed above, for example in connection with FIGs. 1A and 1B, the data generated by using these sequencing techniques may include a series of signal pulses indicative of instances where an amino acid recognition molecule is associated with an amino acid exposed at the terminus of the polypeptide being sequenced. The series of signal pulses may have varying one or more features (e.g., pulse duration, interpulse duration, change in magnitude), depending on the type of amino acid presently at the terminus, over time as the degradation process proceeds in removing successive amino acids. The resulting signal trace may include characteristic patterns, which arise from the varying one or more features, associated with respective amino acids. The computational techniques described herein may be implemented as part of analyzing such data obtained using these sequencing techniques to identify an amino acid sequence.

[0308] Some embodiments may involve obtaining data during a degradation process of a polypeptide, analyzing the data to determine portions of the data corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process, and outputting an amino acid sequence representative of the polypeptide. FIG. 11 is a diagram of an illustrative processing pipeline **1100** for identifying an amino acid sequence by analyzing data obtained using the polypeptide sequencing techniques described herein. As shown in FIG. 11, analyzing sequencing data **1102** may involve using association event identification technique **1104** and amino acid identification technique **1106** to output amino acid sequence(s) **1108**.

[0309] As discussed herein, sequencing data **1102** may be obtained during a degradation process of a polypeptide. In some embodiments, the sequencing data **1102** is indicative of amino acid identity at the terminus of the polypeptide during the degradation process. In some

embodiments, the sequencing data **1102** is indicative of a signal produced by one or more amino acid recognition molecules binding to different types of terminal amino acids at the terminus during the degradation process. Exemplary sequencing data is shown in FIGs. 1A and 1B, which are discussed above.

[0310] Depending on how signals are generated during the degradation process, sequencing data **1102** may be indicative of one or more different types of signals. In some embodiments, sequencing data **1102** is indicative of a luminescent signal generated during the degradation process. For example, a luminescent label may be used to label an amino acid recognition molecule, and luminescence emitted by the luminescent label may be detected as the amino acid recognition molecule associates with a particular amino acid, resulting in a luminescent signal. In some embodiments, sequencing data **1102** is indicative of an electrical signal generated during the degradation process. For example, a polypeptide molecule being sequenced may be immobilized to a nanopore, and an electrical signal (e.g., changes in conductance) may be detected as an amino acid recognition molecule associates with a particular amino acid.

[0311] Some embodiments involve analyzing sequencing data **1102** to determine portions of sequencing data **1102** corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process. As shown in FIG. 11, association event identification technique **1104** may access sequencing data **1102** and analyze sequencing data to identify portions of sequencing data **1102** that correspond to association events. The association events may correspond to characteristic patterns, such as CP_1 and CP_2 shown in FIG. 1B, in the data. In some embodiments, association event identification technique **1104** may involve detecting a series of cleavage events and determining portions of sequencing data **1102** between successive cleavage events. As an example, a cleavage event between CP_1 and CP_2 shown in FIG. 1B may be detected such that a first portion of the data corresponding to CP_1 may be identified as a first association event and a second portion of the data corresponding CP_2 may be identified as a second association event.

[0312] Some embodiments involve identifying a type of amino acid for one or more of the determined portions of sequencing data **1102**. As shown in FIG. 11, amino acid identification technique **1106** may be used to determine a type of amino acid for one or more of the association events identified by association event identification technique **1104**. In some embodiments, the individual portions of data identified by association event identification technique **1104** may include a pulse pattern, and amino acid identification technique **1106** may determine a type of amino acid for one or more of the portions based on its respective pulse pattern. Referring to FIG. 1B, amino acid identification technique **1106** may identify a first type of amino acid for CP_1 and a second type of amino acid for CP_2. In some embodiments, determining the type of amino

acid may include identifying an amount of time within a portion of data, such as a portion identified using association event identification technique **1104**, when the data is above a threshold value and comparing the amount of time to a duration of time for the portion of data. For example, identifying a type of amino acid for CP_1 may include determining an amount of time within CP_1 where the signal is above a threshold value, such as time period, *pd*, where the signal is above M_L, and comparing it to a total duration of time for CP_1. In some embodiments, determining the type of amino acid may involve identifying one or more pulse durations for one or more portions of data identified by association event identification technique **1102**. For example, identifying a type of amino acid for CP_1 may include determining a pulse duration for CP_1, such as time period, *pd*. In some embodiments, determining the type of amino acid may involve identifying one or more interpulse durations for one or more portions of the data identified using association event identification technique **1104**. For example, identifying a type of amino acid for CP_1 may include identifying an interpulse duration, such as *ipd*.

[0313] By identifying a type of amino acid for successive portions of sequencing data **1102**, amino acid identification technique **1106** may output amino acid sequence(s) **1108** representative of the polypeptide. In some embodiments, the amino acid sequence includes a series of amino acids corresponding to the portions of data identified using association event identification technique **1104**.

[0314] FIG. 12 is a flow chart of an illustrative process **1200** for determining an amino acid sequence of a polypeptide molecule, in accordance with some embodiments of the technology described herein. Process **1200** may be performed on any suitable computing device(s) (e.g., a single computing device, multiple computing devices co-located in a single physical location or located in multiple physical locations remote from one another, one or more computing devices part of a cloud computing system, etc.), as aspects of the technology described herein are not limited in this respect. In some embodiments, association event identification technique **1104** and amino acid identification technique **1106** may perform some or all of process **1200** to determine amino acid sequence(s).

[0315] Process **1200** begins at act **1202**, which involves contacting a single polypeptide molecule with one or more terminal amino acid recognition molecules. Next, process **1200** proceeds to act **1104**, which involves detecting a series of signal pulses indicative of association of the one or more terminal amino acid recognition molecules with successive amino acids exposed at a terminus of the single polypeptide while the single polypeptide is being degraded. The series of pulses may allow for sequencing of the single polypeptide molecule, such as by using association event identification technique **1104** and amino acid identification technique **1106**.

[0316] In some embodiments, process **1200** may include act **1206**, which involves identifying a first type of amino acid in the single polypeptide molecule based on a first characteristic pattern in the series of signal pulses, such as by using amino acid identification technique **1106**.

[0317] FIG. 13 is a flow chart of an illustrative process **1300** for determining an amino acid sequence representative of a polypeptide, in accordance with some embodiments of the technology described herein. Process **1300** may be performed on any suitable computing device(s) (e.g., a single computing device, multiple computing devices co-located in a single physical location or located in multiple physical locations remote from one another, one or more computing devices part of a cloud computing system, etc.), as aspects of the technology described herein are not limited in this respect. In some embodiments, association event identification technique **1104** and amino acid identification technique **1106** may perform some or all of process **1300** to determine amino acid sequence(s).

[0318] Process **1300** begins at act **1302**, where data during a degradation process of a polypeptide is obtained. In some embodiments, the data is indicative of amino acid identity at the terminus of the polypeptide during the degradation process. In some embodiments, the data is indicative of a signal produced by one or more amino acid recognition molecules binding to different types of terminal amino acids at the terminus during the degradation process. In some embodiments, the data is indicative of a luminescent signal generated during the degradation process. In some embodiments, the data is indicative of an electrical signal generated during the degradation process.

[0319] Next, process **1300** proceeds to act **1304**, where the data is analyzed to determine portions of the data corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process, such as by using association event identification technique **1104** and amino acid identification technique **1106**. In some embodiments, analyzing the data further comprises detecting a series of cleavage events and determining the portions of the data between successive cleavage events, such as by using association event identification technique **1104**.

[0320] In some embodiments, analyzing the data further comprises determining a type of amino acid for each of the individual portions, such as by using amino acid identification technique **1106**. In some embodiments, each of the individual portions comprises a pulse pattern, and analyzing the data further comprises determining a type of amino acid for one or more of the portions based on its respective pulse pattern. In some embodiments, determining the type of amino acid further comprises identifying an amount of time within a portion when the data is above a threshold value and comparing the amount of time to a duration of time for the portion. In some embodiments, determining the type of amino acid further comprises identifying at least

one pulse duration for each of the one or more portions. In some embodiments, determining the type of amino acid further comprises identifying at least one interpulse duration for each of the one or more portions.

[0321] Next, process **1300** proceeds to act **1306**, where an amino acid sequence representative of the polypeptide is outputted, such as via a user interface. In some embodiments, the amino acid sequence includes a series of amino acids corresponding to the portions.

[0322] An illustrative implementation of a computer system **1400** that may be used in connection with any of the embodiments of the technology described herein is shown in FIG. 14. The computer system **1400** includes one or more processors **1410** and one or more articles of manufacture that comprise non-transitory computer-readable storage media (e.g., memory **1420** and one or more non-volatile storage media **1430**). The processor **1410** may control writing data to and reading data from the memory **1420** and the non-volatile storage device **1430** in any suitable manner, as the aspects of the technology described herein are not limited in this respect. To perform any of the functionality described herein, the processor **1410** may execute one or more processor-executable instructions stored in one or more non-transitory computer-readable storage media (e.g., the memory **1420**), which may serve as non-transitory computer-readable storage media storing processor-executable instructions for execution by the processor **1410**.

[0323] Computing device **1400** may also include a network input/output (I/O) interface **1440** via which the computing device may communicate with other computing devices (e.g., over a network), and may also include one or more user I/O interfaces **1450**, via which the computing device may provide output to and receive input from a user. The user I/O interfaces may include devices such as a keyboard, a mouse, a microphone, a display device (e.g., a monitor or touch screen), speakers, a camera, and/or various other types of I/O devices.

[0324] The above-described embodiments can be implemented in any of numerous ways. For example, the embodiments may be implemented using hardware, software or a combination thereof. When implemented in software, the software code can be executed on any suitable processor (e.g., a microprocessor) or collection of processors, whether provided in a single computing device or distributed among multiple computing devices. It should be appreciated that any component or collection of components that perform the functions described above can be generically considered as one or more controllers that control the above-discussed functions. The one or more controllers can be implemented in numerous ways, such as with dedicated hardware, or with general purpose hardware (e.g., one or more processors) that is programmed using microcode or software to perform the functions recited above.

[0325] In this respect, it should be appreciated that one implementation of the embodiments described herein comprises at least one computer-readable storage medium (e.g., RAM, ROM,

EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or other tangible, non-transitory computer-readable storage medium) encoded with a computer program (i.e., a plurality of executable instructions) that, when executed on one or more processors, performs the above-discussed functions of one or more embodiments. The computer-readable medium may be transportable such that the program stored thereon can be loaded onto any computing device to implement aspects of the techniques discussed herein. In addition, it should be appreciated that the reference to a computer program which, when executed, performs any of the above-discussed functions, is not limited to an application program running on a host computer. Rather, the terms computer program and software are used herein in a generic sense to reference any type of computer code (e.g., application software, firmware, microcode, or any other form of computer instruction) that can be employed to program one or more processors to implement aspects of the techniques discussed herein.

EXAMPLES

Example 1. Solubilizing Linkers for Peptide Surface Immobilization

[0326] Seeking to improve oligopeptide solubility in aqueous buffer, it was determined that peptide fragments could be conjugated with oligonucleotide linkers to both improve aqueous solubility and provide a functional moiety for surface immobilization of peptides at the single molecule level. Different peptide-linker conjugates were synthesized, with example structures depicted in FIG. 15A for a peptide-DNA conjugate and a peptide-PEG conjugate. Linker conjugation was observed to greatly enhance peptide solubility in aqueous solution for each of the different peptide-linker conjugates evaluated.

[0327] The peptide-linker conjugates were evaluated for amino acid cleavage at peptide N-termini by N-terminal aminopeptidases (Table 6, below).

Table 6. Terminal amino acid cleavage of peptide-linker conjugates.

Entry	Peptide	SEQ ID NO.	Class	Linker	Cleaved by Rat APN	Cleaved by PIP
1	KF	223	positive	oligo	No	
2	KKMKKM{LYS(N3)}	224	positive	oligo	No	
3	KKMKKM{LYS(N3)}	225	positive	oligo-PEG	No	
4	KKMKKM{LYS(N3)}	226	positive	PEG4	Yes	
5	DDMDDM{LYS(N3)}	227	negative	oligo	Yes	
6	FFMFFM{LYS(N3)}	228	aromatic	oligo	Yes	

7	AAMAAM{LYS(N3)}	229	hydrophobic	oligo	Yes	
8	FPFPFP{LYS(N3)}	230	aromatic	oligo		Yes
9	DPDPDP{LYS(N3)}	231	negative	oligo		Yes
10	KPKPKP{LYS(N3)}	232	positive	oligo		No
11	KPKPKP{LYS(N3)}	233	positive	PEG4		Yes

[0328] The peptide-linker conjugates shown in Table 6 were incubated with either proline iminopeptidase ("PIP") or rat aminopeptidase N ("Rat APN"), and peptide cleavage was monitored by LCMS. An example of an LCMS demonstrating cleavage of Entry 5 from Table 6 is shown in FIG. 15B. All other cleavage reactions were measured in a similar manner. As shown in Table 6, while positively charged peptide-DNA conjugates ("oligo" and "oligo-PEG" linkers) were not cleaved by the aminopeptidases tested, all other conjugate classes (negatively charged, aromatic, hydrophobic) with DNA oligonucleotide linkers were cleaved. By comparison, the positively charged peptide-PEG conjugates were shown to be cleaved by at least one of the aminopeptidases.

[0329] Using labeled peptide-linker conjugates, it was shown that peptides of different compositions could be immobilized to individual sample well surfaces for single molecule analysis. For these experiments, the DNA linker was labeled with a dye (e.g., as depicted in FIG. 15A for the peptide-DNA conjugate), and loading of different peptide-DNA conjugates into individual sample wells was measured by dye fluorescence. An example loading experiment is shown in FIG. 15C. By measuring fluorescence emission of a labeled peptide-DNA conjugate (50 pM), it was determined that at least 18% of sample wells on a chip were loaded at single occupancy per sample well with a surface-immobilized conjugate. These experiments demonstrated that peptide-linker conjugates display enhanced aqueous solubility compared to non-conjugated peptide counterpart, that conjugated linkers do not prevent terminal amino acid cleavage of peptides by different aminopeptidases, and that peptide-linker conjugates of different compositions can be immobilized to chip surfaces at the single molecule level.

Example 2. Exopeptidase Cleavage of Polypeptide Substrates

[0330] The cleavage capabilities of various aminopeptidases were tested. Cleavage of peptide substrates was assayed using High Performance Liquid Chromatography (HPLC). A summary of amino acid cleavage activities for select exopeptidases is shown in FIG. 16. Specific cleavage activities are shown for the following enzymes: "cVPr" (*V. proteolyticus* aminopeptidase), "yPIP" (*Y. pestis* proline iminopeptidase), "D/E APN" (*L. pneumophila* M1 Aminopeptidase), hTET (*Pyrococcus horikoshii* TET aminopeptidase), and Pfu API ("PfuTET"). Specific activities with respect to terminal amino acids are classified as shown, with single-letter

abbreviations used for amino acids ("XP–" represents any terminal amino acid (X) having an adjacent, or penultimate, proline (P) residue).

Example 3. Terminal Amino Acid Cleavage of Immobilized Peptides at Single Molecule Level

[0331] Assays for on-chip amino acid cleavage of immobilized peptides were developed using labeled peptide conjugates. The assays were designed to provide a method for determining enzymatic recognition and cleavage activity of exopeptidases toward immobilized peptides, which could permit measurement of kinetic binding parameters and general binding affinities.

[0332] To evaluate N-terminal amino acid cleavage of a peptide, a dye labeled peptide was designed and synthesized which contained an N-terminal aspartate that was attached to the dye by way of a PEG spacer. This peptide also contained a proline residue adjacent to the modified aspartate that is recognized specifically by the enzyme proline iminopeptidase (from *Yersinia pestis*, known elsewhere and referred to herein as "yPIP"). The enzyme yPIP should cleave only an N-terminal amino acid upstream from a proline residue.

[0333] After showing that this and other labeled peptides were efficiently cut by yPIP in bulk (e.g., as described in Example 1), an on-chip dye/peptide conjugate assay was developed to observe N-terminal amino acid cleavage at the single molecule level. FIG. 17A shows a general scheme for the dye/peptide conjugate assay (inset panel). As shown, a peptide having a label attached to an N-terminal amino acid via a spacer is immobilized to a surface by way of a linker. After being exposed to peptidase, N-terminal amino acid cleavage results in the removal of the labeled residue from a detectable observation volume and is measured by a concomitant loss in signal from the label. The enzyme-peptide complex to the right of the inset panel generically depicts the N-terminal cleavage site.

[0334] FIG. 17A shows a labeled peptide construct (at bottom) that was designed and synthesized for use in the dye/peptide conjugate assay. In these experiments, a rhodamine dye (ATTO Rho6G) was attached to an N-terminal aspartate residue of a peptide having a penultimate proline residue at the N-terminus. As shown, the peptide was further conjugated to a solubilizing DNA linker with a biotin moiety for surface immobilization.

[0335] The labeled peptide conjugate was loaded onto a glass chip having an array of sample wells. Images of the chip were acquired before and after loading to determine the percent loading of sample wells at single occupancy by rhodamine fluorescence. The enzyme yPIP was then introduced onto the loaded chip and allowed to incubate for two hours at 37 °C. An image of the chip following the introduction of yPIP was taken and the percentage of green dyes lost were calculated to evaluate N-terminal amino acid cleavage. FIG. 17B shows imaging results from an experiment which displayed 6-7% loading in the loading stage and 91% loss of signal in

previously loaded wells after incubation with yPIP, which was indicative of N-terminal amino acid cleavage. FIG. 17C shows representative signal traces from these experiments, which demonstrate a detected increase in dye signal upon loading of labeled peptide and a detected loss in dye signal following exposure to yPIP.

[0336] As further confirmation of N-terminal amino acid cleavage at the single molecule level, on-chip FRET assays were developed to evaluate exopeptidase recognition and cleavage activity. FIG. 18A generically depicts a FRET peptide conjugate assay (panel A) and a FRET enzyme conjugate assay (panel B). In the FRET peptide conjugate assay (panel A), an immobilized peptide construct includes a FRET donor label attached to the linker and a FRET acceptor label attached at the N-terminus. N-terminal amino acid cleavage is detected by a loss in signal from the FRET acceptor label when exposed to peptidase. Additionally, this design permits monitoring loading of the peptide conjugate throughout an experiment by following emission from the FRET donor label.

[0337] In the FRET enzyme conjugate assay (panel B), an immobilized peptide construct includes a first label of a FRET pair attached to the linker and a peptidase is labeled with a second label of the FRET pair. N-terminal amino acid cleavage is detected by an enhancement in fluorescence attributable to FRET interactions, which would occur with sufficient proximity of peptidase to peptide and with sufficient residence time at the N-terminus. Additionally, this assay permits evaluating processive amino acid cleavage by a processive exopeptidase by detecting an increasing FRET signal over time with processive cleavage.

[0338] FIG. 18A also shows a FRET peptide construct under panel A that was designed and synthesized for use in the FRET peptide conjugate assay of panel A. As shown, the FRET peptide construct included a rhodamine dye (ATTO 647N) attached to an N-terminal aspartate residue of a peptide having a penultimate proline residue at the N-terminus. The peptide was further conjugated to a solubilizing DNA linker which was attached to a cyanine dye (Cy3B) for FRET and a biotin moiety for surface immobilization.

[0339] In this experiment, the FRET peptide construct was loaded onto a glass chip having an array of sample wells, and collected light was filtered first by a green filter and then a red filter. Loading of the FRET peptide construct was detected by measuring a signal passing through both the green and red filters. Terminal amino acid cleavage was detected when the signal was measurable only in the green filter, which indicated that the red dye conjugated N-terminal amino acid from the FRET peptide construct was cleaved by yPIP. This detection pattern is illustrated in panel C. As shown, if both dyes are detectable before the addition of yPIP, and only the green dye is visible after incubation with yPIP, it can be reasonably concluded this change in detection pattern is due to cleavage of the peptide and not photobleaching or loss of

the peptide as a whole. Additionally, an increase in fluorescence from the lone green dye would be expected, as its emissions are no longer absorbed by the red dye.

[0340] Following loading of the FRET peptide construct onto the chip, which had been modified by surface passivation using phosphonic acid and silane, yPIP was introduced and images were obtained at several time points. To assess the overall cleaving trend, the ratio of (green)/(green + red) was computed for each experiment. This ratio increases with the extent of cleaving that occurs. FIG. 18B is a plot of FRET emission ratio across all apertures at different time points of incubation with yPIP. As shown, the green dye contribution to the ratio of fluorescence emissions increases over time during incubation with yPIP, indicating that more N-terminal aspartate residues have been cut, leaving behind the truncated peptide with just the green dye.

[0341] Cutting efficiency was then evaluated at different time points by determining at which time points dye fluorescence was observed. This was done with simple thresholding – e.g., if the average dye emission signal was > 2.5 during excitation, the dyes were determined to be present (when each corresponding filter was applied). Apertures exhibiting cutting would then display both green and red dyes during the loading phase of the experiment, but only green dye at time points exposed to yPIP. As shown in FIG. 18C, progressively more cutting was observed as the chip was exposed to longer incubation times with yPIP. Example signal traces showing cutting displayed at each of the three yPIP-treated time points are shown in FIG. 18D.

[0342] Additional experiments were performed with yPIP and other peptidases using chips that had been modified by surface passivation using dextran, which produced similar results showing an increase in terminal amino acid cleavage over time following introduction of peptidase onto chips. FIG. 18E is a plot of FRET emission ratio across loaded apertures at different time points of incubation with yPIP. FIG. 18F is a plot of FRET emission ratio across loaded apertures at different time points of incubation with an aminopeptidase. Overall, the experiments here demonstrate that N-terminal amino acid cleavage is detectable in real-time at the single molecule level using different exopeptidases and different labeling strategies.

Example 4. Terminal Amino Acid Discrimination by Labeled Recognition Molecule

[0343] An adaptor protein involved in proteolytic pathways was identified as a potential candidate for use as a labeled recognition molecule for detecting N-terminal aromatic residues. The adaptor protein, ClpS2 from an α-proteobacterium (*A. tumefaciens*), was expressed and labeled at an exposed cysteine residue. FIG. 19A shows a crystal structure of the ClpS2 protein, with the exposed cysteine residue shown as sticks. The exposed cysteine residue was labeled with a rhodamine dye (ATTO 532).

[0344] Peptides having different N-terminal aromatic residues were prepared to test whether the labeled ClpS2 was capable of N-terminal amino acid discrimination at the single molecule level. Example single molecule intensity traces from these experiments are shown in FIG. 19B. As shown, the signal traces demonstrate residue-specific on-off binding patterns corresponding to the labeled recognition molecule reversibly binding the N-terminus of peptides having either: an N-terminal phenylalanine residue (F, top signal trace), an N-terminal tyrosine residue (Y, middle signal trace), or an N-terminal tryptophan residue (W, bottom signal trace).

[0345] Further analyses of the single molecule trajectories were carried out, with the results shown in FIGs. 19C-19E. FIG. 19C is a plot showing discriminant pulse durations (time duration of signal peaks) among the three N-terminal residues when reversibly bound by labeled ClpS2. FIG. 19D is a plot showing discriminant interpulse durations (time duration between signal pulses) among the three N-terminal residues. FIG. 19E shows plots which further illustrate the discriminant pulse durations among phenylalanine, tyrosine, and tryptophan at peptide N-termini. Mean pulse duration for the different N-terminal residues is visualized by histograms (A)-(B) and layered histogram (C).

[0346] Another adaptor protein, ClpS from *Thermosynochoccus elongatus* (teClpS) was evaluated for use as a labeled recognition molecule for leucine recognition. The data obtained from dwell time analysis, shown in FIGs. 19F-19H, demonstrated that the labeled teClpS protein produces detectable binding interactions with a terminal leucine residue of polypeptides with a mean pulse duration of 0.71 seconds. The amino acid sequence of the teClpS protein used in these experiments is shown in Table 1.

[0347] Similar experiments were carried out to evaluate *A. tumefaciens* ClpS1 and *S. elongatus* ClpS2 as potential reagents for leucine recognition, and GID4 as a potential reagent for proline recognition. FIG. 19I shows example results from dwell time analysis which showed differentiable recognition of phenylalanine, leucine, tryptophan, and tyrosine by *A. tumefaciens* ClpS1. FIG. 19J shows example results from dwell time analysis demonstrating leucine recognition by *S. elongatus* ClpS2. FIGs. 19K-19L show example results from dwell time analysis demonstrating proline recognition by GID4.

[0348] To evaluate the kinetics of one recognition molecule binding different types of terminal amino acids, the binding affinities of *A. Tumefaciens* ClpS2-V1 (atClpS2-V1) for different peptides were determined in a binding polarization assay. Exemplary binding curves with K_D values are shown in FIG. 19M for atClpS2-V1 with peptides having different N-terminal amino acids: phenylalanine (left plot), tyrosine (middle plot), and tryptophan (right plot). Each peptide contained an alanine residue at the adjacent, penultimate position. Based on the ensemble measurements, the K_D values determined with atClpS2-V1 were as follows: 743 nM for the

phenylalanine peptide, 2049 nM for the tyrosine peptide, and 3510 nM for the tryptophan peptide.

Example 5. Polypeptide Sequencing by Recognition During Degradation

[0349] Experiments were conducted to evaluate peptide sequencing by N-terminal amino acid recognition during an ongoing degradation reaction. Example results from these experiments are shown in FIGs. 20A-20D, which show single molecule intensity traces obtained over two independent polypeptide sequencing reactions conducted in real-time using a labeled ClpS2 protein and an aminopeptidase in the same reaction mixture. In each reaction, a polypeptide of sequence YAAWAAFADDDWK (SEQ ID NO: 234) was immobilized to a chip surface through the C-terminal lysine residue by loading the peptide composition (10 pM) onto chips for 20 minutes, and the immobilized peptide was monitored in the presence of a labeled recognition molecule (ATTO 542-labeled *A. Tumefaciens* ClpS2-V1 at 500 nM) and an aminopeptidase cleaving reagent (VPr at 8 μM).

[0350] FIGs. 20A and 20C show signal trace data for two different sequencing runs, with the top panel (panel 1 in FIG. 20A, panel 2 in FIG. 20C) showing a full trace, and the bottom panels (Y, W, F) showing zoomed-in regions corresponding to each of the highlighted regions in the full trace. FIGs. 20B and 20D show pulse duration statistics in histograms for the trace data of the corresponding panels as labeled in FIGs. 20A and 20C, respectively. As shown in the full signal trace of each sequencing run (panels 1, 2), three separate time intervals of signal pulses were observed over the course of the reaction. As highlighted by the zoomed-in regions (panels Y, W, F), the three intervals are visually distinguishable from one another based on an observable difference in pattern of signal pulses.

[0351] To further analyze the signal pulse data, pulse duration statistics were determined for each time interval (FIGs. 20B and 20D). The differences in pulse duration distribution were determined to correspond to those observed for these amino acids individually in steady-state on-chip binding assays with ClpS2, and the signal pulse information was phenotypically consistent between intervals from sequencing runs and the individual amino acid binding assays.

[0352] As confirmed by the analysis of signal pulse information, the three time intervals of signal pulses observed over the progression of each sequencing run correspond to recognition patterns of Y, W, and F, respectively (panels 1, 2). The intervening time period between signal pulse patterns is due to the selectivity of ClpS2-V1, which does not bind to N-terminal alanine residues. As illustrated by the full signal trace, the first interval corresponds to Y recognition, which is followed by a pause as VPr peptidase cuts Y and two alanine residues, followed by the second interval corresponding to W recognition, which is followed by another pause as VPr

peptidase cuts W and two alanine residues, and finally the third interval corresponding to F recognition before VPr peptidase cuts off the F and stops at the remaining ADDDWK peptide. These results show that pulse duration information, which was obtained by terminal amino acid recognition during an ongoing degradation reaction, can be used to determine characteristic patterns that discriminate between different types of terminal amino acids.

Example 6. Terminal Amino Acid Identification and Cleavage by Labeled Exopeptidase

[0353] Studies were performed to investigate the potential for a single reagent that is capable of both identifying a terminal amino acid of a peptide and cleaving the terminal amino acid from the peptide. As a single reagent, an exopeptidase must be able to bind to the peptide while retaining cleavage activity toward a terminal residue. Accordingly, an initial approach employing traditional labeling strategies was carried out by targeting the native surface-exposed amino acids of different exopeptidases. In these experiments, surface-exposed cysteine (–SH) or lysine ($–NH_2$) residues were labeled with fluorescent dyes, which proved to be a robust methodology for exopeptidase labeling. In certain cases, however, this approach produced a heterogeneous population of proteins that are labeled with one or more dyes.

[0354] In order to more precisely control where labeling occurs on exopeptidases and ensure that each exopeptidase molecule is labeled with a single fluorescent dye (as well as eliminate off-target reactivity of the dye), a new labeling strategy was investigated. In these experiments, labeled exopeptidases were prepared using a site-specific labeling strategy in which an unnatural amino acid containing a reactive functional group is introduced into the exopeptidase (see, e.g., Chin, J.W., et al. J Am Chem Soc. 2002 Aug 7; 124(31):9026-9027).

[0355] The proline iminopeptidase from Yersinia pestis (yPIP) was modified by mutation of a lysine residue at position 287 to a residue having a para-azidophenylalanine (pAzF) side chain. FIG. 21A shows a crystal structure of yPIP, with the mutation indicated by the chemical structure of pAzF shown with the K287 sidechain shown as sticks. This mutation site was selected based on the stability provided by the alpha helix at this position and to ensure that the new azido functional group is solvent exposed.

[0356] A pEVOL plasmid containing the mutant amino tRNA synthetase and the mutant tRNA necessary to incorporate pAzF into the amino acid chain was obtained. The amber stop codon (TAG), which is necessary for the specific incorporation of pAzF, was then introduced into the cDNA using the QuickChange II mutagenesis kit. The cDNA was then sequenced and the TAG codon position was confirmed. This was followed by co-transfection of both the pET21b+ plasmid containing the yPIP amber mutant and the pEVOL plasmid containing the cellular machinery to charge the tRNA for the amber codon with pAzF. The co-transfected cells were

then grown to 0.8 ODU, induced with 0.02% arabinose and 1 mM IPTG in the presence of 2 mM pAzF in 2 L of LB, and harvested using chemolysis. Purification was carried out using a 5 mL affinity chromatography column, and the protein was eluted in 100 mM imidazole. The resulting protein was then dialyzed and concentrated into 50 mM HEPES pH 8.0 and 0.5 M KCl, aliquoted, and flash frozen prior to storage at -20 °C.

[0357] To confirm the presence of the azido group in the purified protein, DBCO-Cy3 (2 mM) was reacted with the pAzF-yPIP variant (220 μM) (Reaction Conditions: 50 mM HEPES pH 8.0, 0.5 mM KCl, 20% DMSO; 10 hours at 37 °C, 48 hours at room temperature). The protein reaction product was purified by size-exclusion chromatography, and it was determined that the resulting protein was 100% labeled with the azide-reactive DBCO-Cy3 reagent (FIG. 21B), indicating robust incorporation of the unnatural amino acid.

[0358] Protein labeling and purity of the final product was confirmed by SDS-PAGE analysis of the unlabeled and labeled pAzF variant. FIG. 21C shows a picture of SDS-PAGE gel confirming Cy3-labeling of pAzF-yPIP (overexposed image of gel shown in FIG. 21D to show ladder). FIG. 21E shows a picture of Coomassie-stained gel confirming that both dye and protein co-migrate and are pure.

[0359] The dye-labeled pAzF-yPIP variant was used in an activity assay to confirm that the enzyme was still active after labeling and purification. As shown in FIG. 21F, Cy3-pAzF-yPIP was able to hydrolyze 100% of the peptide substrate in 1 hour using 1000-fold excess substrate, as measured by HPLC. These experiments demonstrate a methodology which allows site-specific modification and labeling of an exopeptidase with minimal perturbation of the native protein structure/function.

Example 7. Recognition of Modified Amino Acids in Polypeptide Sequencing

[0360] Experiments were performed to evaluate recognition of amino acids containing specific post-translation modifications. A triple-mutant variant (T8V, S10A, K15L) of the Src Homology 2 (SH2) domain from Fyn, a tyrosine kinase, was tested as a potential recognition molecule for phosphorylated tyrosine residues in peptide sequencing. The variant protein was immobilized to the bottom of sample wells, and single-molecule signal traces were collected upon addition of a fluorescently-labeled peptide containing N-terminal phospho-tyrosine. Peptide binding by the immobilized protein was detected during these experiments, as shown by the representative traces in FIG. 22A. Pulse duration data collected during these experiments is shown in FIG. 22B (top, middle, and bottom plots corresponding to the top, middle, and bottom traces of FIG. 22A, respectively). Pulse duration and interpulse duration statistics are shown in FIG. 22C (top and bottom panels, respectively).

[0361] Control experiments were performed to confirm that the Fyn protein was specific for the phosphorylated tyrosine. The experiments were repeated for each of three different peptides: a first peptide containing N-terminal unmodified tyrosine (Y; FIG. 22D), a second peptide containing N-terminal and penultimate unmodified tyrosines (YY; FIG. 22E), and a third peptide containing N-terminal phospho-serine (FIG. 22F). As shown, binding was not detected with any of the peptides used in the negative control experiments.

Example 8. Recognition of Penultimate Amino Acids in Polypeptide Sequencing

[0362] Experiments were performed to determine the effects of penultimate amino acids on pulse duration for *A. Tumefaciens* ClpS2-V1. Forty-nine different fluorescently-labeled peptides were prepared containing unique dipeptide sequences at the N-terminus, where the N-terminal amino acid was F, W, or Y, and the penultimate position was one of the 20 natural amino acids. For each experiment, ClpS2-V1 was immobilized at the bottom of sample wells, and single-molecule signal traces were collected for 10-20 minutes upon addition of one of the fluorescently-labeled peptides. Pulse duration data was collected for a minimum of 50 sample wells for each peptide.

[0363] FIG. 23 shows the median pulse duration for each of the 50 peptides, with data points grouped by penultimate amino acid (x-axis) and N-terminal amino acids represented with different symbols.

Example 9. Simultaneous Amino Acid Recognition with Multiple Recognition Molecules

[0364] Single-molecule peptide recognition experiments were performed to demonstrate terminal amino acid recognition of an immobilized peptide by more than one labeled recognition molecule. Single peptide molecules containing N-terminal phenylalanine (FYPLPWPDDDY (SEQ ID NO: 235)) were immobilized in sample wells of a chip. Buffer containing 500 nM each of atClpS1 (*Agrobacterium tumifaciens* ClpS1; sequence provided in Table 1) and atClpS2-V1 (*Agrobacterium tumifaciens* ClpS2 variant 1; sequence provided in Table 1) was added, where atClpS1 and atClpS2-V1 were labeled with Cy3 and Cy3B, respectively. Since the intensity of Cy3B is higher than Cy3, atClpS2-V1 binding events were readily distinguishable from atClpS1 binding events.

[0365] FIGs. 24A-24C shows the results of the experiments showing single-molecule peptide recognition with differentially labeled recognition molecules. A representative trace is displayed in FIG. 24A. The pulse duration distributions were distinct for each binder (FIG. 24B) and corresponded to their kinetic profiles as observed in single-binder experiments. Mean pulse duration was 1.3 seconds for atClpS1 and 1.0 seconds for atClpS2-V1 (FIG. 24C). Pulse rate

was also distinct: 8.1 pulses/min for atClpS1 and 14.1 pulses/min for atClp2-V1 (FIG. 24C). Thus, when more than one recognition molecule is included for dynamic recognition of immobilized peptides, the binding characteristics of each recognition molecule (including pulse duration, interpulse duration, and pulse rate) can simultaneously provide information about peptide sequence.

Example 10. Enhancing Photostability with Recognition Molecule Linkers

[0366] Experiments were performed to evaluate the photostability of immobilized peptides during single-molecule sequencing. The dye-labeled atClpS2-V1 described in Example 4 was added to sample wells containing immobilized peptide substrates in the presence of excitation light at 532 nm to monitor recognition by emission from ATTO 532. A representative trace is shown in FIG. 25A. As shown in the top panel, recognition was observed to cease at approximately 600 seconds into the experiment. The bottom panel is a zoomed view showing signal pulses at approximately 180-430 seconds into the reaction.

[0367] FIG. 25B shows a visualization of the crystal structure of the ClpS2 protein used in these experiments. As shown, the cysteine residue that serves as the dye conjugation site is approximately 2 nm from the terminal amino acid binding site. It was hypothesized that photodamage to the peptide was caused by proximity of the dye to the N-terminus of peptide during binding. To mitigate the potential photodamaging effects of dye proximity, the ClpS2 protein was dye-labeled through a linker that increased distance between the dye and N-terminus of peptide by more than 10 nm. The linker included streptavidin and a double-stranded nucleic acid; the double-stranded nucleic acid was labeled with two Cy3B dye molecules and attached to streptavidin through a bis-biotin moiety, and a ClpS2 protein was attached to each of the remaining two binding sites on streptavidin through a biotin moiety. A representative trace using this dye-shielded ClpS2 molecule is shown in FIG. 25C. As shown in the top panel, recognition time was extended to approximately 6,000 seconds into the experiment. The bottom panel is a zoomed view showing signal pulses at approximately 750-930 seconds into the reaction.

[0368] A DNA-streptavidin recognition molecule was generated with a linker containing a double-stranded nucleic acid labeled with two Cy3B dye molecules and attached to streptavidin through a bis-biotin moiety, and a single ClpS2 protein attached to the remaining two binding sites on streptavidin through a bis-biotin moiety. This construct was used in a single-molecule peptide sequencing reaction, and representative traces from these experiments are shown in FIGs. 26A-26D.

[0369] The sequencing experiments described in Example 5 were repeated, with the reaction conditions changed as follows: the DNA-streptavidin ClpS2 recognition molecule was used in

combination with hTET amino acid cleaving reagent. A representative signal trace is shown in FIG. 27.

Example 11. Sequencing by Recognition During Degradation by Multiple Exopeptidases

[0370] Experiments were performed to evaluate the use of multiple types of exopeptidases with differential cleavage specificities in a single-molecule peptide sequencing reaction mixture. Single peptide molecules (YAAWAAFADDDWK (SEQ ID NO: 234)) were immobilized through a C-terminal lysine residue in sample wells of a chip. Buffer containing atClpS2-V1 for amino acid recognition and hTET for amino acid cleavage was added. A representative trace is displayed in FIG. 28A, with expanded views of pulse pattern regions shown in FIG. 28B.

[0371] An experiment was carried out to evaluate sequencing reactions in the presence of two types of exopeptidases with differential specificities. Single peptide molecules (FYPLPWPDDDYK (SEQ ID NO: 236)) were immobilized through a C-terminal lysine residue in sample wells of a chip. Buffer containing atClpS2-V1 for amino acid recognition, and both hTET and yPIP for amino acid cleavage was added. A representative trace is displayed in FIG. 28C, with expanded views of pulse pattern regions shown in FIG. 28D. Additional representative traces from these reaction conditions are shown in FIG. 28E.

[0372] Further experiments were carried out to evaluate sequencing reactions in the presence of two types of exopeptidases with differential specificities. Single peptide molecules (YPLPWPDDDYK (SEQ ID NO: 237)) were immobilized through a C-terminal lysine residue in sample wells of a chip. In one experiment, buffer containing atClpS2-V1 for amino acid recognition, and both hTET and yPIP for amino acid cleavage was added. A representative trace is displayed in FIG. 28F, with expanded views of pulse pattern regions shown in FIG. 28G. Additional representative traces from these reaction conditions are shown in FIG. 28H. In a further experiment, buffer (50 mM MOPS, 60 mM KOAc, 200 μM $Co(OAc)_2$) containing atClpS2-V1 for amino acid recognition, and both PfuTET and yPIP for amino acid cleavage was added. A representative trace is displayed in FIG. 28I, with expanded views of pulse pattern regions shown in FIG. 28J.

Example 12. Identification and Evaluation of New ClpS Proteins

[0373] ClpS proteins that bind to N-terminal phenylalanine (F), tryptophan (W), tyrosine (Y), and leucine (L) have been reported. To search for potentially new ClpS homologs with previously unknown N-terminal amino acid binding properties, a highly diverse panel of uncharacterized ClpS proteins from approximately 60 species encompassing all of the sequence diversity present in this protein family was designed based on homology analysis.

[0374] The panel of ClpS proteins were screened using a high-throughput expression and purification workflow. ClpS proteins were overexpressed in *E. coli* cells at 100-mL scale, biotinylated *in vivo* by co-expressed biotin ligase, released by cell lysis, complexed with streptavidin, and purified by cobalt affinity chromatography. Analysis by SDS-PAGE chromatography showed >85% pure ClpS protein (FIG. 29A).

[0375] The amino acid binding profile of each new homolog was evaluated by biolayer interferometry. ClpS proteins with known binding profiles (atClpS2-V1, atClpS1, and teClpS) were included as controls. In these assays, peptides containing an N-terminal amino acid of interest were immobilized on a biolayer interferometry sensor surface, allowed to bind to ClpS protein, and incubated in buffer to allow the ClpS molecules to dissociate. The binding response (nm) and dissociation off-rate (kdis, s^{-1}) were measured using a kinetic model. Measurements with each ClpS protein were carried out at 30 °C for the amino acids I, V, M, F, Y, and W, along with D, E, A, and R as negative controls. All peptides contained a penultimate alanine (A) and consisted of the sequence XAKLDEESILKQK (SEQ ID NO: 238).

[0376] Results from the biolayer interferometry screening of 61 ClpS homologs (58 new homologs, 3 reference homologs) are shown in FIG. 29B. Select results for 14 ClpS homologs, are summarized in the plots shown in FIG. 29C. Response (y-axis) is plotted against the inverse of dissociation rate for eight different dipeptides (data not shown for negative controls DA, AA, and RA). In these plots, higher values on the y-axis correspond to stronger association response and higher values on the x-axis indicate slower dissociation rates. Among the ClpS homologs that were screened, the homolog "PS372" stood out as a binder of leucine, isoleucine, and valine. Response curves for PS372 with LA, IA, and VA are shown in FIG. 29D. PS372 did not show binding to the F, Y, or W peptides. The results indicated that PS372 binds IA and LA with kinetic parameters that are typical of binders that display detectable on-and-off binding for sequencing reactions.

[0377] ClpS homologs were evaluated by fluorescence polarization to measure the interaction of a FITC-labeled peptide (XAKLDEESILKQK-FITC (SEQ ID NO: 238)) with ClpS:streptavidin complexes. Measurements of polarization response (millipolarization, mP) and total intensity for I, V, L, F, Y, and W were performed in a high-throughput 384-well plate format, using 480 nm excitation and 530 nm emission wavelength with readings collected after 30 minute incubation at 23 °C. The results with in-solution binding validated that PS372 binds with leucine, isoleucine, and valine with strength of response in the order LA>IA>VA (FIG. 29E). Binding responses of atClpS2-V1 were FA>YA>WA, and those of teClpS were LA>FA>YA>WA.

[0378] The effect of penultimate amino acids on isoleucine and valine binding by PS372 were evaluated by measuring binding to peptides with N-terminal IR, IQ, and VR by biolayer

interferometry (peptide sequence: XXKLDEESILKQK (SEQ ID NO: 291)). Measurements were also obtained with teClpS for comparison. The results showed that PS372 binds IQ and IA with higher affinity relative to IR (FIG. 29F), whereas teClpS displayed a minor response with very fast dissociation for IR peptide only. PS372 binds well with VA and has negligible response with VR, and teClpS showed no response with V peptides.

[0379] Binder-on-chip experiments were performed for PS372 with LF, LA, IR, IA, VR, and VA peptides to evaluate recognition by signal pulse detection in single-molecule assays. PS372 displayed long mean pulse widths with LF and LA peptides (1207 ms and 876 ms, respectively). For comparison, the mean pulse width observed with the L-binder teClpS was 768 ms (~37% shorter than PS372). Pulsing with short pulse durations was observed for PS372 with both IR (83 ms) and IA (70 ms) peptides. FIG. 29G shows histograms of pulse widths for PS372 with IR peptide (top) and LF peptide (bottom), with representative traces in a 5-minute window shown to the right of each histogram.

[0380] These experiments demonstrate the identification and biochemical characterization of the first ClpS protein known to display strong inherent affinity for isoleucine and valine. Based on analysis of binding kinetics using biolayer interferometry and fluorescence polarization, PS372 recognizes N-terminal leucine, isoleucine, and valine, unlike all other known ClpS proteins which are limited to recognition of W, F, Y, and L. Additionally, observable binding of PS372 with I and L at the single-molecule level was confirmed in binder-on-chip assays, and signal pulse data showed advantageous properties for sequencing reactions.

Example 13. Labeling of New ClpS Homolog and Use in Sequencing

[0381] A large scale (5L) batch of the ClpS homolog PS372 and BirA plasmids were transformed into *E. coli* and expressed overnight with 0.4 mM IPTG and 160 μM biotin (for in-vivo biotinylation) at 14 °C. Harvested cells were purified over 5 mL Cobalt affinity chromatography column. Biotinylation efficiency was evaluated using SDS-PAGE (FIG. 30A) and was determined to be 100% based on complete shift of PS372 band when mixed with 6 times excess of Streptavidin (SV).

[0382] Biotinylated PS372 (PS372-Bt) was transferred into an appropriate buffer for PEGylation using dialysis. The protein was then PEGylated using mPEG4-NHS ester at room temperature for 2 hours followed by overnight dialysis to remove excess mPEG4-NHS. PEGylated and biotinylated PS372 (PS372-Bt-mPEG4) was concentrated and conjugated with a pre-formed 1:1 complex of SV and a biotinylated Cy3B-labeled oligonucleotide and purified over HPLC column. The HPLC profile showed two major peaks (FIG. 30B), with Peak 3 as the peak of interest (1:1 (PS372:SV Dye) complex).

[0383] Following HPLC purification, the resulting peaks were concentrated to appropriate concentrations and tested for any free biotin binding sites on the labeled SV, which could interfere in on-chip assays due to non-specific binding to the surface of the chip. Free biotin labeled with AttoRho6 was spiked in all samples and run on SDS-PAGE along with controls as shown in FIG. 30C. The Cy3 channel (panels (a) and (c)) showed presence of 1:1 species (PS372:SV-Dye) in lanes 6 and 7, and absence of AttoRho6 (panels (b) and (c)) confirmed no free biotin binding sites on the labeled protein.

[0384] The labeled protein was tested in peptide-on-chip recognition assays and showed binding to N-terminal isoleucine peptide (IAALAAVAADDDW (SEQ ID NO: 239)) with mean pulse width of 78 ms and to N-terminal leucine peptide (LAAIAAFAADDDW (SEQ ID NO: 292)) with mean pulse width of 957 ms. Dynamic sequencing assay results in the presence of exopeptidase (hTET) showed a clear transition from I to L for N-terminal isoleucine peptide, and from L to I for N-terminal leucine peptide (FIG. 30D).

[0385] Experiments were conducted to evaluate peptide sequencing by N-terminal amino acid recognition with dye-labeled PS372 during an ongoing degradation reaction. Real-time dynamic peptide sequencing assays were carried out by monitoring a surface-immobilized polypeptide (IAALAAVAADDDW (SEQ ID NO: 239)) in the presence of dye-labeled PS372 recognition molecule and an aminopeptidase cleaving reagent (hTET). Example data from a real-time dynamic peptide sequencing assay are shown in FIG. 30E, with mean pulse width listed below each cluster of pulses. As shown, the differences in pulse width distribution readily distinguished isoleucine and leucine recognition, demonstrating recognition of these amino acids by dye-labeled PS372 in a dynamic sequencing reaction.

Example 14. Engineering New Methionine-Binding ClpS Proteins

[0386] A library of ClpS mutants with a potential diversity of 160,000 variants was created starting from a thermostable variant of *Agrobacterium tumefaciens* ClpS2 (*see* Tullman et al. 2020, *Biochem. Enj. J.* 154:107438). The library was created by homologous recombination in yeast using degenerate primers to PCR amplify the mutated ClpS gene such that the protein can be displayed on the surface of the yeast. Using this method, each protein variant is displayed in approximately 1,000 copies on the surface of a yeast cell, and by isolating the clone one can determine the genetic sequence of the protein that is displayed. This maintains the genotype-phenotype link necessary to determine which protein variant has the properties of interest.

[0387] After obtaining over 1 million clones, selections were performed via fluorescence-activated cell sorting (FACS) by combining the yeast displaying the variant library with a peptide containing the sequence MRFVGECK-biotin (SEQ ID NO: 240), and the fluorophores

streptavidin-PE and anti-myc AlexaFluor647. Cells that bind to the peptide and contain the myc-tag at the C-terminus of the ClpS protein are expected to be retained in Quadrant 2 (upper-right) of the plot obtained by FACS, and the cell-sorter sorts these into a culture tube to be grown for another round of selection and/or sequencing to identify the variants that have been captured. Two rounds of FACS were performed on this library resulting in an improvement seen in the plots shown in FIG. 31A (left plot: 1st round with 0.5 μM MR peptide; middle plot: 2nd round with 0.5 μM MR peptide; right plot: 2nd round with 0.05 μM MR peptide for comparison, showing little binding as expected).

[0388] After two rounds of selection, single clones of yeast were isolated by plating and sequenced. Two candidates emerged with the sequences of the 4 residues that were mutated as "TMR L" (PS490) and "TAF K" (PS489). Upon further characterization, it was shown using yeast display that the proteins bound MR, MQ, and IR, but did not bind IQ or EG peptides. These two clones were subcloned into expression vectors to facilitate purification and streptavidin labeling for biolayer interferometry assays. The proteins were expressed, purified, and assayed to measure the binding kinetics with different N-terminal amino acids alongside the original thermostable atClpS2 variant (PS023). The data shown in FIGs. 31B-31F confirmed binding of PS489 to the MR peptide, as well as MA, LA, and FA peptides. Additionally, PS490 binds MR, LA, and FA.

Example 15. Designed Mutational Variant ClpS Recognition Molecule

[0389] A variant of atClpS1 was created with active-site substitutions rationally designed to achieve longer pulse widths which can be favorable in single-molecule sequencing reactions. The designed variant (PS218) contains the following mutations relative to wild-type atClpS1: M51F, E52Q, M73T, T70W. Binder-on-chip experiments were performed for PS218 with FA, LF, WA, and YA peptides to evaluate recognition by signal pulse detection in single-molecule assays. As shown by the results in Table 7, the substitutions relative to the wild-type protein result in longer pulse widths with each peptide tested.

Table 7. Binder-on-chip results for PS218 and Wild-Type atClpS1.

Binder	Peptide	# of wells	Mean pulse width	Median PW
PS218	FA	212	1.28	0.78
PS218	LF	265	0.56	0.32
PS218	WA	180	1.00	0.58
PS218	YA	185	1.64	1.06
atClpS1 (WT)	FA	387	0.66	0.36

atClpS1 (WT)	LF	291	0.30	0.18
atClpS1 (WT)	WA	240	0.39	0.24
atClpS1 (WT)	YA	404	0.77	0.40

Example 16. Evaluation of UBR-Box Domain Homologs for Use in Sequencing

[0390] The inventors found that the UBR-box domain from yeast UBR exhibited strong binding affinity to R only when followed by L or I. To identify UBR-box domains with a wider range of high affinity R binding in the presence of penultimate amino acids, as is useful in single-molecule peptide sequencing, a screen of UBR-box domain homologs was conducted. The amino acid binding properties of UBR-box domain homologs were evaluated by biolayer interferometry. Peptides containing an N-terminal arginine (R), followed by different amino acids in the penultimate position, were immobilized on a biolayer interferometry sensor surface, allowed to bind to UBR-box protein, and incubated in buffer to allow the UBR-box molecules to dissociate. Example response trajectories are displayed for UBR-box homologs PS535 (FIG. 32A) and PS522 (FIG. 32B) binding with 14 polypeptides containing N-terminal R followed by different amino acids in the penultimate position. FIG. 32C is a heatmap showing results measured for 24 UBR-box homologs against 14 polypeptides containing N-terminal R followed by different amino acids in the penultimate position. The results demonstrate that homologs PS535, PS522, PS528, and PS505 maintain high binding affinity to R in the presence of a wide range of penultimate amino acids.

[0391] UBR proteins bind to the basic N-terminal residues R, K, and H as part of the conserved N-end rule pathway. To identify high-affinity UBR proteins with properties favorable for use in sequencing, a diverse panel of uncharacterized UBR-box domain homologs from 44 species (PS501-PS544) was designed. UBR-box domain homologs were expressed, purified, and evaluated for binding to N-terminal amino acids by biolayer interferometry. Measurements were carried out at 30 °C for each protein for R, K and H binding. The amino-acid binding profile of each new homolog was first evaluated by biolayer interferometry using a single-point screening assay. All peptides in this screen contained a penultimate alanine (A) and consisted of the sequence XAKLDEESILKQK (SEQ ID NO: 238), where X is R, K, or H.

[0392] The biolayer interferometry response measurements for RA/KA/HA peptide binding are summarized in FIG. 32D. The homologs PS522, PS528, PS535, and PS538 displayed favorable binding trajectories to RA and KA peptides with notably high response for RA peptide. PS528 showed a slower dissociation trajectory and displayed binding to HA peptide. Based on these results, a second panel of 15 UBR-box domain homologs (PS614-PS629) was designed,

including an expanded set of homologs related to *K. lactis*. PS621 showed favorable binding characteristics with results that were similar to PS528 (FIG. 32D).

[0393] To evaluate the effect of penultimate amino acids on arginine binding, UBR box-domain homologs were evaluated for binding to 14 different peptides with N-terminal RX (peptide sequence: RXGGGDDDFFK (SEQ ID NO: 241)). The selected RX dipeptides are the 14 most frequently found in the human proteome. The full set of biolayer interferometry response measurements for 40 UBR box proteins is summarized in FIG. 32E. The results showed that PS535, PS522, PS528, and PS621 have an extensive binding range for RX dipeptides. The mammalian origin UBR candidates PS522 and PS535 displayed a wide range of recognition for R dipeptides with faster dissociation rates (not shown). The homologs PS528 and PS621 from the *Kluveromyces* yeast family also displayed wide recognition of R dipeptides.

[0394] Single point fluorescence polarization assays were then performed for selected candidates. These assays measured the interaction of labeled peptide (XAKLDEESILKQK-FITC (SEQ ID NO: 238)) with UBR-streptavidin complex in a format in which these molecules are free in solution. Measurements of polarization response (millipolarization, mP) and total intensity for RA, KA, and HA peptides were performed. The results showed that PS528 and PS621 bind to RA peptides with strong responses (FIG. 32F). The response trajectories also indicated that PS528 and PS621 are strong RA binders (not shown) and bind with kinetic parameters that are typical of binders that display detectable on-and-off binding in dynamic sequencing reactions.

[0395] The binding affinity of selected UBR-box domain candidates was then evaluated by measuring K_D (dissociation constant) values for N-terminal R and K peptides. Assays were performed for RA and KA peptides at increasing concentration of UBR protein (FIG. 32G, top plots). K_D values were determined from the binding titration curves. As shown in FIG. 32G (bottom bar chart), binding was observed for RA and KA peptides, and PS621 and PS528 showed relatively high binding affinities for RA peptide (K_D values of 420 nM and 460 nM, respectively).

[0396] Peptide-on-chip recognition assays for N-terminal arginine peptide were performed with PS621 and PS528. Both PS528 and PS621 recognized N-terminal R, and binding to dye-labeled RL peptide at the single-molecule level was confirmed in these recognition assays, with average pulse width of approximately 60 ms. A representative trace for PS621 in peptide-on-chip recognition assay with dye-labeled RL peptide is shown in FIG. 32H.

[0397] Dynamic sequencing runs were also performed using a 3-binder system with the UBR proteins PS621 or PS528 combined with atClpS2-V1 and PS557. These sets of binders were shown to recognize R, L, I, and F of a natural peptide fragment from human ubiquitin (sequence:

DQQRLIFAGK (SEQ ID NO: 242)). Example sequencing traces for this peptide from a 3-binder dynamic sequencing run using PS528, atClpS2-V1, and PS557 are shown in FIG. 32I.

Example 17. Identification and Evaluation of PS372 Homologs for Use in Sequencing

[0398] To identify ClpS proteins capable of isoleucine and valine recognition in single-molecule peptide sequencing, a screen of PS372-homologous proteins was conducted. The amino acid binding properties of PS372 homologs were evaluated by biolayer interferometry as described in Example 12. Example response trajectories are displayed for PS372 (FIG. 33A) and homologs PS545 (FIG. 33B), PS551 (FIG. 33C), PS557 (FIG. 33D), and PS558 (FIG. 33E) binding with 4 polypeptides containing different N-terminal amino acids (I, V, L, F) followed by alanine in the penultimate position. FIG. 33F is a heatmap showing results measured for 34 PS372 homologs.

Example 18. Engineering Multivalent Amino Acid Binders

[0399] Multivalent amino acid binders were designed as tandem fusion molecules expressed from a single coding sequence containing segments encoding two copies of atClpS2-V1 joined end-to-end by a segment encoding a flexible peptide linker. Expression of the single coding sequence produced a single full-length polypeptide having two ClpS proteins oriented in tandem (Bis-atClpS2-V1). Three Bis-atClpS2-V1 binders, each having a different linker, were designed and expressed: PS609 ("Linker 1" sequence: GGGSGGGSGGGSG (SEQ ID NO: 243)); PS610 ("Linker 2" sequence: GSAGSAAGSGEF (SEQ ID NO: 244)); and PS611 ("Linker 3" sequence: GSAGSAAGSGEFGSAGSAAGSGEFGSAGSAAGSGEF (SEQ ID NO: 245)).

[0400] The PS610 polypeptide was biotinylated and conjugated with dye-labeled streptavidin, as described above. Dye-labeled PS610 was tested in peptide-on-chip recognition assays and showed binding to N-terminal phenylalanine peptide (FAAAYP (SEQ ID NO: 246)). For comparative purposes, the monovalent binder (dye-labeled atClpS2-V1) and the tandem binder PS610 were each tested in the recognition assays using the same binder concentration (500 nM). Representative single-molecule pulsing trajectories and pulse statistics from these experiments are shown in FIG. 34A (left panel: monovalent binder, right panel: PS610). As shown, the median interpulse duration was approximately 5.8 times shorter and the observed rate of pulsing was 2.3 times faster using the tandem binder, while the median pulse widths using the different binders were consistent.

[0401] Dynamic recognition of the N-terminus of immobilized insulin B-chain (sequence: FVNQHLCGSHLVEALYLVCGERGFFYTPKA (SEQ ID NO: 247)) was examined independently for dye-labeled atClpS2-V1 and PS610 at different binder concentrations. Single-molecule pulsing trajectories were acquired for 20 minutes. The results from these experiments

(shown in Table 8) were used to generate a plot of mean pulse rate as a function of binder concentration (FIG. 34B). As shown, the tandem binder PS610 displayed a higher mean rate of on-off binding (pulse rate) at each concentration.

Table 8. Dynamic recognition results for multivalent binders.

Binder	Num. traces	Binder conc. (nM)	Pulse Rate (/min)
PS610	1696	100	6
PS610	519	50	4.8
PS610	594	100	8.4
PS610	1625	250	6.6
atClpS2V1	461	100	3.6
atClpS2V1	105	50	2.4
atClpS2V1	399	100	4.2
atClpS2V1	2314	250	3.6

[0402] Experiments were conducted to evaluate peptide sequencing by N-terminal amino acid recognition with dye-labeled PS610 during an ongoing degradation reaction. Real-time dynamic peptide sequencing assays were carried out by monitoring a surface-immobilized peptide fragment of Glucagon-like peptide 1 (sequence: EFIAWLVK (SEQ ID NO: 248)) in the presence of dye-labeled PS610 (100 nM), dye-labeled PS372 (250 nM), and aminopeptidase cleaving reagents (hTET and pfuTET). PS372 was labeled with a distinguishable dye and was provided for I and L recognition, and iterative cleavage of N-terminal amino acids of the immobilized peptides was performed by the aminopeptidases. Example traces from a real-time sequencing assay are shown in FIG. 34C. Regions of pulsing corresponding to characteristic F and W recognition by PS610 were correctly identified using an automated analysis workflow.

[0403] Binder-on-chip experiments were performed for PS610, or monovalent binder for comparative purposes, with dye-labeled peptide containing N-terminal FA. Representative traces are shown in FIG. 34D for the monovalent binder (top trace) and for the tandem binder PS610 (bottom trace). As shown, immobilized PS610, consisting of tandem copies of atClpS2-V1, displays a pulse-over-pulse pattern in binding to freely-diffusing dye-labeled peptide. These results demonstrate that the two linked monomers of PS610 are each capable of independent and simultaneous binding to their target peptides.

[0404] The amino acid binding profile of each Bis-atClpS2-V1 binder (PS609, PS610, and PS611) was evaluated by biolayer interferometry as described above. The monovalent atClpS2-V1 was run as a control. Measurements with each binder were carried out for the amino acids F, Y, and W, along with I, L, M, and V as negative controls. All peptides contained a penultimate alanine (A). Response curves for the monovalent binder are shown in FIG. 35A. Response

curves for the tandem binders PS609, PS610, and PS611 are shown in FIGs. 35B, 35C, and 35D, respectively.

[0405] The multivalent binder PS651 was designed as a tandem fusion molecule expressed from a single coding sequence containing segments encoding three copies of atClpS2-V1 (Tris-atClpS2-V1) joined end-to-end by segments encoding a flexible peptide linker (Linker 2). Fluorescence polarization studies were performed with PS651, the Bis-atClpS2-V1 binders (PS609, PS610, PS611), and the monomeric binder atClpS2-V1, to determine the binding affinity of each for a peptide having an N-terminal phenylalanine residue with an alanine residue at the penultimate position. The K_D values obtained from these studies are reported below in Table 9. These results demonstrate that tandem atClpS2-V1 constructs containing different linkers retain binding to F, Y, and W.

Table 9. Binding affinities of tandem recognizers for N-terminal phenylalanine peptide.

Binder	FA Peptide (K_D, nM)
atClpS2-V1	185
PS609 (Bis-atClpS2-V1, linker 1)	221
PS610 (Bis-atClpS2-V1, linker 2)	271
PS611 (Bis-atClpS2-V1, linker 3)	137
PS651 (Tris-atClpS2-V1)	256

[0406] Additional multivalent amino acid binders were designed and evaluated for N-terminal recognition. Table 10 includes a list of different tandem binders that were designed and the corresponding polypeptide sequence. Each expression construct included a C-terminal His/bis-biotinylation tag from Table 3.

Table 10. Non-limiting examples of multivalent binders.

Name	SEQ ID NO:	Sequence
PS609 (Bis-atClpS2-V1, Linker 1)	249	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGGGSGGGSGGGSGMSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMS FVTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLG KEAGFPLMFTTEPEEGHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEW HEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS610 (Bis-atClpS2-V1, Linker 2)	250	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSF VTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGK EAGFPLMFTTEPEEGHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWH EGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS611	251	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE

(Bis-atClpS2-V1, Linker 3)		EGSAGSAAGSGEFGSAGSAAGSGEFGSAGSAAGSGEFMSDSPVDLKPKPKV KPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSEDTGRRVMMTAHRFG SAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPEEGHHHHHHHHHHGG GSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWH E
PS612 (atClpS2-V1 + PS372, Linker 2)	252	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIE MLQKIFGFPPEKGFQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPY LGRPCSGSMTCVIEPAVGGSHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQ KIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS613 (Bis-PS372, Linker 2)	253	MAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEMLQKIFGFPPEKG FQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPYLGRPCSGSMTCVI EPAVGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTP MSFVTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATD LGKEAGFPLMFTTEPEEGHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKI EWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS614 (PS372 + atClpS2-V1, Linker 2)	254	MAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEMLQKIFGFPPEKG FQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPYLGRPCSGSMTCVI EPAVGSAGSAAGSGEFMAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRY VIEMLQKIFGFPPEKGFQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGP DPYLGRPCSGSMTCVIEPAVGGSHHHHHHHHHHGGGSGGGSGGGSGLNDFF EAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS637 (Bis-PS557, Linker 1)	255	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGGGSGGGSGGGSGMPTAASATESAIEDTPAPARP EVDGRTKPKRQPRYHVVLWNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEV DTQGRVIVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEAEEGGSH HHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLN DFFEAQKIEWHE
PS638 (Bis-PS557, Linker 2)	256	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMPTAASATESAIEDTPAPARPE VDGRTKPKRQPRYHVVLWNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEVD TQGRVIVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEAEEGGSHH HHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLND FFEAQKIEWHE
PS639 (Bis-PS557, Linker 3)	257	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFGSAGSAAGSGEFGSAGSAAGSG EFMPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQY VVVMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGY DRLLARSKGSMKASIEAEEGGSHHHHHHHHHHGGGSGGGSGGGSGLNDFFE AQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS640 (atClpS2-V1 + PS557, Linker 2)	258	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVV LWNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAE LKRDQIHAFGYDRLLARSKGSMKASIEAEEGGSHHHHHHHHHHGGGSGGGS GGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS641 (PS557 + atClpS2-V1, Linker 2)	259	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPK LYKVMLLNDDYTPMSFVTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCE RDIAETKAKEATDLGKEAGFPLMFTTEPEEGGSHHHHHHHHHHGGGSGGGS GGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE

PS651 (3×atClpS2-V1, Linker 2)	260	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSF VTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGK EAGFPLMFTTEPEEGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYK VMLLNDDYTPMSFVTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDI AETKAKEATDLGKEAGFPLMFTTEPEEGHHHHHHHHHHGGGSGGGSGGGSG LNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS652 (4×atClpS2-V1, Linker 2)	261	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSF VTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGK EAGFPLMFTTEPEEGSAGSAAGSGEFMSDSPVDLKPKPKVKPKLERPKLYK VMLLNDDYTPMSFVTVVLKAVFRMSEDTGRRVMMTAHRFGSAVVVVCERDI AETKAKEATDLGKEAGFPLMFTTEPEEGSAGSAAGSGEFMSDSPVDLKPKP KVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSEDTGRRVMMTAHR FGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPEEGHHHHHHHHHH GGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIE WHE
PS653 (3×PS372, Linker 2)	262	MAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEMLQKIFGFPPEKG FQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPYLGRPCSGSMTCVI EPAVGSAGSAAGSGEFMAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRY VIEMLQKIFGFPPEKGFQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGP DPYLGRPCSGSMTCVIEPAVGSAGSAAGSGEFMAFPARGKTAPKNEVRRQP PYNVILLNDDDHTYRYVIEMLQKIFGFPPEKGFQIAEEVDRTGRVILLTTS KEHAELKQDQVHSYGPDPYLGRPCSGSMTCVIEPAVGGSHHHHHHHHHHGG GSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWH E
PS654 (4×PS372, Linker 2)	263	MAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEMLQKIFGFPPEKG FQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPYLGRPCSGSMTCVI EPAVGSAGSAAGSGEFMAFPARGKTAPKNEVRRQPPYNVILLNDDDHTYRY VIEMLQKIFGFPPEKGFQIAEEVDRTGRVILLTTSKEHAELKQDQVHSYGP DPYLGRPCSGSMTCVIEPAVGSAGSAAGSGEFMAFPARGKTAPKNEVRRQP PYNVILLNDDDHTYRYVIEMLQKIFGFPPEKGFQIAEEVDRTGRVILLTTS KEHAELKQDQVHSYGPDPYLGRPCSGSMTCVIEPAVGSAGSAAGSGEFMAF PARGKTAPKNEVRRQPPYNVILLNDDDHTYRYVIEMLQKIFGFPPEKGFQI AEEVDRTGRVILLTTSKEHAELKQDQVHSYGPDPYLGRPCSGSMTCVIEPA VGGSHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGG GSGLNDFFEAQKIEWHE
PS655 (3×PS557, Linker 2)	264	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMPTAASATESAIEDTPAPARPE VDGRTKPKRQPRYHVVLWNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEVD TQGRVIVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEAEEGSAGS AAGSGEFMPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDD HTYQYVVVMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQI HAFGYDRLLARSKGSMKASIEAEEGGSHHHHHHHHHHGGGSGGGSGGGSGL NDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS656 (4×PS557, Linker 2)	265	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMPTAASATESAIEDTPAPARPE VDGRTKPKRQPRYHVVLWNDDDHTYQYVVVMLQSLFGHPPERGYRLAKEVD TQGRVIVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEAEEGSAGS AAGSGEFMPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDD HTYQYVVVMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQI HAFGYDRLLARSKGSMKASIEAEEGSAGSAAGSGEFMPTAASATESAIEDT PAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVVVMLQSLFGHPPERGY

		RLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDRLLARSKGSMKASIEA EEGGSHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSG GGSGLNDFFEAQKIEWHE
PS690 (Bis-PS621, Linker 1)	266	MHSKFSHAGRICGAKFKVGEPIYRCKECSFDDTCVLCVNCFNPKDHTGHHV YTTICTEFNNGICDCGDKEAWNHTLFCKAEEGGGGSGGGSGGGSGMHSKFS HAGRICGAKFKVGEPIYRCKECSFDDTCVLCVNCFNPKDHTGHHVYTTICT EFNNGICDCGDKEAWNHTLFCKAEEGGGSHHHHHHHHHGGGSGGGSGGGS GLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS691 (Bis-PS621, Linker 2)	267	MHSKFSHAGRICGAKFKVGEPIYRCKECSFDDTCVLCVNCFNPKDHTGHHV YTTICTEFNNGICDCGDKEAWNHTLFCKAEEGGSAGSAAGSGEFMHSKFSH AGRICGAKFKVGEPIYRCKECSFDDTCVLCVNCFNPKDHTGHHVYTTICTE FNNGICDCGDKEAWNHTLFCKAEEGGGSHHHHHHHHHGGGSGGGSGGGSG LNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS692 (Bis-PS621, Linker 3)	268	MHSKFSHAGRICGAKFKVGEPIYRCKECSFDDTCVLCVNCFNPKDHTGHHV YTTICTEFNNGICDCGDKEAWNHTLFCKAEEGGSAGSAAGSGEFGSAGSAA GSGEFGSAGSAAGSGEFMHSKFSHAGRICGAKFKVGEPIYRCKECSFDDTC VLCVNCFNPKDHTGHHVYTTICTEFNNGICDCGDKEAWNHTLFCKAEEGGG SHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSG LNDFFEAQKIEWHE
PS693 (Bis-PS528, Linker 1)	269	MHSKFNHAGRICGAKFRVGEPIYRCKECSFDDTCVLCVNCFNPKDHVGHHV YTSICTEFNNGICDCGDKEAWNHELNCKGAEDGGGSGGGSGGGSGMHSKFN HAGRICGAKFRVGEPIYRCKECSFDDTCVLCVNCFNPKDHVGHHVYTSICT EFNNGICDCGDKEAWNHELNCKGAEDGGSHHHHHHHHHGGGSGGGSGGGS GLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS694 (Bis-PS528, Linker 2)	270	MHSKFNHAGRICGAKFRVGEPIYRCKECSFDDTCVLCVNCFNPKDHVGHHV YTSICTEFNNGICDCGDKEAWNHELNCKGAEDGSAGSAAGSGEFMHSKFNH AGRICGAKFRVGEPIYRCKECSFDDTCVLCVNCFNPKDHVGHHVYTSICTE FNNGICDCGDKEAWNHELNCKGAEDGGSHHHHHHHHHGGGSGGGSGGGSG LNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS695 (Bis-PS528, Linker 3)	271	MHSKFNHAGRICGAKFRVGEPIYRCKECSFDDTCVLCVNCFNPKDHVGHHV YTSICTEFNNGICDCGDKEAWNHELNCKGAEDGSAGSAAGSGEFGSAGSAA GSGEFGSAGSAAGSGEFMHSKFNHAGRICGAKFRVGEPIYRCKECSFDDTC VLCVNCFNPKDHVGHHVYTSICTEFNNGICDCGDKEAWNHELNCKGAEDGG SHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSG LNDFFEAQKIEWHE

[0407] The multivalent binder PS612 is a polypeptide that contains tandem copies of two different ClpS protein monomers (atClpS2-V1 and PS372). The amino acid binding profile of PS612 was evaluated by biolayer interferometry as described above. The monovalent binders atClpS2-V1 and PS372 were separately run as controls. FIG. 36A shows response curves for atClpS2-V1 (left plot) and PS372 (right plot). As previously observed for each monomeric binder, atClpS2-V1 binds N-terminal F, W, and Y, and PS372 binds N-terminal I, L, and V. Response curves for the tandem binder PS612 are shown in FIG. 36B. As shown, binding was observed for PS612 with N-terminal F, W, Y, I, L, and V. These results demonstrate that a single polypeptide consisting of tandem copies of two different ClpS monomers (atClpS2-V1 and PS372 in this example) exhibits the full amino acid binding capability of each of the parent ClpS proteins.

[0408] The multivalent binder PS614 is a polypeptide that contains two copies of the ClpS binder PS372 oriented in tandem. The amino acid binding profile of PS614 was evaluated by

biolayer interferometry as described above. FIG. 36C shows response curves for the monovalent binder PS372, which confirmed previous observations that PS372 binds N-terminal I, L, and V. Response curves for the tandem binder PS614 are shown in FIG. 36D. As shown, binding was observed for PS614 with N-terminal I, L, and V. These results demonstrate that a tandem PS372 construct retains strong binding to L, I, and V.

[0409] The multivalent binders PS637, PS638, and PS639 are polypeptides that each contain two copies of the ClpS binder PS557 oriented in tandem and separated by Linker 1, Linker 2, and Linker 3, respectively. The amino acid binding profiles of each binder was evaluated by biolayer interferometry as described above. FIG. 36E shows response curves for the monovalent binder PS557, which confirmed previous observations that PS557 binds N-terminal I, L, and V. Response curves for the tandem binders PS637, PS638, and PS639 are shown in FIGs. 36F, 36G, and 36H, respectively. As shown, binding was observed for tandem binder with N-terminal I, L, and V. These results demonstrate that tandem PS557 constructs containing different linkers retain binding to I, L, and V.

[0410] The above results showed that different multivalent ClpS binders retain binding to terminal amino acids including F, Y, W, L, I, and V. Additional multivalent amino acid binders were designed as tandem fusion molecules expressed from a single coding sequence containing segments encoding two copies of a UBR-box binder (PS621 or PS528) joined end-to-end by a segment encoding a flexible peptide linker. Expression of the single coding sequence produced a single full-length polypeptide having two UBR-box homologs oriented in tandem (Bis-PS621, Bis-PS528).

[0411] Three different constructs were designed and expressed for each multivalent UBR-box binder: PS690, PS691, and PS692 (Bis-PS621 binders having Linker 1, Linker 2, and Linker 3, respectively); and PS693, PS694, and PS695 (Bis-PS528 binders having Linker 1, Linker 2, and Linker 3, respectively). Fluorescence polarization studies were performed to evaluate the binding affinity of each of the multivalent UBR-box binders, and the corresponding monomeric binders, for a peptide having an N-terminal arginine residue with an alanine residue at the penultimate position. The K_D values obtained from these studies are reported below in Table 11.

Table 11. Binding affinities of tandem recognizers for N-terminal arginine peptide.

Binder	RA Peptide (K_D, nM)
PS528	403
PS621	321
PS690 (Bis-PS621, linker 1)	354
PS691 (Bis-PS621, linker 2)	369
PS692 (Bis-PS621, linker 3)	478

PS693 (Bis-PS528, linker 1)	263
PS694 (Bis-PS528, linker 2)	353
PS695 (Bis-PS528, linker 3)	317

[0412] The results from these studies demonstrate that multivalent binders can be obtained as tandem fusion molecules expressed from a single coding sequence containing segments encoding multiple ClpS or UBR-box protein homologs. Additionally, these multivalent amino acid binders were shown to retain binding to terminal amino acids including F, Y, W, L, I, V, and R.

Example 19. Shielded Recognition Molecule Fusions for Photodamage Mitigation

[0413] As described in Example 10, it was shown that the use of a DNA-streptavidin shielding element resulted in enhanced photostability of immobilized peptides during a dynamic sequencing reaction. Additional shielded recognition molecules were designed as tandem fusion molecules expressed from a single coding sequence containing segments encoding an amino acid binding protein and one or two copies of a protein shield joined end-to-end by a segment encoding a flexible peptide linker. In this way, the binding component and the shielding component of a shielded recognition molecule can be produced from a single expression construct. Table 12 provides a list of different fusion constructs that were designed and the corresponding polypeptide sequence.

Table 12. Non-limiting examples of fusion constructs.

Name	SEQ ID NO:	Sequence
PS696 (PS557-SNAP TAG fusion)	272	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMDKDCEMKRTTLDSPLGKLELS GCEQGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQP EAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNP AATAAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHR LGKPGLGGSAGSAAGSGEFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQK IEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS697 (PS557-2x SNAP TAG fusion)	273	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMDKDCEMKRTTLDSPLGKLELS GCEQGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQP EAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNP AATAAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHR LGKPGLGMDKDCEMKRTTLDSPLGKLELSGCEQGLHEIKLLGKGTSAADAV EVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFT RQVLWKLLKVVKFGEVISYQQLAALAGNPAATAAVKTALSGNPVPILIPCH RVVSSSGAVGGYEGGLAVKEWLLAHEGHRLGKPGLGGSAGSAAGSGEFHHH HHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDF FEAQKIEWHE

PS698 (PS557-SNAP TAG fusion 2)	274	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIE WHEGGGSGGGSGGGSGLNDFFEAQKIEWHEGSAGSAAGSGEFMDKDCEMKR TTLDSPLGKLELSGCEQGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLM QATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEV ISYQQLAALAGNPAATAAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGL AVKEWLLAHEGHRLGKPGLG
PS699 (PS557-EF fusion)	275	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMIKKYTGDFETTTDLNDCRVWS WGVCDIDNVDNITFGLEIDSFFEWCEMQGSTDIYFHDEKFDGEFMLSWLFK NGFKWCKEAKEERTFSTLISNMGQWYALEICWNVKCTTTKTGKTKKEKQRT IIYDSLKKYPFPVKEIAEAFNFPIKKGEIDYTKERPIGYNPTDDEWDYLKN DIQIMAMALKIQFDQGLTRMTRGSDALGDYQDWVKTTYGKSRFKQWFPVLS LGFDKDLRKAYKGGFTWVNKVFQGKEIGEGIVFDVNSLYPSQMYVRPLPYG TPLFYEGEYKENIDYPLYIQNIKVRFRLKERHIPTIQVKQSSLFIQNEYLE SSVNKLGVDELIDLTLTNVDLDLFFEHYDILEIHYTYGYMFKASCDMFKGW IDKWIEVKNTTEGARKANAKGMLNSLYGKFGTNPDITGKVPYMGEDGIVRL TLGEEELRDPVYVPLASFVTAWGRYTTITTAQRCFDNIIYCDTDSIHLTGT EVPEAIEHLVDSKKLGYWKHESTFQRAKFIRQKTYVEEIDGELNVKCAGMP DRIKELVTFDNFEVGFSSYGKLLPKRTQGGVVLVDTMFTIKGSAGSAAGSG EFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGS GLNDFFEAQKIEWHE
PS700 (PS557-MBP fusion)	276	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMKIEEGKLVIWINGDKGYNGLA EVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQ SGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLP NPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG KYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMT INGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELA KEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKG EIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNN NNLGIEGRGSAGSAAGSGEFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQ KIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS701 (PS557-GST fusion)	277	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMSPILGYWKIKGLVQPTRLLLE YLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIR YIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLS KLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPK LVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRGS PGIHRDGSAGSAAGSGEFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKI EWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS702 (PS557-GFP fusion)	278	MPTAASATESAIEDTPAPARPEVDGRTKPKRQPRYHVVLWNDDDHTYQYVV VMLQSLFGHPPERGYRLAKEVDTQGRVIVLTTTREHAELKRDQIHAFGYDR LLARSKGSMKASIEAEEGSAGSAAGSGEFMSKGEELFTGVVPILVELDGDV NGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSR YPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIE LKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS VQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTA AGITHGMDELYKGSAGSAAGSGEFHHHHHHHHHHGGGSGGGSGGGSGLNDF FEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS703	279	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMDKDCEMKRTTLDSPLGKLELSGCEQGLHEIKLLGKGT

(atClpS2-V1-SNAP TAG fusion)		SAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVF QQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNPAATAAVKTALSGNPVP ILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHRLGKPGLGGSAGSAAGS GEFHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGG SGLNDFFEAQKIEWHE
PS704 (atClpS2-V1-2x SNAP TAG fusion)	280	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMDKDCEMKRTTLDSPLGKLELSGCEQGLHEIKLLGKGT SAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIEEFPVPALHHPVF QQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNPAATAAVKTALSGNPVP ILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHRLGKPGLGMDKDCEMKR TTLDSPLGKLELSGCEQGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLM QATAWLNAYFHQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEV ISYQQLAALAGNPAATAAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGL AVKEWLLAHEGHRLGKPGLGGSAGSAAGSGEFHHHHHHHHHGGGSGGGSG GGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS705 (atClpS2-V1-SNAP TAG fusion 2)	281	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSG LNDFFEAQKIEWHEGSAGSAAGSGEFMDKDCEMKRTTLDSPLGKLELSGCE QGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAI EEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNPAAT AAVKTALSGNPVPILIPCHRVVSSSGAVGGYEGGLAVKEWLLAHEGHRLGK PGLG
PS706 (atClpS2-V1-EF fusion)	282	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMIKKYTGDFETTTDLNDCRVWSWGVCDIDNVDNITFGL EIDSFFEWCEMQGSTDIYFHDEKFDGEFMLSWLFKNGFKWCKEAKEERTFS TLISNMGQWYALEICWNVKCTTTKTGKTKKEKQRTIIYDSLKKYPFPVKEI AEAFNFPIKKGEIDYTKERPIGYNPTDDEWDYLKNDIQIMAMALKIQFDQG LTRMTRGSDALGDYQDWVKTTYGKSRFKQWFPVLSLGFDKDLRKAYKGGFT WVNKVFQGKEIGEGIVFDVNSLYPSQMYVRPLPYGTPLFYEGEYKENIDYP LYIQNIKVRFRLKERHIPTIQVKQSSLFIQNEYLESSVNKLGVDELIDLTL TNVDLDLFFEHYDILEIHYTYGYMFKASCDMFKGWIDKWIEVKNTTEGARK ANAKGMLNSLYGKFGTNPDITGKVPYMGEDGIVRLTLGEEELRDPVYVPLA SFVTAWGRYTTITTAQRCFDNIIYCDTDSIHLTGTEVPEAIEHLVDSKKLG YWKHESTFQRAKFIRQKTYVEEIDGELNVKCAGMPDRIKELVTFDNFEVGF SSYGKLLPKRTQGGVVLVDTMFTIKGSAGSAAGSGEFHHHHHHHHHGGGS GGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGSGLNDFFEAQKIEWHE
PS707 (atClpS2-V1-MBP fusion)	283	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTV EHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDK LYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKEL KAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKA GLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVN YGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAV RTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNNNNLGIEGRGSAGSAAG SGEFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGG GSGLNDFFEAQKIEWHE
PS708 (atClpS2-V1-GST fusion)	284	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEG DKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKER AEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHK TYLNGDHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDK

		YLKSSKYIAWPLQGWQATFGGGDHPPKSDLVPRGSPGIHRDGSAGSAAGSG EFHHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGGGSGGGS GLNDFFEAQKIEWHE
PS709 (atClpS2-V1-GFP fusion)	285	MSDSPVDLKPKPKVKPKLERPKLYKVMLLNDDYTPMSFVTVVLKAVFRMSE DTGRRVMMTAHRFGSAVVVVCERDIAETKAKEATDLGKEAGFPLMFTTEPE EGSAGSAAGSGEFMSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDAT YGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKQHDFFKSAMP EGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHK LEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDG PVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYKGSAG SAAGSGEFHHHHHHHHHGGGSGGGSGGGSGLNDFFEAQKIEWHEGGGSGG GSGGGSGLNDFFEAQKIEWHE

[0414] As shown in Table 12, each fusion construct included an amino acid binding protein (PS557 or atClpS2-V1) fused to one of the following protein shields: a SNAP-tag protein fused via Linker 2 (denoted as "SNAP TAG fusion"); two copies of a SNAP-tag protein oriented in tandem (denoted as "2x SNAP TAG fusion"); a SNAP-tag protein fused via a segment including Linker 2 and a His/bis-biotinylation tag (denoted as "SNAP TAG fusion 2"); a DNA polymerase (denoted as "EF fusion"); a maltose-binding protein (denoted as "MBP fusion"); a glutathione S-transferase protein (denoted as "GST fusion"); or a green fluorescent protein (denoted as "GFP fusion").

[0415] Stopped-flow assays were performed to measure the association (k_{on}) and dissociation (k_{off}) rates for PS557, atClpS2-V1, and fusion proteins derived by C-terminal addition of protein shields. To measure the on-rate constant by stopped-flow, the binder was rapidly mixed with FITC labeled peptide, and the reaction was followed in real time. A schematic illustrating the assay design is shown in FIG. 37A (top panel). N-terminal LA peptide was used for PS557 derivatives and N-terminal FA peptide was used for atClpS2-V1 derivatives. In this assay, the fluorescence signal decreased due to quenching upon protein binding. The averaged traces obtained at multiple concentrations of binder were fitted with a decay equation to derive association rates (FIG. 37A, middle panel). A linear slope from the plot of association rates against different binder concentrations gave the k_{on} rate constant (FIG. 37A, bottom panel).

[0416] To measure the dissociation rate (k_{off}) by stopped-flow, complexes of binder bound to labeled peptide substrate at an optimal concentration were rapidly mixed with excess unlabeled trap peptide. A schematic illustrating the assay design is shown in FIG. 37B (top panel). Binder dissociation from the peptide N-terminus resulted in an increase in the fluorescence signal due to the reversal of quenching. The averaged raw traces were fitted with an exponential equation to determine the k_{off} (FIG. 37B, bottom panel).

[0417] The rates determined by stopped-flow assays are shown in Table 13 (PS557 and PS557-derived fusions) and Table 14 (atClpS2-V1 and atClpS2-V1-derived fusions).

Table 13. LA Peptide k_{on} rate constants and k_{off} rates for PS557 fusions.

Binder	LA kon (nM/s)	LA koff (/s)
PS557	0.0018	0.22
PS702 (PS557 GFP fusion)	0.0025	0.1
PS696 (PS557-SNAP TAG fusion)	–	0.15
PS698 (PS557-SNAP TAG fusion 2)	0.004	0.19

Table 14. FA Peptide k_{on} rate constants and k_{off} rates for atClpS2-V1 fusions.

Binder	FA kon (nM/s)	FA koff (/s)
atClpS2-V1	0.0015	1.187
PS706 (atClpS2-V1-EF fusion)	0.0014	0.58909
PS703 (atClpS2-V1-SNAP TAG fusion)	–	0.76741
PS704 (atClpS2-V1-2xSNAP TAG fusion)	–	0.6774
PS705 (atClpS2-V1-SNAP TAG fusion 2)	–	1.2655
PS709 (atClpS2-V1-GFP fusion)	–	0.9344

[0418] The results demonstrate that C-terminal fusion of protein shield constructs to ClpS proteins generates fully active N-terminal recognizers with similar kinetic profiles to ClpS controls that are favorable for use in single-molecule peptide sequencing assays.

EQUIVALENTS AND SCOPE

[0419] In the claims articles such as "a," "an," and "the" may mean one or more than one unless indicated to the contrary or otherwise evident from the context. Claims or descriptions that include "or" between one or more members of a group are considered satisfied if one, more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process unless indicated to the contrary or otherwise evident from the context. The invention includes embodiments in which exactly one member of the group is present in, employed in, or otherwise relevant to a given product or process. The invention includes embodiments in which more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process.

[0420] Furthermore, the invention encompasses all variations, combinations, and permutations in which one or more limitations, elements, clauses, and descriptive terms from one or more of the listed claims is introduced into another claim. For example, any claim that is dependent on another claim can be modified to include one or more limitations found in any other claim that is dependent on the same base claim. Where elements are presented as lists, e.g., in Markush group format, each subgroup of the elements is also disclosed, and any element(s) can be removed from the group. It should it be understood that, in general, where the invention, or aspects of the invention, is/are referred to as comprising particular elements and/or features, certain

embodiments of the invention or aspects of the invention consist, or consist essentially of, such elements and/or features. For purposes of simplicity, those embodiments have not been specifically set forth *in haec verba* herein.

[0421] 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.

[0422] 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.

[0423] 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.

[0424] It should also be understood that, unless clearly indicated to the contrary, in any methods claimed herein that include more than one step or act, the order of the steps or acts of the method is not necessarily limited to the order in which the steps or acts of the method are recited.

[0425] 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. It should be appreciated that embodiments described in this document using an open-ended transitional phrase (e.g., "comprising") are also contemplated, in alternative embodiments, as "consisting of" and "consisting essentially of" the feature described by the open-ended transitional phrase. For example, if the application describes "a composition comprising A and B," the application also contemplates the alternative embodiments "a composition consisting of A and B" and "a composition consisting essentially of A and B."

[0426] Where ranges are given, endpoints are included. Furthermore, unless otherwise indicated or otherwise evident from the context and understanding of one of ordinary skill in the art, values that are expressed as ranges can assume any specific value or sub-range within the stated ranges in different embodiments of the invention, to the tenth of the unit of the lower limit of the range, unless the context clearly dictates otherwise.

[0427] This application refers to various issued patents, published patent applications, journal articles, and other publications, all of which are incorporated herein by reference. If there is a conflict between any of the incorporated references and the instant specification, the specification shall control. In addition, any particular embodiment of the present invention that falls within the prior art may be explicitly excluded from any one or more of the claims. Because such embodiments are deemed to be known to one of ordinary skill in the art, they may be excluded even if the exclusion is not set forth explicitly herein. Any particular embodiment of the invention can be excluded from any claim, for any reason, whether or not related to the existence of prior art.

[0428] Those skilled in the art will recognize or be able to ascertain using no more than routine experimentation many equivalents to the specific embodiments described herein. The scope of the present embodiments described herein is not intended to be limited to the above Description,

but rather is as set forth in the appended claims. Those of ordinary skill in the art will appreciate that various changes and modifications to this description may be made without departing from the spirit or scope of the present invention, as defined in the following claims.

[0429] The recitation of a listing of chemical groups in any definition of a variable herein includes definitions of that variable as any single group or combination of listed groups. The recitation of an embodiment for a variable herein includes that embodiment as any single embodiment or in combination with any other embodiments or portions thereof. The recitation of an embodiment herein includes that embodiment as any single embodiment or in combination with any other embodiments or portions thereof.

CLAIMS

What is claimed is:

1. A recombinant amino acid binding protein having an amino acid sequence that is at least 80% identical to a sequence selected from Table 1 or Table 2 and comprising one or more labels.

2. The recombinant amino acid binding protein of claim 1, wherein the one or more labels comprise a luminescent label or a conductivity label.

3. The recombinant amino acid binding protein of claim 2, wherein the luminescent label comprises at least one fluorophore dye molecule.

4. The recombinant amino acid binding protein of claim 2 or 3, wherein the luminescent label comprises 20 or fewer fluorophore dye molecules.

5. The recombinant amino acid binding protein of any one of claims 2-4, wherein the luminescent label comprises at least one FRET pair comprising a donor label and an acceptor label.

6. The recombinant amino acid binding protein of claim 5, wherein the ratio of the donor label to the acceptor label is 1:1, 2:1, 3:1, 4:1, or 5:1.

7. The recombinant amino acid binding protein of claim 5, wherein the ratio of the acceptor label to the donor label is 1:1, 2:1, 3:1, 4:1, or 5:1.

8. The recombinant amino acid binding protein of any one of claims 1-7, wherein the conductivity label comprises a charged polymer.

9. The recombinant amino acid binding protein of any one of claims 1-8, wherein the one or more labels comprise a tag sequence.

10. The recombinant amino acid binding protein of claim 9, wherein the tag sequence comprises one or more of a purification tag, a cleavage site, and a biotinylation sequence.

11. The recombinant amino acid binding protein of claim 10, wherein the biotinylation sequence comprises at least one biotin ligase recognition sequence.

12. The recombinant amino acid binding protein of claim 10 or 11, wherein the biotinylation sequence comprises two biotin ligase recognition sequences oriented in tandem.

13. The recombinant amino acid binding protein of any one of claims 1-12, wherein the one or more labels comprise a biotin moiety.

14. The recombinant amino acid binding protein of claim 13, wherein the biotin moiety comprises at least one biotin molecule.

15. The recombinant amino acid binding protein of claim 13 or 14, wherein the biotin moiety is a bis-biotin moiety.

16. The recombinant amino acid binding protein of claim 14 or 15, wherein the label comprises at least one biotin ligase recognition sequence having the at least one biotin molecule attached thereto.

17. The recombinant amino acid binding protein of any one of claims 1-16, wherein the one or more labels comprise one or more polyol moieties.

18. The recombinant amino acid binding protein of claim 17, wherein the one or more polyol moieties comprise dextran, polyvinylpyrrolidone, polyethylene glycol, polypropylene glycol, polyoxyethylene glycol, polyvinyl alcohol, or a combination or variation thereof.

19. The recombinant amino acid binding protein of any one of claims 1-18, wherein the recombinant amino acid binding protein comprises one or more unnatural amino acids having the one or more labels attached thereto.

20. The recombinant amino acid binding protein of any one of claims 1-19, wherein the amino acid sequence is 80-90%, 90-95%, or at least 95% identical to a sequence selected from Table 1.

21. A composition comprising a recombinant amino acid binding protein of any one of claims 1-20.

22. A polypeptide sequencing reaction composition comprising two or more amino acid recognition molecules, wherein at least one of the two or more amino acid recognition molecules is a recombinant amino acid binding protein of any one of claims 1-20.

23. The polypeptide sequencing reaction composition of claim 22, wherein the composition comprises at least one type of cleaving reagent.

24. A method of polypeptide sequencing, the method comprising:

contacting a polypeptide with a polypeptide sequencing reaction composition according to claim 22 or 23; and

detecting a series of interactions of the polypeptide with at least one amino acid recognition molecule while the polypeptide is being degraded, thereby sequencing the polypeptide.

25. A polypeptide sequencing reaction mixture comprising an amino acid binding protein and a peptidase, wherein the molar ratio of the labeled amino acid binding protein to the peptidase is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1, and wherein the amino acid binding protein comprises one or more labels.

26. The polypeptide sequencing reaction mixture of claim 25, wherein the molar ratio is between about 1:100 and about 1:1 or between about 1:1 and about 10:1.

27. The polypeptide sequencing reaction mixture of claim 25 or 26, wherein the molar ratio is about 1:1,000, about 1:500, about 1:200, about 1:100, about 1:10, about 1:5, about 1:2, about 1:1, about 5:1, about 10:1, about 50:1, about 100:1.

28. The polypeptide sequencing reaction mixture of any one of claims 25-27, wherein the amino acid binding protein is a synthetic or recombinant protein.

29. The polypeptide sequencing reaction mixture of any one of claims 25-28, wherein the amino acid binding protein is a degradation pathway protein, an inactivated peptidase, an

antibody, an aminotransferase, a tRNA synthetase, or an SH2 domain-containing protein or fragment thereof.

30. The polypeptide sequencing reaction mixture of any one of claims 25-29, wherein the amino acid binding protein is a ClpS protein, a Gid protein, a UBR-box protein or UBR-box domain-containing fragment thereof, or a p62 protein or ZZ domain-containing fragment thereof.

31. The polypeptide sequencing reaction mixture of any one of claims 25-30, wherein the amino acid binding protein is a ClpS protein.

32. The polypeptide sequencing reaction mixture of claim 31, wherein the ClpS protein is ClpS1 or ClpS2 from *A. tumifaciens*, *C. crescentus*, *E. coli*, *S. elongatus*, *P. falciparum*, *T. elongatus*, or a homologous species thereof.

33. The polypeptide sequencing reaction mixture of any one of claims 25-32, wherein the amino acid binding protein is a protein having an amino acid sequence that is at least 80%, 80-90%, 90-95%, or at least 95% identical to a sequence selected from Table 1 or Table 2.

34. The polypeptide sequencing reaction mixture of any one of claims 25-33, wherein the amino acid binding protein is a recombinant amino acid binding protein according to any one of claims A1-A19.

35. The polypeptide sequencing reaction mixture of any one of claims 25-34, wherein the peptidase is an exopeptidase.

36. The polypeptide sequencing reaction mixture of any one of claims 25-35, wherein the peptidase is an aminopeptidase or a carboxypeptidase.

37. The polypeptide sequencing reaction mixture of any one of claims 25-36, wherein the peptidase is a non-specific exopeptidase that cleaves more than one type of amino acid from a terminal end of a polypeptide.

38. The polypeptide sequencing reaction mixture of any one of claims 25-37, wherein the peptidase is a proline aminopeptidase, a proline iminopeptidase, a glutamate/aspartate-specific aminopeptidase, a methionine-specific aminopeptidase, or a zinc metalloprotease.

39. The polypeptide sequencing reaction mixture of any one of claims 25-38, wherein the peptidase is an enzyme having an amino acid sequence that is at least 80%, 80-90%, 90-95%, or at least 95% identical to a sequence selected from Table 4 or Table 5.

40. The polypeptide sequencing reaction mixture of any one of claims 25-39, wherein the reaction mixture comprises more than one amino acid binding protein.

41. The polypeptide sequencing reaction mixture of any one of claims 25-40, wherein the reaction mixture comprises more than one peptidase.

42. The polypeptide sequencing reaction mixture of any one of claims 25-41, wherein the reaction mixture comprises a polypeptide molecule immobilized to a surface.

50. A polypeptide sequencing reaction mixture comprising a single polypeptide molecule, at least one peptidase molecule, and at least three amino acid recognition molecules.

51. The polypeptide sequencing reaction mixture of claim 50 comprising at least 1 and up to 10 peptidase molecules.

52. The polypeptide sequencing reaction mixture of claim 50 or 51 comprising at least 1 and up to 5 peptidase molecules.

53. The polypeptide sequencing reaction mixture of any one of claims 50-52 comprising at least 1 and up to 3 peptidase molecules.

54. The polypeptide sequencing reaction mixture of any one of claims 50-53 comprising at least 3 and up to 30 amino acid recognition molecules.

55. The polypeptide sequencing reaction mixture of claim 54 comprising up to 20, up to 10, or up to 5 amino acid recognition molecules.

56. A substrate comprising an array of sample wells, wherein at least one sample well of the array comprises a polypeptide sequencing reaction mixture in accordance with any one of claims 50-55.

57. The substrate of claim 56, wherein the at least one sample well comprises a bottom surface, and wherein the single polypeptide molecule is immobilized to the bottom surface.

58. A substrate comprising an array of sample wells, wherein at least one sample well of the array comprises a single polypeptide molecule, a cleaving means, and a binding means, wherein the binding means and the cleaving means are configured to achieve at least 10 association events between the binding means and a terminal amino acid on the polypeptide prior to removal of the terminal amino acid from the polypeptide by the cleaving means.

59. An amino acid recognition molecule comprising a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end, wherein the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids.

60. The amino acid recognition molecule of claim 59, wherein the first and second amino acid binding proteins are the same.

61. The amino acid recognition molecule of claim 59, wherein the first and second amino acid binding proteins are different.

62. The amino acid recognition molecule of any one of claims 59-61, wherein the first and second amino acid binding proteins each independently has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

63. The amino acid recognition molecule of any one of claims 59-62, wherein the linker comprises up to 100 amino acids.

64. The amino acid recognition molecule of any one of claims 59-63, wherein the linker comprises between about 5 and about 50 amino acids.

65. An amino acid recognition molecule comprising a polypeptide of Formula (**I**):

$$(Z^1–X^1)_n–Z^2$$

(**I**),

wherein:

Z^1 and Z^2 are independently amino acid binding proteins;

X^1 is a linker comprising at least two amino acids, wherein the amino acid binding proteins are joined end-to-end by the linker; and

n is an integer from 1 to 5, inclusive.

66. The amino acid recognition molecule of claim 65, wherein Z^1 and Z^2 comprise amino acid binding proteins of the same type.

67. The amino acid recognition molecule of claim 65, wherein Z^1 and Z^2 comprise different types of amino acid binding proteins.

68. The amino acid recognition molecule of any one of claims 65-67, wherein Z^1 and Z^2 are independently optionally associated with a label component comprising at least one detectable label.

69. The amino acid recognition molecule of any one of claims 65-68, wherein the polypeptide further comprises a tag sequence.

70. The amino acid recognition molecule of claim 69, wherein the tag sequence is at or near a terminus of the polypeptide.

71. The amino acid recognition molecule of claim 69 or 70, wherein the tag sequence comprises at least one biotin ligase recognition sequence.

72. The amino acid recognition molecule of any one of claims 69-71, wherein the tag sequence comprises two biotin ligase recognition sequences oriented in tandem.

73. The amino acid recognition molecule of any one of claims 65-72, wherein Z^1 and Z^2 each independently has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

74. The amino acid recognition molecule of any one of claims 65-73, wherein X^1 comprises up to 100 amino acids.

75. The amino acid recognition molecule of any one of claims 65-74, wherein X^1 comprises between about 5 and about 50 amino acids.

76. An amino acid recognition molecule comprising a polypeptide having an amino acid binding protein and a labeled protein joined end-to-end, wherein the amino acid binding protein and the labeled protein are separated by a linker comprising at least two amino acids.

77. The amino acid recognition molecule of claim 76, wherein the linker comprises up to 100 amino acids.

78. The amino acid recognition molecule of claim 76 or 77, wherein the linker comprises between about 5 and about 50 amino acids.

79. The amino acid recognition molecule of any one of claims 76-78, wherein the labeled protein has a molecular weight of at least 10 kDa.

80. The amino acid recognition molecule of any one of claims 76-79, wherein the labeled protein has a molecular weight of between about 10 kDa and about 150 kDa.

81. The amino acid recognition molecule of any one of claims 76-80, wherein the labeled protein has a molecular weight of between about 15 kDa and about 100 kDa.

82. The amino acid recognition molecule of any one of claims 76-81, wherein the labeled protein comprises at least 50 amino acids.

83. The amino acid recognition molecule of any one of claims 76-82, wherein the labeled protein comprises between about 50 and about 1,000 amino acids.

84. The amino acid recognition molecule of any one of claims 76-83, wherein the labeled protein comprises between about 100 and about 750 amino acids.

85. The amino acid recognition molecule of any one of claims 76-84, wherein the labeled protein comprises a protein selected from the group consisting of: a DNA polymerase, a maltose-binding protein, a glutathione S-transferase, a green fluorescent protein, and a SNAP-tag.

86. The amino acid recognition molecule of any one of claims 76-85, wherein the labeled protein comprises a luminescent label.

87. The amino acid recognition molecule of claim 86, wherein the luminescent label comprises at least one fluorophore dye molecule.

88. The amino acid recognition molecule of any one of claims 76-87, wherein the amino acid binding protein is a Gid protein, a UBR-box protein or UBR-box domain-containing fragment thereof, a p62 protein or ZZ domain-containing fragment thereof, or a ClpS protein.

89. The amino acid recognition molecule of any one of claims 76-88, wherein the amino acid binding protein has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

90. A method of polypeptide sequencing, the method comprising:

contacting a single polypeptide molecule in a reaction mixture with a composition comprising a binding means and a cleaving means,

wherein the binding means and the cleaving means are configured to achieve at least 10 association events between the binding means and a terminal amino acid on the polypeptide prior to removal of the terminal amino acid from the polypeptide by the cleaving means.

91. The method of claim 90, wherein the binding means and the cleaving means are configured to achieve at least 10 and up to 1,000 association events prior to the removal of the terminal amino acid.

92. The method of claim 90 or 91, wherein the binding means and the cleaving means are configured to achieve up to 500, up to 200, or up to 100 association events prior to the removal of the terminal amino acid.

93. The method of any one of claims 90-92, wherein the terminal amino acid was exposed at the polypeptide terminus in a cleavage event prior to the at least 10 association events.

94. The method of claim 93, wherein the at least 10 association events occur after the cleavage event.

95. The method of any one of claims 90-94, wherein the binding means and the cleaving means are configured to achieve a time interval of at least 1 minute between cleavage events.

96. The method of claim 95, wherein the time interval is between about 1 minute and about 20 minutes, between about 5 minutes and about 15 minutes, or between about 1 minute and about 10 minutes.

97. The method of any one of claims 90-96, wherein the binding means comprise one or more amino acid recognition molecules, and wherein the cleaving means comprise one or more peptidase molecules.

98. The method of claim 97, wherein the molar ratio of an amino acid recognition molecule to a peptidase molecule is configured to achieve the at least 10 association events prior to the removal of the terminal amino acid.

99. The method of claim 98, wherein the molar ratio of the amino acid recognition molecule to the peptidase molecule is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1.

100. The method of claim 98, wherein the molar ratio of the amino acid recognition molecule to the peptidase molecule is between about 1:100 and about 1:1 or between about 1:1 and about 10:1.

101. The method of claim 98, wherein the molar ratio of the amino acid recognition molecule to the peptidase molecule is about 1:1,000, about 1:500, about 1:200, about 1:100, about 1:10, about 1:5, about 1:2, about 1:1, about 5:1, about 10:1, about 50:1, about 100:1.

102. A method comprising:

obtaining data during a degradation process of a polypeptide;

analyzing the data to determine portions of the data corresponding to amino acids that are sequentially exposed at a terminus of the polypeptide during the degradation process, wherein each of the individual portions comprises a pulse pattern having at least one pulse duration, wherein the pulse pattern comprises a mean pulse duration of between about 1 millisecond and about 10 seconds; and

outputting an amino acid sequence representative of the polypeptide.

103. The method of claim 102, wherein the mean pulse duration is between about 50 milliseconds and about 2 seconds.

104. The method of claim 102 or 103, wherein the mean pulse duration is between about 50 milliseconds and about 500 milliseconds or between about 500 milliseconds and about 2 seconds.

105. The method of any one of claims 102-104, wherein the pulse pattern of a first type of amino acid is different from the pulse pattern of a second type of amino acid by a mean pulse duration of at least 10 milliseconds.

106. The method of any one of claims 102-104, wherein the pulse pattern of a first type of amino acid is different from the pulse pattern of a second type of amino acid by a mean pulse duration of between about 10 milliseconds and about 100 milliseconds or between about 100 milliseconds and about 10 seconds.

107. A method of polypeptide sequencing, the method comprising:

contacting a single polypeptide molecule with one or more amino acid recognition molecules; and

detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with successive amino acids exposed at a terminus of the single polypeptide while the single polypeptide is being degraded, thereby sequencing the single polypeptide molecule,

wherein association of the one or more amino acid recognition molecules with each type of amino acid exposed at the terminus produces a characteristic pattern in the series of signal pulses that is different from other types of amino acids exposed at the terminus, wherein signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds.

108. A method of sequencing a polypeptide, the method comprising:

contacting a single polypeptide molecule in a reaction mixture with a composition comprising one or more amino acid recognition molecules and a cleaving reagent; and

detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with a terminus of the single polypeptide molecule in the presence of the

cleaving reagent, wherein the series of signal pulses is indicative of a series of amino acids exposed at the terminus over time as a result of terminal amino acid cleavage by the cleaving reagent,

wherein association of the one or more amino acid recognition molecules with each type of amino acid exposed at the terminus produces a characteristic pattern in the series of signal pulses that is different from other types of amino acids exposed at the terminus, wherein signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds.

109. A method of polypeptide sequencing, the method comprising:

a) identifying a first amino acid at a terminus of a single polypeptide molecule;

b) removing the first amino acid to expose a second amino acid at the terminus of the single polypeptide molecule; and

c) identifying the second amino acid at the terminus of the single polypeptide molecule,

wherein (a)-(c) are performed in a single reaction mixture comprising one or more amino acid recognition molecules,

wherein the first and second amino acids are identified by detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with the terminus of the single polypeptide molecule,

wherein association of the one or more amino acid recognition molecules with the first amino acid produces a characteristic pattern in the series of signal pulses that is different from the second amino acid, wherein signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds.

110. A method of identifying an amino acid of a polypeptide, the method comprising:

contacting a single polypeptide molecule with one or more amino acid recognition molecules that bind to the single polypeptide molecule;

detecting a series of signal pulses indicative of association of the one or more amino acid recognition molecules with the single polypeptide molecule under polypeptide degradation conditions; and

identifying a first type of amino acid in the single polypeptide molecule based on a characteristic pattern in the series of signal pulses, wherein signal pulses of the characteristic pattern comprise a mean pulse duration of between about 1 millisecond and about 10 seconds.

111. The method of any one of claims 107-110, wherein the mean pulse duration is between about 50 milliseconds and about 2 seconds.

112. The method of any one of claims 107-111, wherein the mean pulse duration is between about 50 milliseconds and about 500 milliseconds or between about 500 milliseconds and about 2 seconds.

113. The method of any one of claims 107-112, wherein the characteristic pattern comprises at least 10 signal pulses.

114. The method of any one of claims 107-113, wherein the characteristic pattern of at least one type of amino acid comprises between about 50 and about 200 signal pulses.

115. The method of any one of claims 107-114, wherein the characteristic pattern of at least one type of amino acid comprises between about 25 and about 100 signal pulses.

116. The method of any one of claims 107-115, wherein the characteristic pattern of a first type of amino acid is different from the characteristic pattern of a second type of amino acid by a mean pulse duration of at least 10 milliseconds.

117. The method of any one of claims 107-116, wherein the characteristic pattern of a first type of amino acid is different from the characteristic pattern of a second type of amino acid by a mean pulse duration of between about 10 milliseconds and about 100 milliseconds or between about 100 milliseconds and about 1 second.

118. The method of any one of claims 107-117, wherein the method is performed in a reaction mixture with a composition that comprises the one or more amino acid recognition molecules and one or more cleaving reagents.

119. The method of claim 118, wherein the molar ratio of an amino acid recognition molecule to a cleaving reagent in the reaction mixture is between about 1:1,000 and about 1:1 or between about 1:1 and about 100:1.

120. The method of claim 118, wherein the molar ratio of an amino acid recognition molecule to a cleaving reagent in the reaction mixture is between about 1:100 and about 1:1 or between about 1:1 and about 10:1.

121. The method of claim 118, wherein the molar ratio of an amino acid recognition molecule to a cleaving reagent in the reaction mixture is about 1:1,000, about 1:500, about 1:200, about 1:100, about 1:10, about 1:5, about 1:2, about 1:1, about 5:1, about 10:1, about 50:1, about 100:1.

122. The method of any one of claims 107-121, wherein the one or more amino acid recognition molecules comprise one or more terminal amino acid recognition molecules.

123. The method of any one of claims 107-122, wherein each of the one or more amino acid recognition molecules comprises an amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

124. A system comprising:

at least one hardware processor; and

at least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by the at least one hardware processor, cause the at least one hardware processor to perform the method of any of claims 107-123.

125. At least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by at least one hardware processor, cause the at least one hardware processor to perform the method of any of claims 107-123.

126. A composition comprising a soluble amino acid recognition molecule of Formula (**II**):

$$A–(Y)_n–D$$

(**II**),

wherein:

A is an amino acid binding component comprising at least one amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2;

each instance of Y is a polymer that forms a covalent or non-covalent linkage group;

n is an integer from 1 to 10, inclusive; and

D is a label component comprising at least one detectable label, wherein D is less than 200 Å in diameter.

127. The composition of claim 126, wherein the soluble amino acid recognition molecule comprises a polypeptide having A and Y^1 joined end-to-end, wherein A and Y^1 are separated by a linker comprising at least two amino acids.

128. The composition of claim 127, wherein the linker comprises up to 100 amino acids.

129. The composition of claim 127 or 128, wherein the linker comprises between about 5 and about 50 amino acids.

130. The composition of any one of claims 126-129, wherein Y^1 is a protein having a molecular weight of at least 10 kDa.

131. The composition of any one of claims 126-130, wherein Y^1 is a protein having a molecular weight of between about 10 kDa and about 150 kDa.

132. The composition of any one of claims 126-131, wherein Y^1 is a protein having a molecular weight of between about 15 kDa and about 100 kDa.

133. The composition of any one of claims 126-132, wherein Y^1 is a protein comprising at least 50 amino acids.

134. The composition of any one of claims 126-133, wherein Y^1 is a protein comprising between about 50 and about 1,000 amino acids.

135. The composition of any one of claims 126-134, wherein Y^1 is a protein comprising between about 100 and about 750 amino acids.

136. The composition of any one of claims 126-135, wherein Y^1 is a protein selected from the group consisting of: a DNA polymerase, a maltose-binding protein, a glutathione S-transferase, a green fluorescent protein, and a SNAP-tag.

137. The composition of any one of claims 126-136, wherein A comprises a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end, wherein the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids.

138. The composition of claim 137, wherein the first and second amino acid binding proteins are the same.

139. The composition of claim 137 or 138, wherein the first and second amino acid binding proteins are different.

140. The composition of any one of claims 137-139, wherein the first and second amino acid binding proteins each independently has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

141. The composition of any one of claims 137-140, wherein the linker comprises up to 100 amino acids.

142. The composition of any one of claims 137-141, wherein the linker comprises between about 5 and about 50 amino acids.

143. An amino acid recognition molecule of Formula (**III**):

$$A–Y^1–D$$

(**III**),

wherein:

A is an amino acid binding component comprising at least one amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2;

Y^1 is a nucleic acid or a polypeptide;

D is a label component comprising at least one detectable label;

provided that when Y^1 is a nucleic acid, the nucleic acid forms a covalent or non-covalent linkage group; and

provided that when Y^1 is a polypeptide, the polypeptide forms a non-covalent linkage group characterized by a dissociation constant (K_D) of less than 50×10^{-9} M.

144. The amino acid recognition molecule of claim 143, wherein A comprises a polypeptide having at least a first amino acid binding protein and a second amino acid binding protein joined end-to-end, wherein the first and second amino acid binding proteins are separated by a linker comprising at least two amino acids.

145. The amino acid recognition molecule of claim 144, wherein the first and second amino acid binding proteins are the same.

146. The amino acid recognition molecule of claim 144 or 145, wherein the first and second amino acid binding proteins are different.

147. The amino acid recognition molecule of any one of claims 144-146, wherein the first and second amino acid binding proteins each independently has an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2.

148. The amino acid recognition molecule of any one of claims 144-147, wherein the linker comprises up to 100 amino acids.

149. The amino acid recognition molecule of any one of claims 144-148, wherein the linker comprises between about 5 and about 50 amino acids.

150. An amino acid recognition molecule comprising:

a nucleic acid;

at least one amino acid recognition molecule attached to a first attachment site on the nucleic acid, wherein the at least one amino acid recognition molecule comprises an amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2; and

at least one detectable label attached to a second attachment site on the nucleic acid,

wherein the nucleic acid forms a covalent or non-covalent linkage group between the at least one amino acid recognition molecule and the at least one detectable label.

151. An amino acid recognition molecule comprising:

a multivalent protein comprising at least two ligand-binding sites;

at least one amino acid recognition molecule attached to the protein through a first ligand moiety bound to a first ligand-binding site on the protein, wherein the at least one amino acid

recognition molecule comprises an amino acid binding protein having an amino acid sequence that is at least 80% identical to an amino acid sequence selected from Table 1 or Table 2; and

at least one detectable label attached to the protein through a second ligand moiety bound to a second ligand-binding site on the protein.

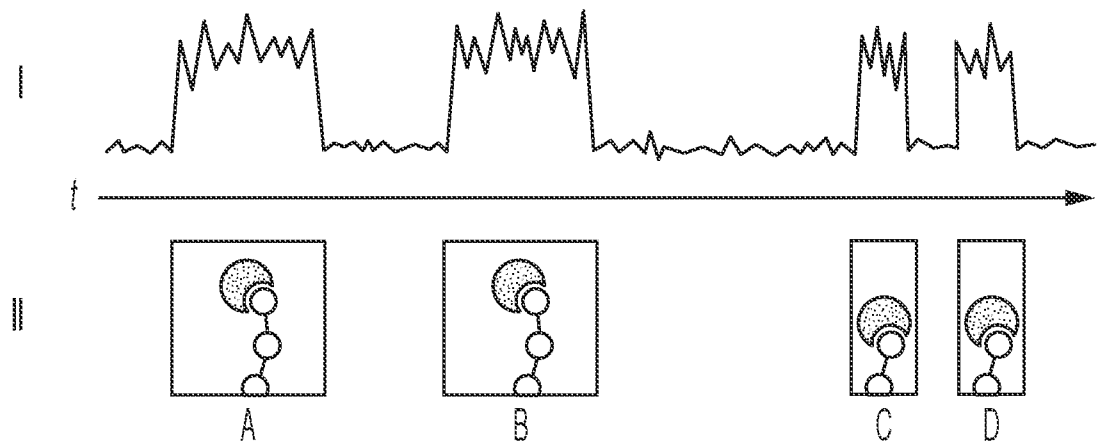


FIG. 1A

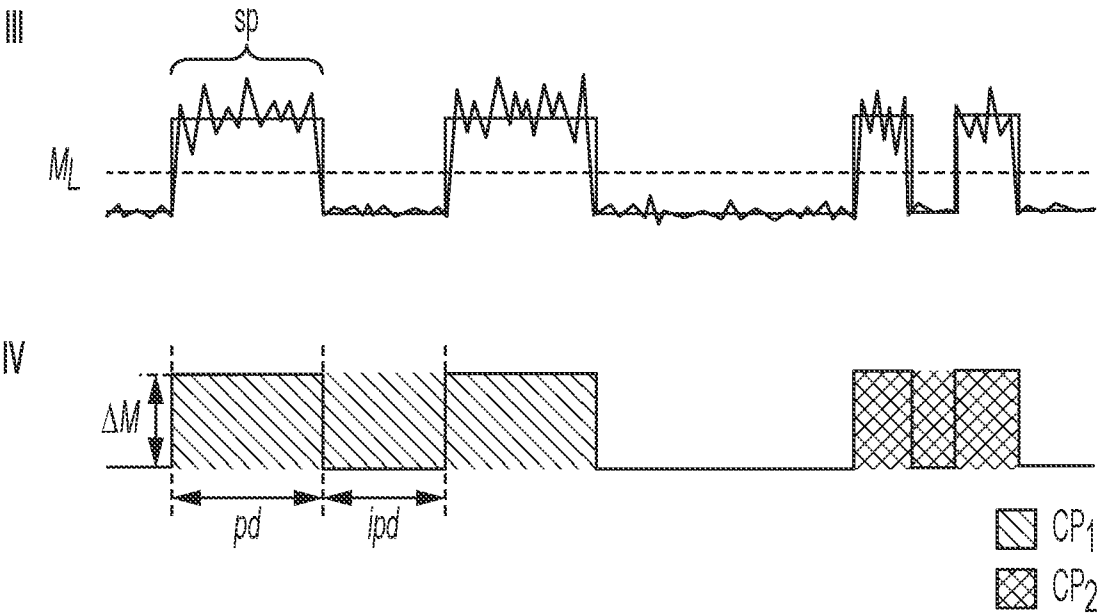


FIG. 1B

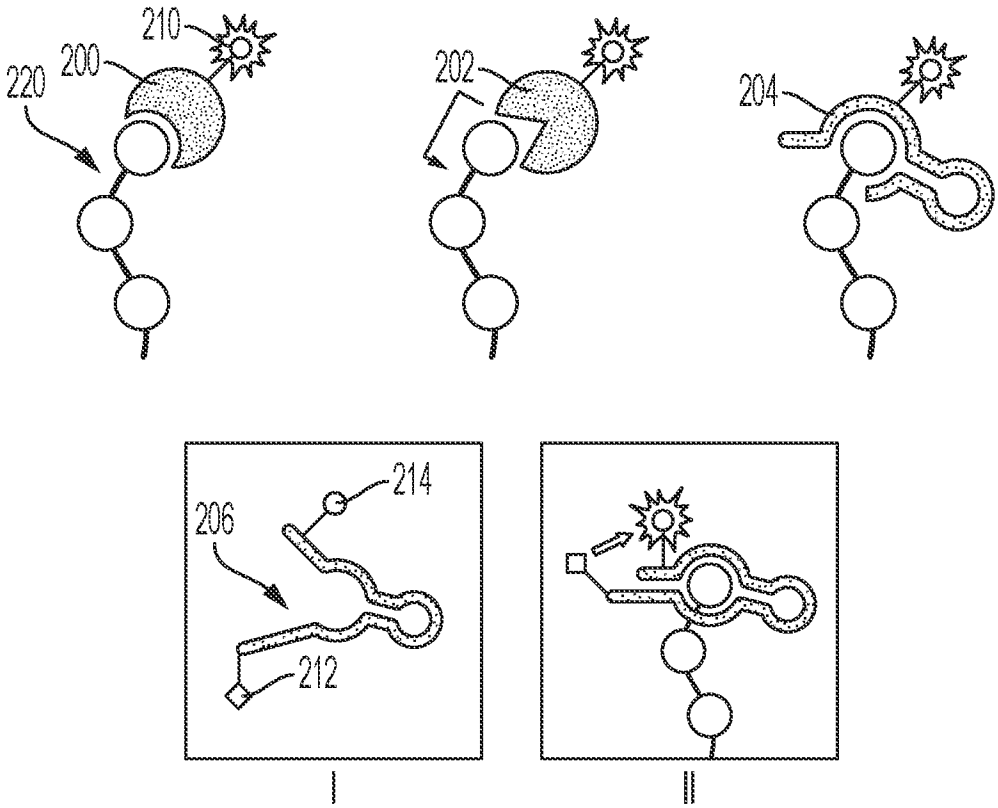


FIG. 2

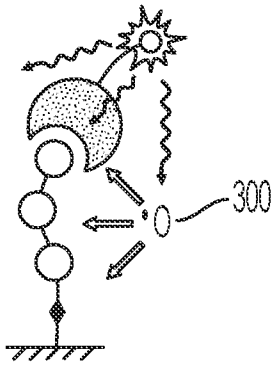


FIG. 3A

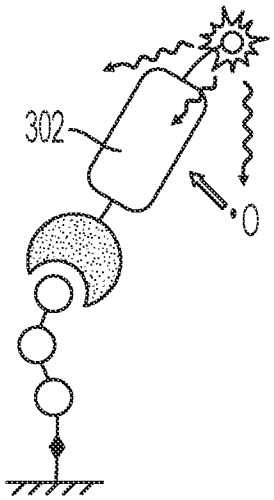


FIG. 3B

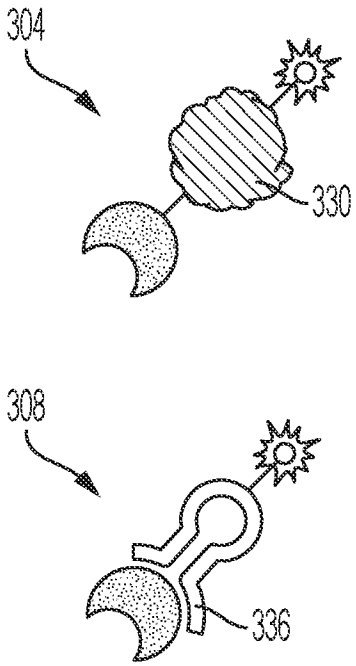


FIG. 3C

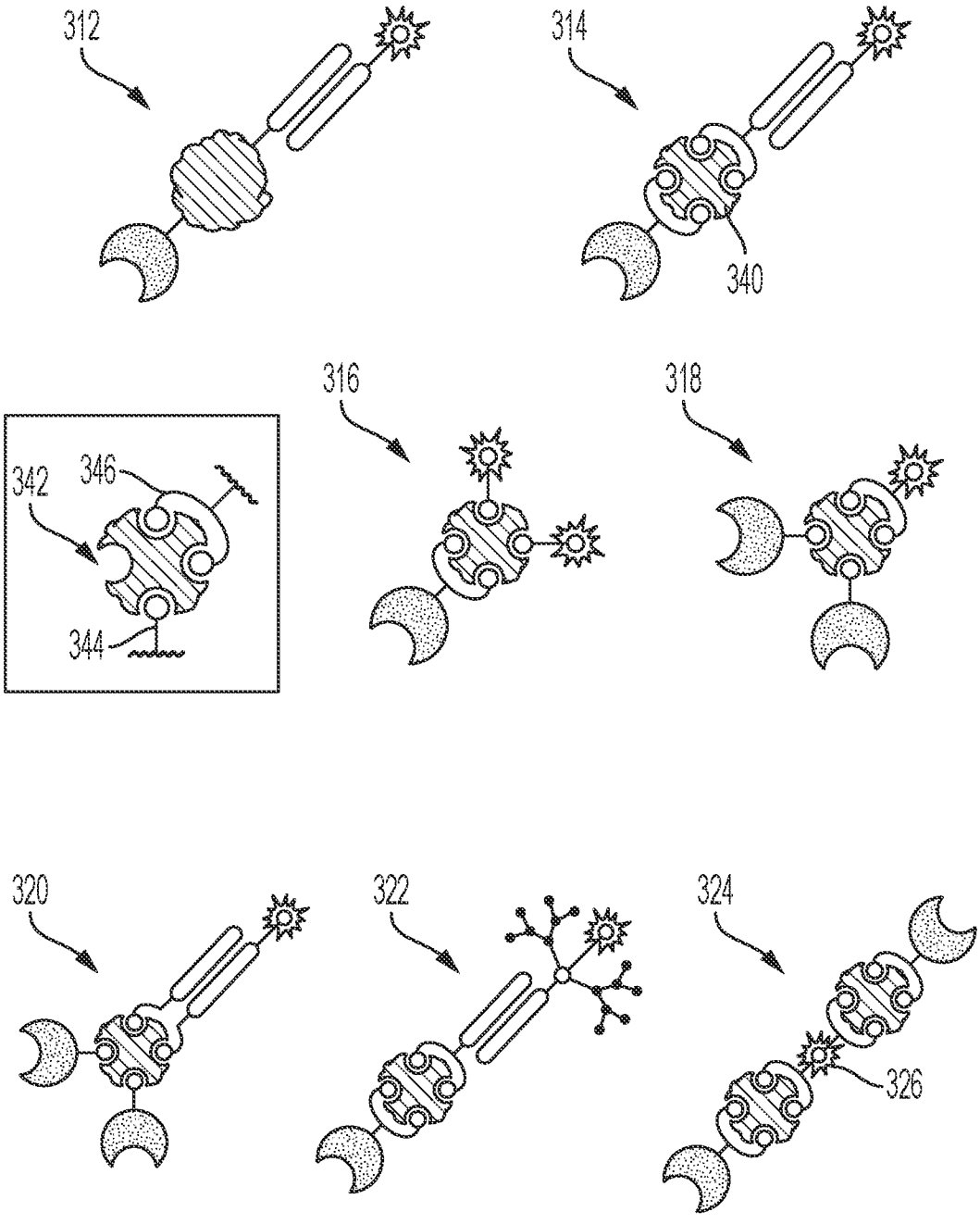


FIG. 3D

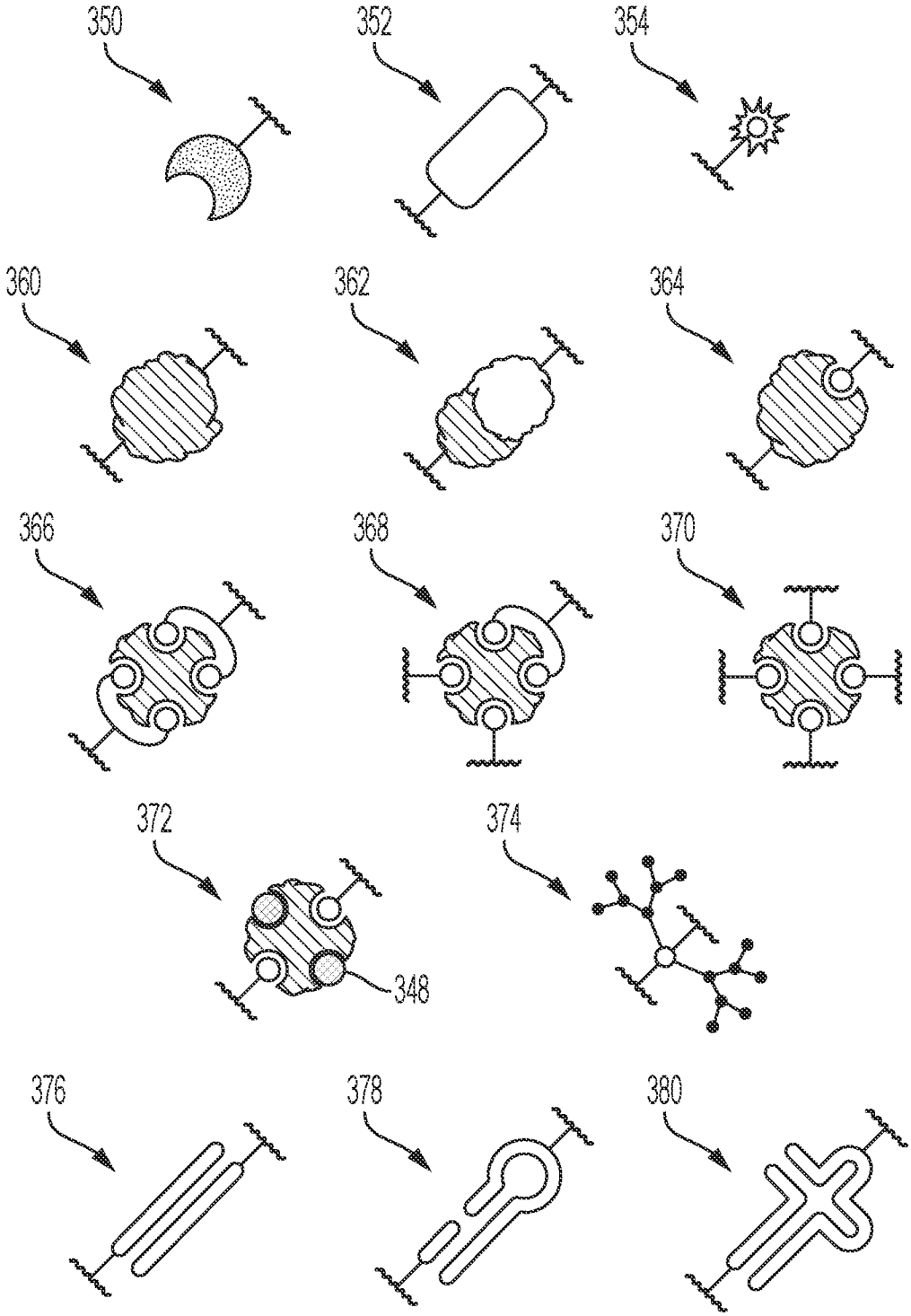


FIG. 3E

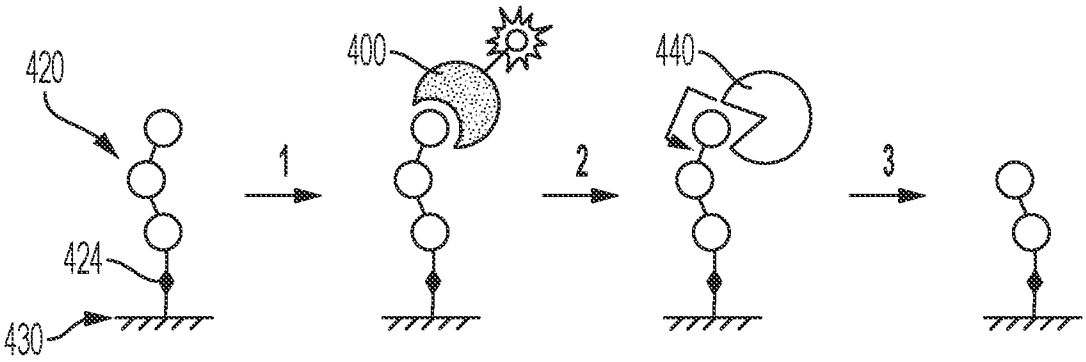

420
400
440
1
2
3
424
430

FIG. 4

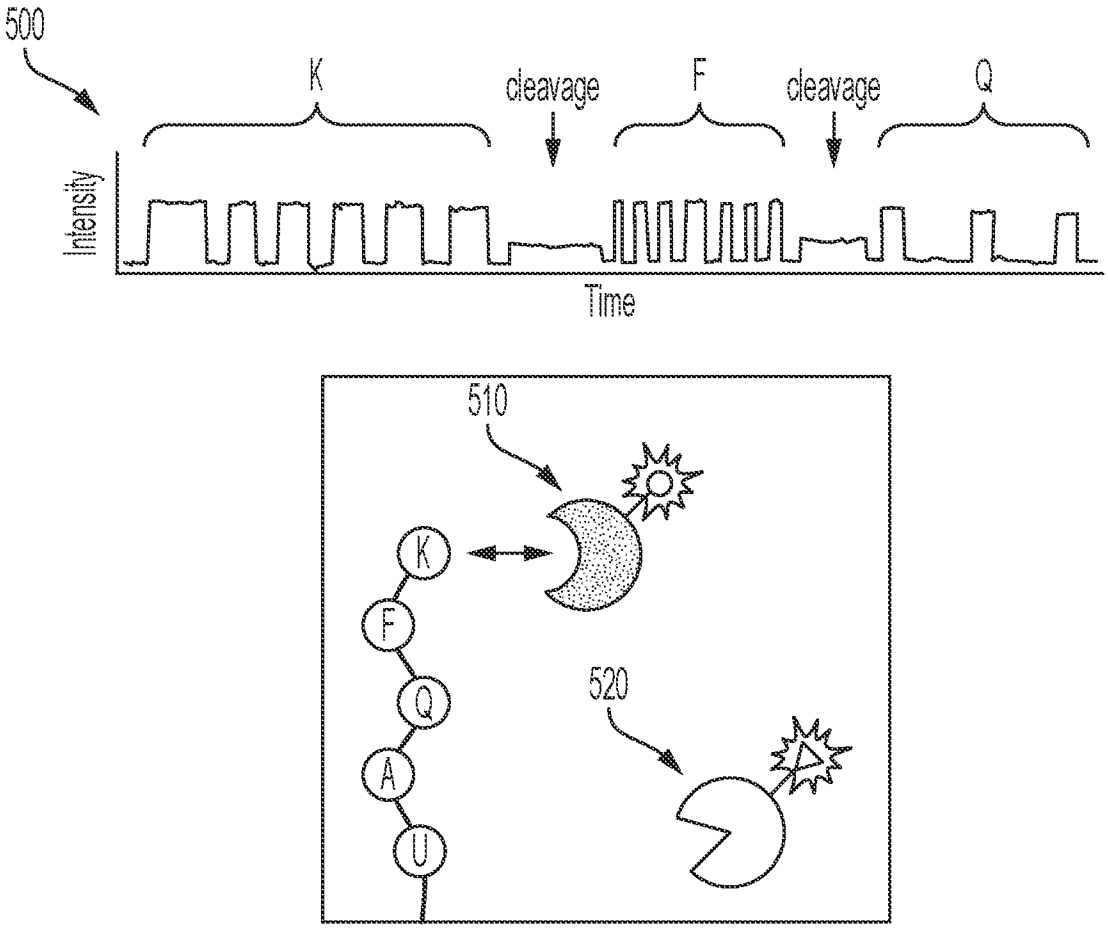

500
K
cleavage
F
cleavage
Q
Intensity
Time
510
K
F
Q
A
U
520

FIG. 5

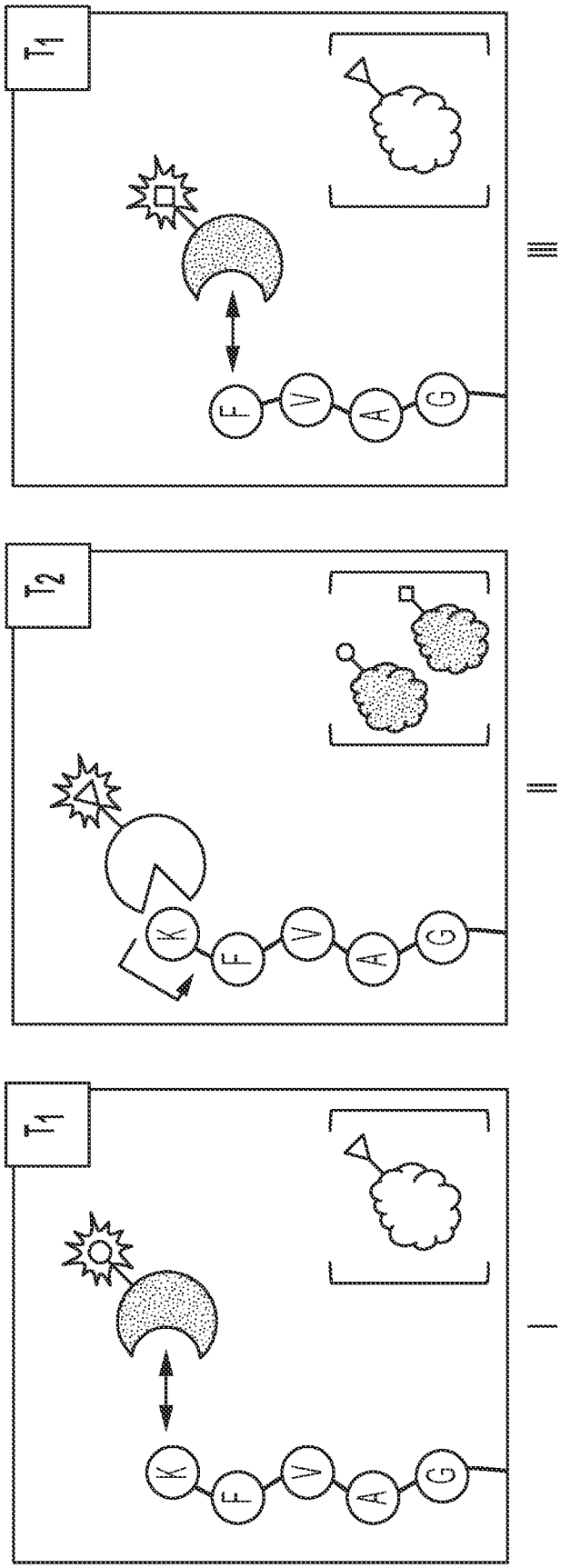


FIG. 6

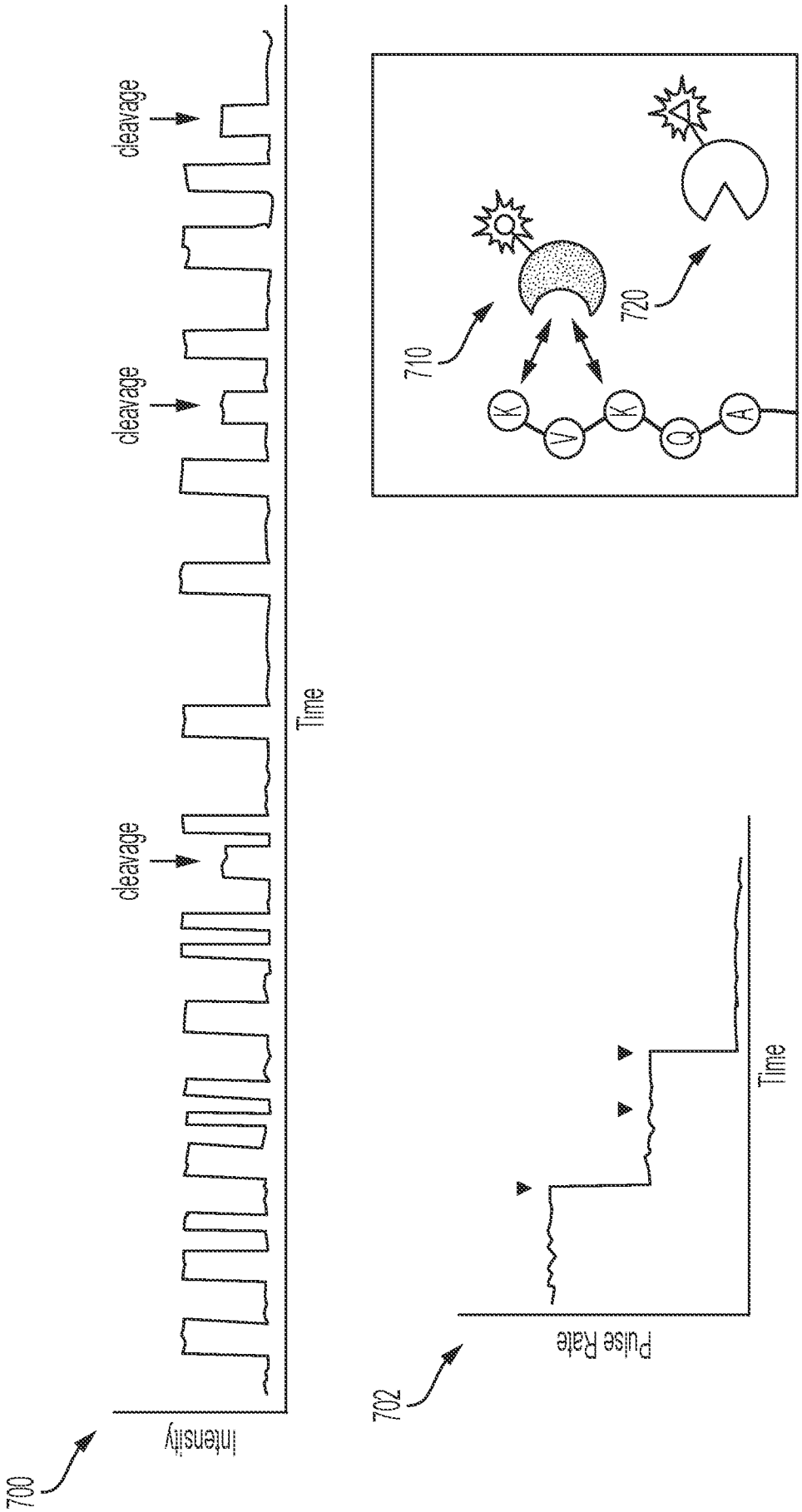

700
Intensity
Time
cleavage
cleavage
cleavage
710
720
K
V
K
Q
A
702
Pulse Rate
Time

FIG. 7

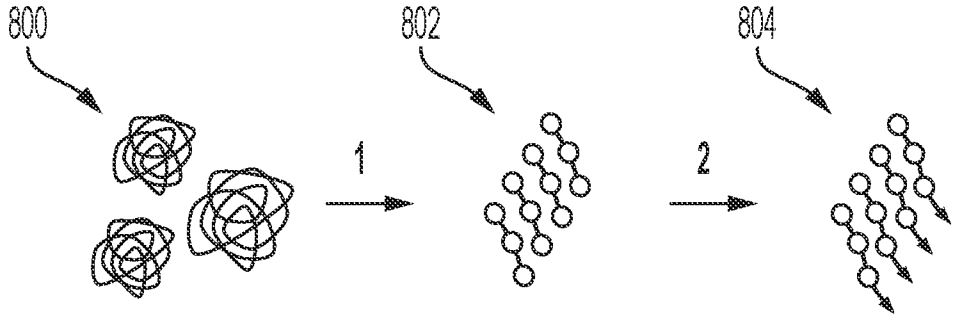


FIG. 8

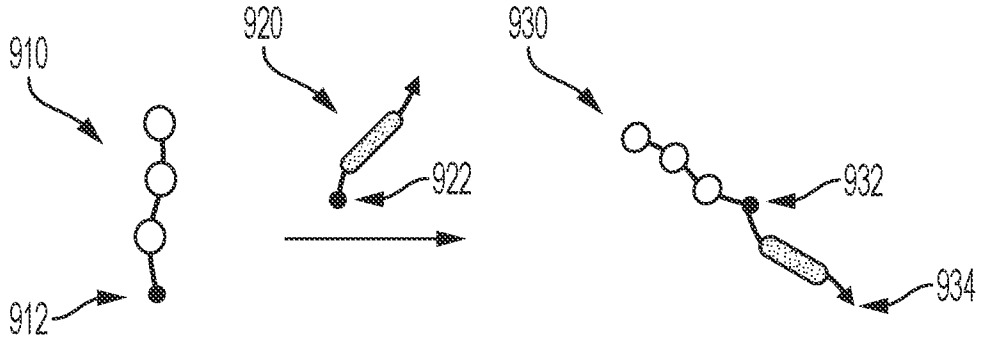


FIG. 9

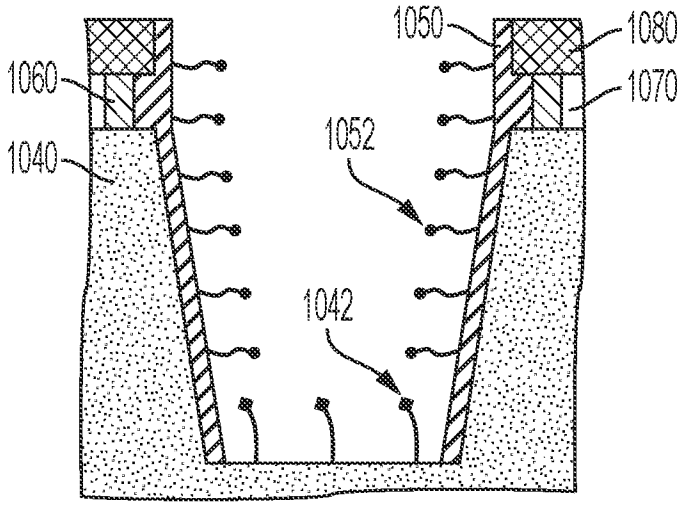


FIG. 10

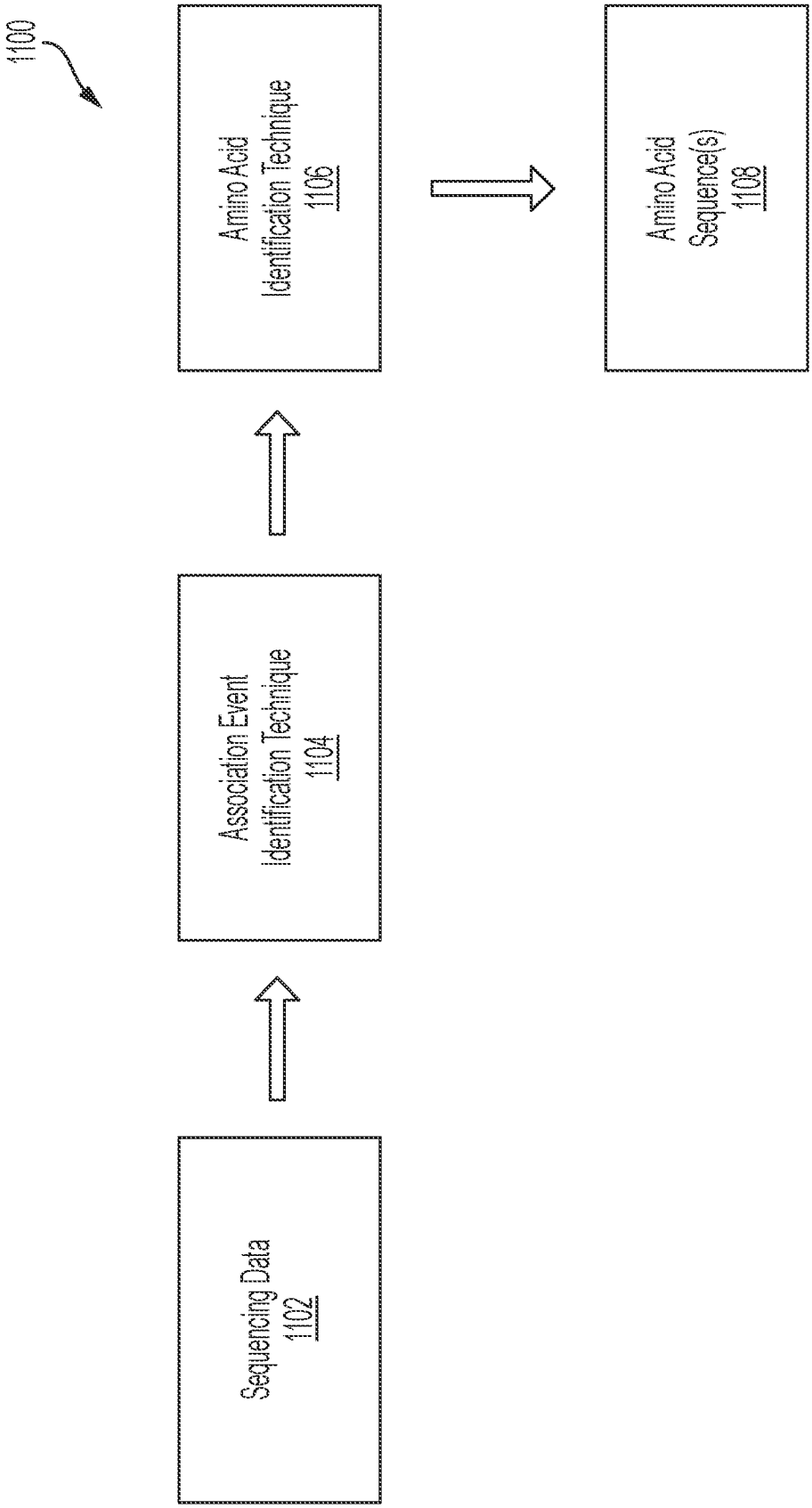

1100
Sequencing Data
1102
Association Event
Identification Technique
1104
Amino Acid
Identification Technique
1106
Amino Acid
Sequence(s)
1108

FIG. 11

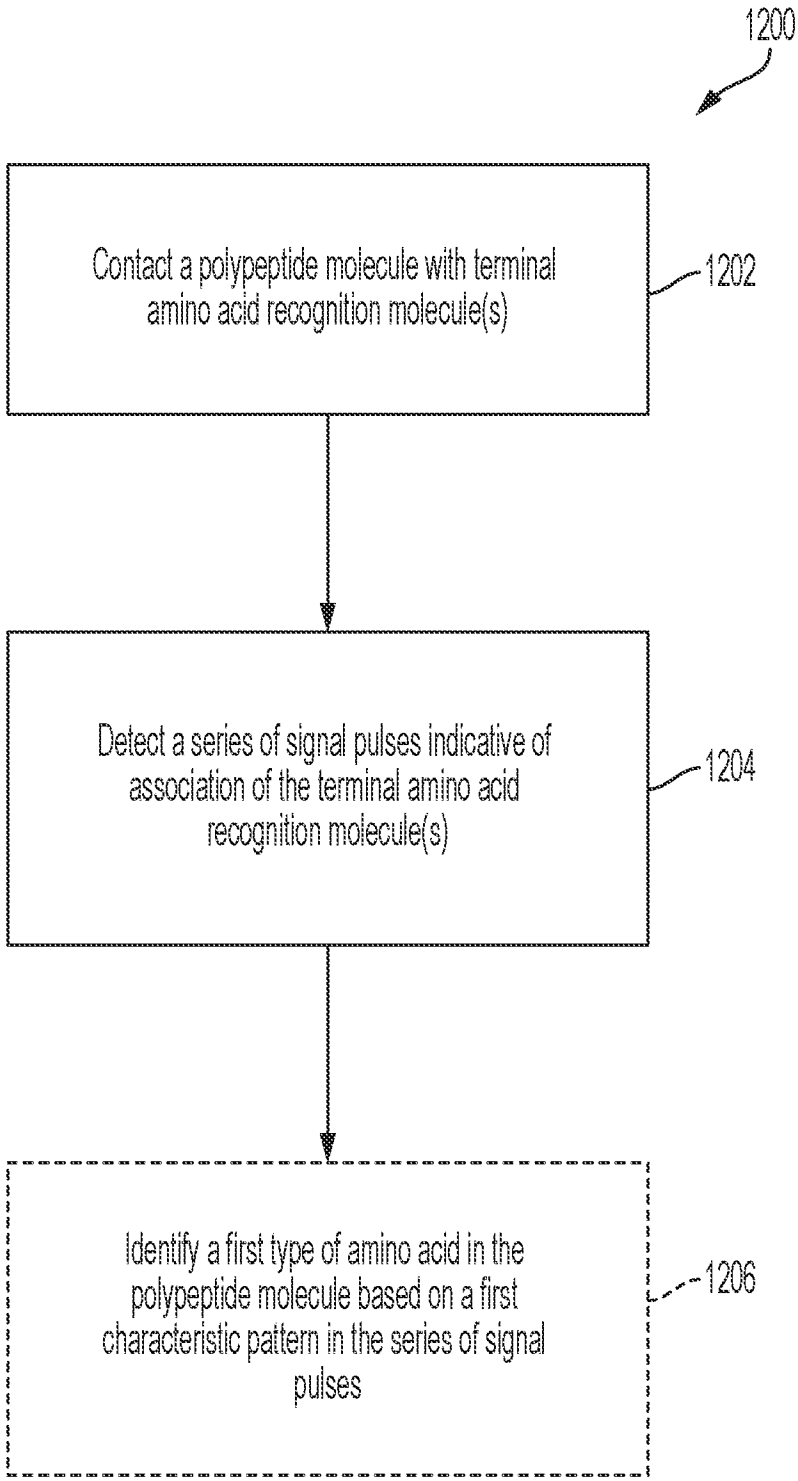

1200
Contact a polypeptide molecule with terminal amino acid recognition molecule(s)
1202
Detect a series of signal pulses indicative of association of the terminal amino acid recognition molecule(s)
1204
Identify a first type of amino acid in the polypeptide molecule based on a first characteristic pattern in the series of signal pulses
1206

FIG. 12

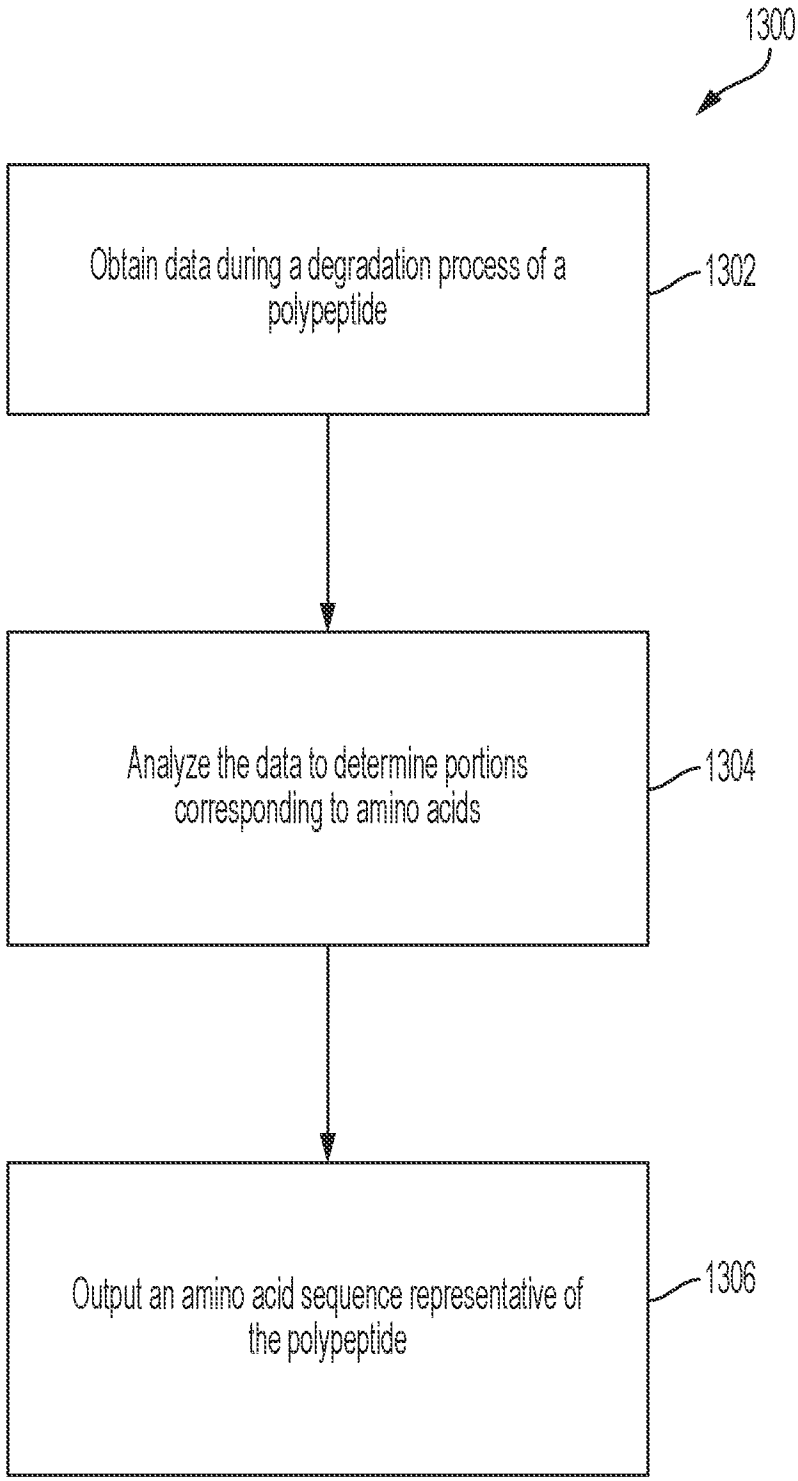

1300
Obtain data during a degradation process of a polypeptide
1302
Analyze the data to determine portions corresponding to amino acids
1304
Output an amino acid sequence representative of the polypeptide
1306

FIG. 13

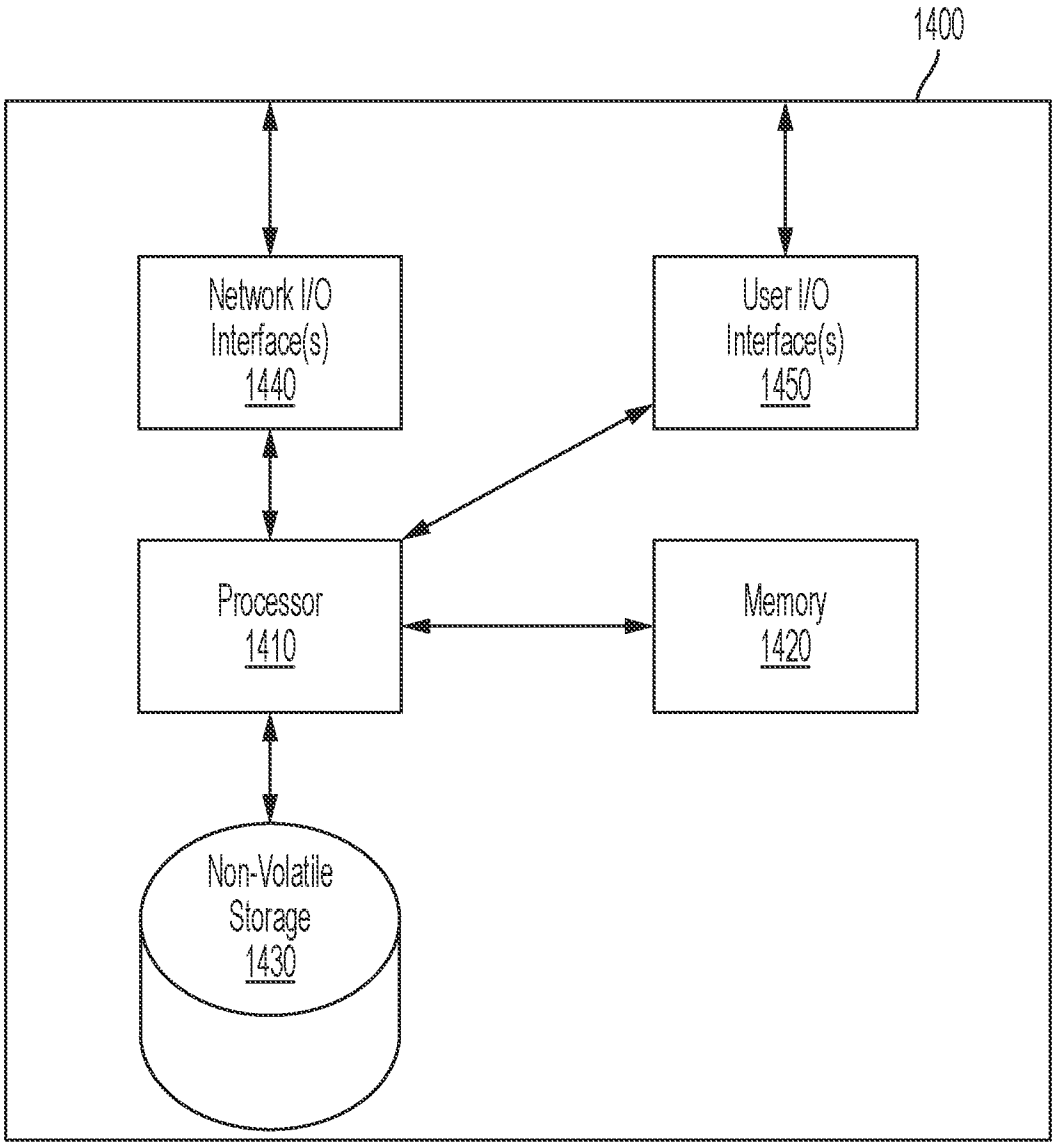

1400
Network I/O Interface(s) 1440
User I/O Interface(s) 1450
Processor 1410
Memory 1420
Non-Volatile Storage 1430

FIG. 14

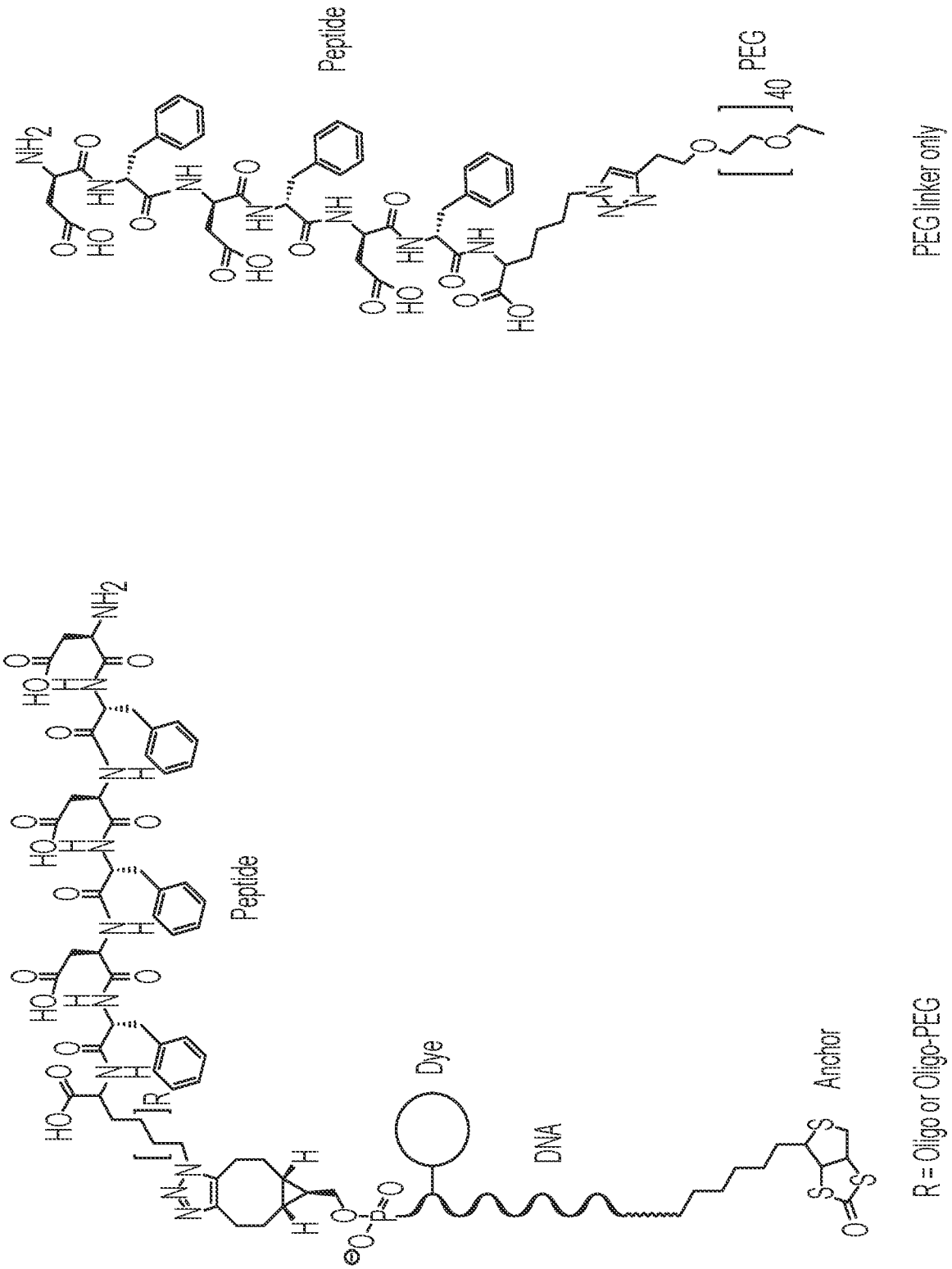

Peptide
PEG
40
PEG linker only
Peptide
Dye
DNA
Anchor
R = Oligo or Oligo-PEG

FIG. 15A

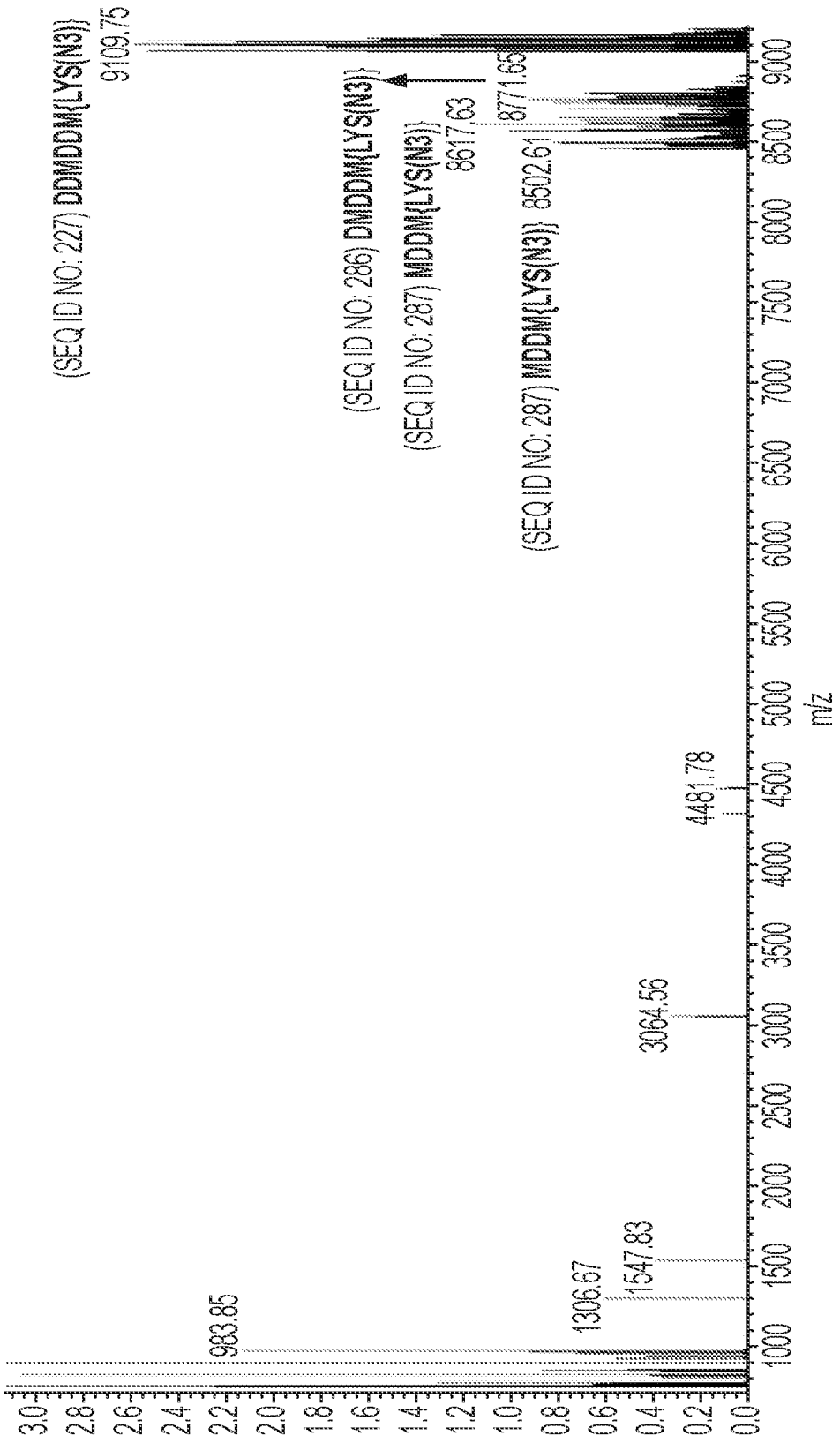

(SEQ ID NO: 227) DDMDDIM(LYS(N3))
9109.75
(SEQ ID NO: 286) DMDDIM(LYS(N3))
(SEQ ID NO: 287) MDDIM(LYS(N3))
8617.63
(SEQ ID NO: 287) MDDIM(LYS(N3)) 8502.61
8771.65
983.85
1306.67
1547.83
3064.56
4481.78
m/z

FIG. 15B

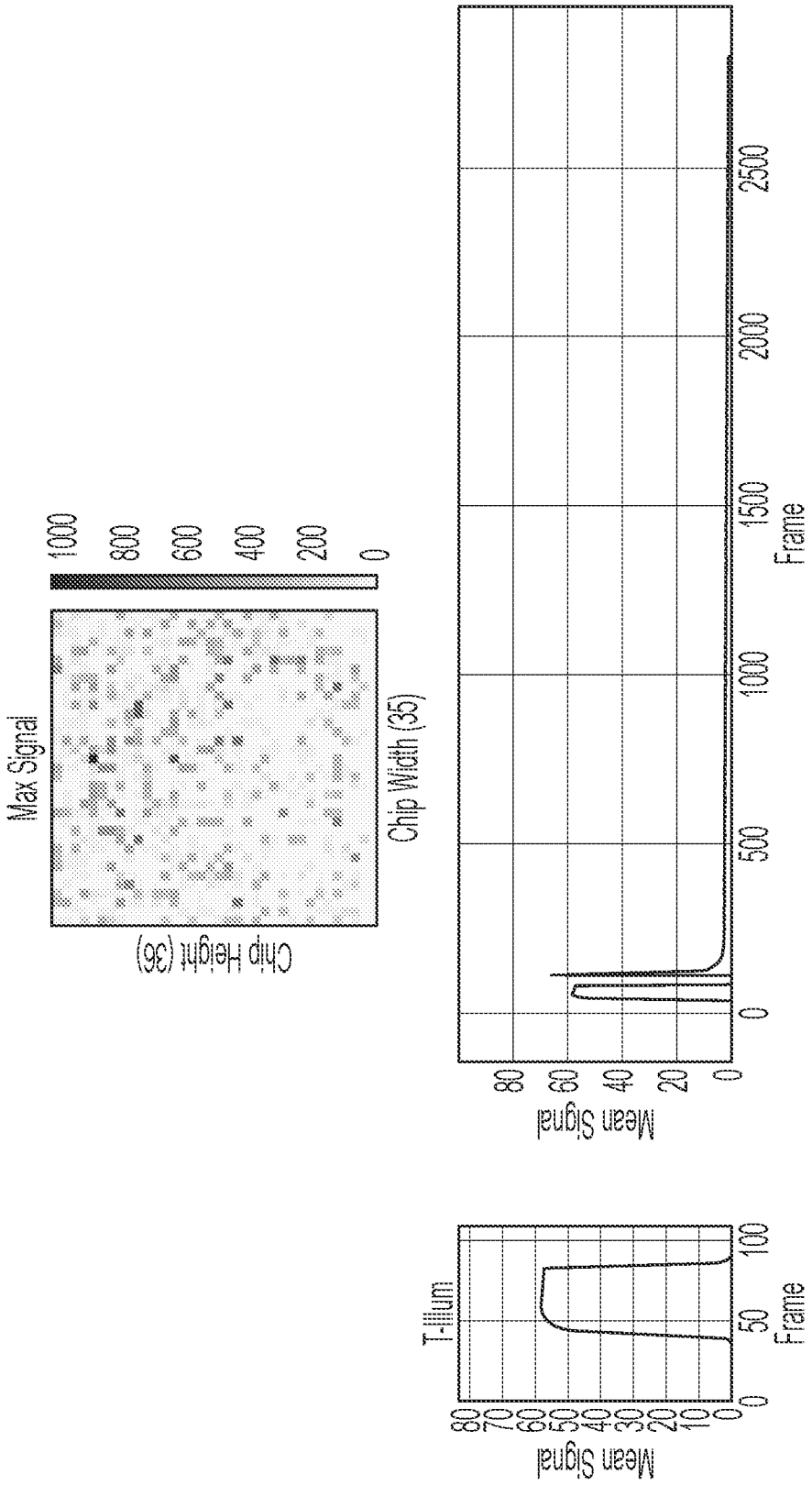

Max Signal
1000
800
600
400
200
0
Chip Height (36)
Chip Width (35)
Mean Signal
80
60
40
20
0
0
500
1000
1500
2000
2500
Frame
T-Illum
Mean Signal
80
70
60
50
40
30
20
10
0
0
50
100
Frame

FIG. 15C

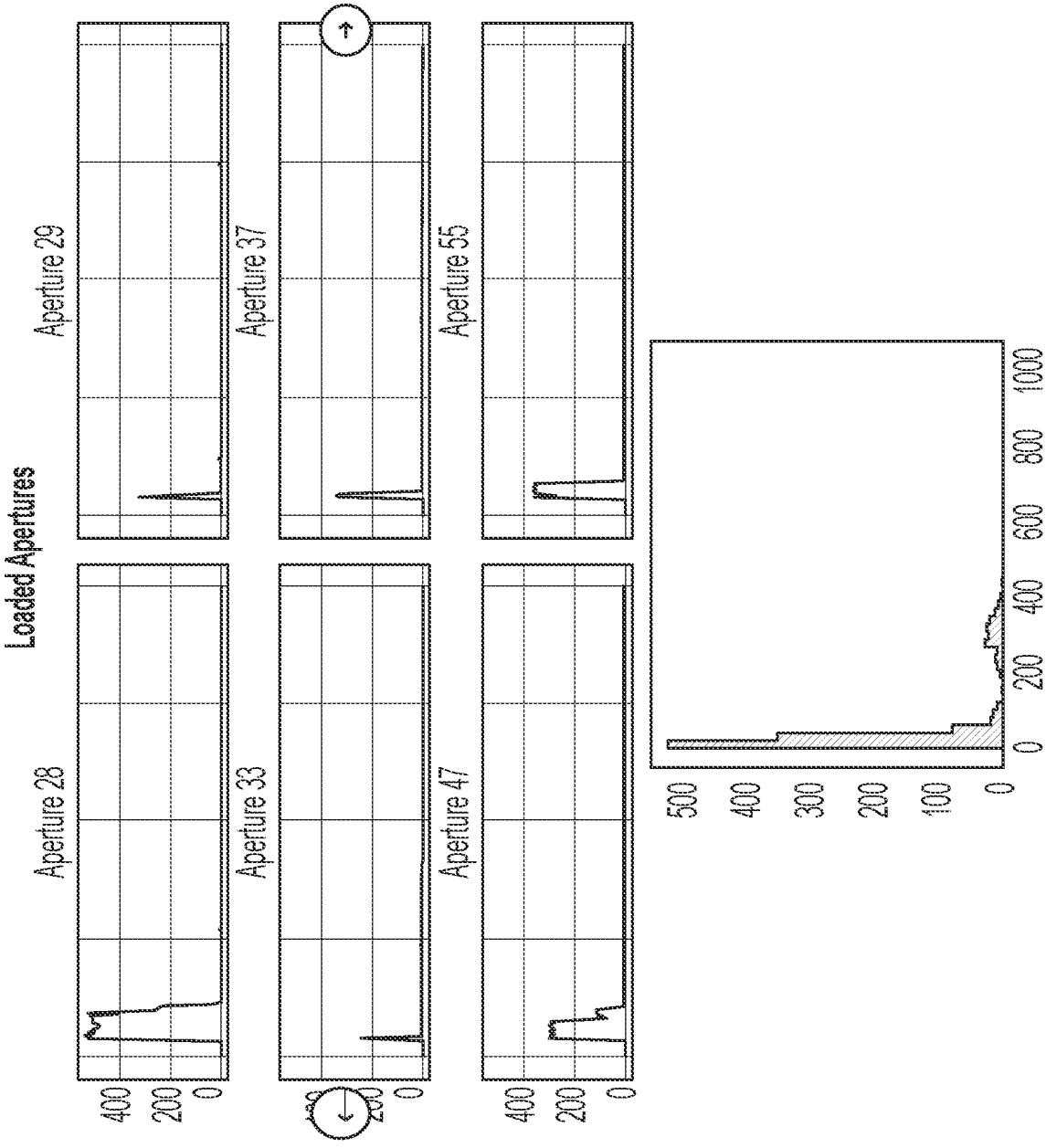


FIG. 15C
CONTINUED

PepSeq Cutter Cocktail

yPIP

hTET

PfuTET

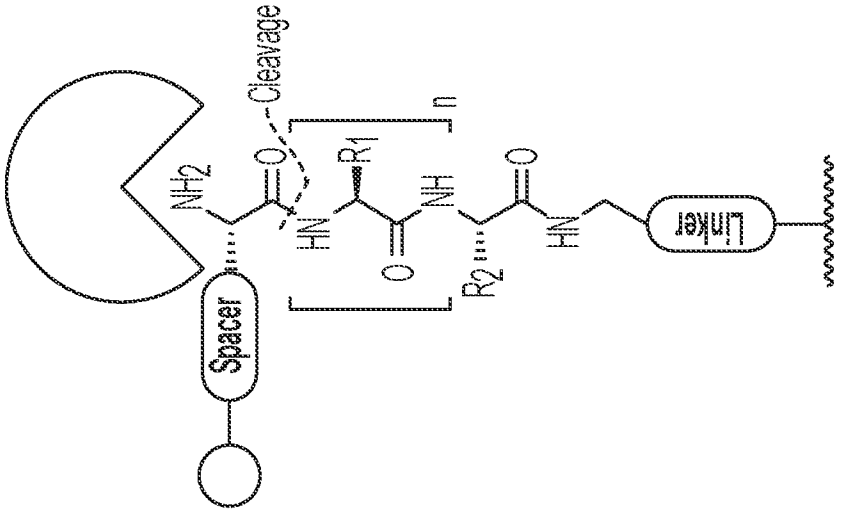
Only cuts XP-

Does not cut: R, K, E, M, XP-

Does not cut: V, I, G, XP-

Enzyme:	Residue(s) with High Specific Activity (X ≥ 75% Cleavage in 30 min at 1mM Substrate with 1 uM Enzyme at 25C)	Moderate Specific Activity (75% ≤ X ≥25%)	Low Specific Activity (25% ≤ X >0%)	No Activity
cvPr	R, A, F, V, I, L, T, W, M	K, Y, Q, S	N, G, H	D, E, P (when followed by two Ps), C
yPIP		P (when followed by two Ps)		R, A, F, V, I, L, T, W, M, K, Y, Q, S, N, G, H, D, E, C
D/EAPN		E, I	D	R, A, F, V, L, T, W, M, K, Y, Q, S, N, G, H, C, P
hTET	A, F, Y, V, I, L, S, T	N, Q, C, H, G	D, W	R, K, E, M
PfuAPI	R, A, K, Y, D, E, N, Q, S, T, M, L	W, F, C	H	V, I, G

FIG. 16


Cleavage
n
NH2
O
R1
HN
NH
O
HN
Linker
R2
Spacer

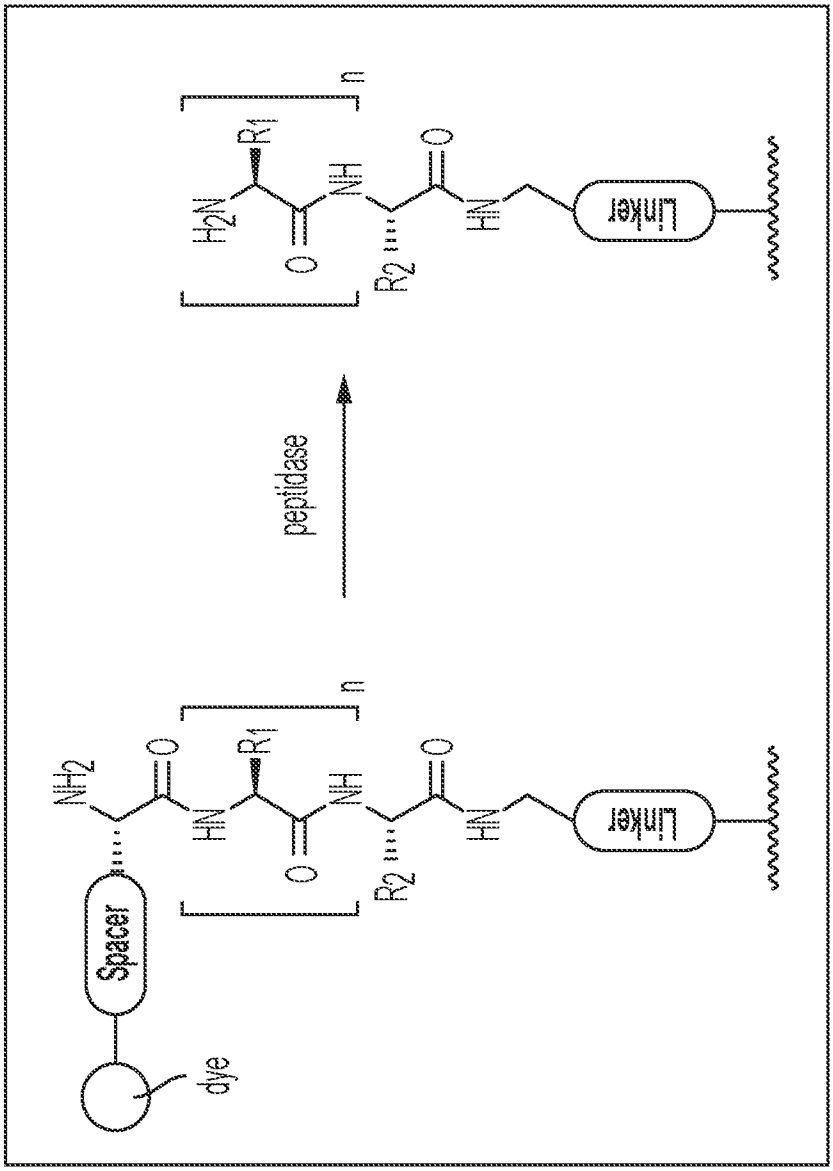

n
R1
H2N
NH
O
HN
Linker
R2
peptidase
n
NH2
O
R1
HN
NH
O
HN
Linker
R2
Spacer
dye

FIG. 17A

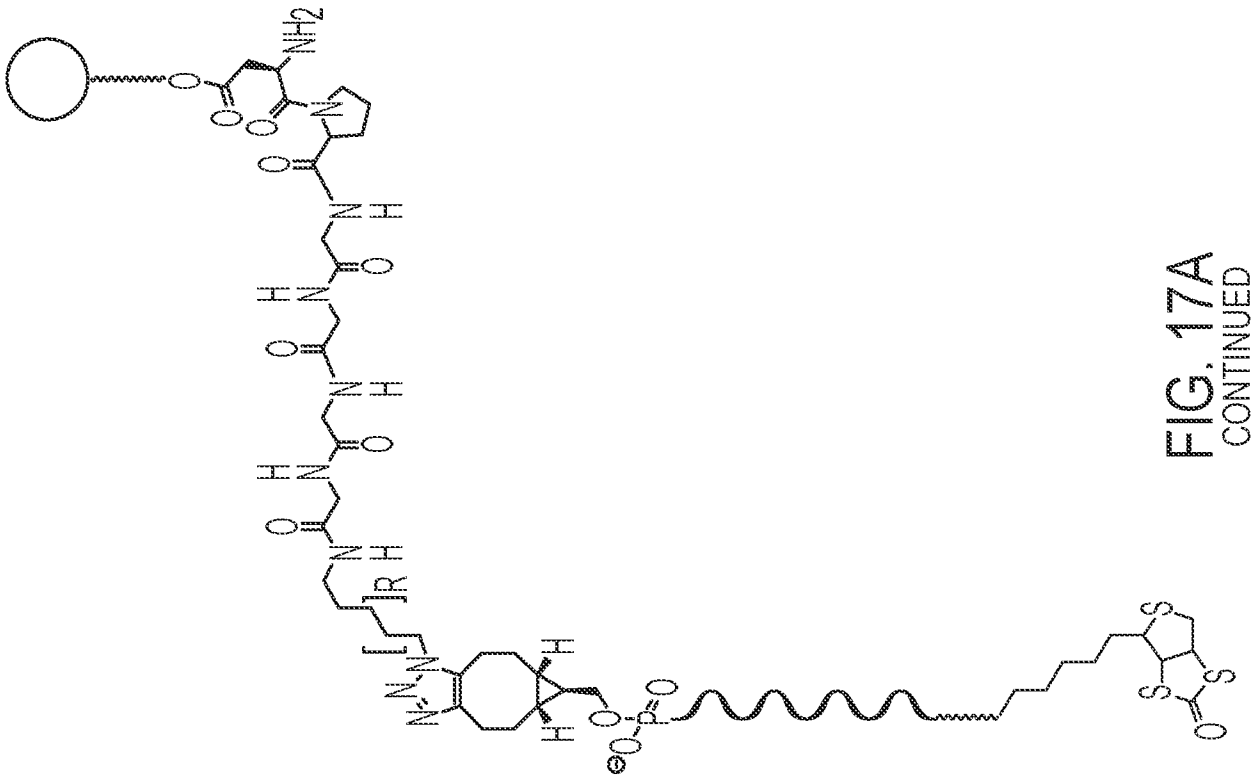


FIG. 17A
CONTINUED

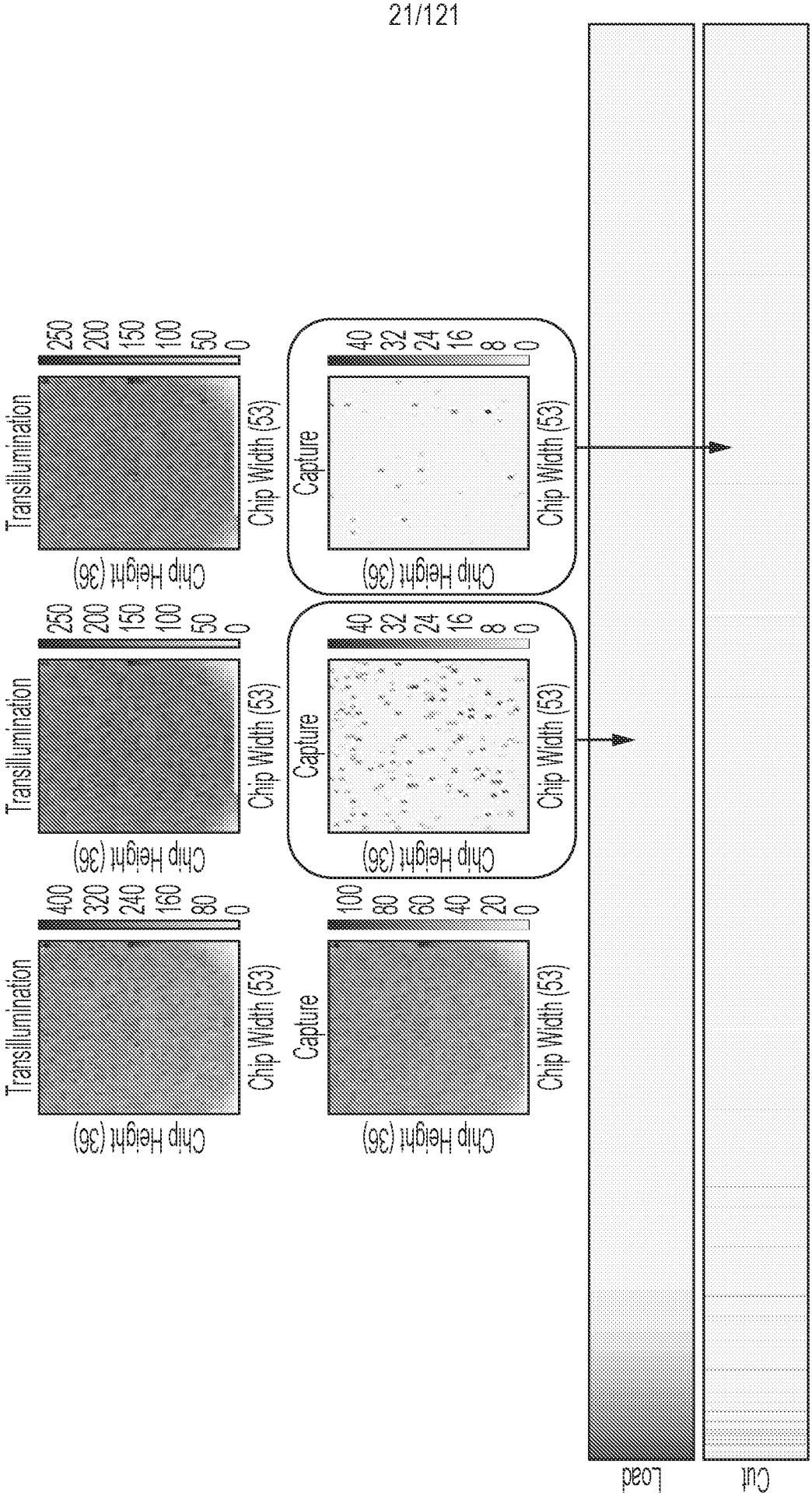

Transillumination
Capture
Chip Width (53)
Chip Height (36)
250
200
150
100
50
0
40
32
24
16
8
400
320
240
160
80
100
60
20
Load
Out

FIG. 17B

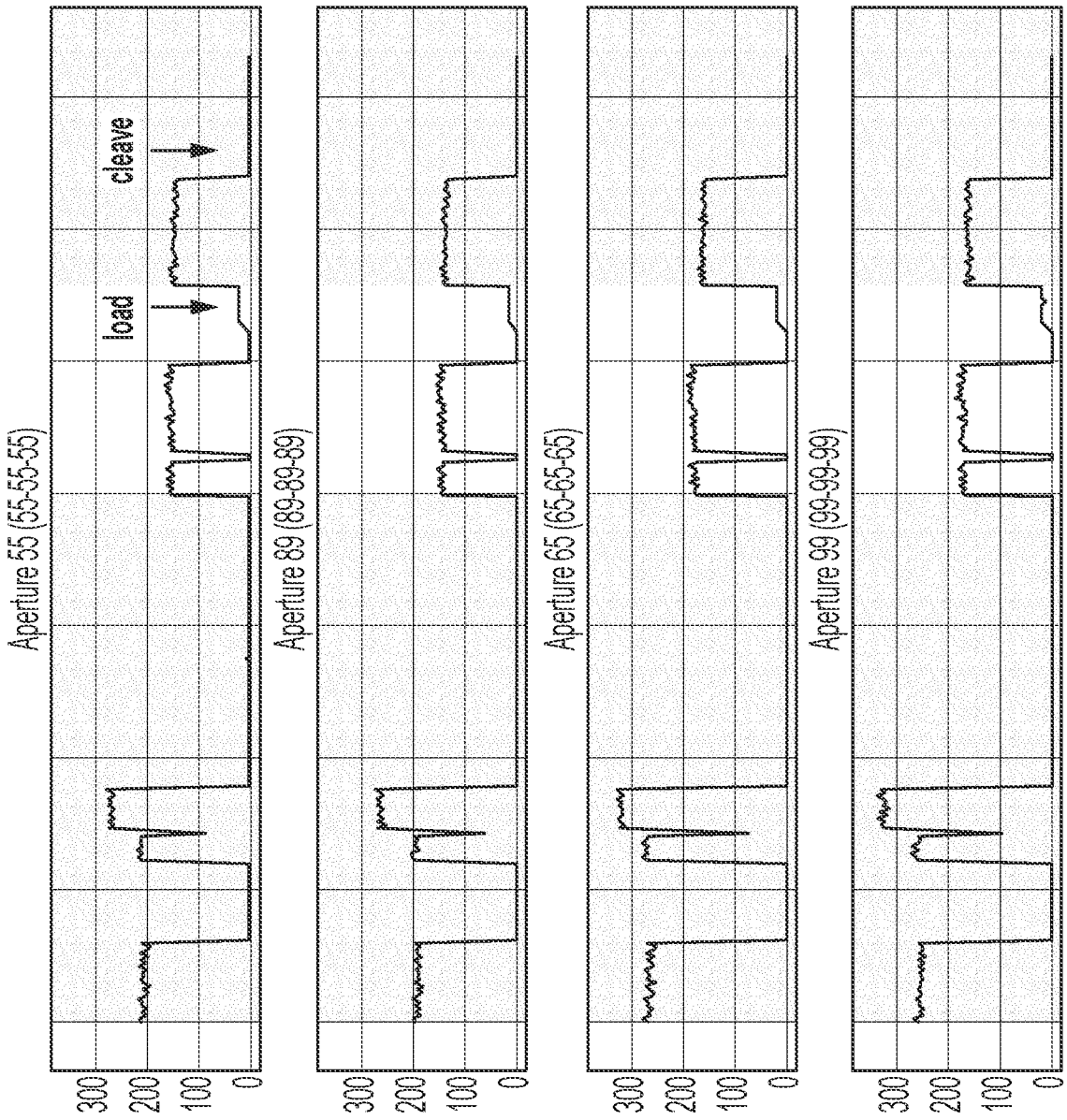
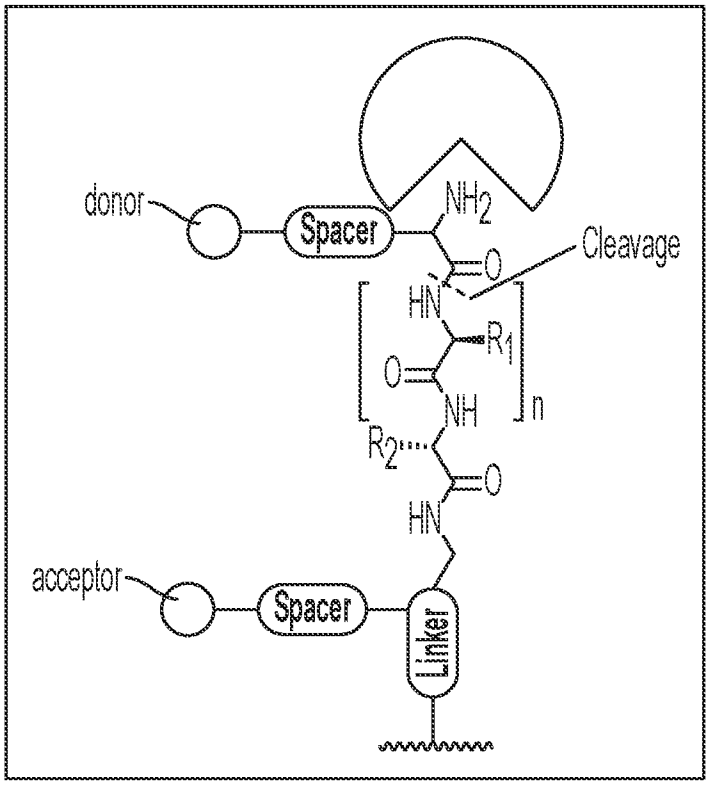
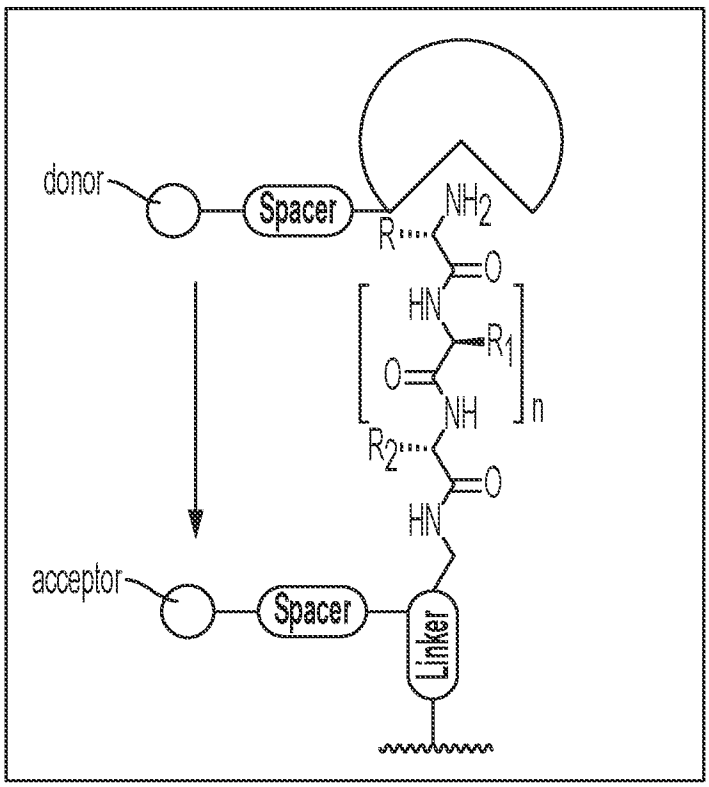

Aperture 55 (55-55-55)
load
cleave
300
200
100
0
Aperture 89 (89-89-89)
300
200
100
0
Aperture 65 (65-65-65)
300
200
100
0
Aperture 99 (99-99-99)
300
200
100
0

FIG. 17C

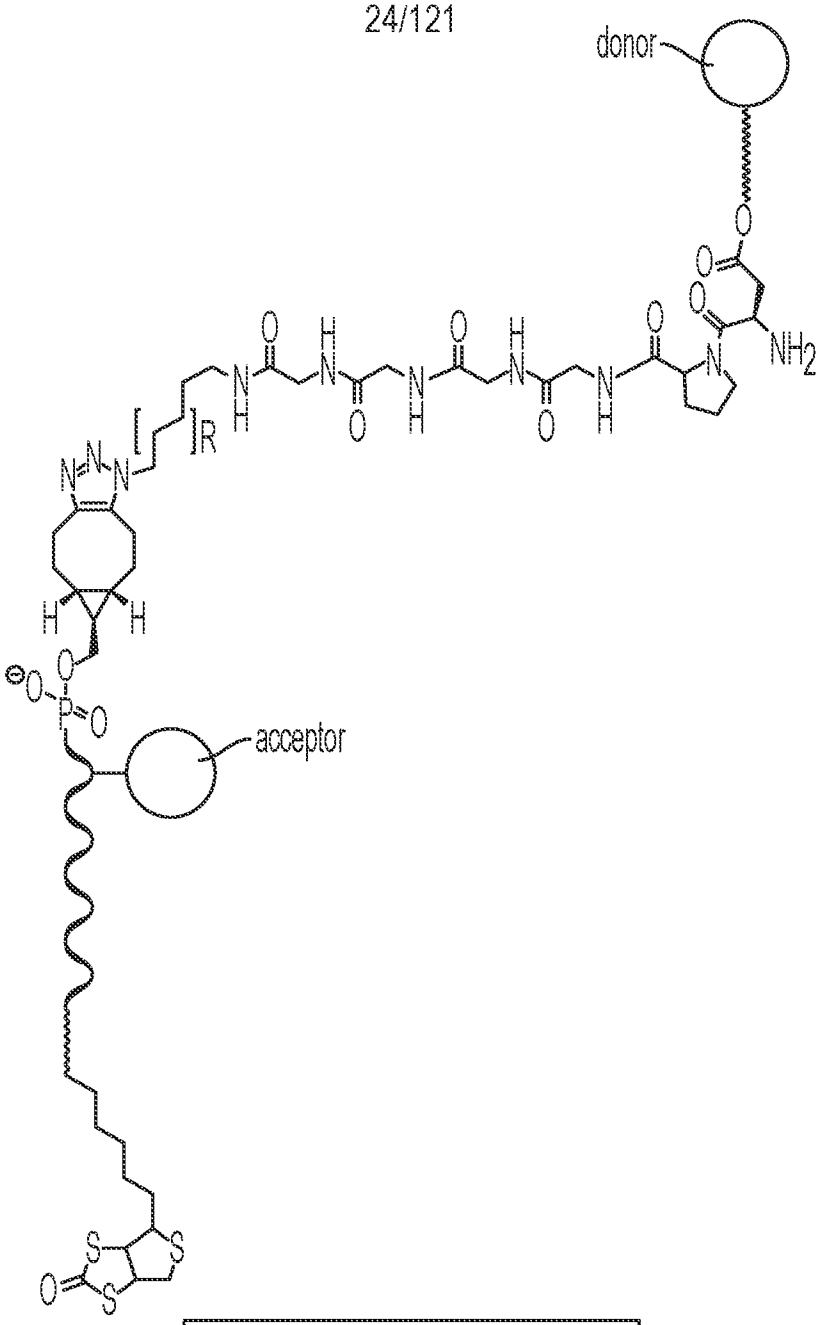

donor
Spacer
NH2
Cleavage
HN
R1
O
NH
n
R2
O
HN
acceptor
Spacer
Linker

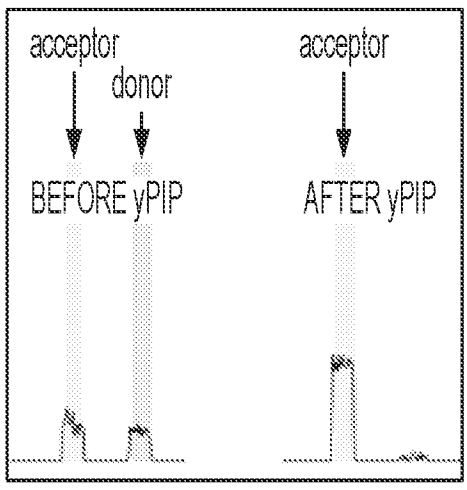
A


donor
Spacer
R
NH2
O
HN
R1
O
NH
n
R2
O
HN
acceptor
Spacer
Linker

B

FIG. 18A





C

FIG. 18A
CONTINUED

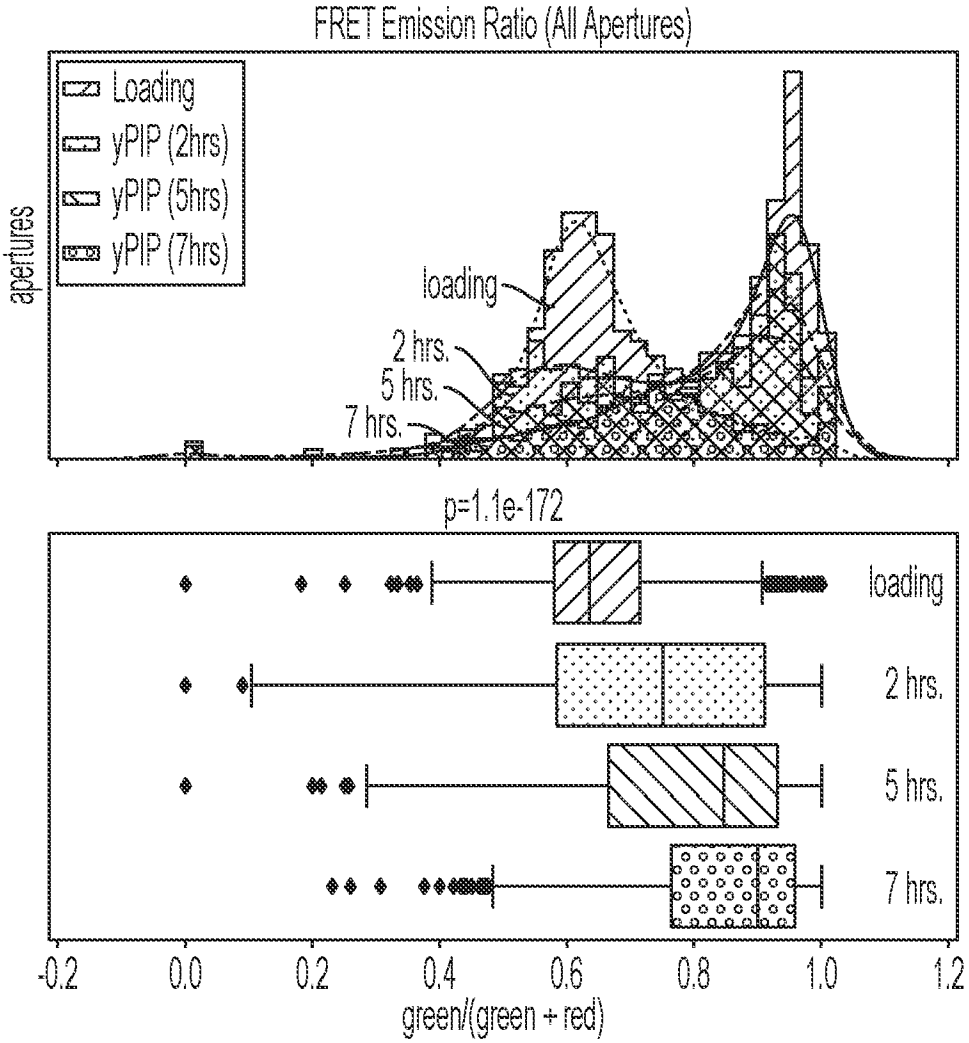


FIG. 18B

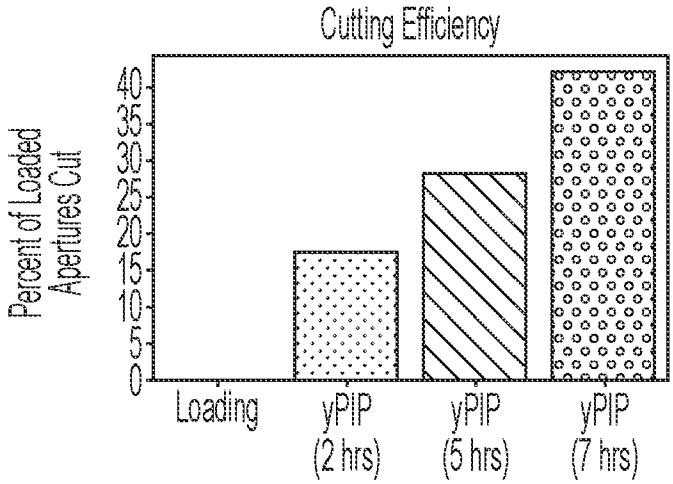


FIG. 18C

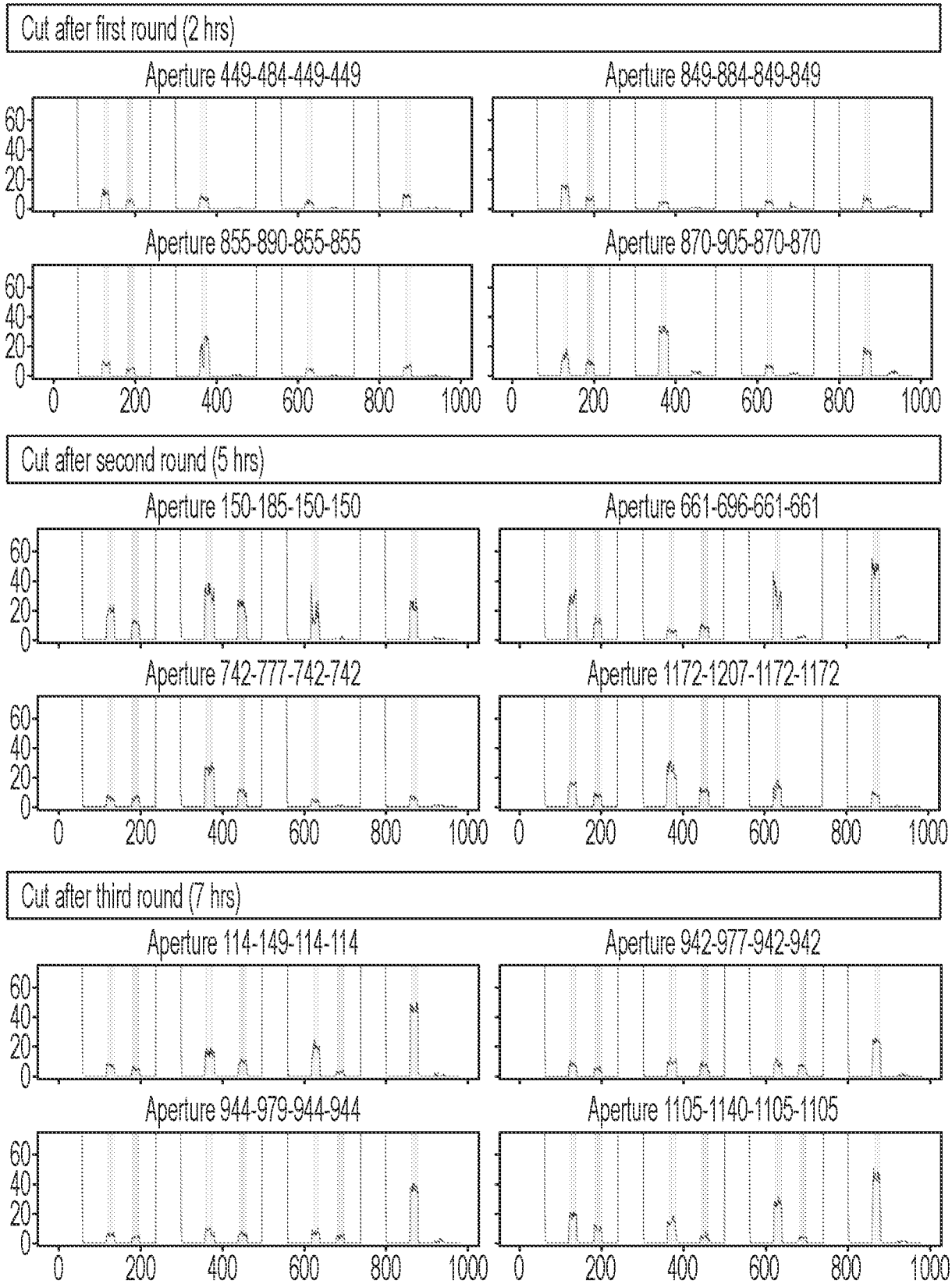

Cut after first round (2 hrs)
Aperture 449-484-449-449
Aperture 849-884-849-849
Aperture 855-890-855-855
Aperture 870-905-870-870
Cut after second round (5 hrs)
Aperture 150-185-150-150
Aperture 661-696-661-661
Aperture 742-777-742-742
Aperture 1172-1207-1172-1172
Cut after third round (7 hrs)
Aperture 114-149-114-114
Aperture 942-977-942-942
Aperture 944-979-944-944
Aperture 1105-1140-1105-1105
0
20
40
60
0
200
400
600
800
1000

FIG. 18D

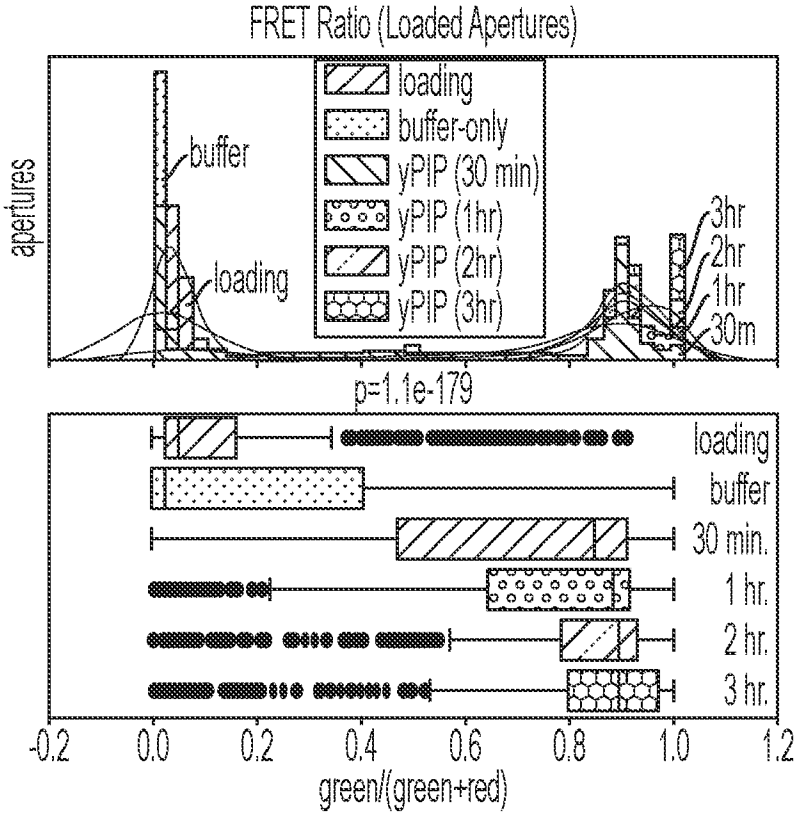


FIG. 18E

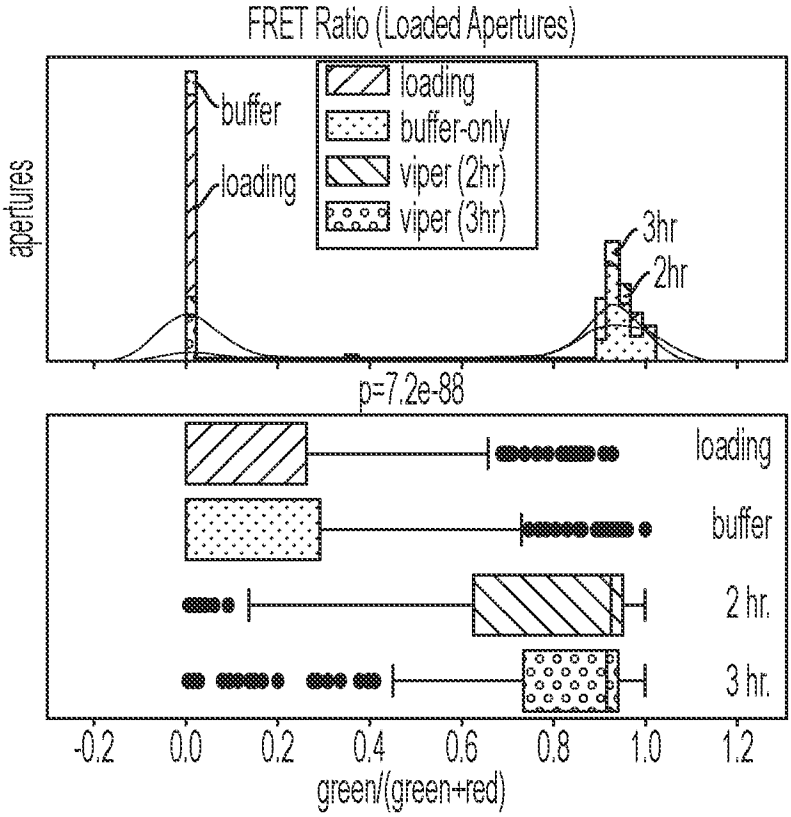


FIG. 18F

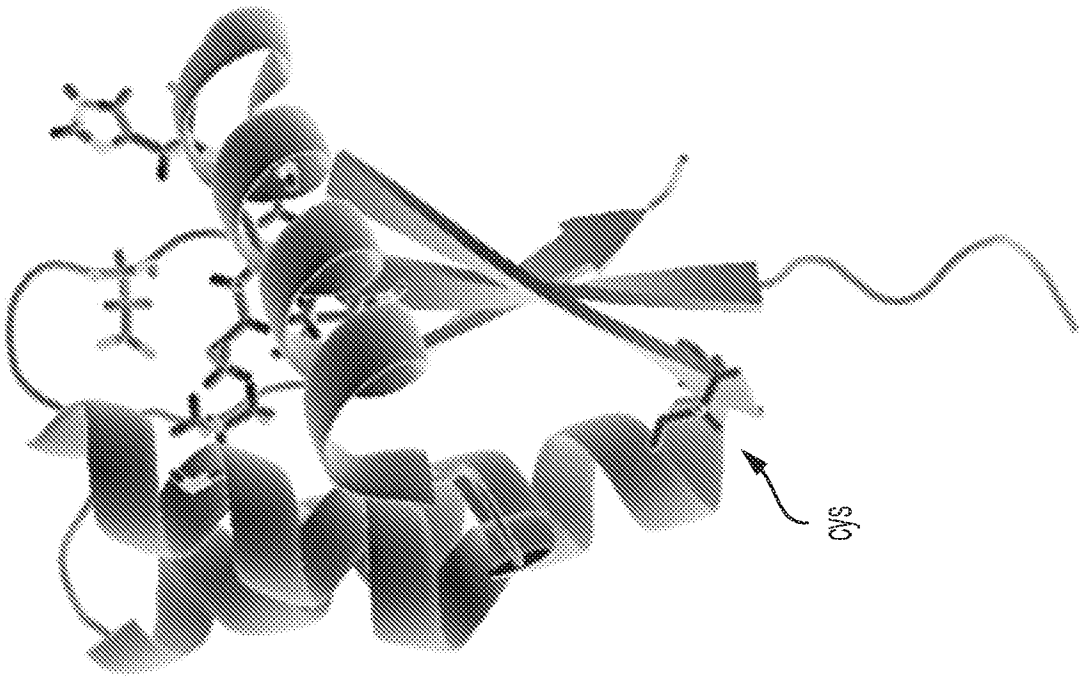


FIG. 19A

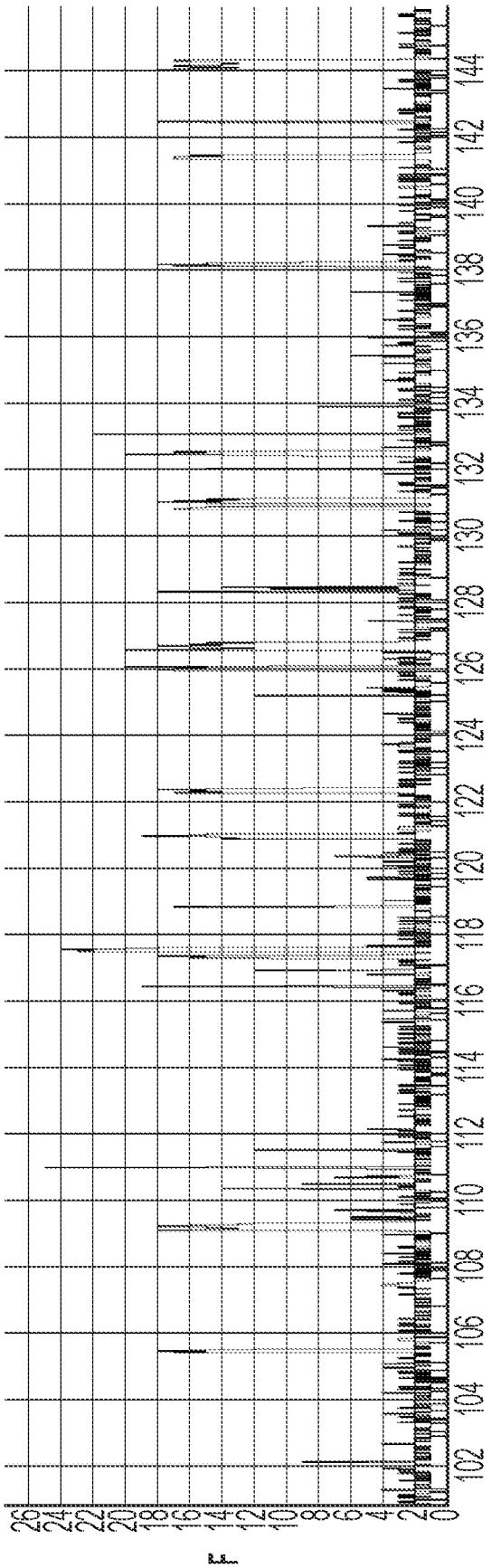

102
104
106
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
0
2
4
6
8
10
12
14
16
18
20
22
24
26

FIG. 19B

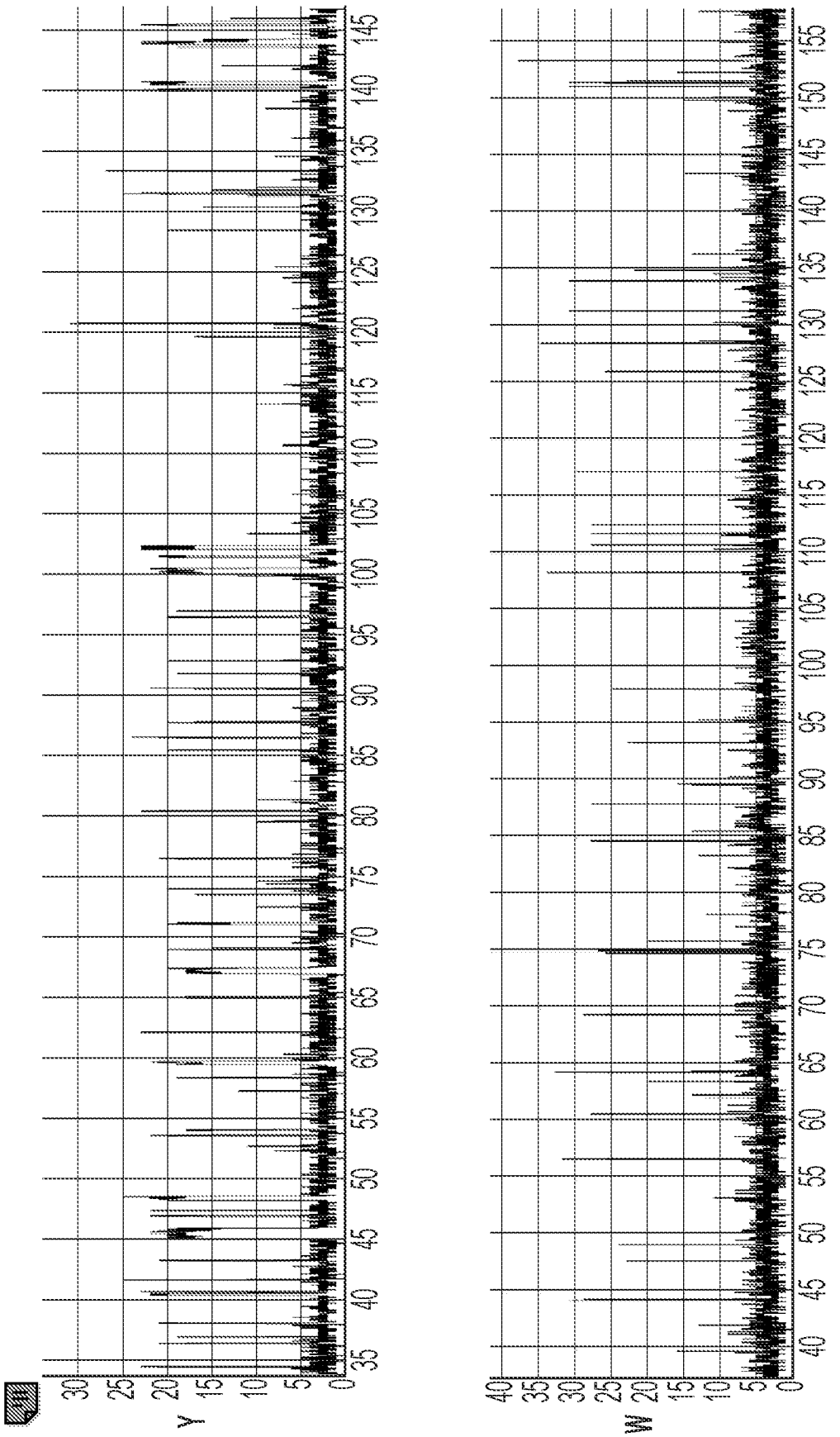


FIG. 19B
CONTINUED

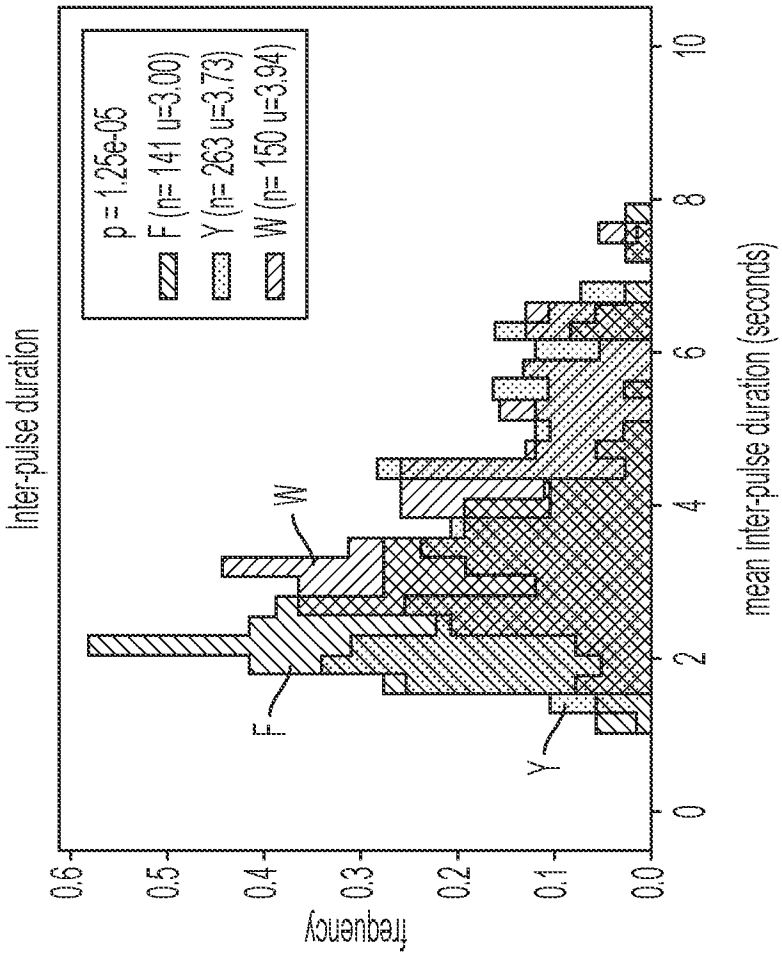


FIG. 19D

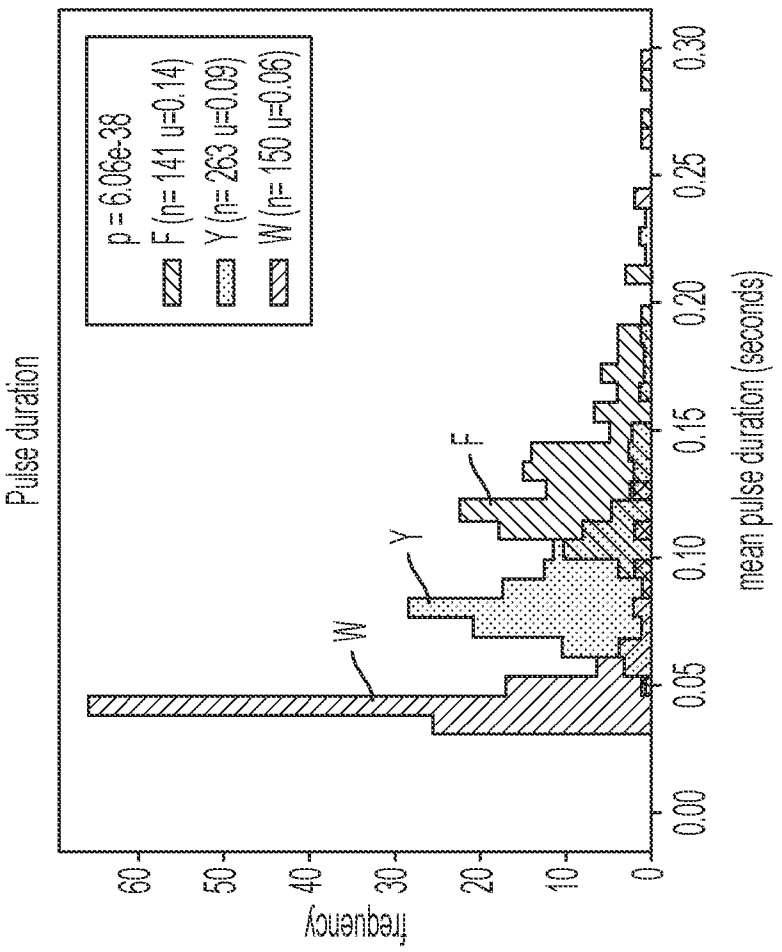


FIG. 19C

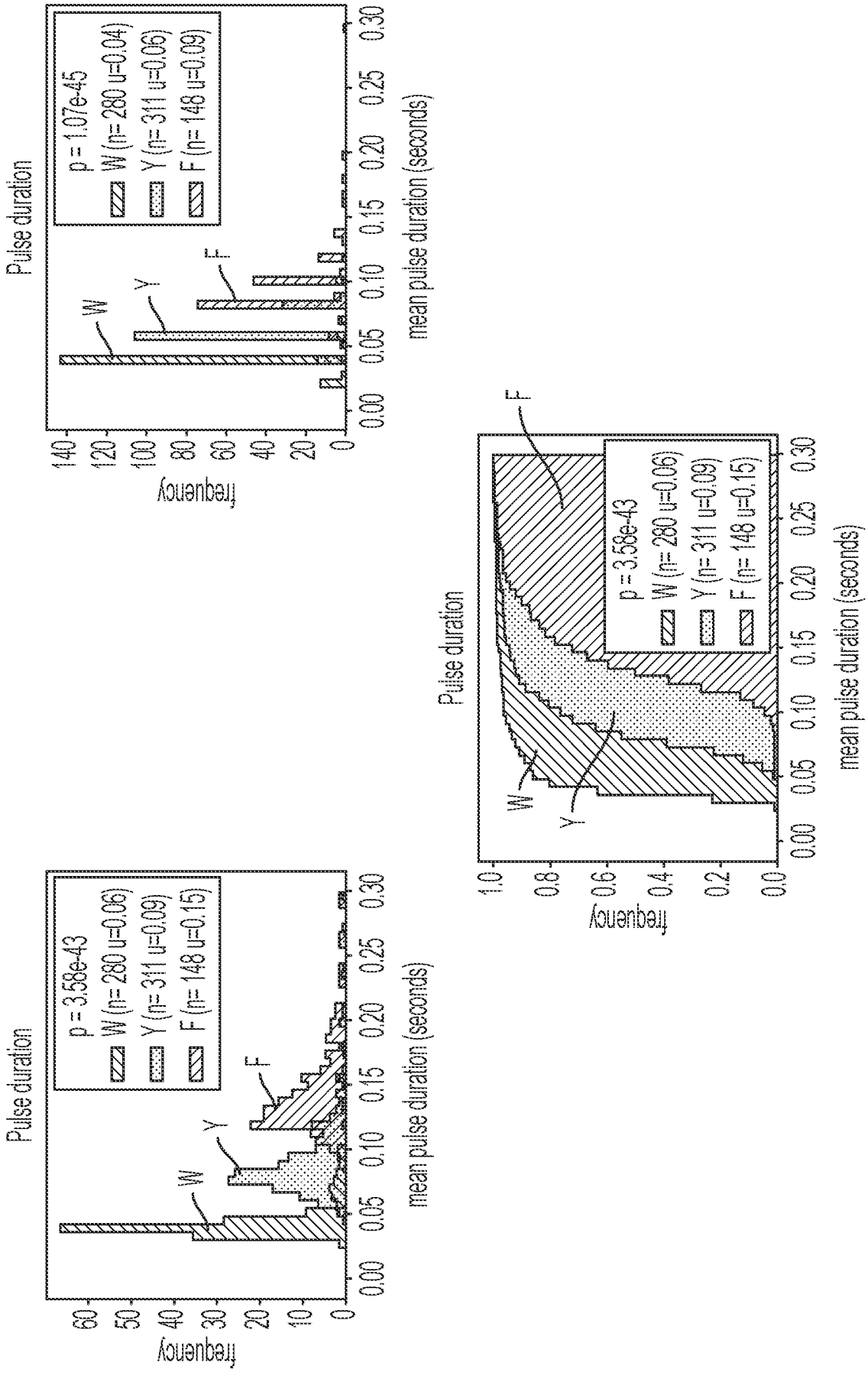

Pulse duration
p = 1.07e-45
W (n= 280 u=0.04)
Y (n= 311 u=0.06)
F (n= 148 u=0.09)
W
Y
F
Frequency
mean pulse duration (seconds)
Pulse duration
p = 3.58e-43
W (n= 280 u=0.06)
Y (n= 311 u=0.09)
F (n= 148 u=0.15)
W
Y
F
Frequency
mean pulse duration (seconds)
Pulse duration
p = 3.58e-43
W (n= 280 u=0.06)
Y (n= 311 u=0.09)
F (n= 148 u=0.15)
W
Y
F
Frequency
mean pulse duration (seconds)

FIG. 19E

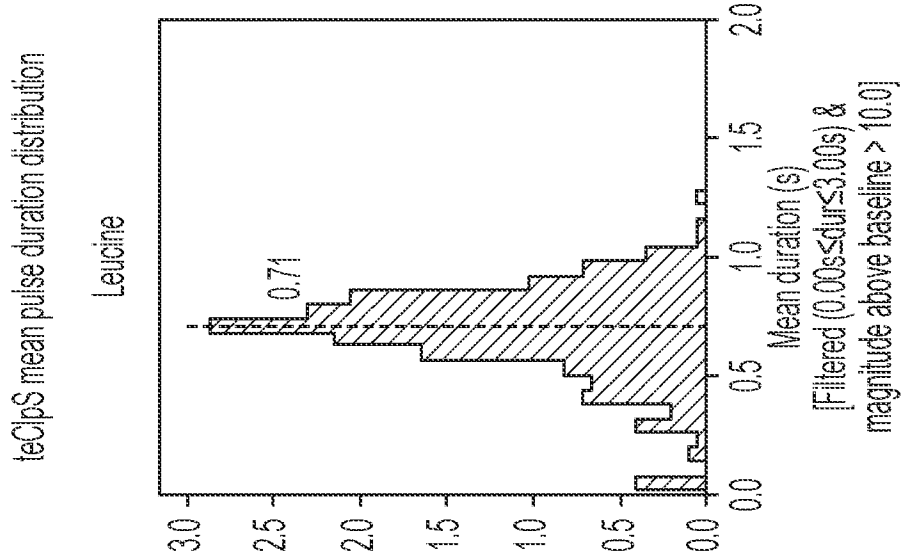

teClpS mean pulse duration distribution
Leucine
0.71
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
Mean duration (s)
[Filtered (0.00s≤dur≤3.00s) &
magnitude above baseline > 10.0]

FIG. 19F

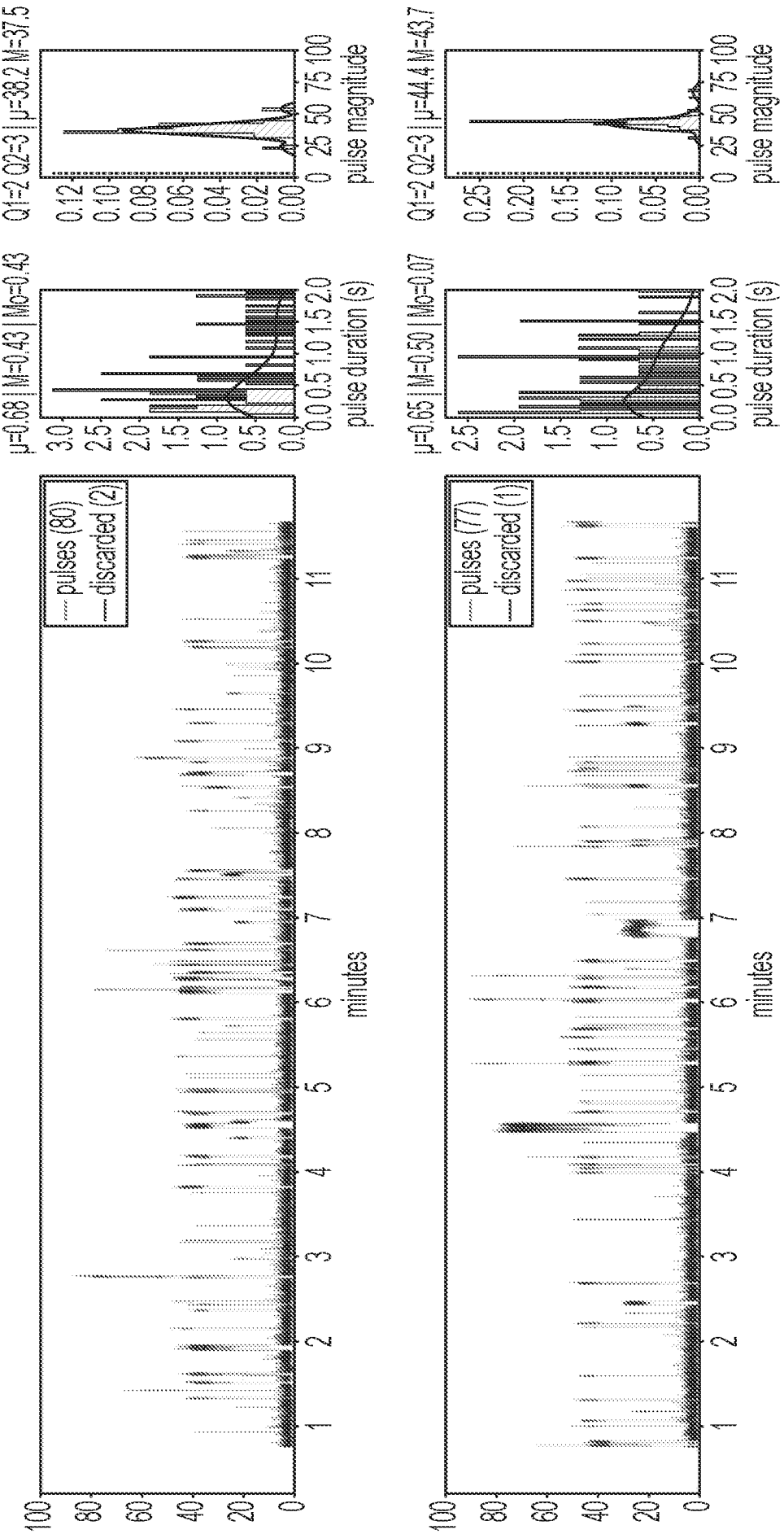

pulses (80)
discarded (2)
minutes
μ=0.68 | M=0.43 | Mo=0.43
pulse duration (s)
Q1=2 Q2=3 | μ=38.2 M=37.5
pulse magnitude
pulses (77)
discarded (1)
minutes
μ=0.65 | M=0.50 | Mo=0.07
pulse duration (s)
Q1=2 Q2=3 | μ=44.4 M=43.7
pulse magnitude

FIG. 19G

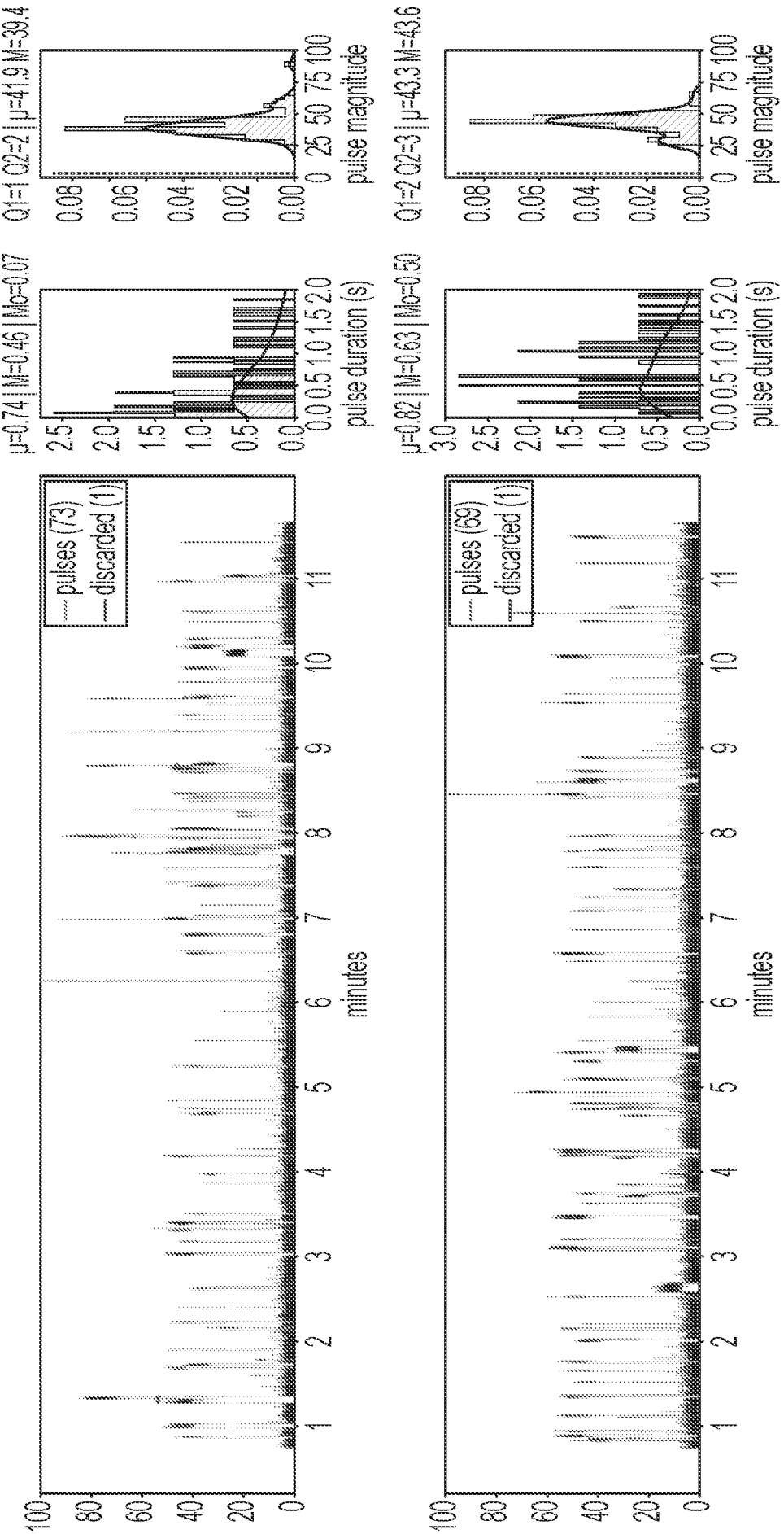

Q1=1 Q2=2 | μ=41.9 M=39.4
pulse magnitude
μ=0.74 | M=0.46 | Mo=0.07
pulse duration (s)
pulses (73)
discarded (11)
minutes
Q1=2 Q2=3 | μ=43.3 M=43.6
pulse magnitude
μ=0.82 | M=0.63 | Mo=0.50
pulse duration (s)
pulses (69)
discarded (11)
minutes

FIG. 19H

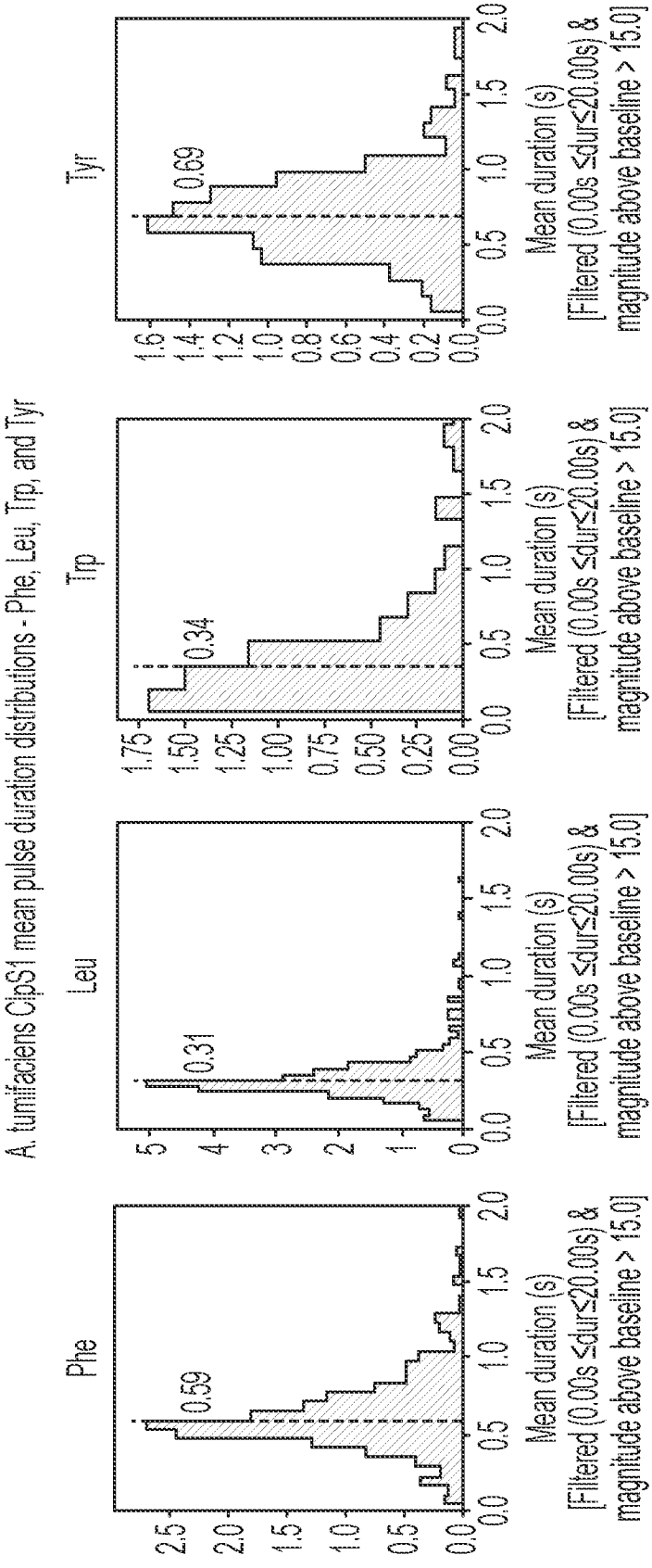

A. tumifaciens OjpS1 mean pulse duration distributions - Phe, Leu, Trp, and Tyr
Phe
0.59
Leu
0.31
Trp
0.34
Tyr
0.69
Mean duration (s)
[Filtered (0.00s ≤dur≤20.00s) & magnitude above baseline > 15.0]

FIG. 19I

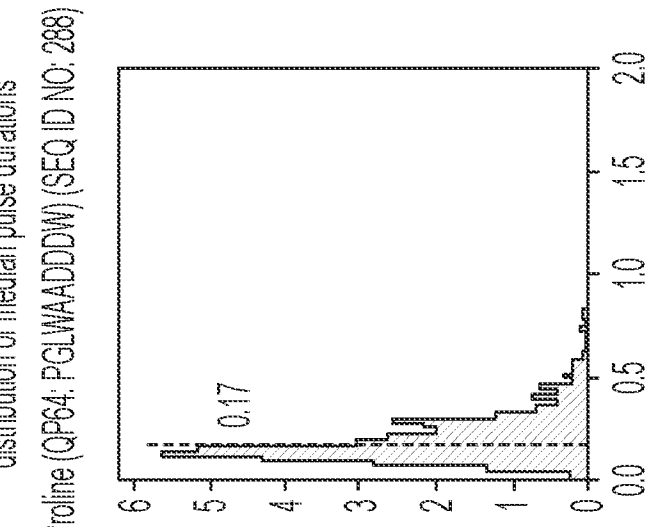


FIG. 19K

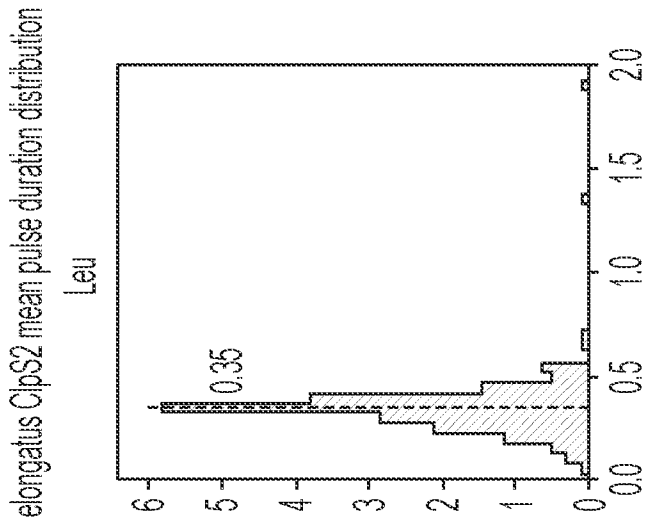


FIG. 19J

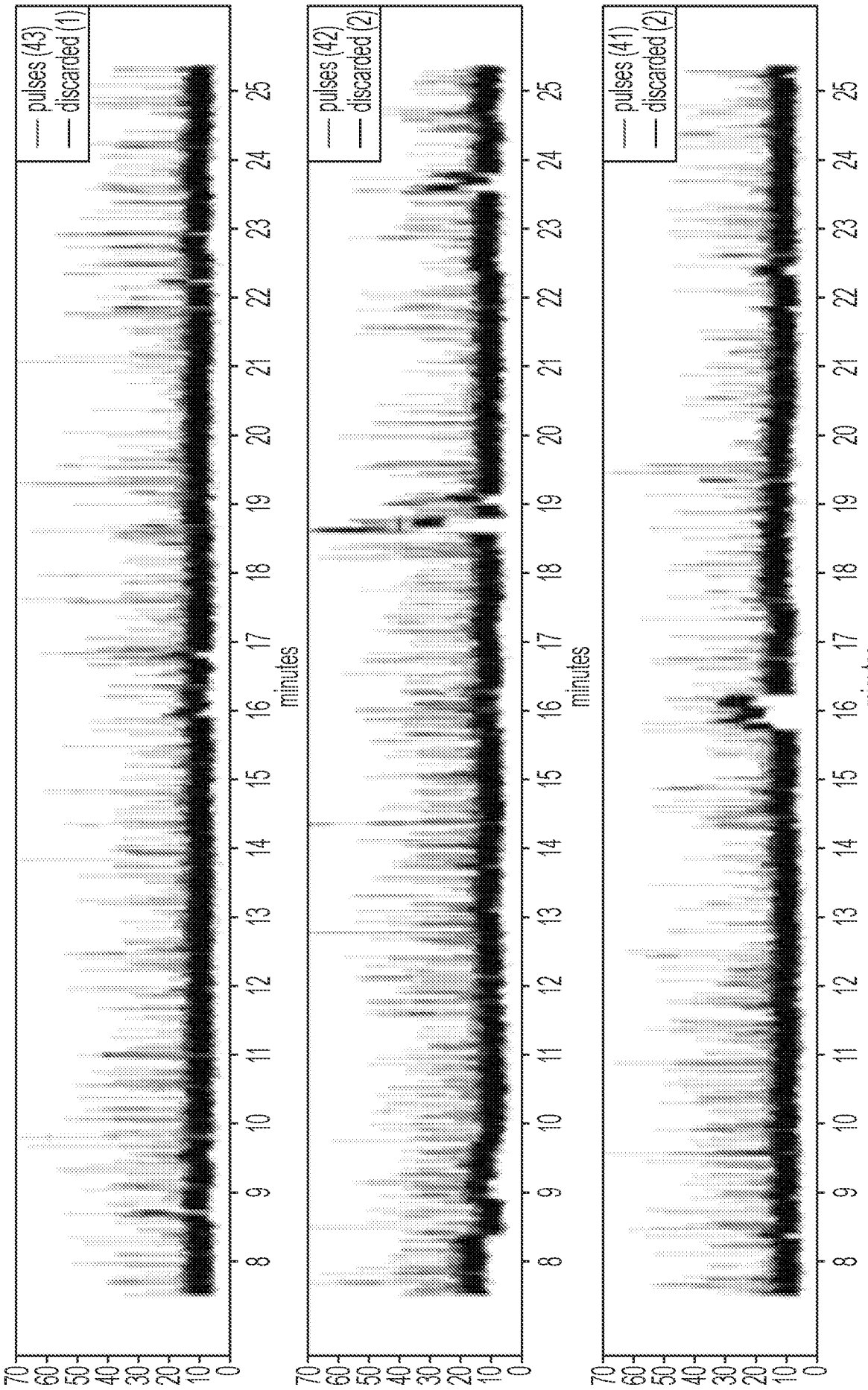

pulses (43)
discarded (1)
pulses (42)
discarded (2)
pulses (41)
discarded (2)
minutes
minutes
minutes
0
10
20
30
40
50
60
70
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

FIG. 19L

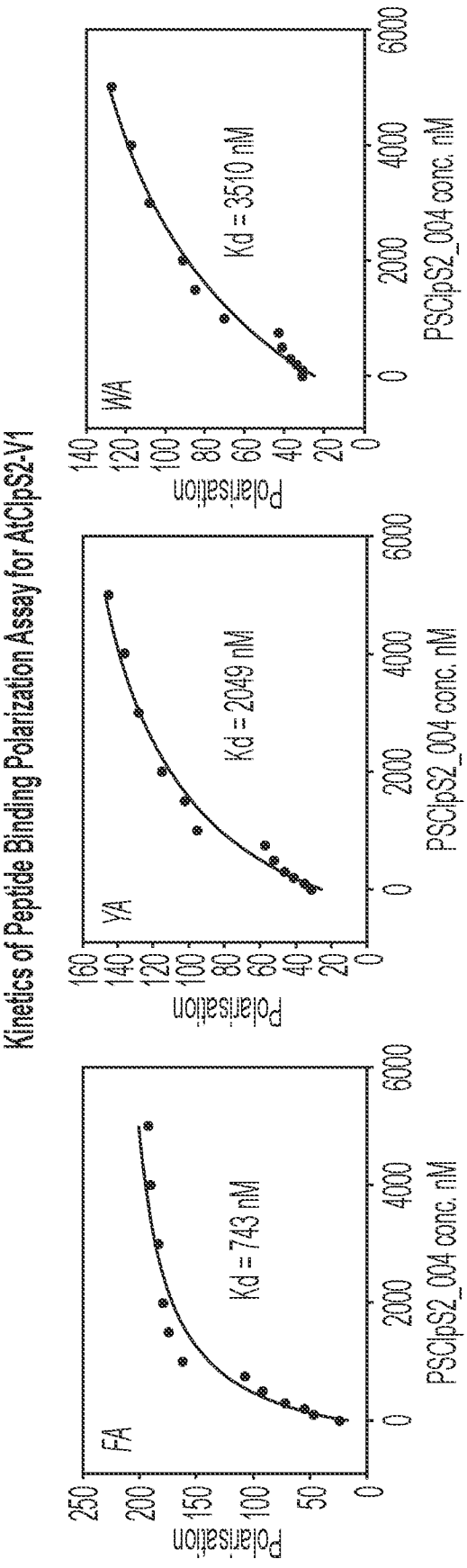

Kinetics of Peptide Binding Polarization Assay for AtClpS2-V1
FA
Kd = 743 nM
YA
Kd = 2049 nM
WA
Kd = 3510 nM
Polarisation
PSClpS2_004 conc. nM

FIG. 19M

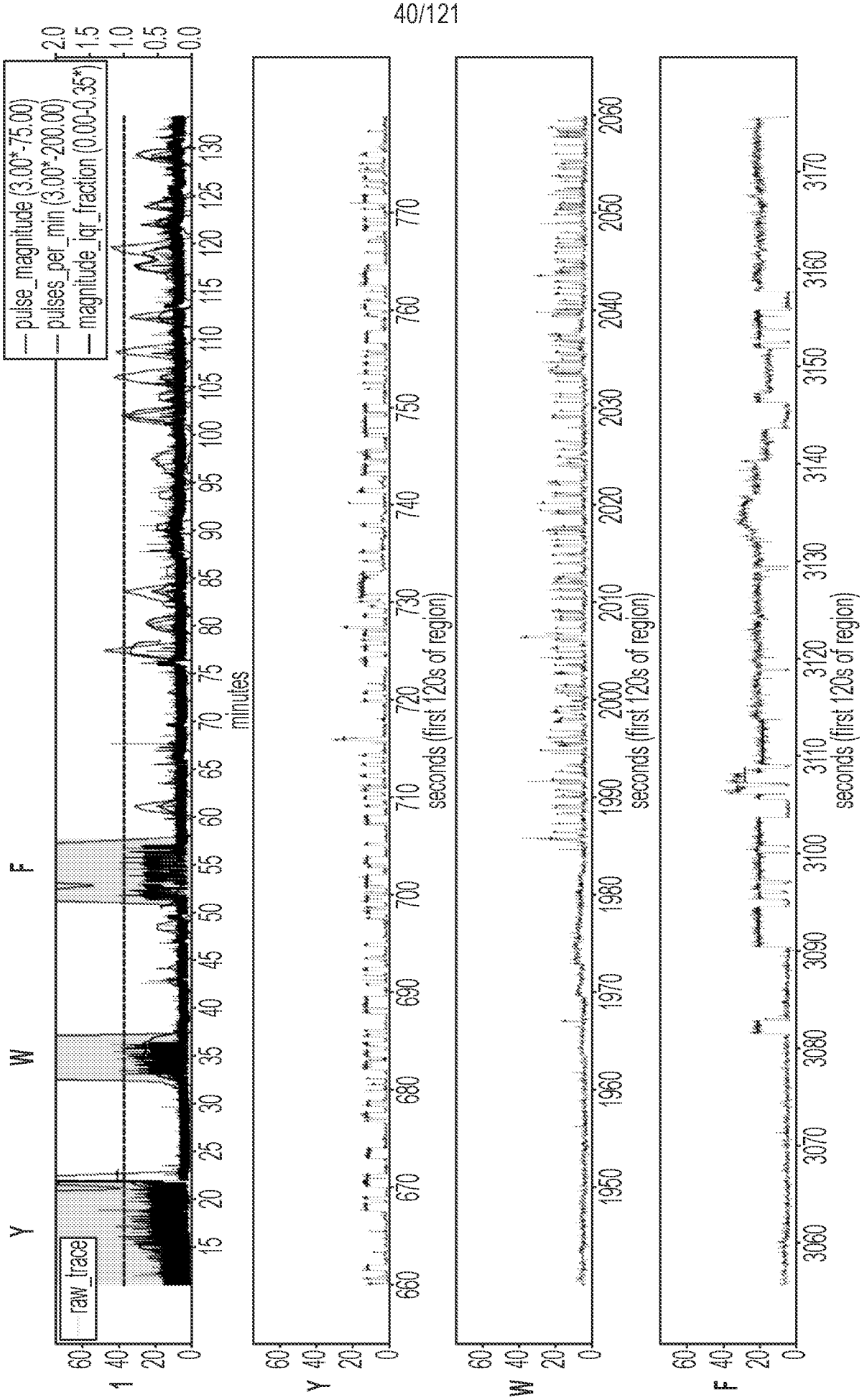

raw_trace
pulse_magnitude (3.00*-75.00)
pulses_per_min (3.00*-200.00)
magnitude_iqr_fraction (0.00-0.35*)
minutes
seconds (first 120s of region)
seconds (first 120s of region)
seconds (first 120s of region)

FIG. 20A

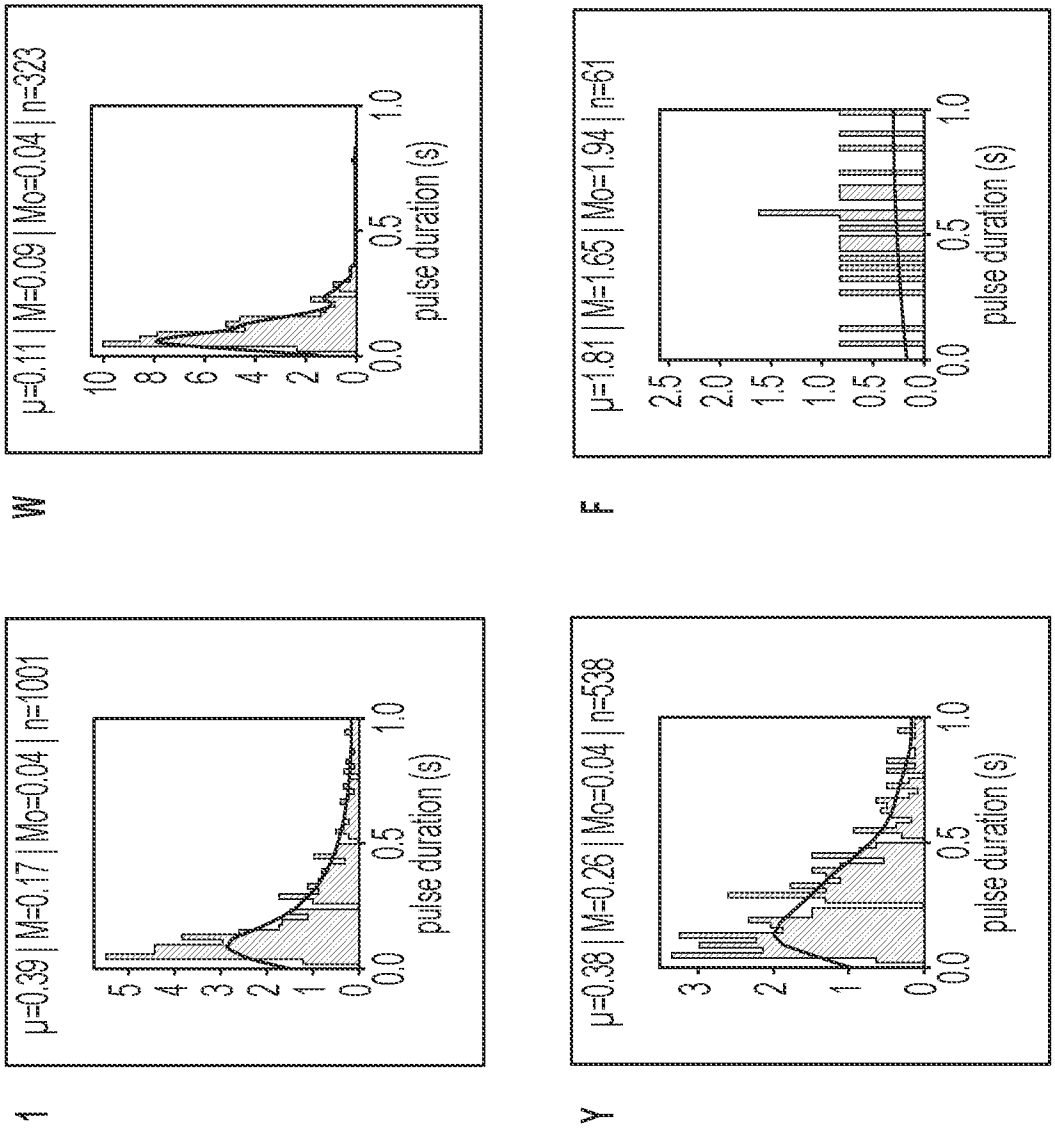

1
μ=0.39 | M=0.17 | Mo=0.04 | n=1001
pulse duration (s)
W
μ=0.11 | M=0.09 | Mo=0.04 | n=323
pulse duration (s)
Y
μ=0.38 | M=0.26 | Mo=0.04 | n=538
pulse duration (s)
F
μ=1.81 | M=1.65 | Mo=1.94 | n=61
pulse duration (s)

FIG. 20B

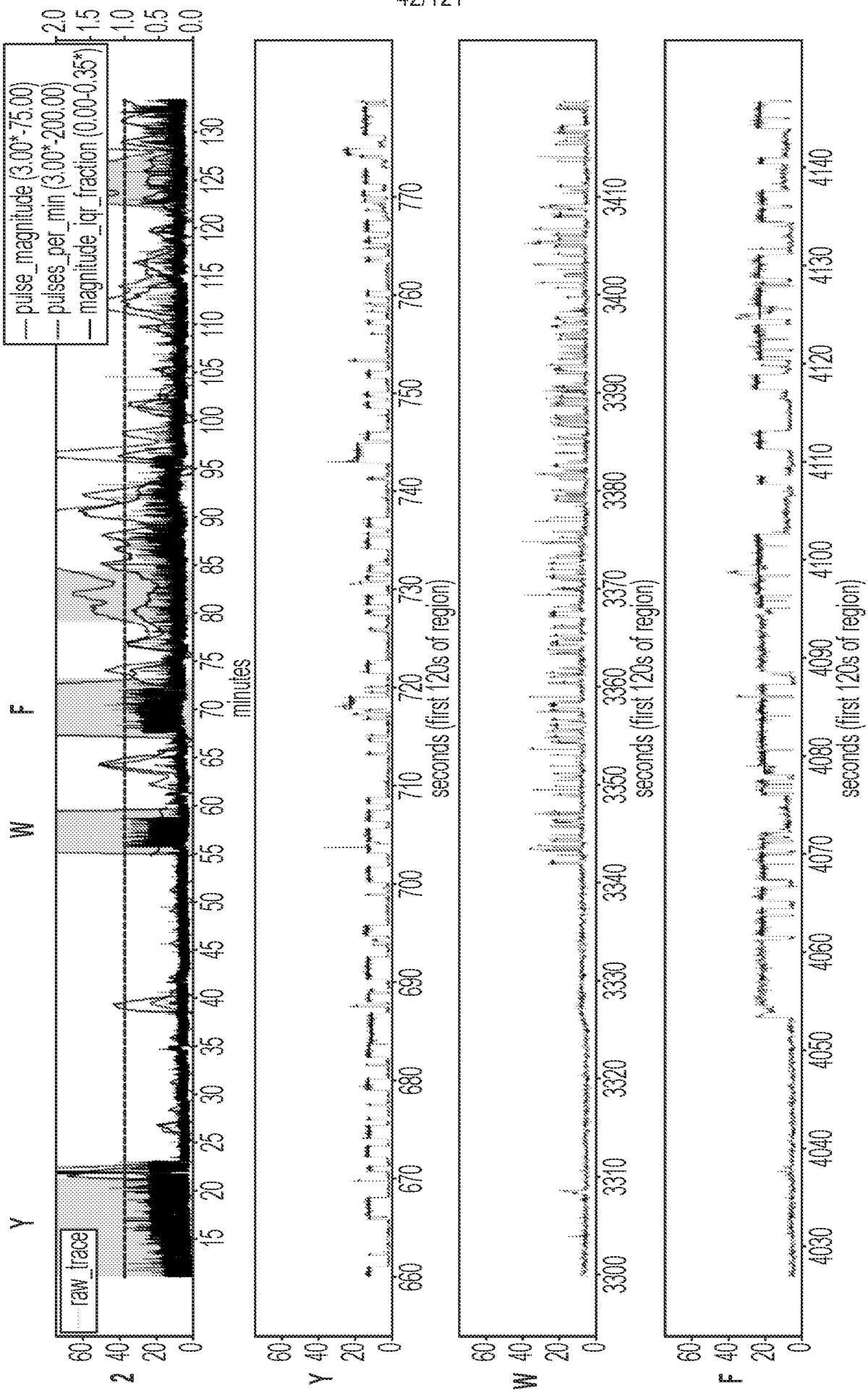

raw_trace
pulse_magnitude (3.00*-75.00)
pulses_per_min (3.00*-200.00)
magnitude_iqr_fraction (0.00-0.35*)
minutes
seconds (first 120s of region)
seconds (first 120s of region)
seconds (first 120s of region)

FIG. 20C

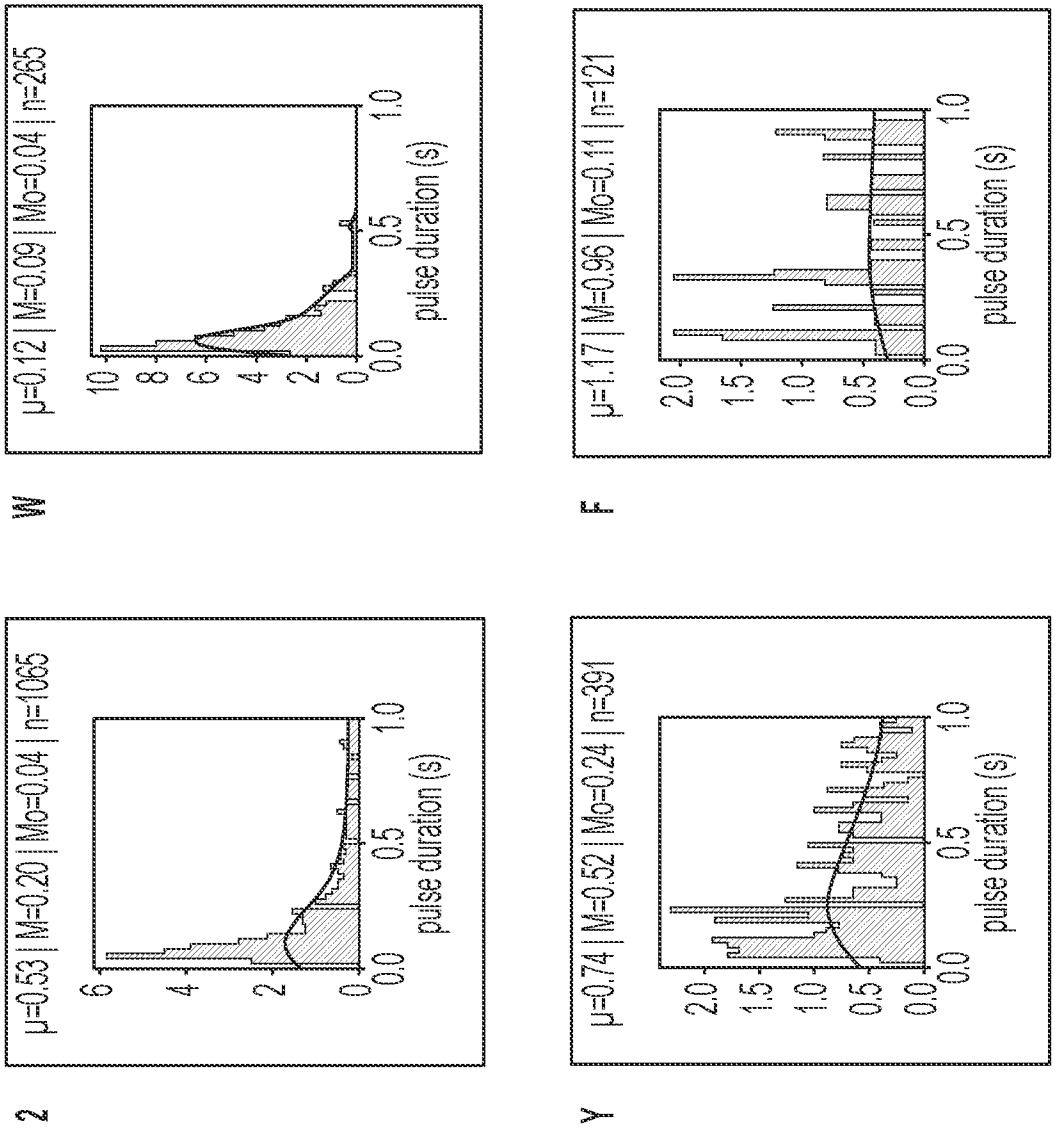

W
μ=0.12 | M=0.09 | Mo=0.04 | n=265
F
μ=1.17 | M=0.96 | Mo=0.11 | n=121
Z
μ=0.53 | M=0.20 | Mo=0.04 | n=1065
Y
μ=0.74 | M=0.52 | Mo=0.24 | n=391
pulse duration (s)

FIG. 20D

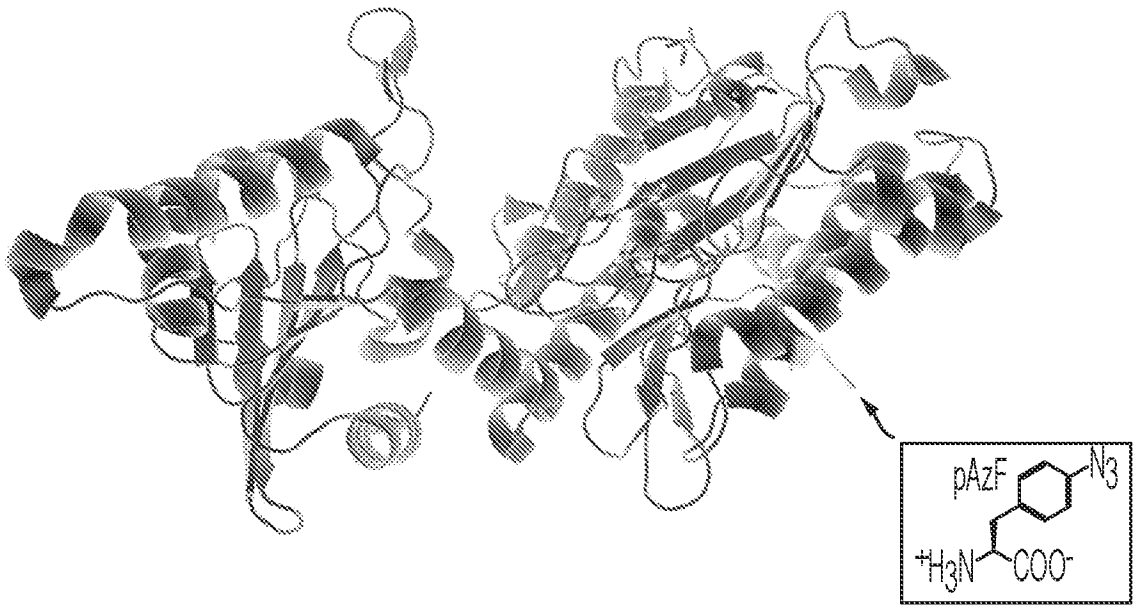


FIG. 21A

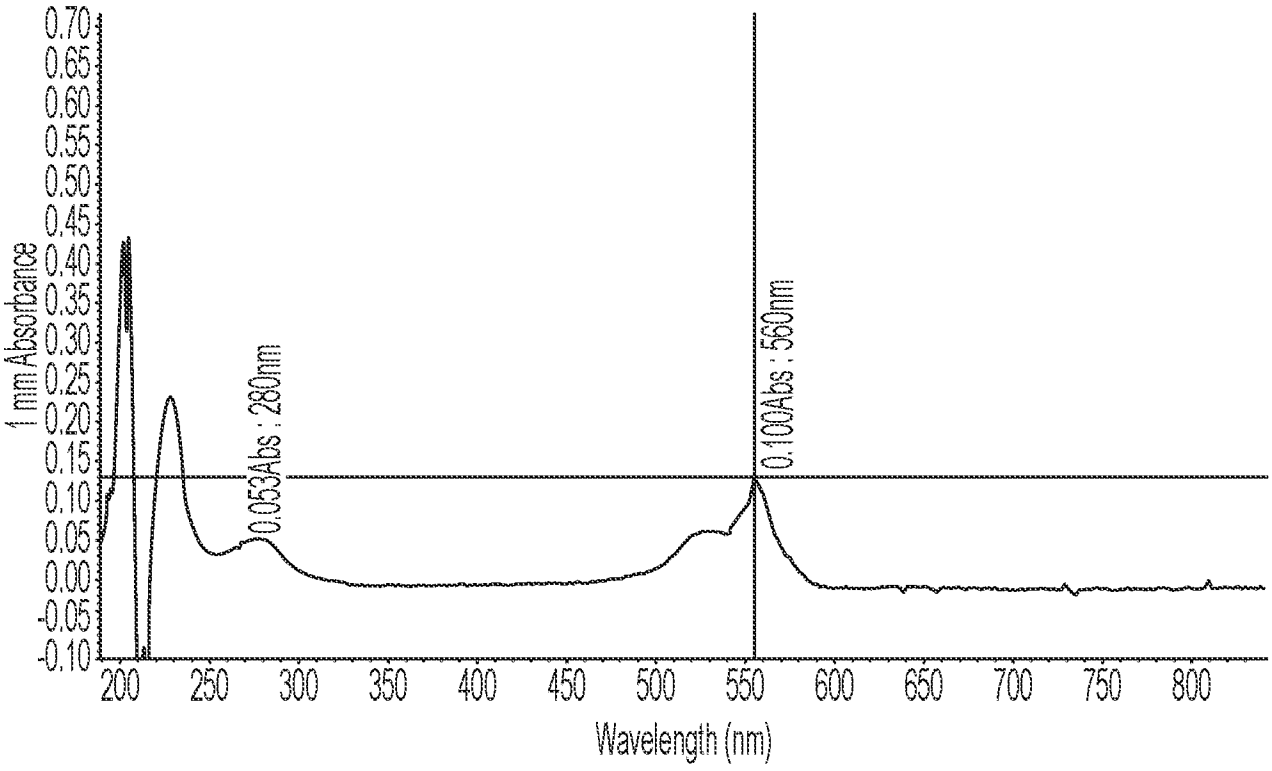


FIG. 21B

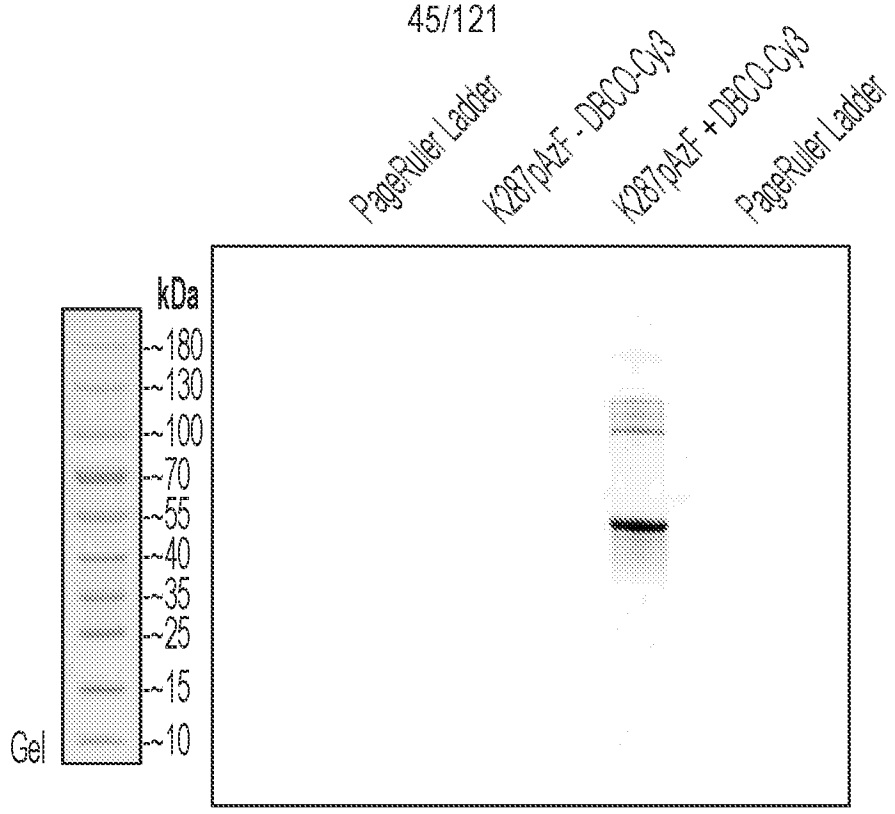


FIG. 21C

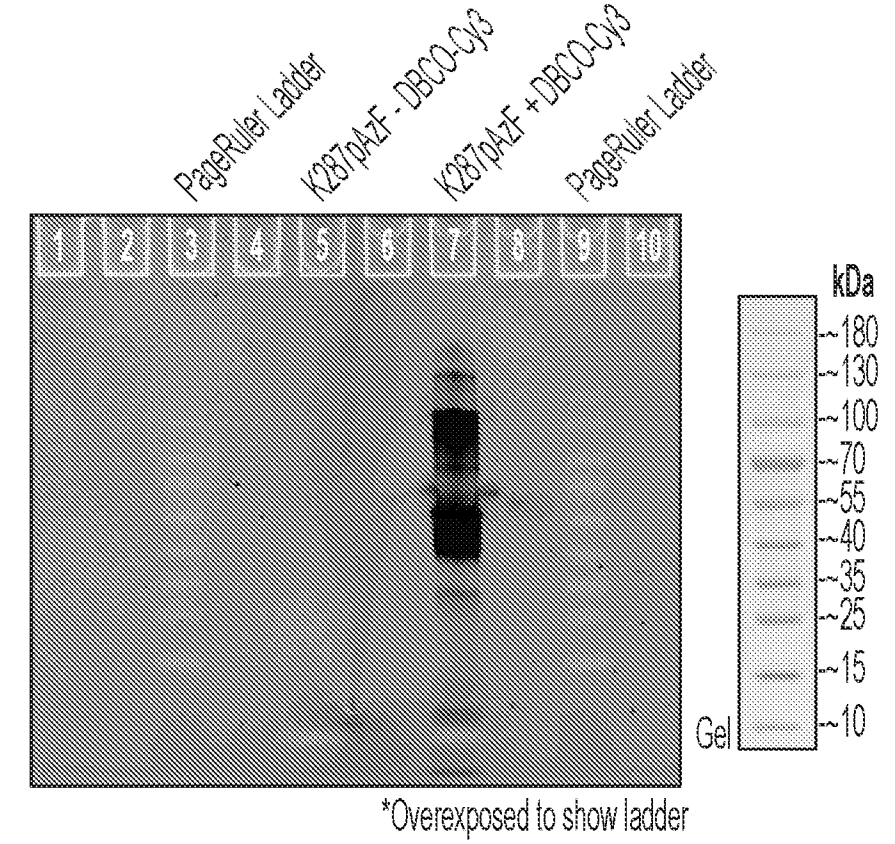


FIG. 21D

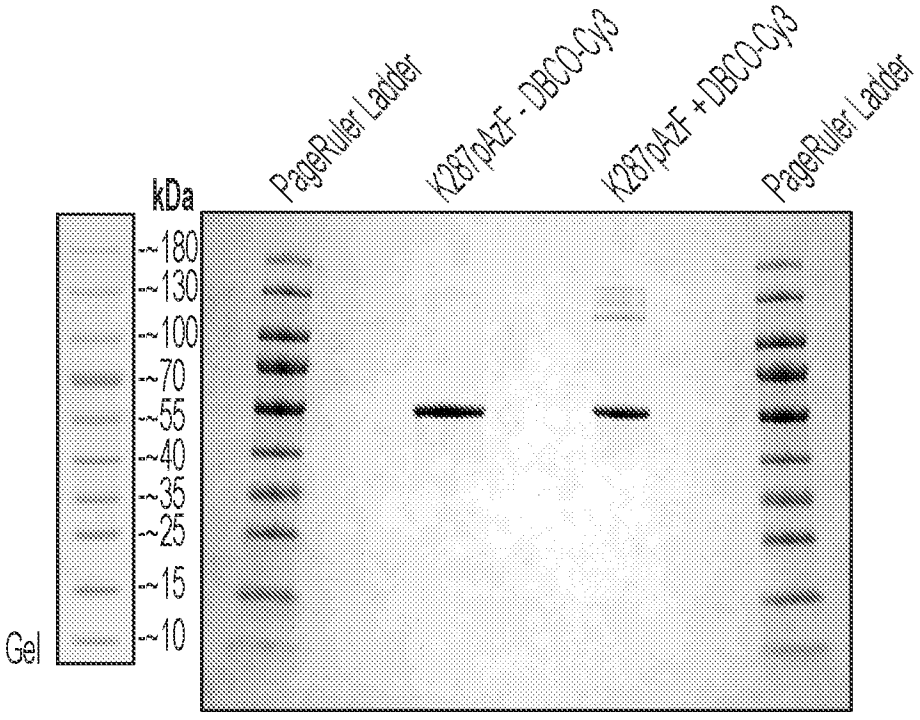


FIG. 21E

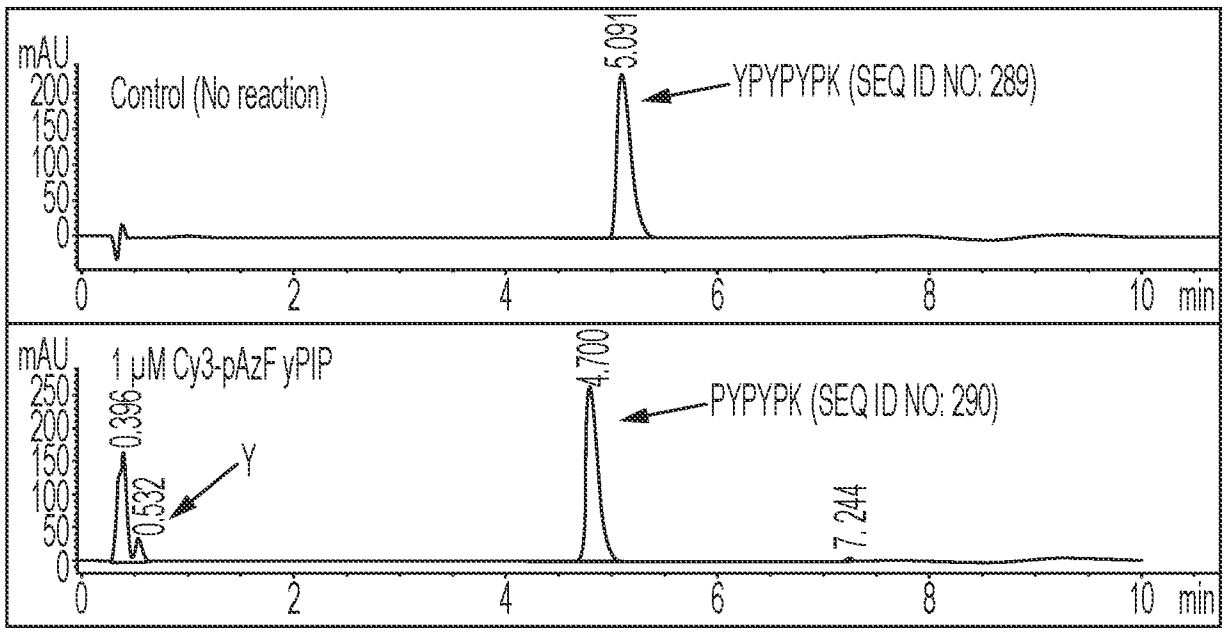


FIG. 21F

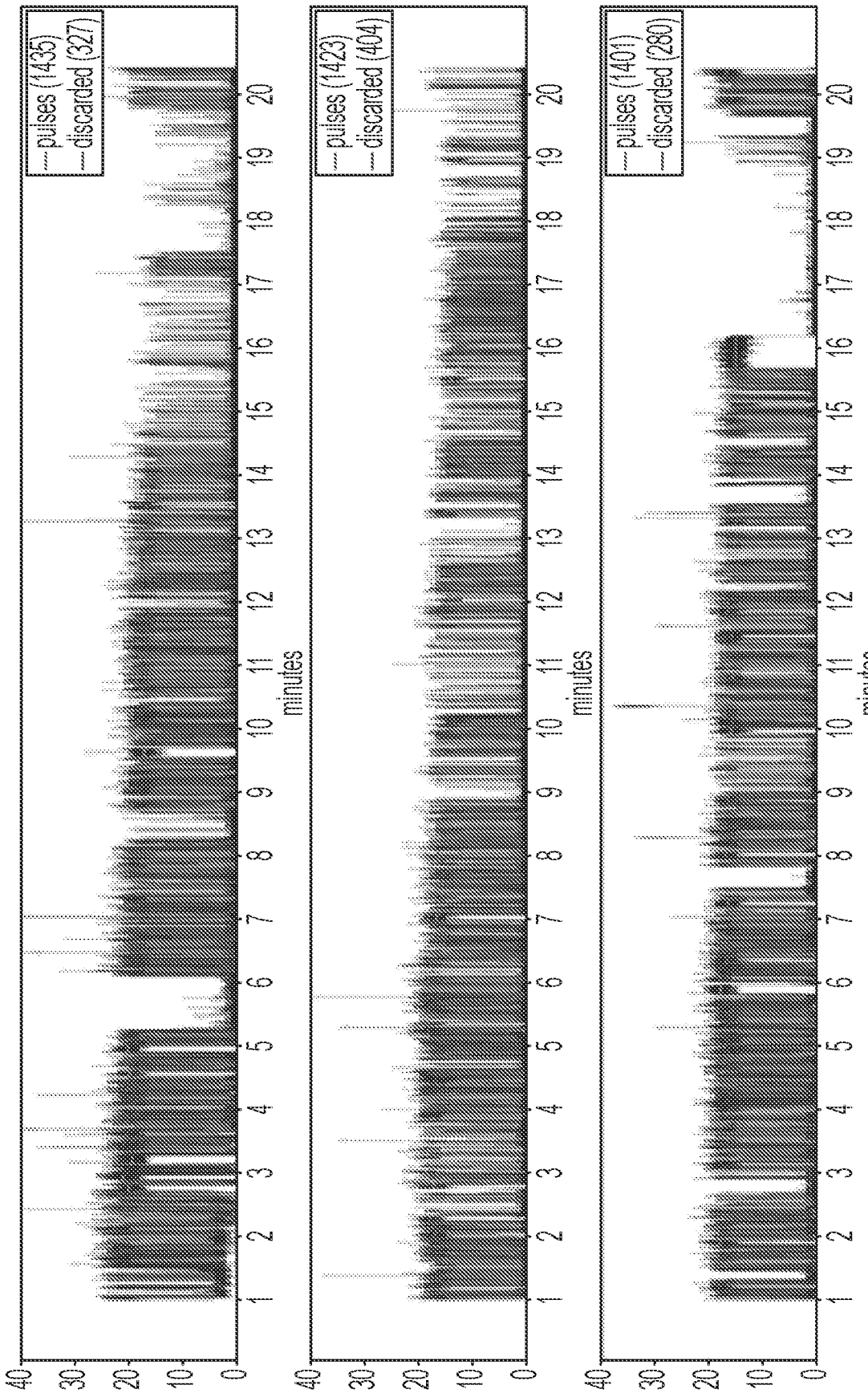

pulses (1435)
discarded (327)
pulses (1423)
discarded (404)
pulses (1401)
discarded (280)
minutes
minutes
minutes

FIG. 22A

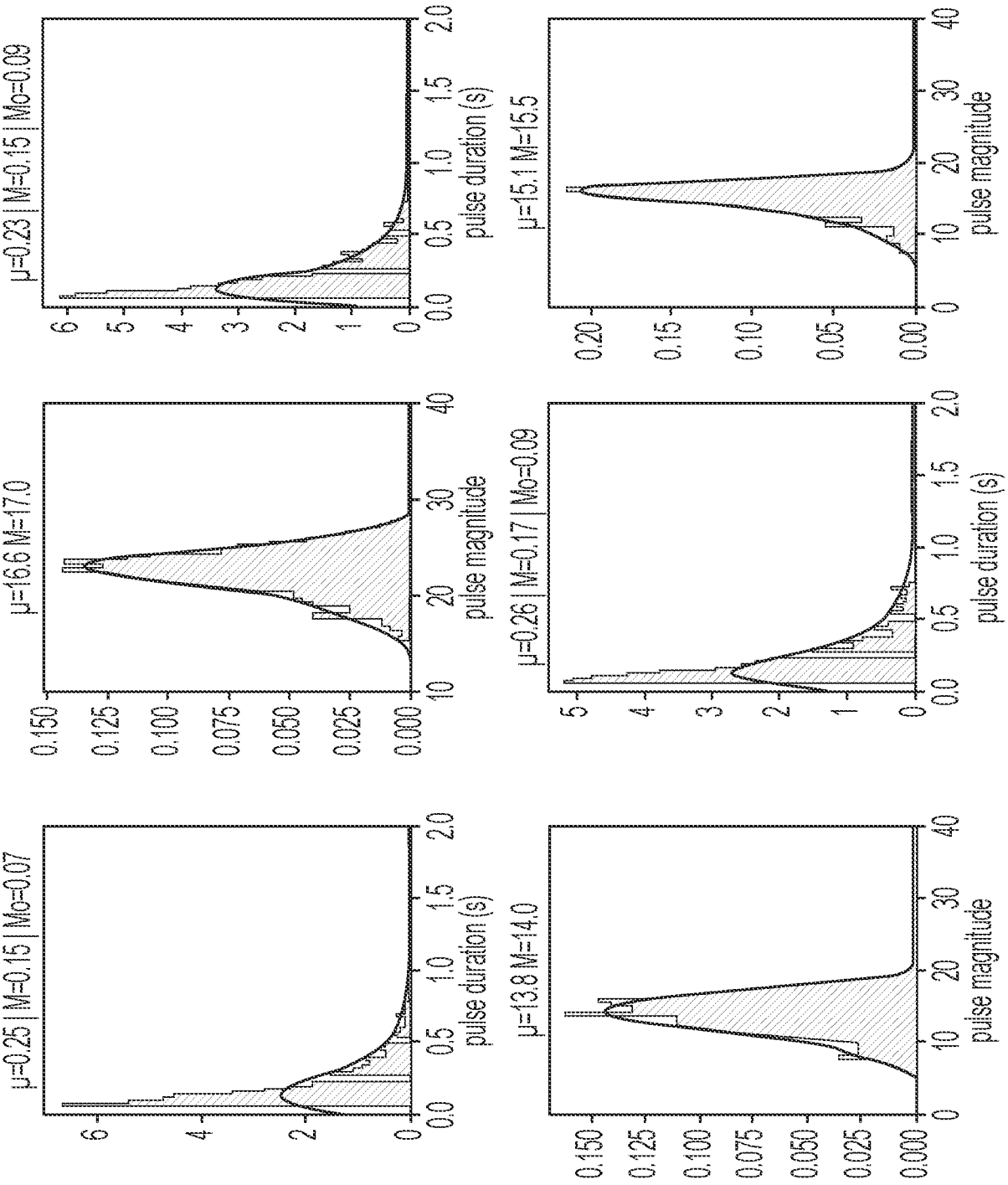

μ=0.25 | M=0.15 | Mo=0.07
μ=16.6 M=17.0
μ=0.23 | M=0.15 | Mo=0.09
μ=13.8 M=14.0
μ=0.26 | M=0.17 | Mo=0.09
μ=15.1 M=15.5
pulse duration (s)
pulse magnitude

FIG. 22B

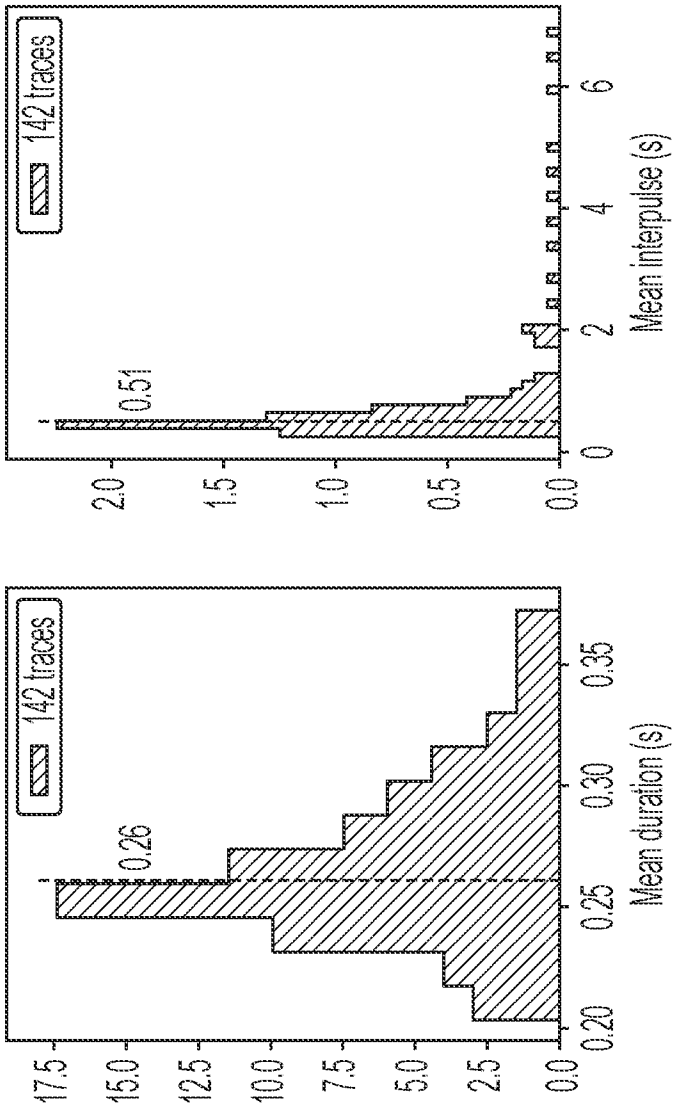

142 traces
0.51
2.0
1.5
1.0
0.5
0.0
0
2
4
6
Mean interpulse (s)
142 traces
0.26
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.20
0.25
0.30
0.35
Mean duration (s)

FIG. 22C

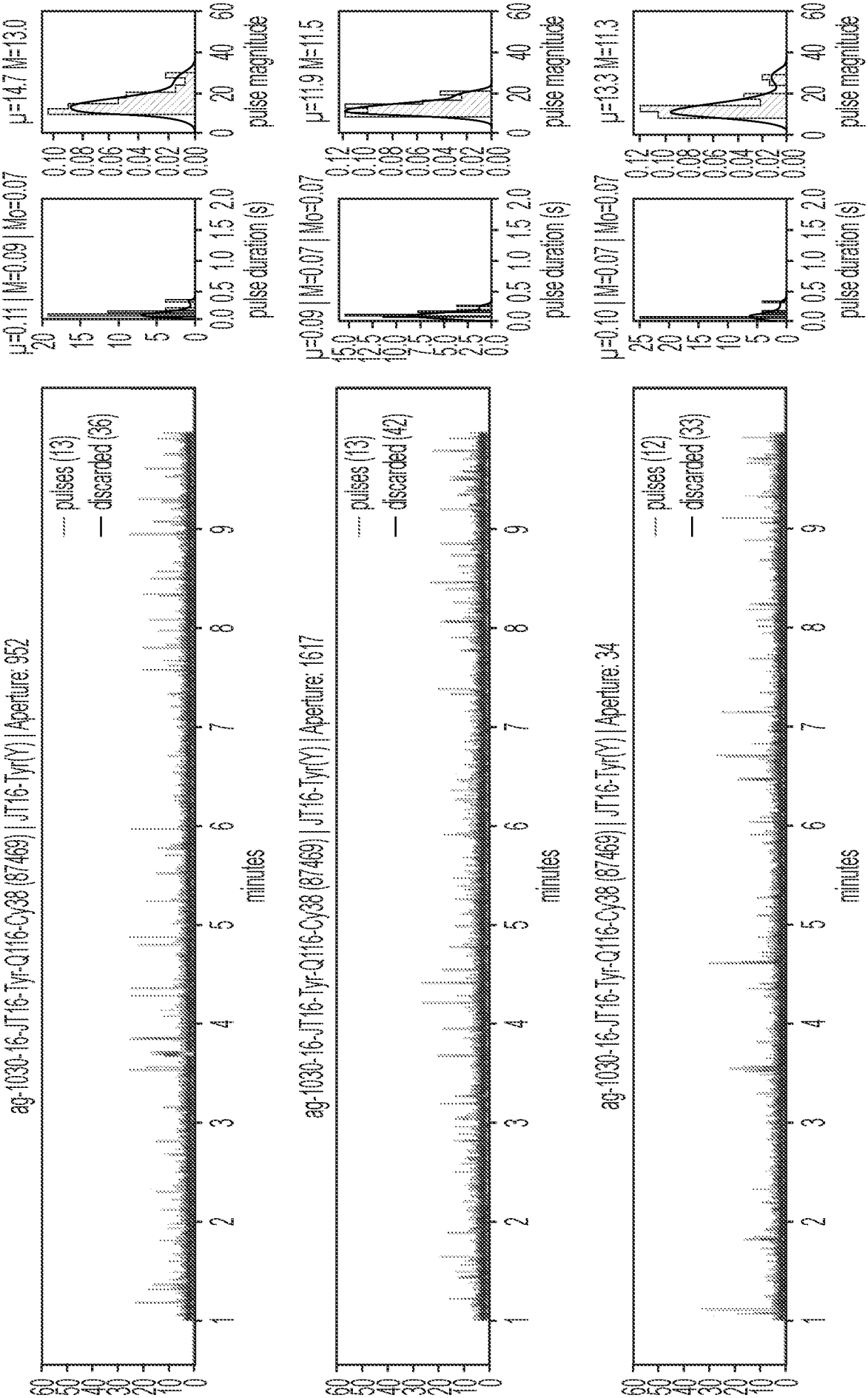

ag-1030-16-JT16-Tyr-Q116-Cy38 (87469) | JT16-Tyr(Y) | Aperture: 952
pulses (13)
discarded (36)
minutes
μ=0.11 | M=0.09 | Mo=0.07
pulse duration (s)
μ=14.7 M=13.0
pulse magnitude
ag-1030-16-JT16-Tyr-Q116-Cy38 (87469) | JT16-Tyr(Y) | Aperture: 1617
pulses (13)
discarded (42)
minutes
μ=0.09 | M=0.07 | Mo=0.07
pulse duration (s)
μ=11.9 M=11.5
pulse magnitude
ag-1030-16-JT16-Tyr-Q116-Cy38 (87469) | JT16-Tyr(Y) | Aperture: 34
pulses (12)
discarded (33)
minutes
μ=0.10 | M=0.07 | Mo=0.07
pulse duration (s)
μ=13.3 M=11.3
pulse magnitude

FIG. 22E

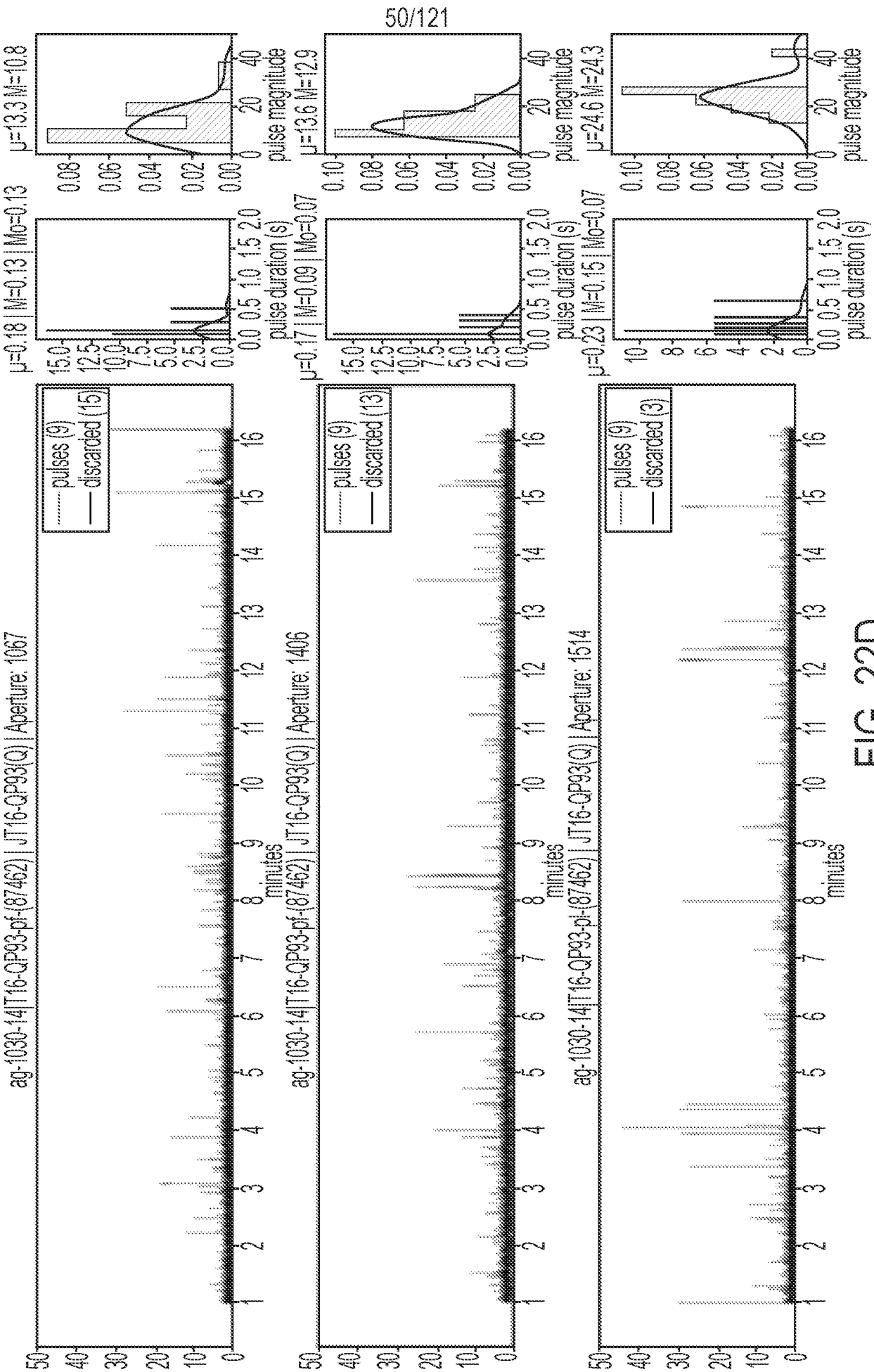

ag-1030-14|T16-QP93-pf-(87462)| JT16-QP93(Q)| Aperture: 1067
ag-1030-14|T16-QP93-pf-(87462)| JT16-QP93(Q)| Aperture: 1406
ag-1030-14|T16-QP93-pl-(87462)| JT16-QP93(Q)| Aperture: 1514
pulses (9)
discarded (15)
discarded (13)
discarded (3)
minutes
μ=0.18 | M=0.13 | Mo=0.13
μ=0.17 | M=0.09 | Mo=0.07
μ=0.23 | M=0.15 | Mo=0.07
pulse duration (s)
μ=13.3 M=10.8
μ=13.6 M=12.9
μ=24.6 M=24.3
pulse magnitude

FIG. 22D

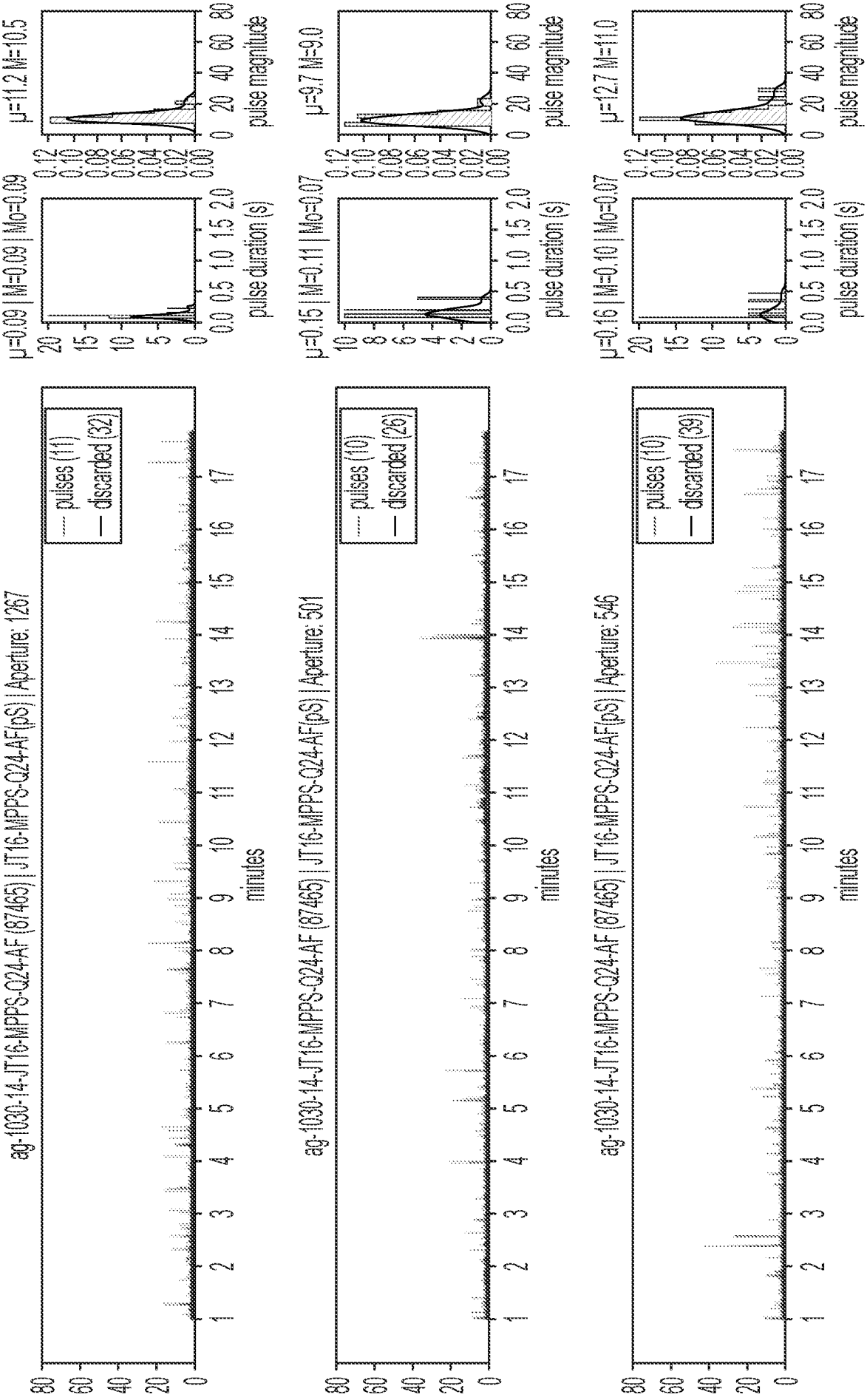

ag-1030-14-JT16-MPPS-Q24-AF (87465) | JT16-MPPS-Q24-AF(pS) | Aperture: 1267
pulses (11)
discarded (32)
minutes
μ=0.09 | M=0.09 | Mo=0.09
pulse duration (s)
μ=11.2 M=10.5
pulse magnitude
ag-1030-14-JT16-MPPS-Q24-AF (87465) | JT16-MPPS-Q24-AF(pS) | Aperture: 501
pulses (10)
discarded (26)
minutes
μ=0.15 | M=0.11 | Mo=0.07
pulse duration (s)
μ=9.7 M=9.0
pulse magnitude
ag-1030-14-JT16-MPPS-Q24-AF (87465) | JT16-MPPS-Q24-AF(pS) | Aperture: 546
pulses (10)
discarded (39)
minutes
μ=0.16 | M=0.10 | Mo=0.07
pulse duration (s)
μ=12.7 M=11.0
pulse magnitude

FIG. 22F

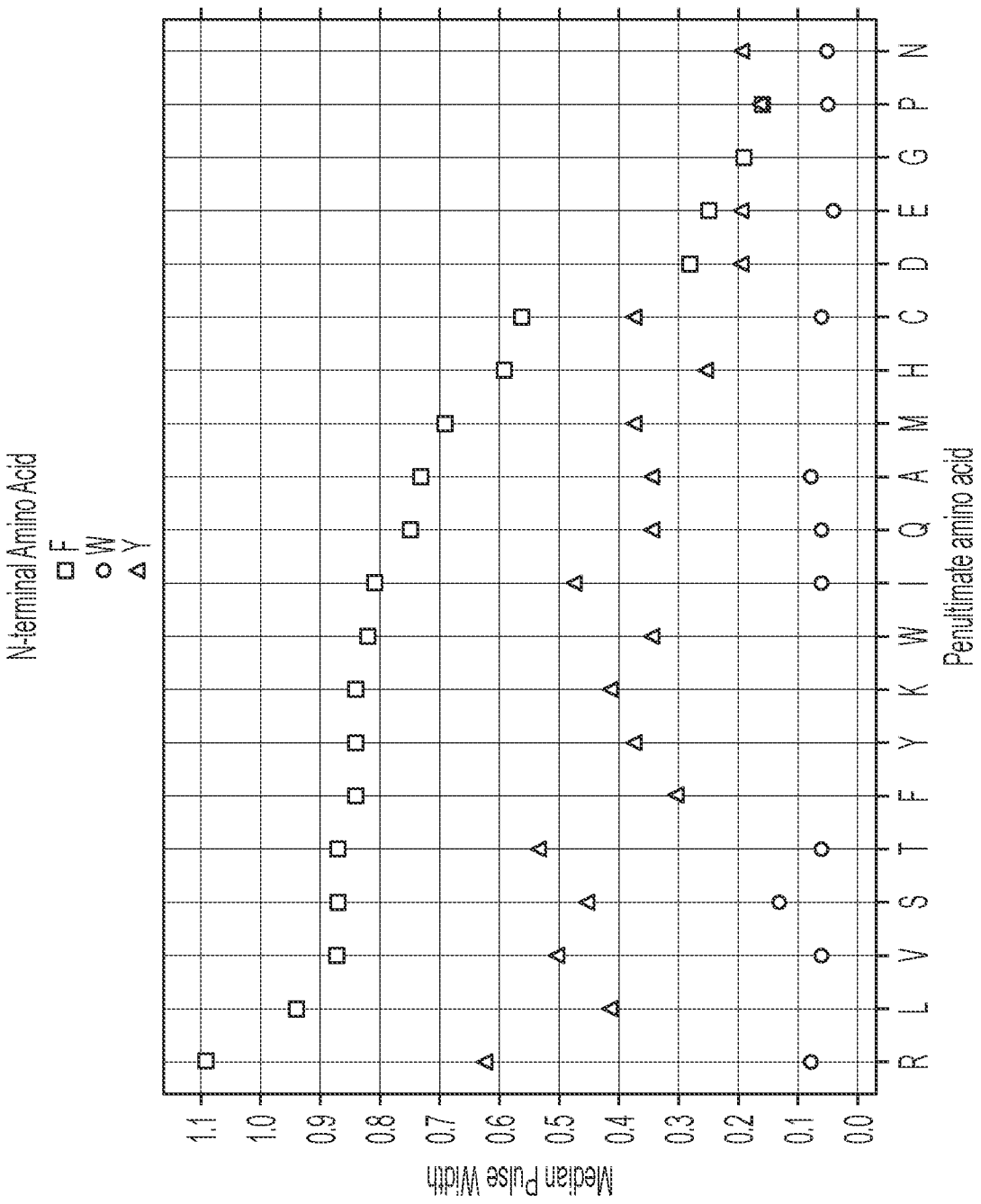

N-terminal Amino Acid
F
W
Y
Median Pulse Width
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
R L V S T F Y K W I Q A M H C D E G P N
Penultimate amino acid

FIG. 23

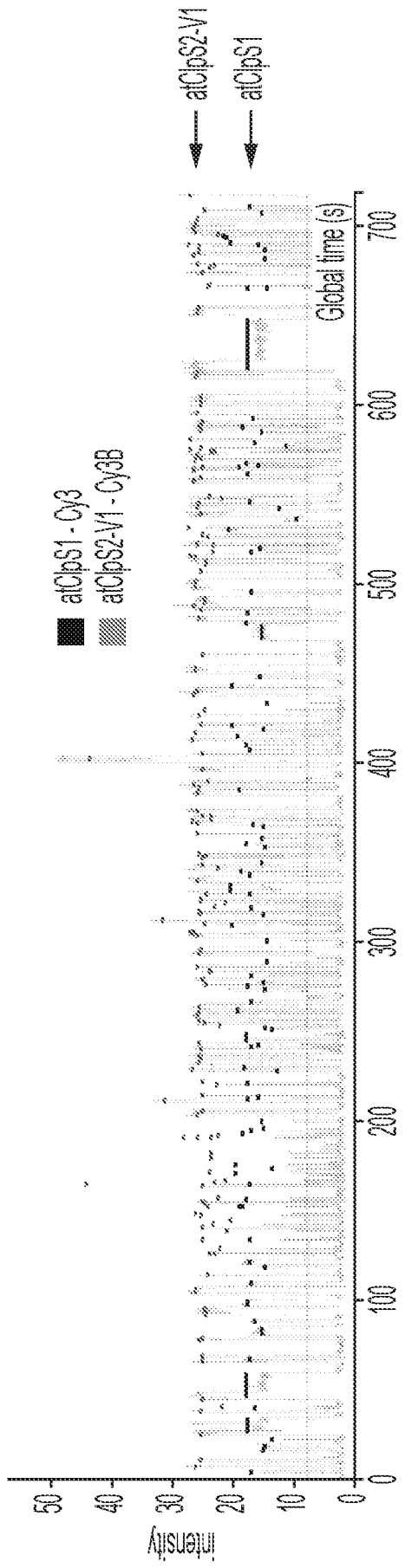

atClpS2-V1
atClpS1
atClpS1 - Cy3
atClpS2-V1 - Cy3B
Global time (s)
intensity
0
10
20
30
40
50
0
100
200
300
400
500
600
700

FIG. 24A

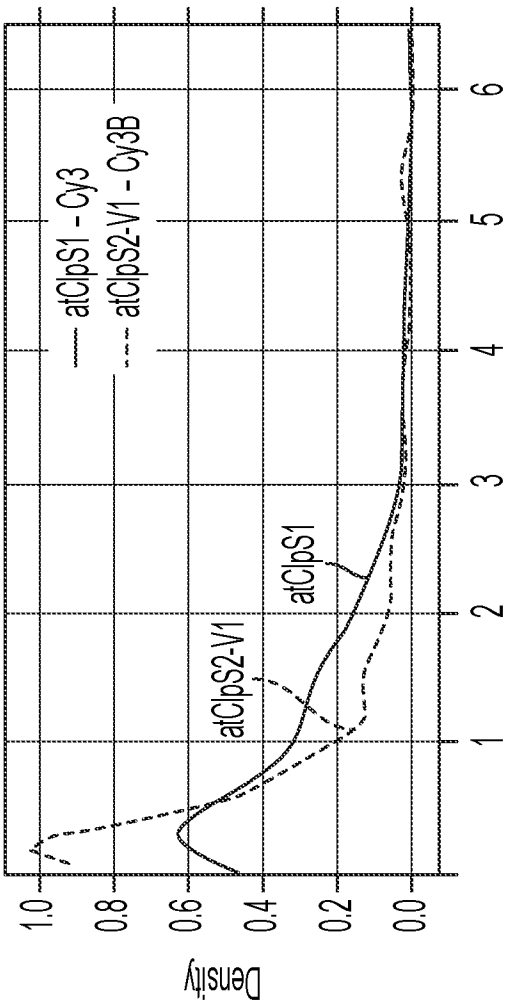


FIG. 24B

	N pulses	mean pulse width (s)	median pulse width (s)	pulse rate (pulses/min)
atClpS1	98	1.3	0.63	8.1
atClpS2-V1	169	1.0	0.35	14.1

FIG. 24C

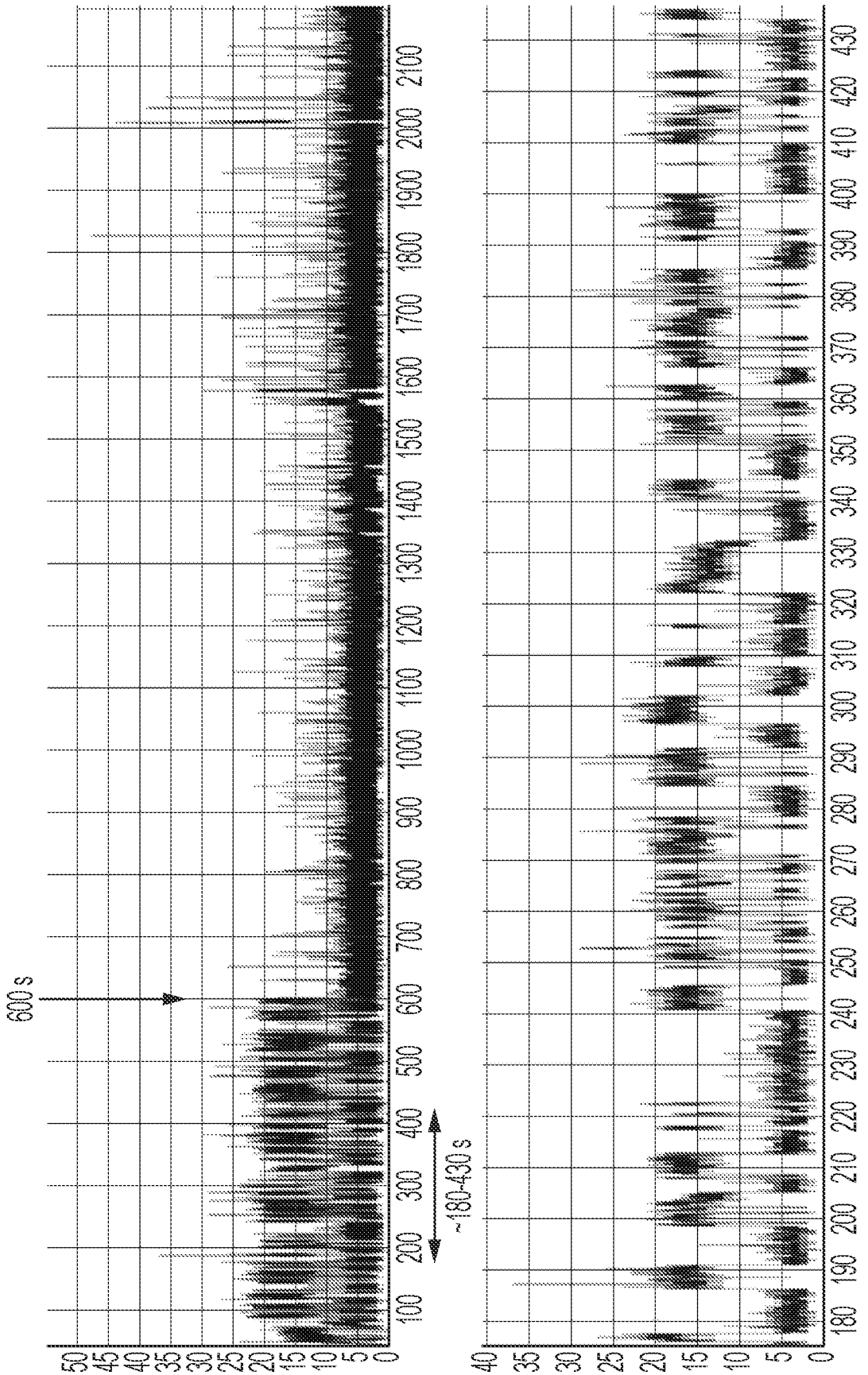


FIG. 25A

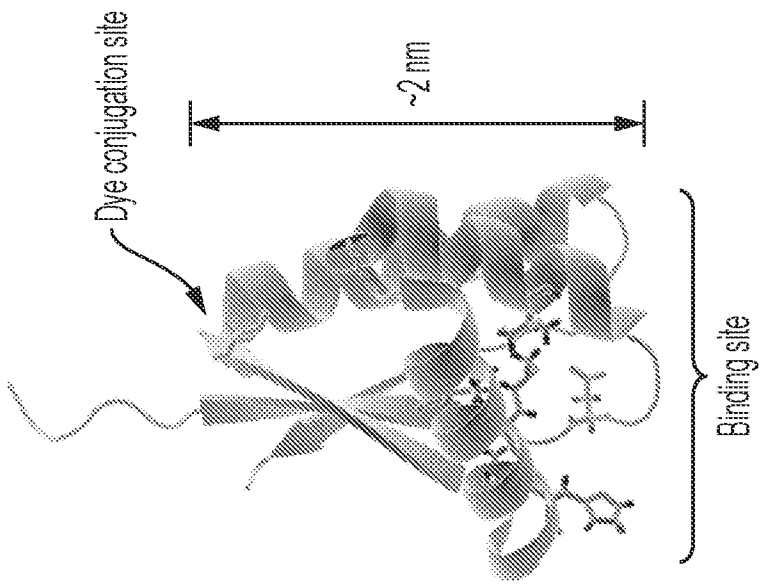


FIG. 25B

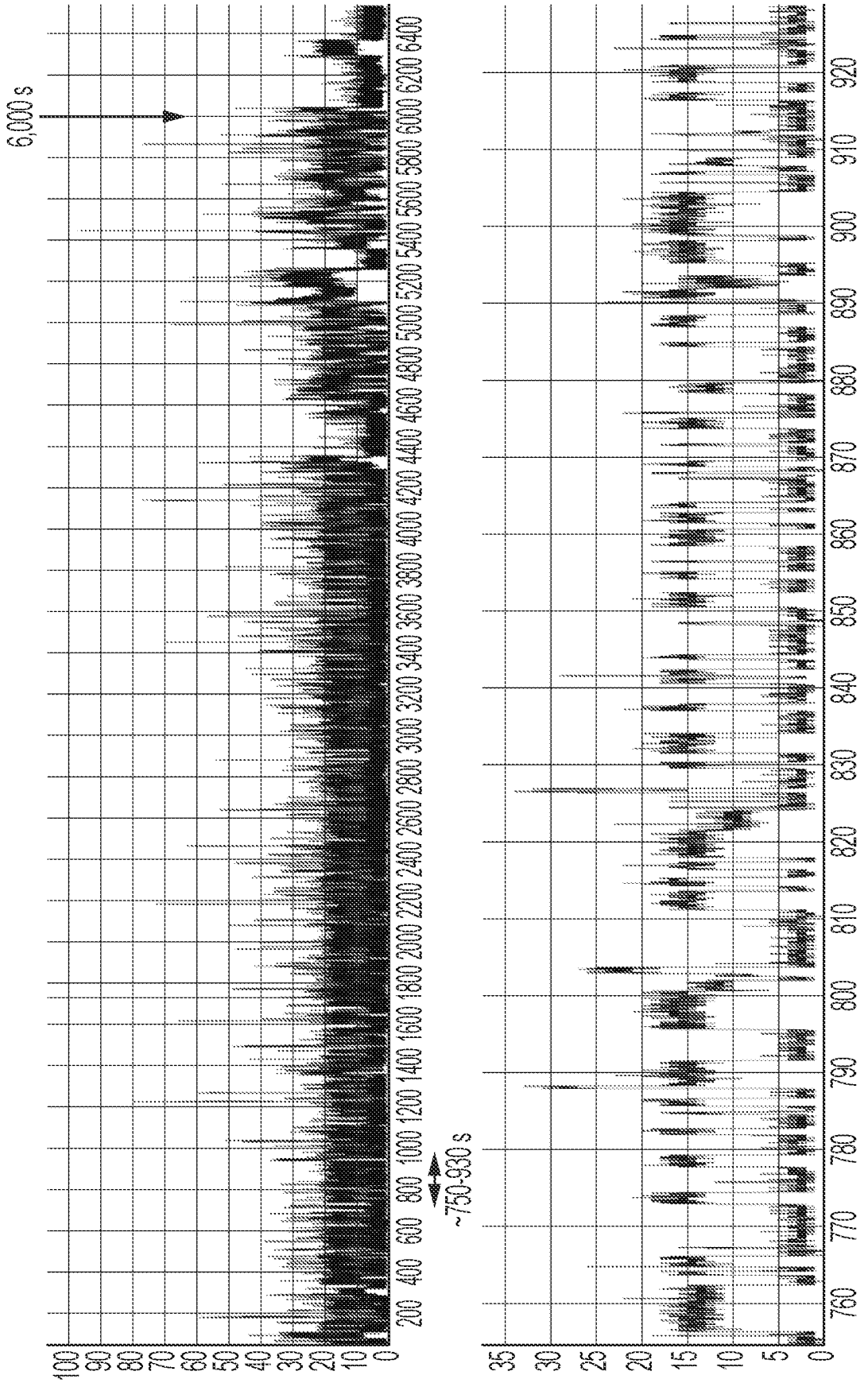

6,000 s
~750-930 s
0 10 20 30 40 50 60 70 80 90 100
200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000 4200 4400 4600 4800 5000 5200 5400 5600 5800 6000 6200 6400
0 5 10 15 20 25 30 35
760 770 780 790 800 810 820 830 840 850 860 870 880 890 900 910 920

FIG. 25C

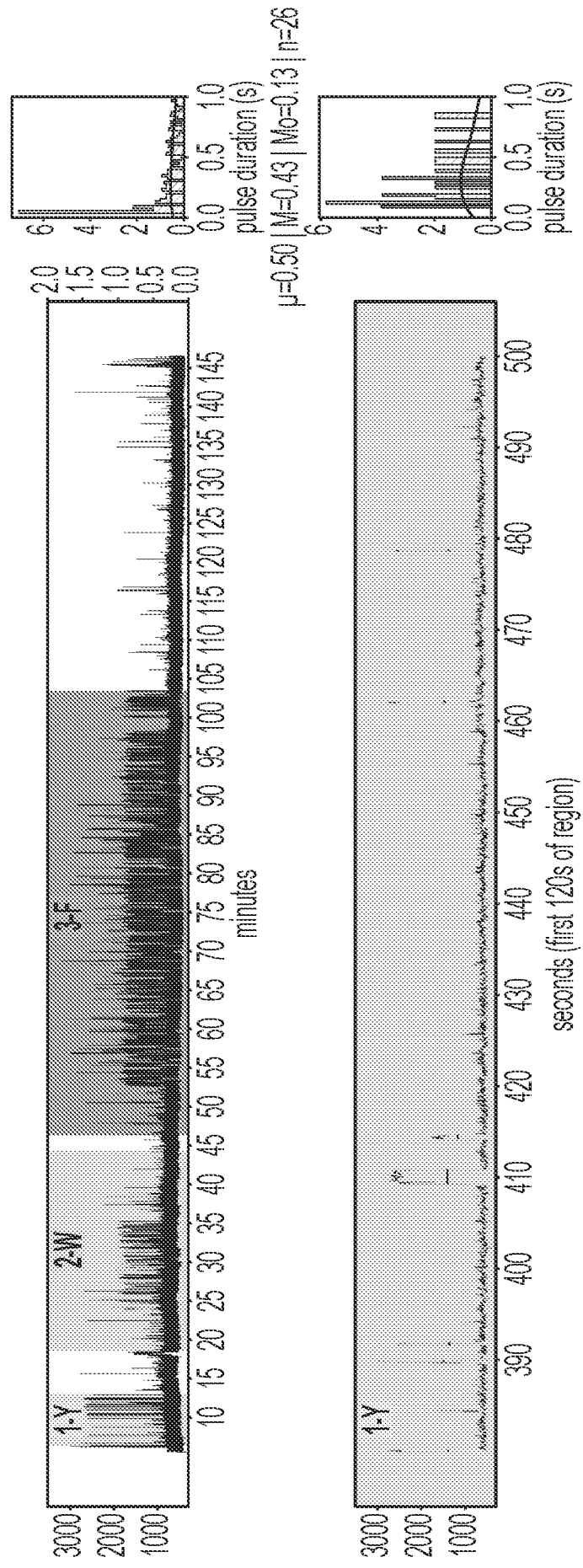

μ=0.50 | M=0.43 | Mo=0.13 | n=26
pulse duration (s)
pulse duration (s)
1-Y
2-W
3-F
minutes
1-Y
seconds (first 120s of region)

FIG. 26A

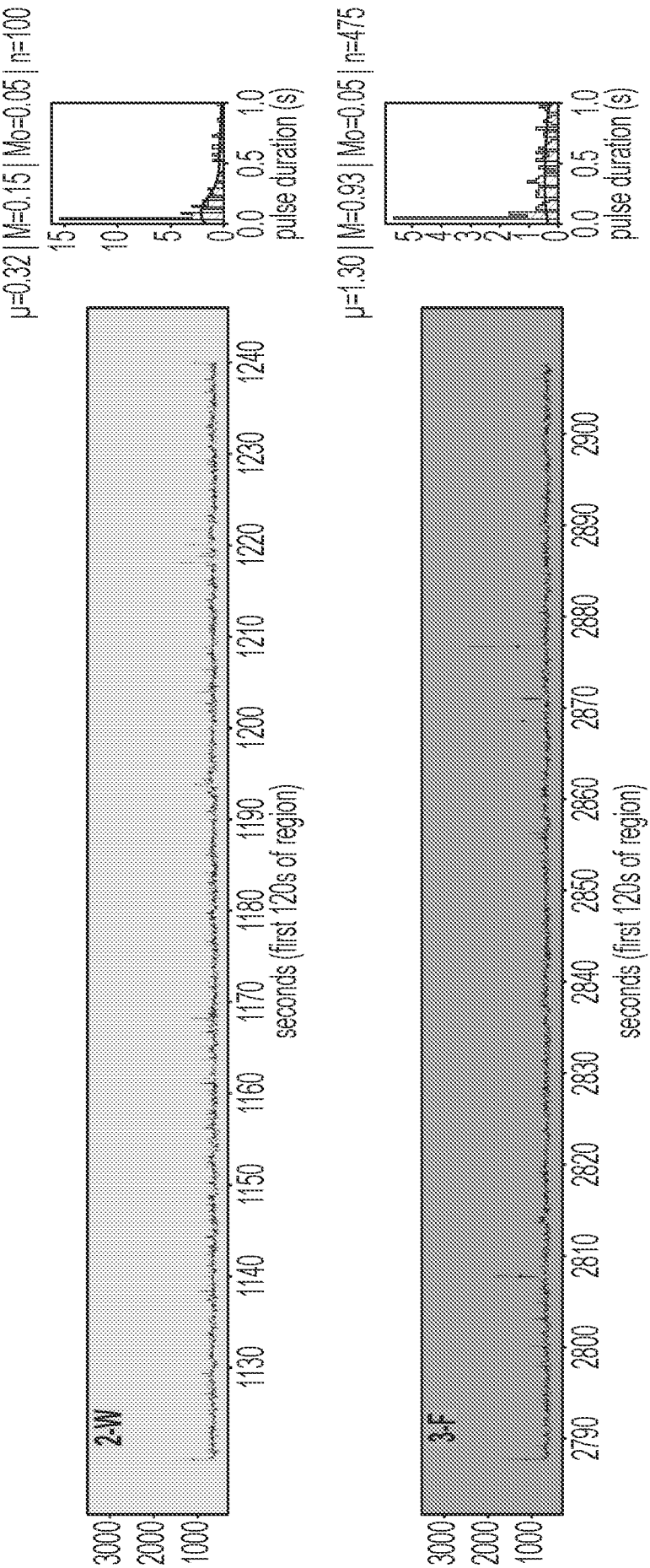


FIG. 26A
CONTINUED

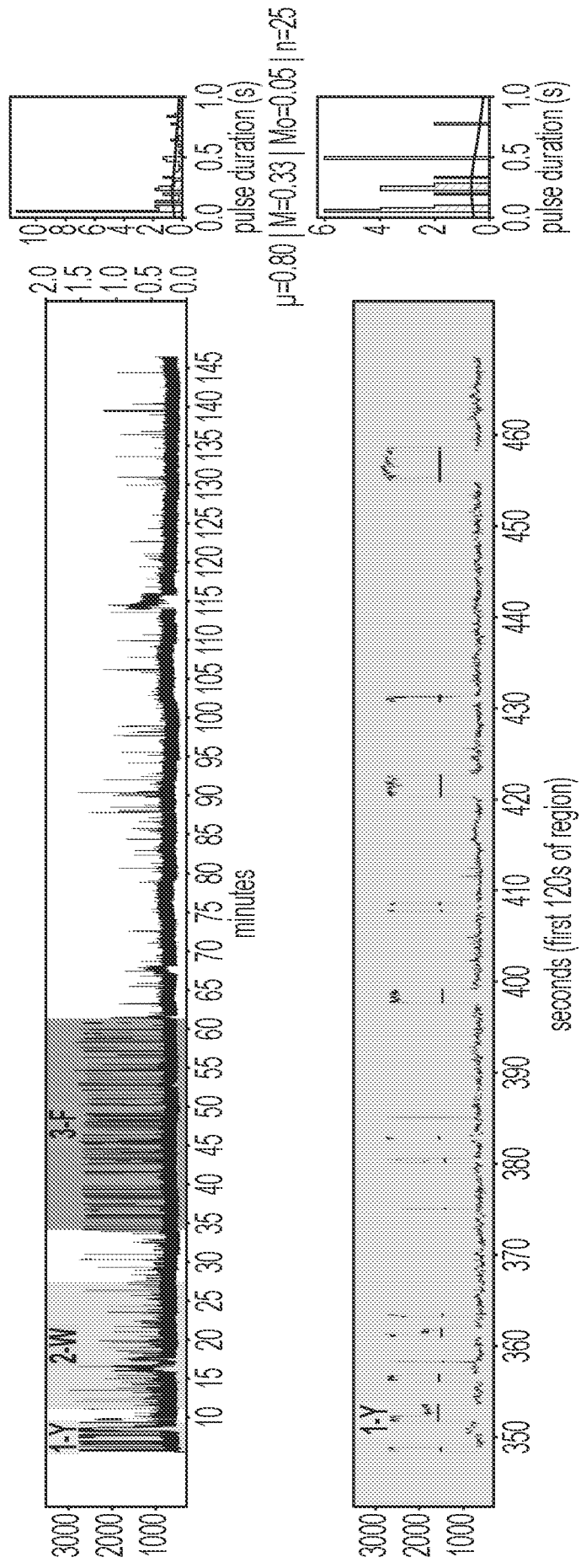

μ=0.80 | M=0.33 | Mo=0.05 | n=25
pulse duration (s)
pulse duration (s)
1-Y
2-W
3-F
minutes
1-Y
seconds (first 120s of region)

FIG. 26B

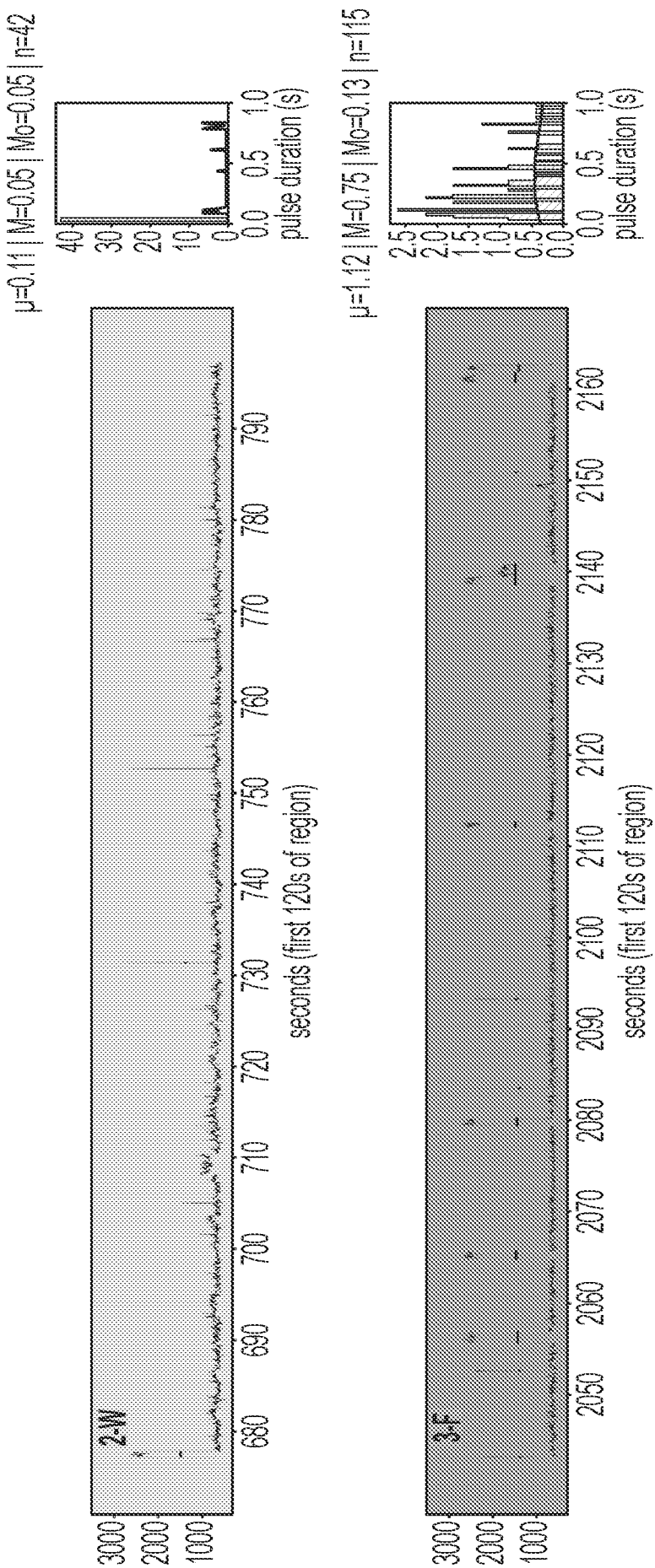


FIG. 26B
CONTINUED

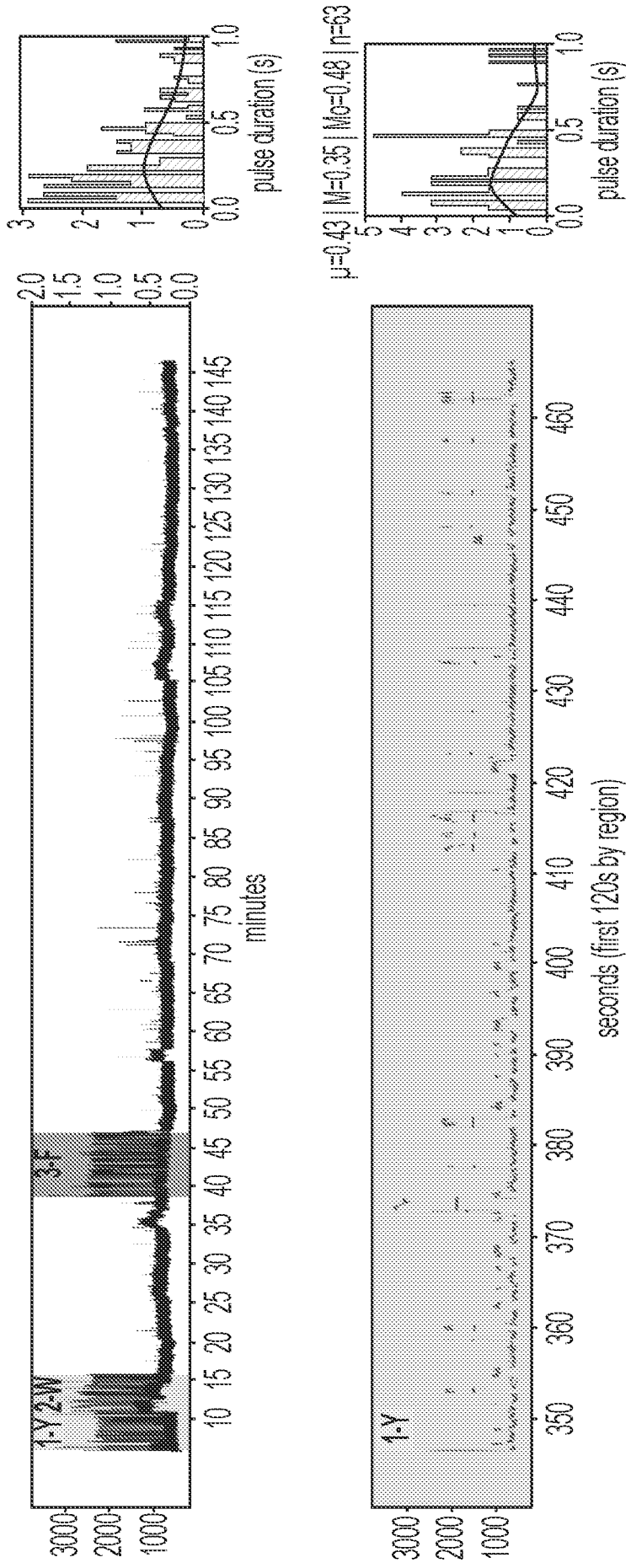

pulse duration (s)
μ=0.43 | M=0.35 | Mo=0.48 | n=63
pulse duration (s)
1-Y 2W
3-F
minutes
1-Y
seconds (first 120s by region)

FIG. 26C

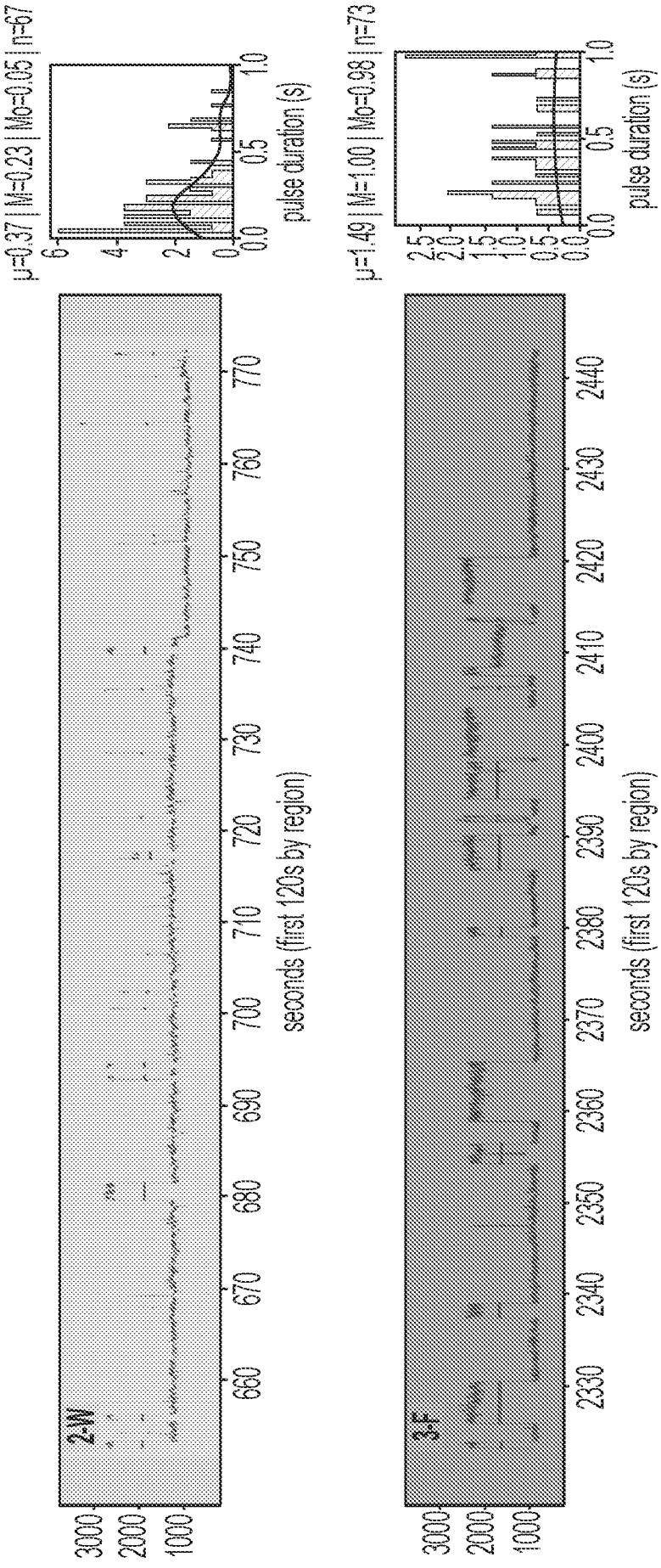


FIG. 26C
CONTINUED

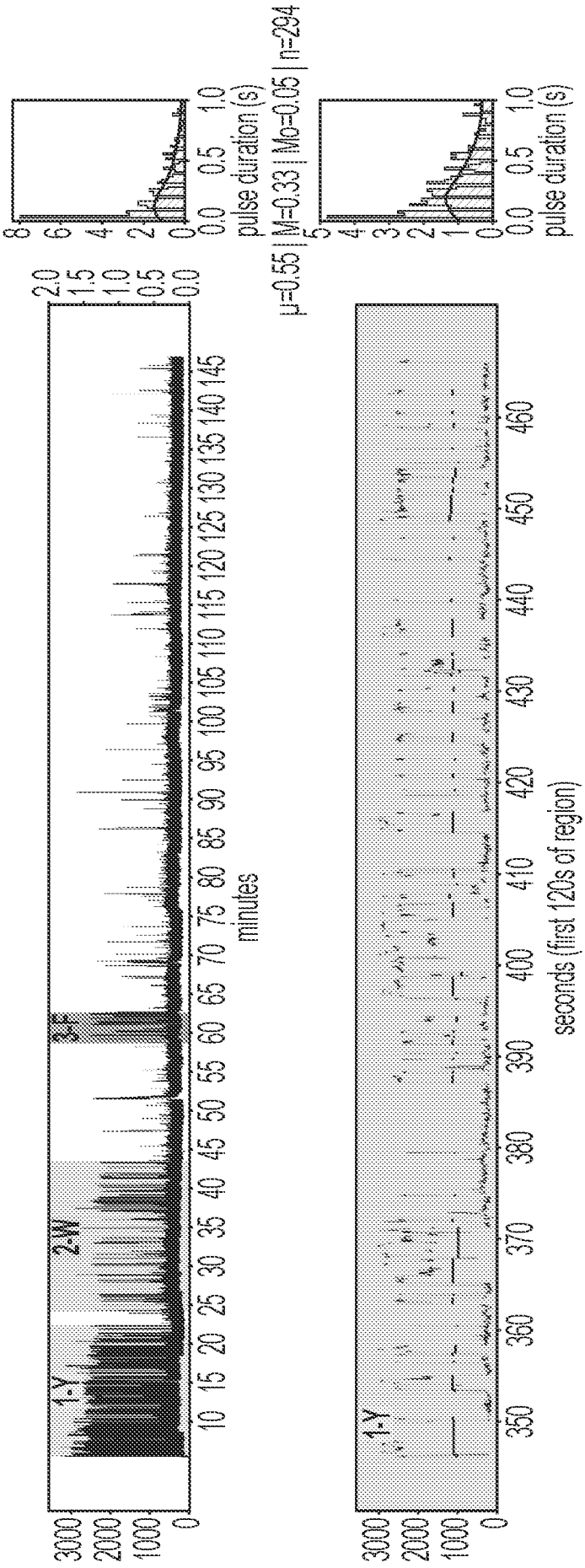

μ=0.55 | M=0.33 | Mo=0.05 | n=294
pulse duration (s)
pulse duration (s)
1-Y
2-W
3-F
minutes
1-Y
seconds (first 120s of region)

FIG. 26D

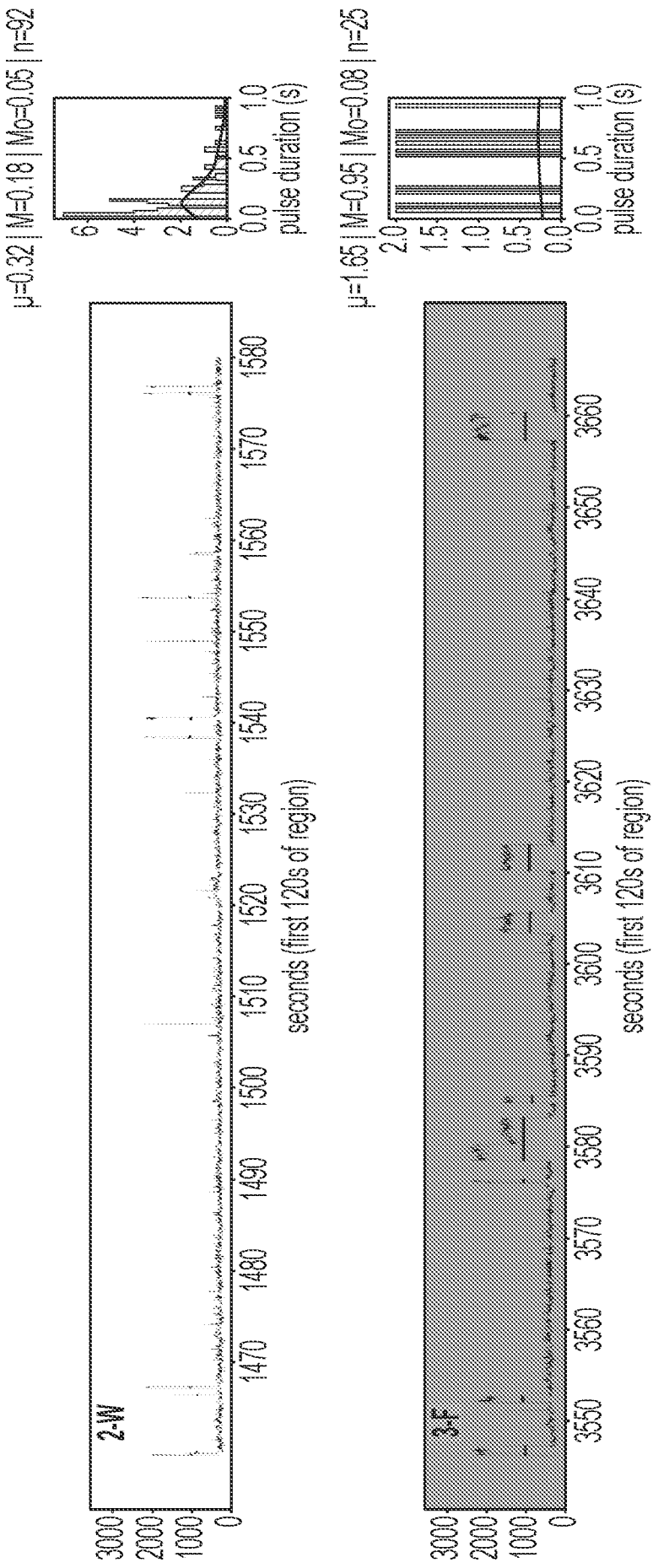


FIG. 26D
CONTINUED

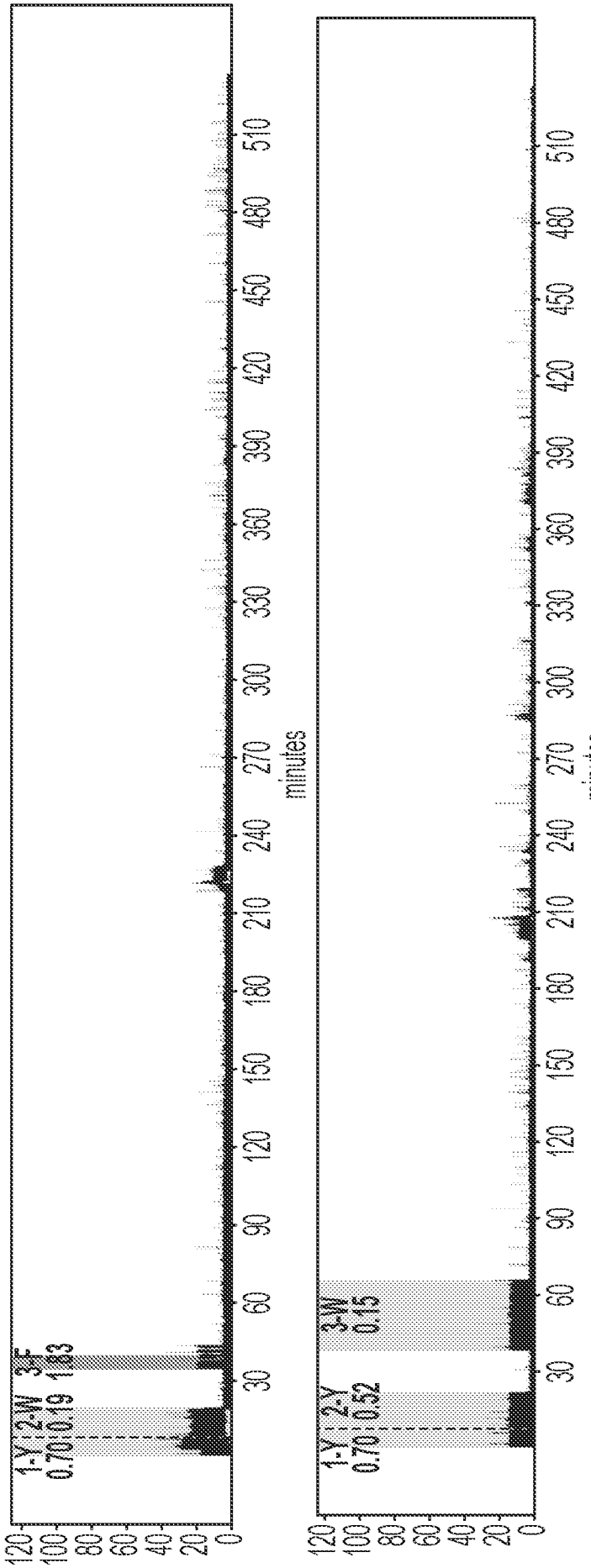

1-Y 2-W 3-F
0.70 0.19 1.83
1-Y 2-Y 3-W
0.70 0.52 0.15
minutes
minutes
0
20
40
60
80
100
120
30
60
90
120
150
180
210
240
270
300
330
360
390
420
450
480
510

FIG. 27

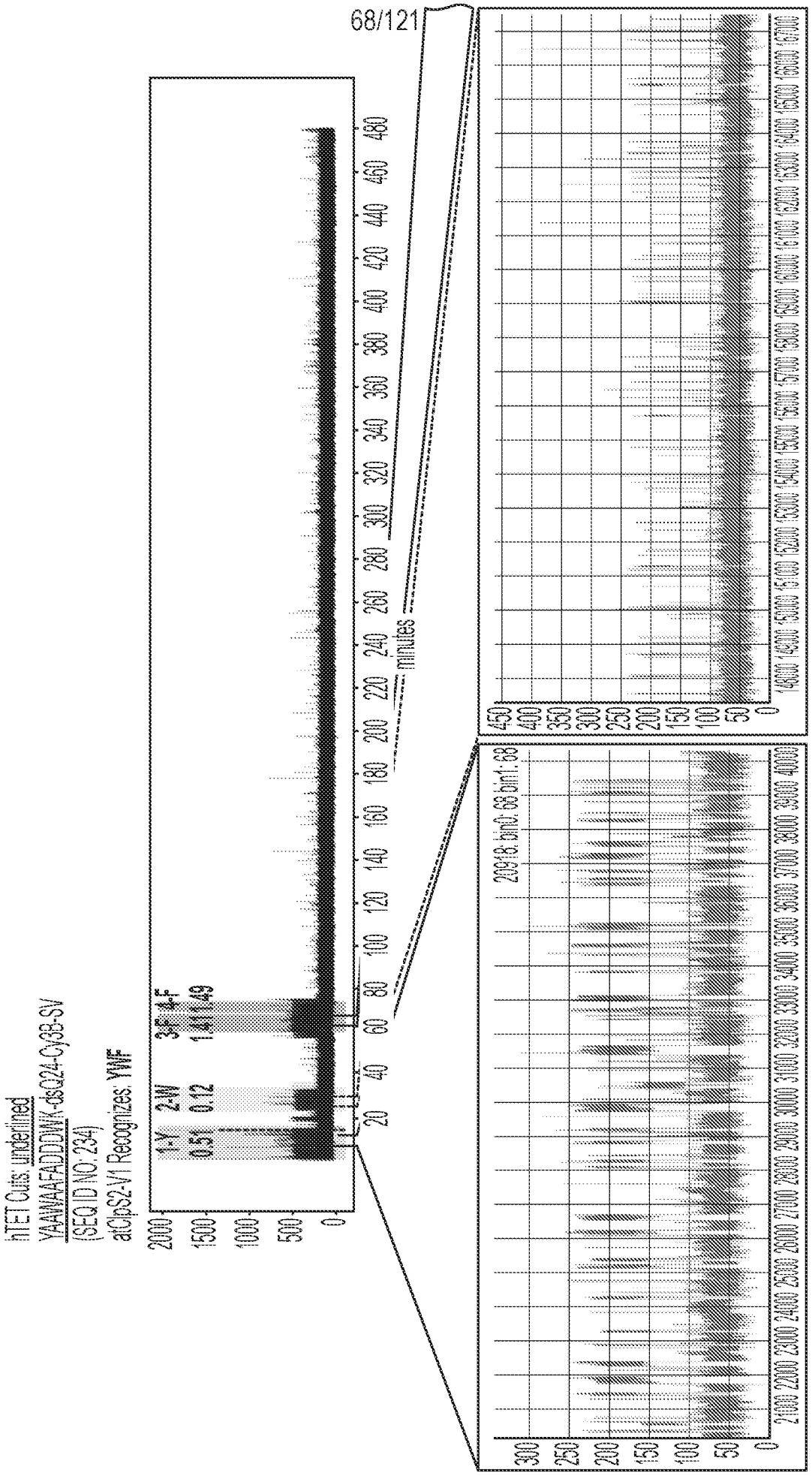

68/121
hTET Cuts underlined
YAAWAAFADDDWK-dsQ24-Cy3B-SV
(SEQ ID NO: 234)
a0bS2-V1 Recognizes: YWF
1-Y 2-W 3-F 4-F
0.51 0.12 1.41149
minutes
Y
W

FIG. 28A

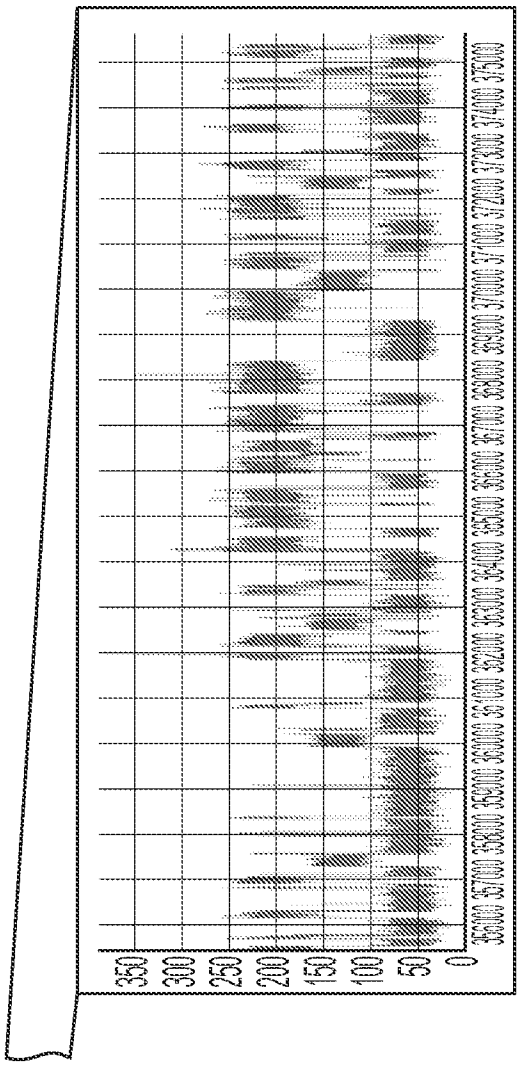


FIG. 28A
CONTINUED

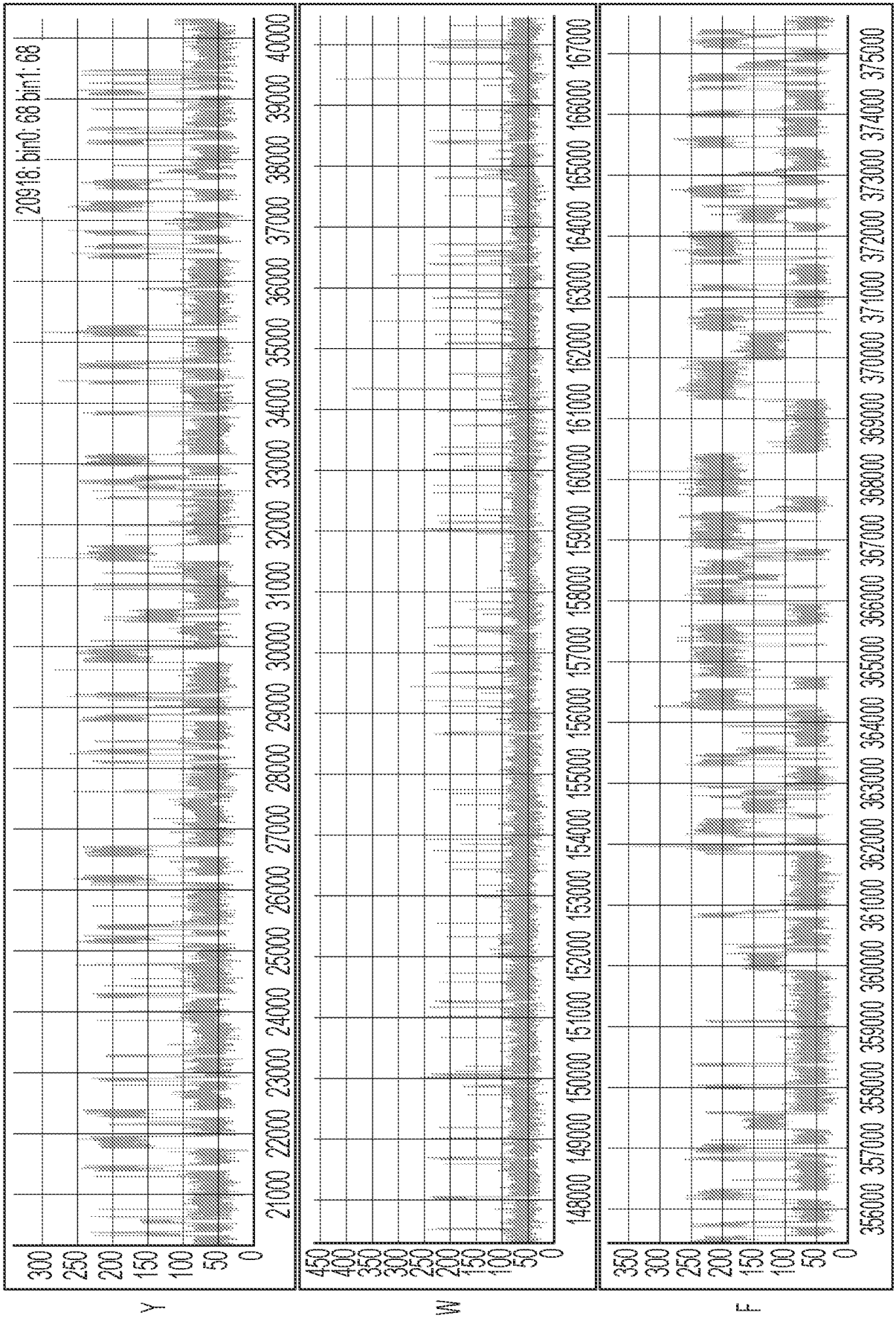


FIG. 28B

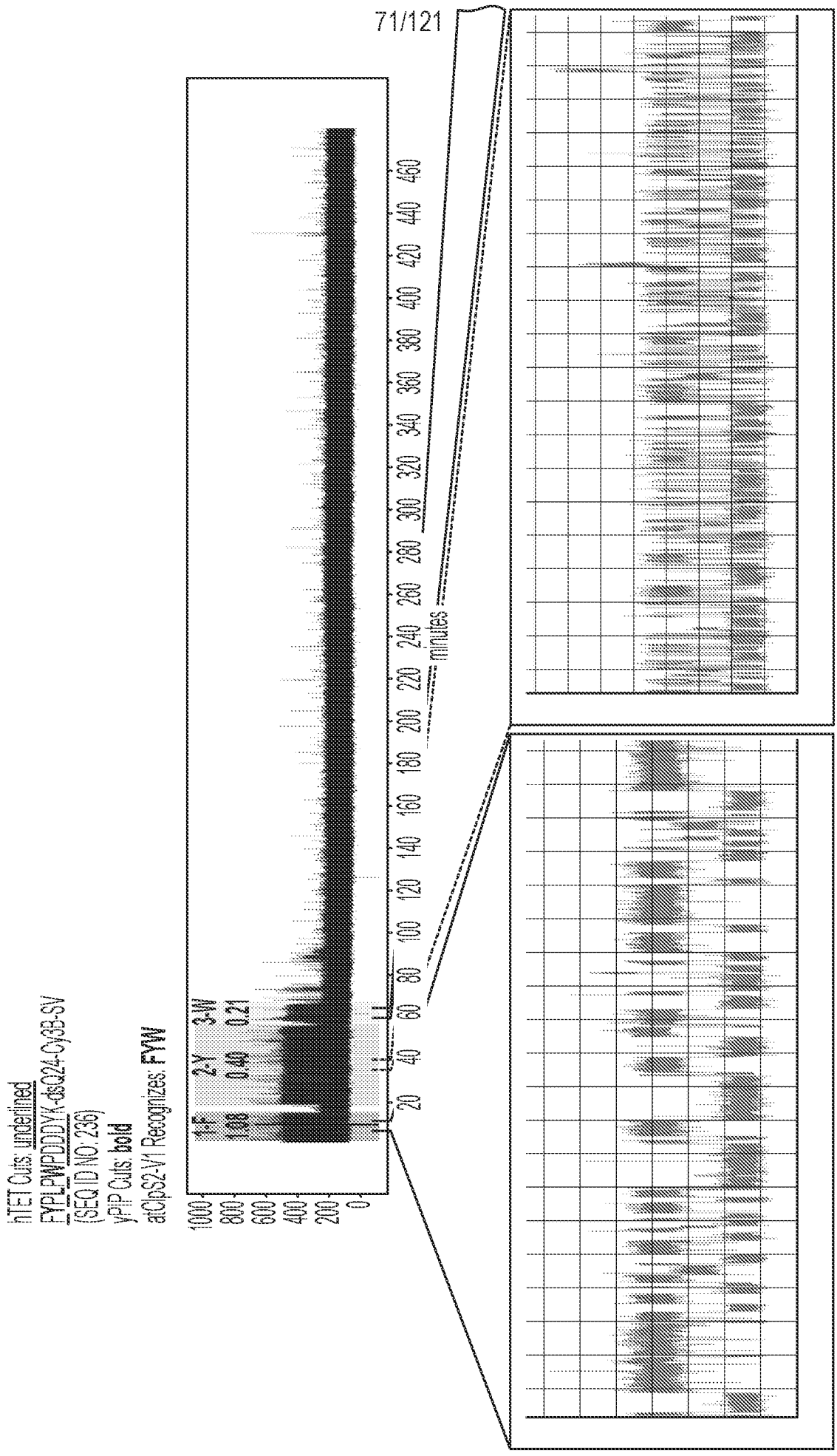

hTET Cuts: underlined
FYPLPWPDDDYK-ds024-Cy3B-SV
(SEQ ID NO: 236)
yPIP Cuts: bold
atOlbS2-V1 Recognizes: FYW
1-F
1.08
2-Y
0.40
3-W
0.21
minutes
Y
F

FIG. 28C

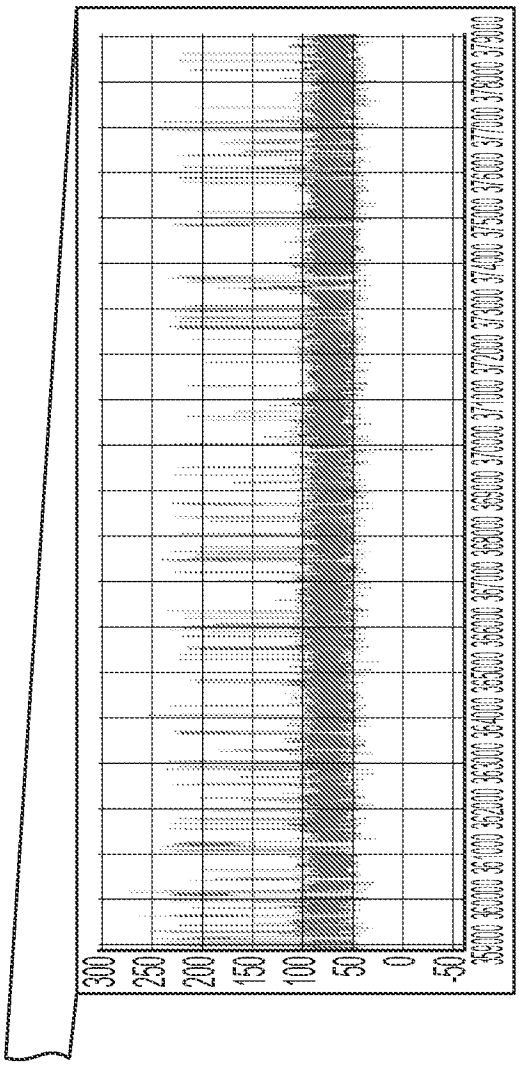


FIG. 28C
CONTINUED

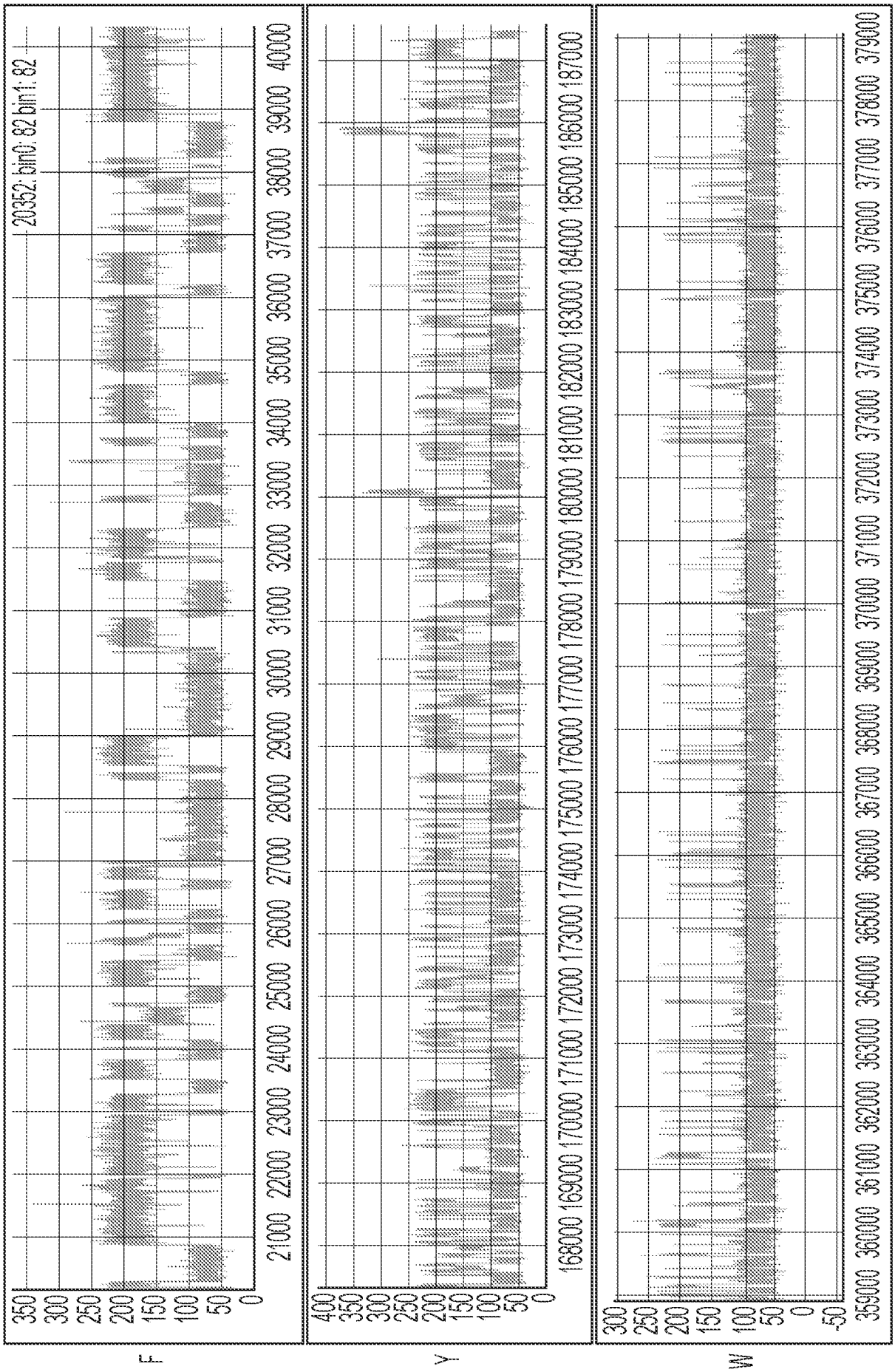

E
20352: bin0: 82 bin1: 82
0 50 100 150 200 250 300 350
21000 22000 23000 24000 25000 26000 27000 28000 29000 30000 31000 32000 33000 34000 35000 36000 37000 38000 39000 40000
Y
0 50 100 150 200 250 300 350 400
168000 169000 170000 171000 172000 173000 174000 175000 176000 177000 178000 179000 180000 181000 182000 183000 184000 185000 186000 187000
W
-50 0 50 100 150 200 250 300
359000 360000 361000 362000 363000 364000 365000 366000 367000 368000 369000 370000 371000 372000 373000 374000 375000 376000 377000 378000 379000

FIG. 28D

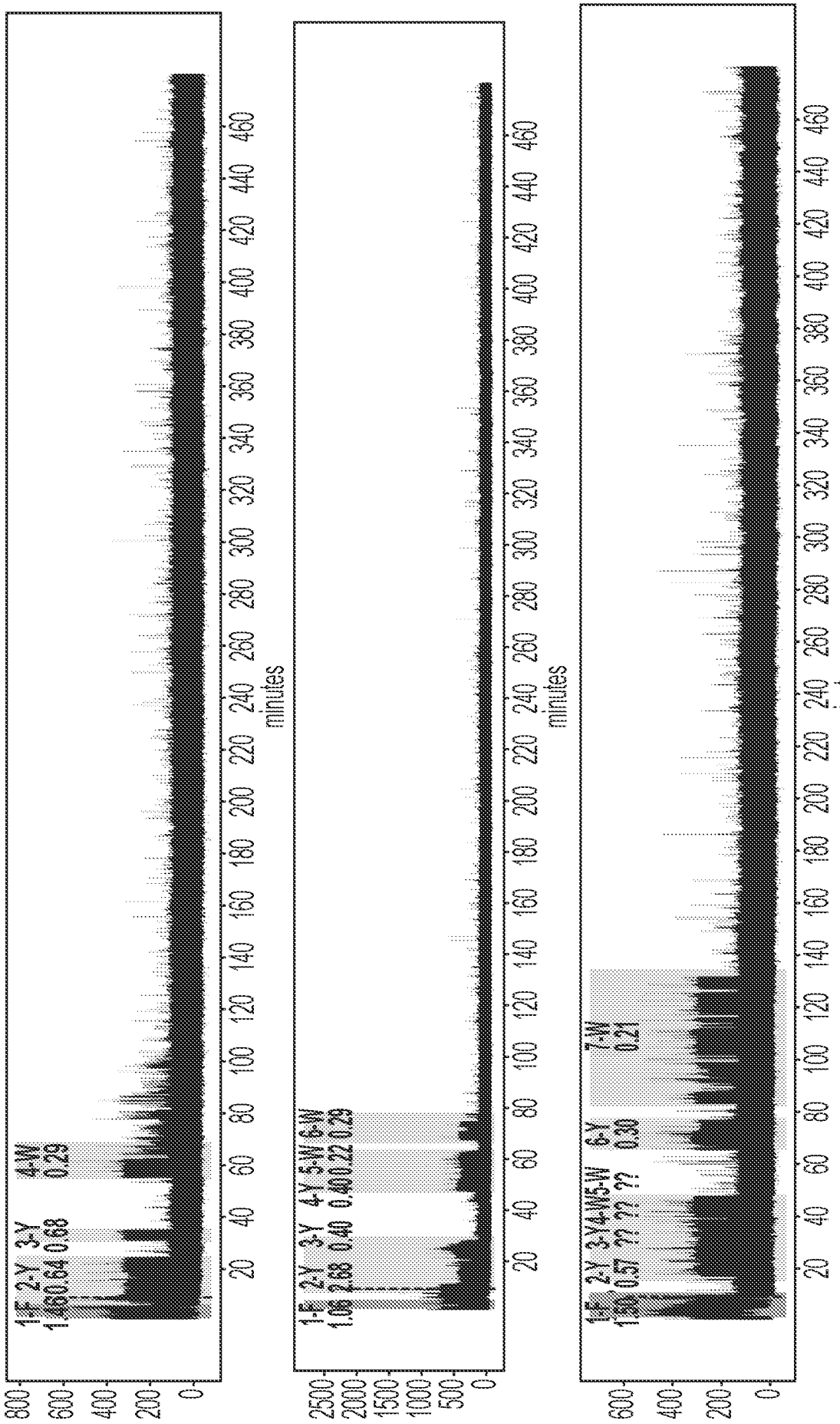

1-E 2-Y 3-Y 4-W
1.46 0.64 0.68 0.29
minutes
1-E 2-Y 3-Y 4-Y 5-W 6-W
1.06 2.68 0.40 0.40 0.22 0.29
minutes
1-E 2-Y 3-Y 4-W 5-W 6-Y 7-W
1.30 0.57 ?? ?? ?? ?? 0.30 0.21
minutes

FIG. 28E

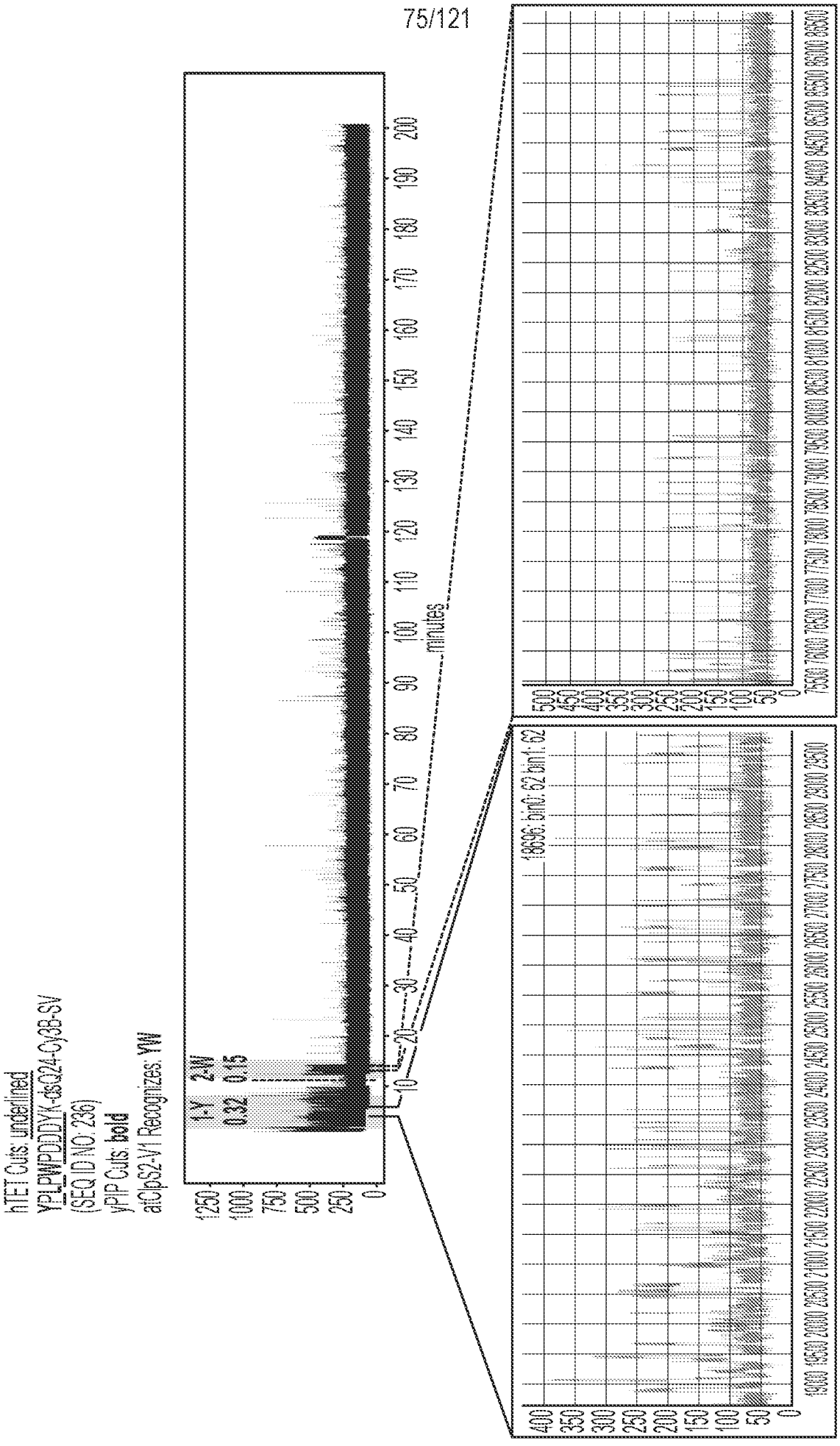

hTET Cuts: underlined
YPLPWPDDDYK-dsQ24-Cy3B-SV
(SEQ ID NO: 236)
yPIP Cuts: bold
aiOpS2-V1 Recognizes: YW
1-Y 2-W
0.32 0.15
minutes
Y
W

FIG. 28F

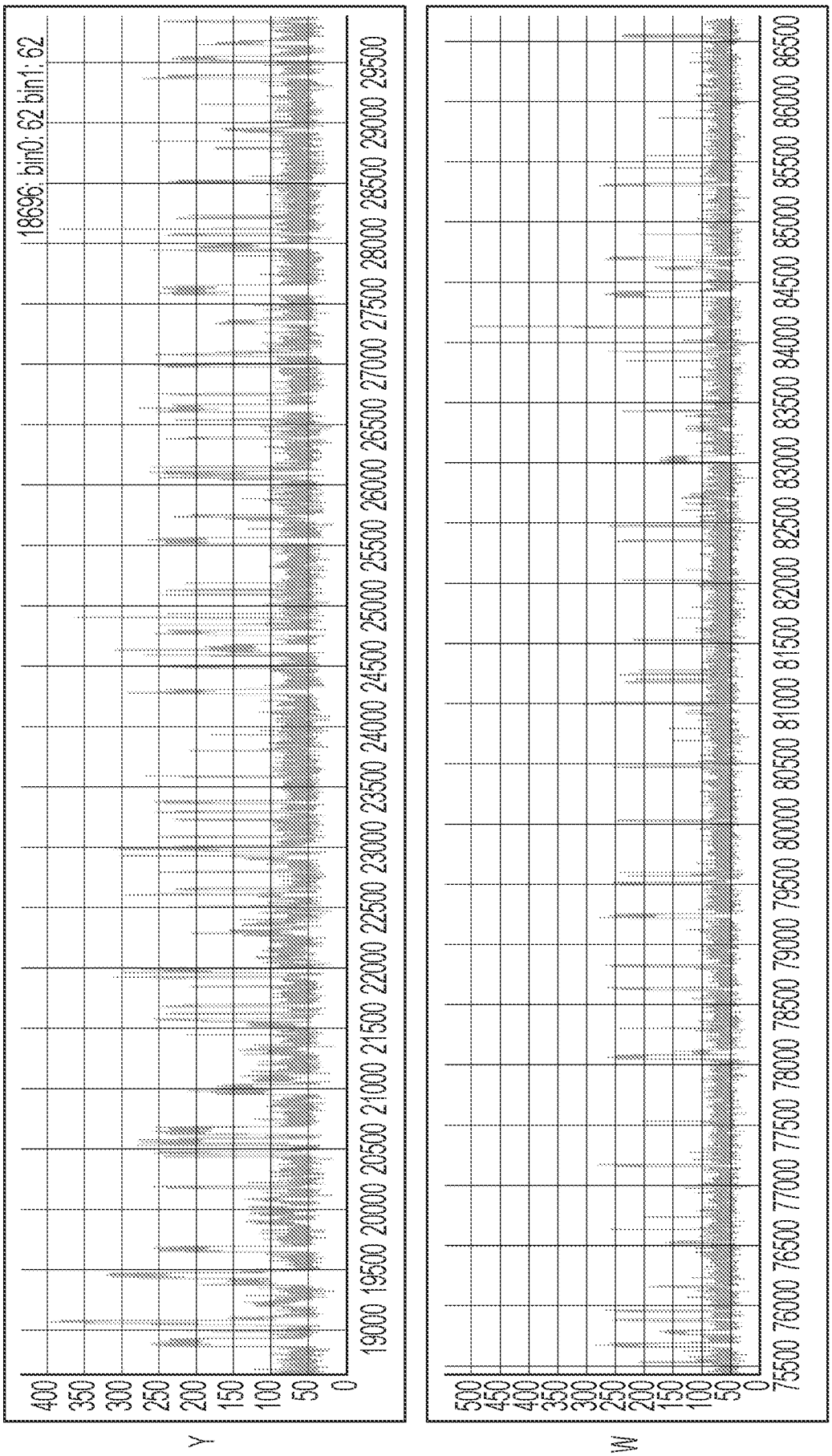


FIG. 28G

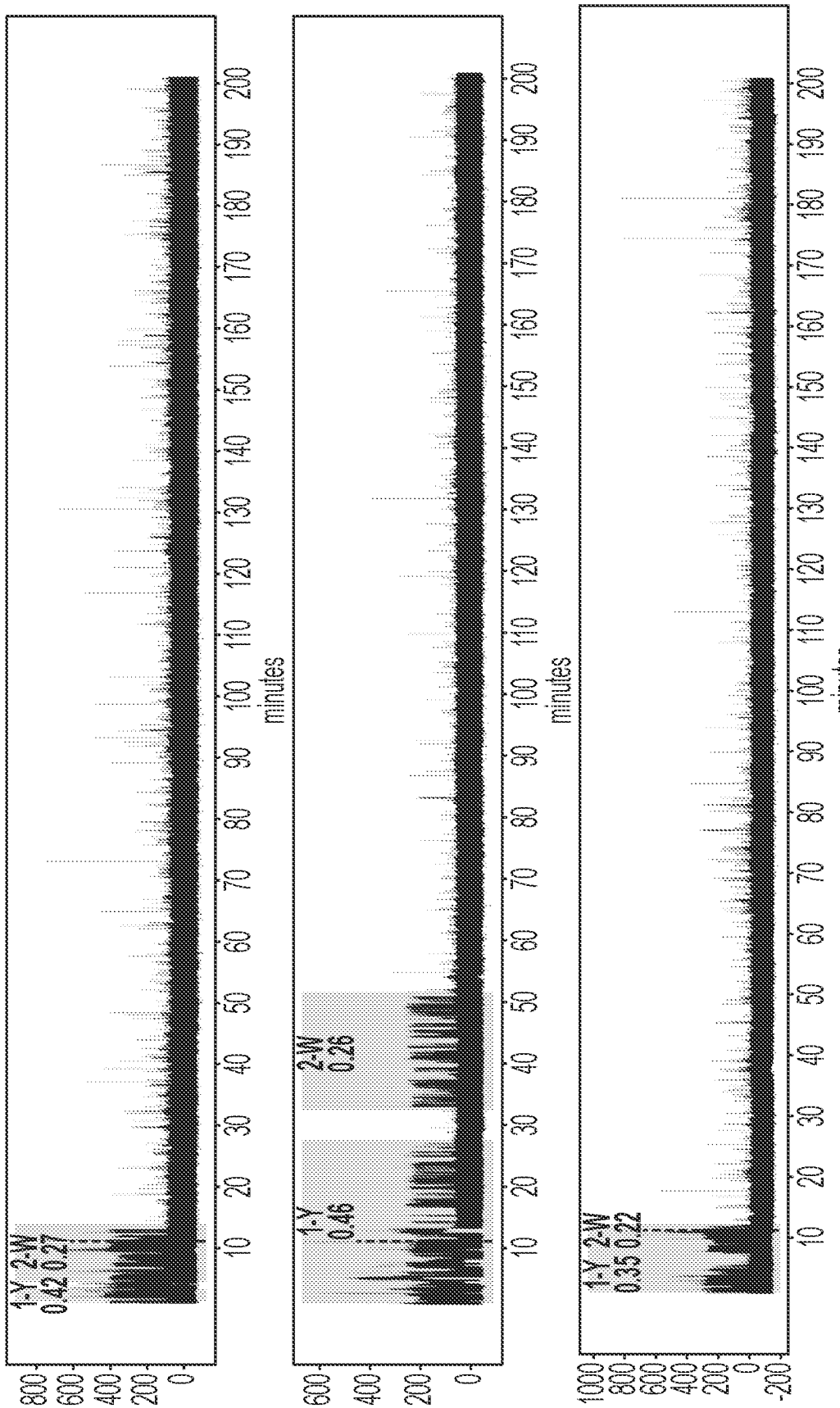

1-Y 2-W
0.42 0.27
1-Y
0.46
2-W
0.26
1-Y 2-W
0.35 0.22
minutes
minutes
minutes

FIG. 28H

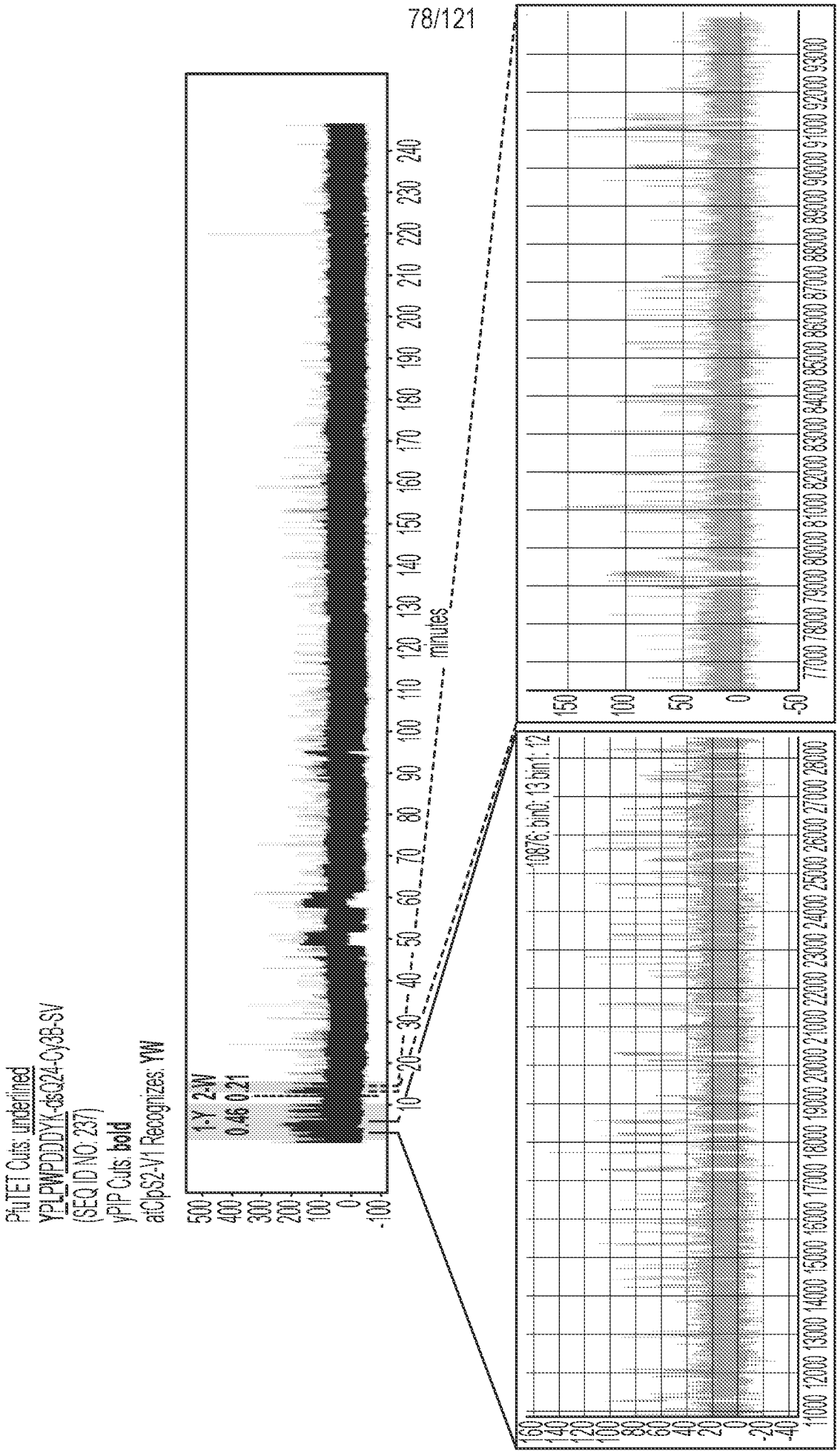

PiuTEI Cuts: underlined
YPLPWPDDDYK-dsQ24-Cy3B-SV
(SEQ ID NO: 237)
yPIP Cuts: bold
aiOpS2-V1 Recognizes: YW
1:Y 2:W
0.46 0.21
minutes
Y
W

FIG. 28I

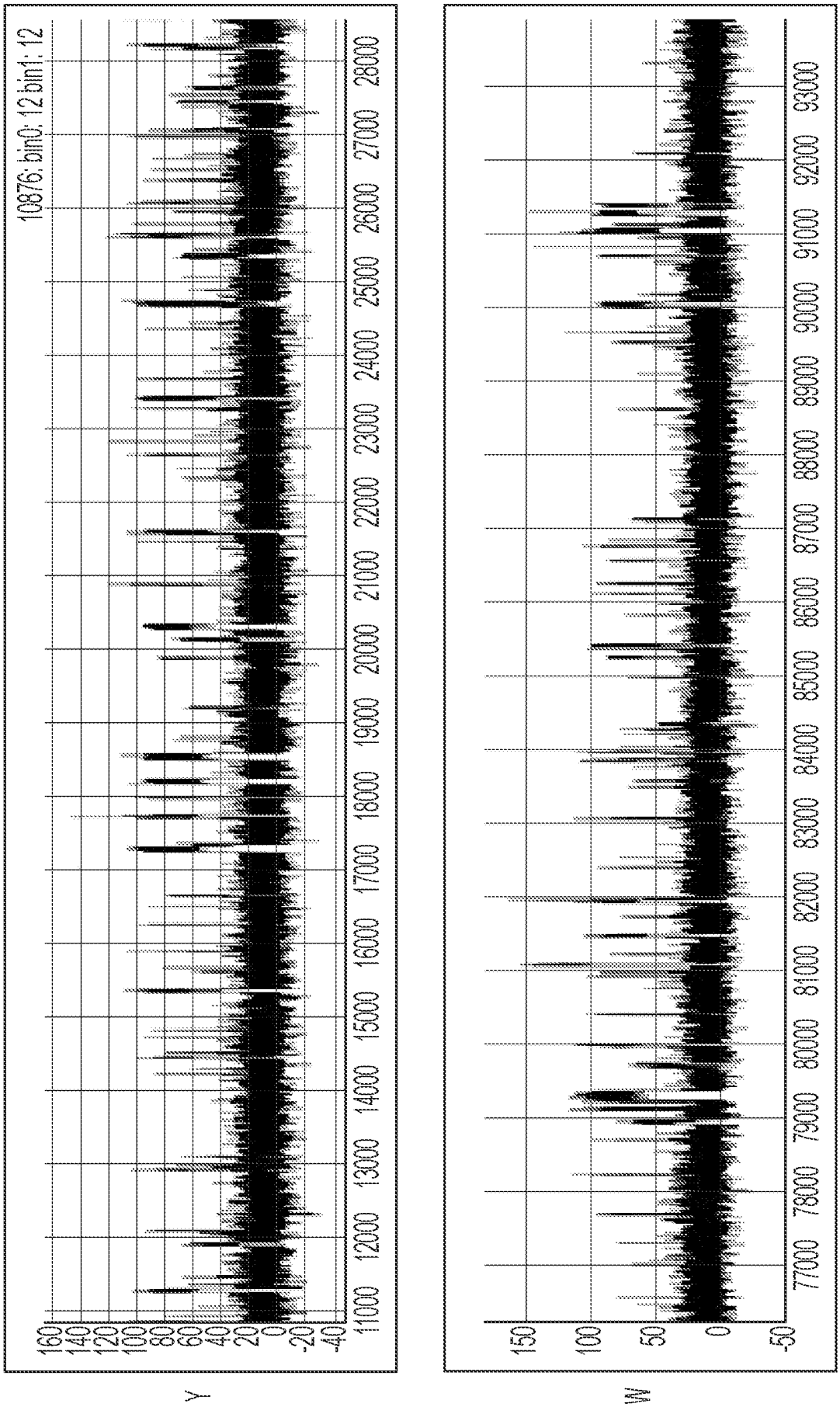


FIG. 28J

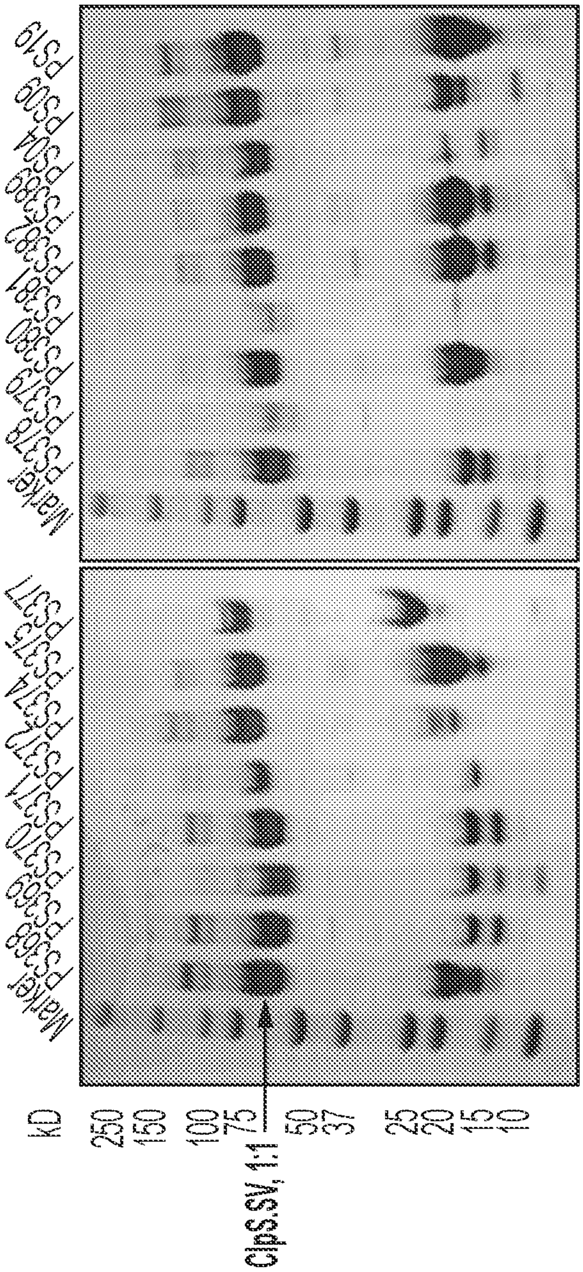

Marker
pS378
pS379
pS380
pS381
pS382
pS389
pS04
pS09
pS19
Marker
pS368
pS369
pS370
pS371
pS372
pS374
pS375
pS377
kD
250
150
100
75
50
37
25
20
15
10
ClpS.SV, 1:1

FIG. 29A

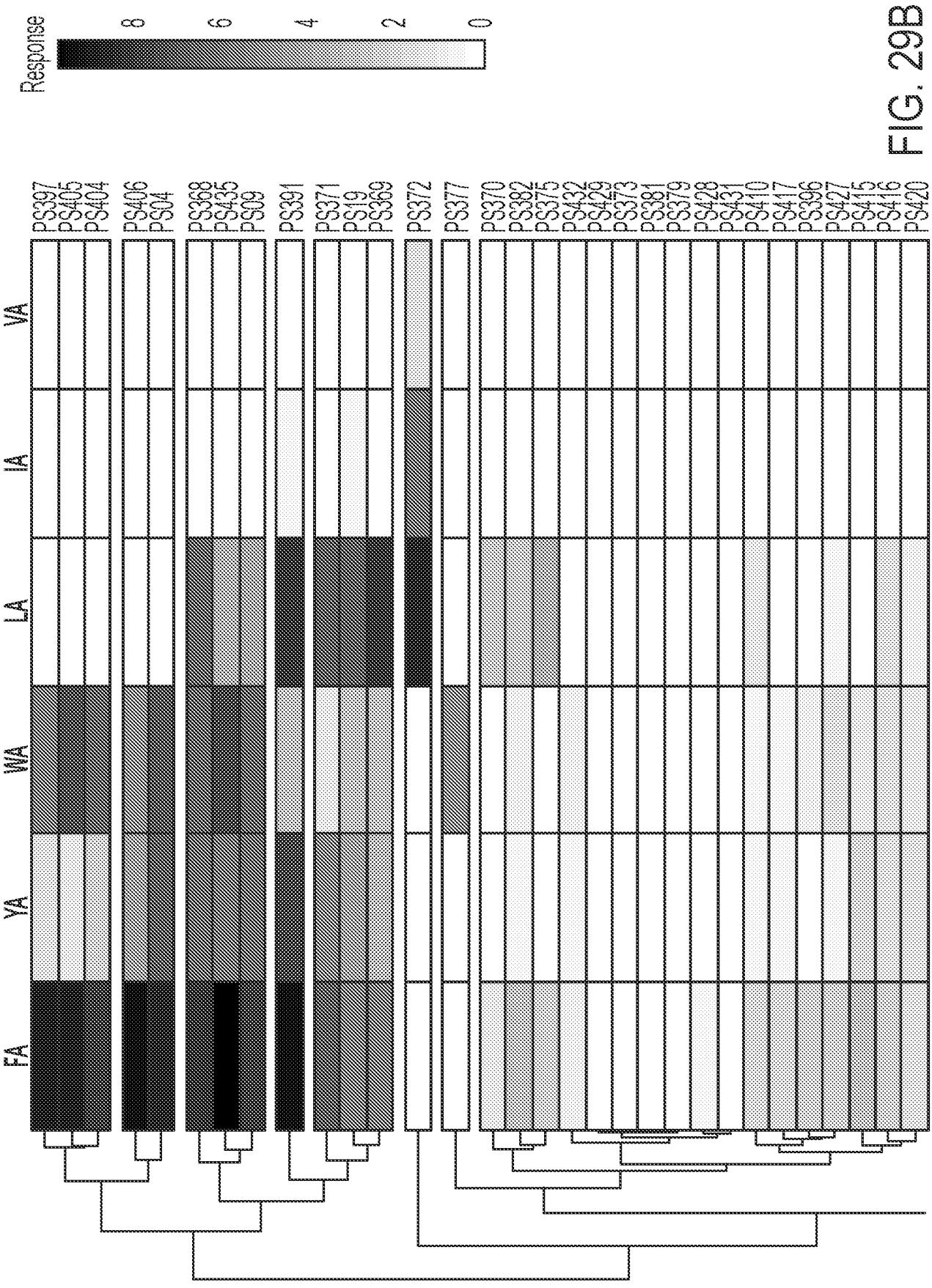

Response
8
6
4
2
0
PS397
PS405
PS404
PS406
PS04
PS368
PS435
PS09
PS391
PS371
PS19
PS369
PS372
PS377
PS370
PS382
PS375
PS432
PS429
PS373
PS381
PS379
PS428
PS431
PS410
PS417
PS396
PS427
PS415
PS416
PS420
VA
IA
LA
WA
YA
FA

FIG. 29B

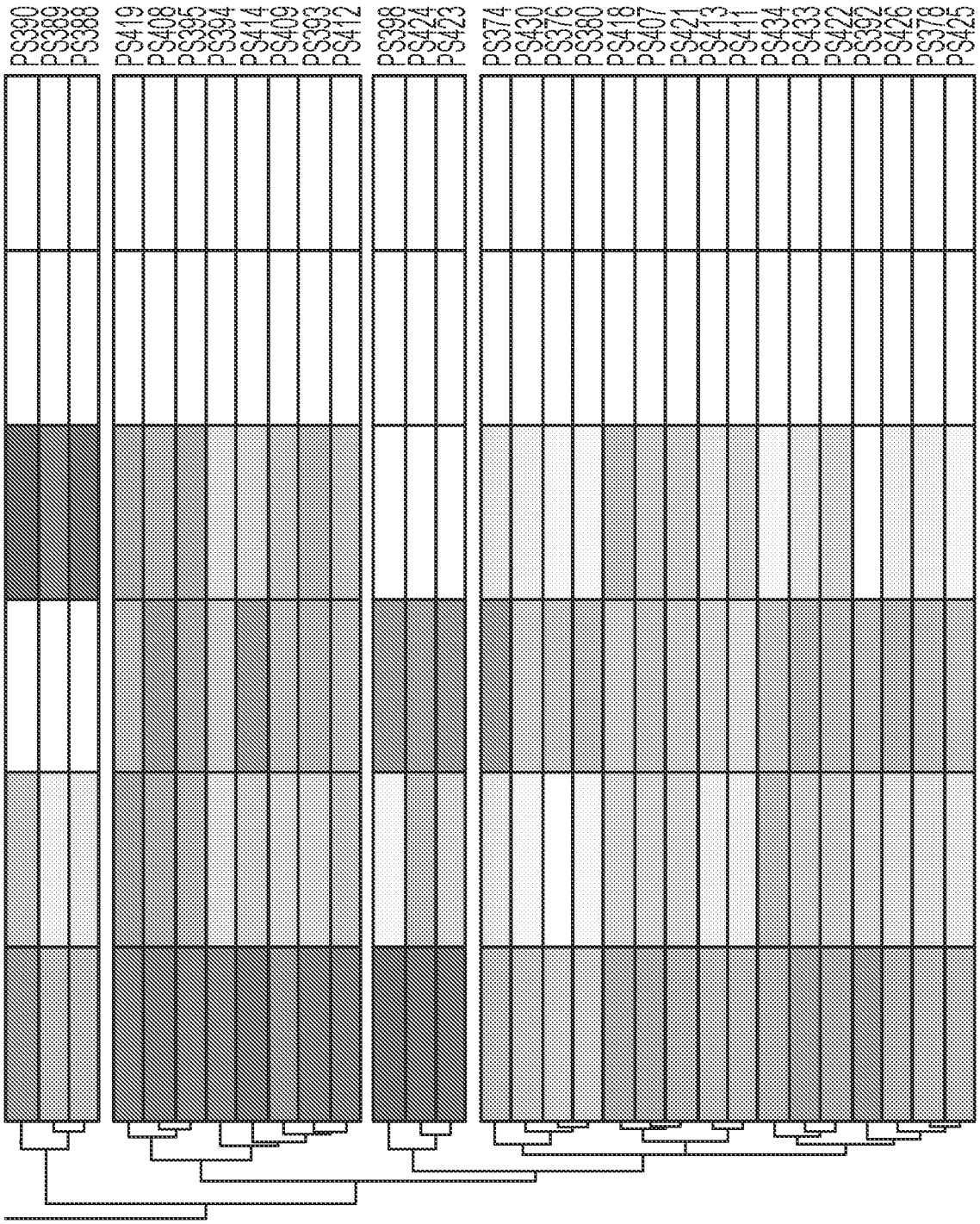


FIG. 29B
CONTINUED

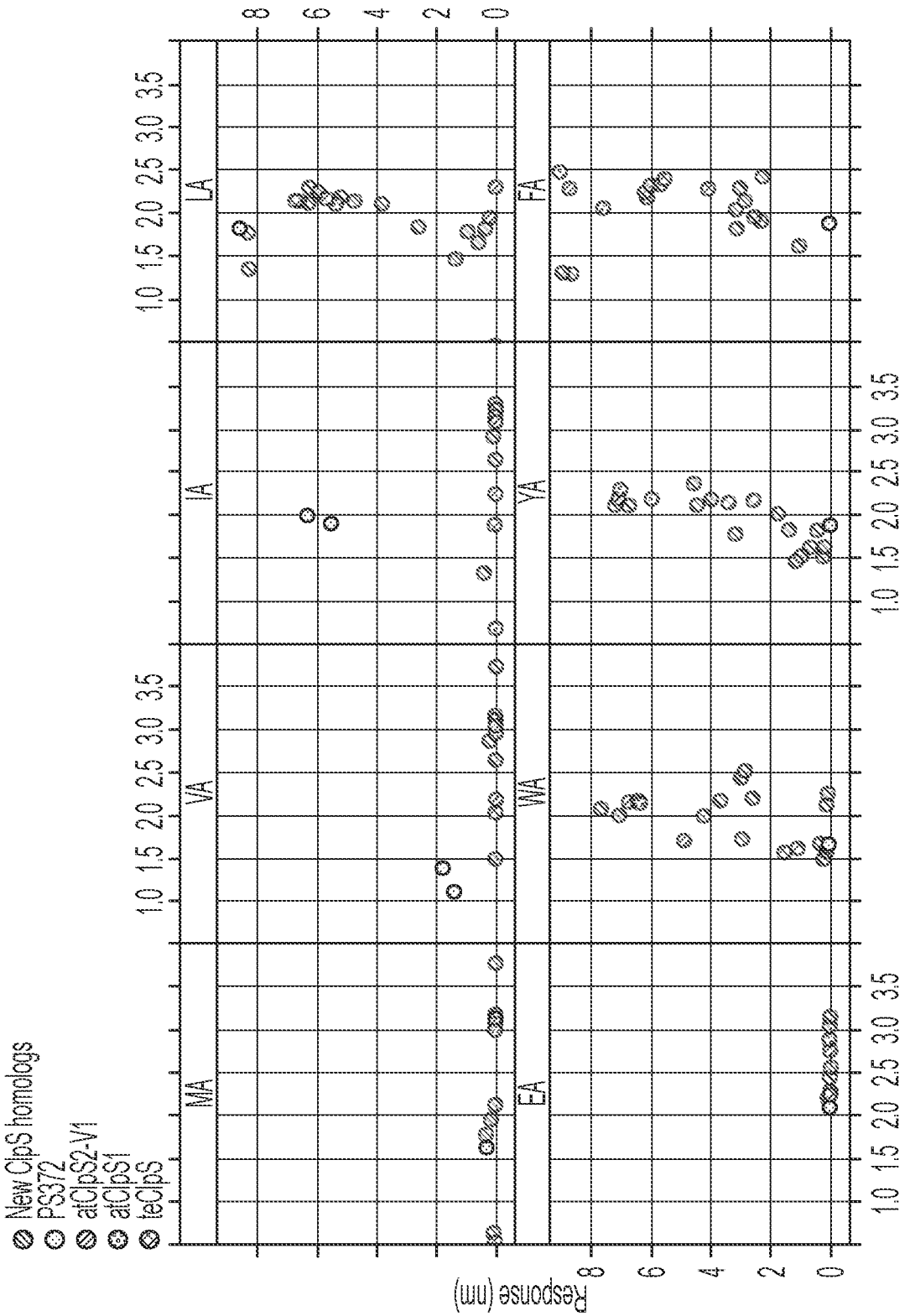

New ClpS homologs
PS372
atClpS2-V1
atClpS1
teClpS
LA
IA
VA
MA
FA
YA
WA
EA
Response (nm)
log10 1/kdis (s)

FIG. 29C

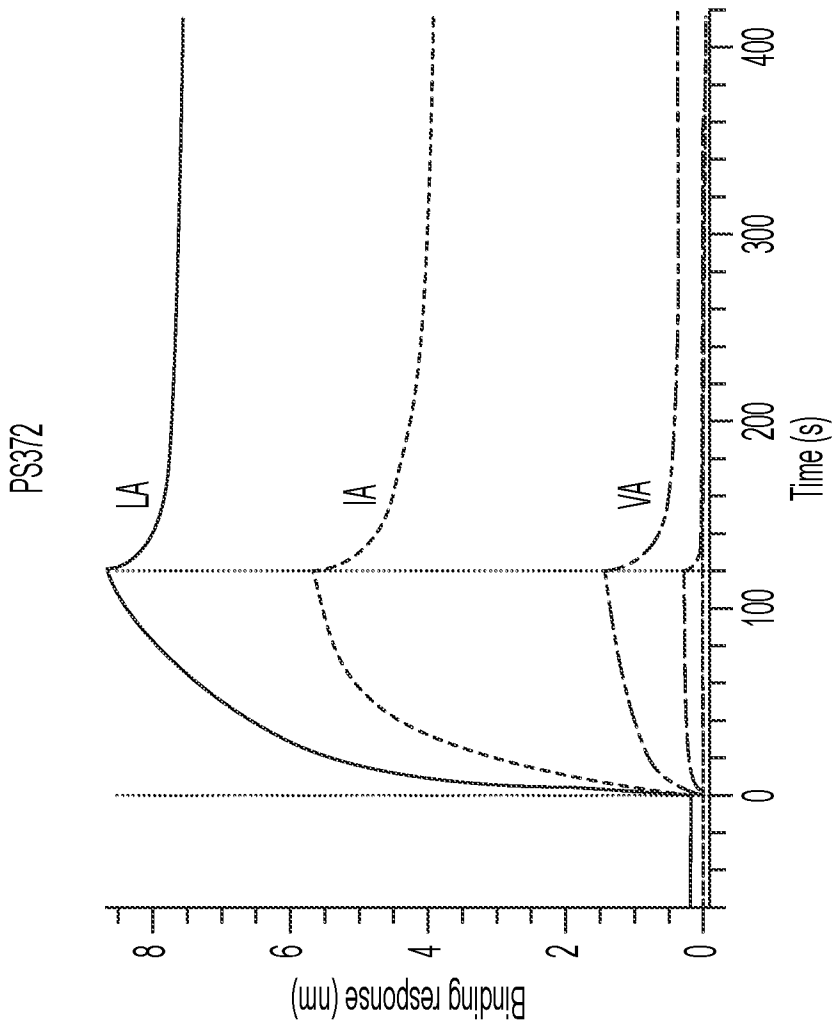

PS372
LA
IA
VA
Binding response (nm)
0
2
4
6
8
0
100
200
300
400
Time (s)

FIG. 29D

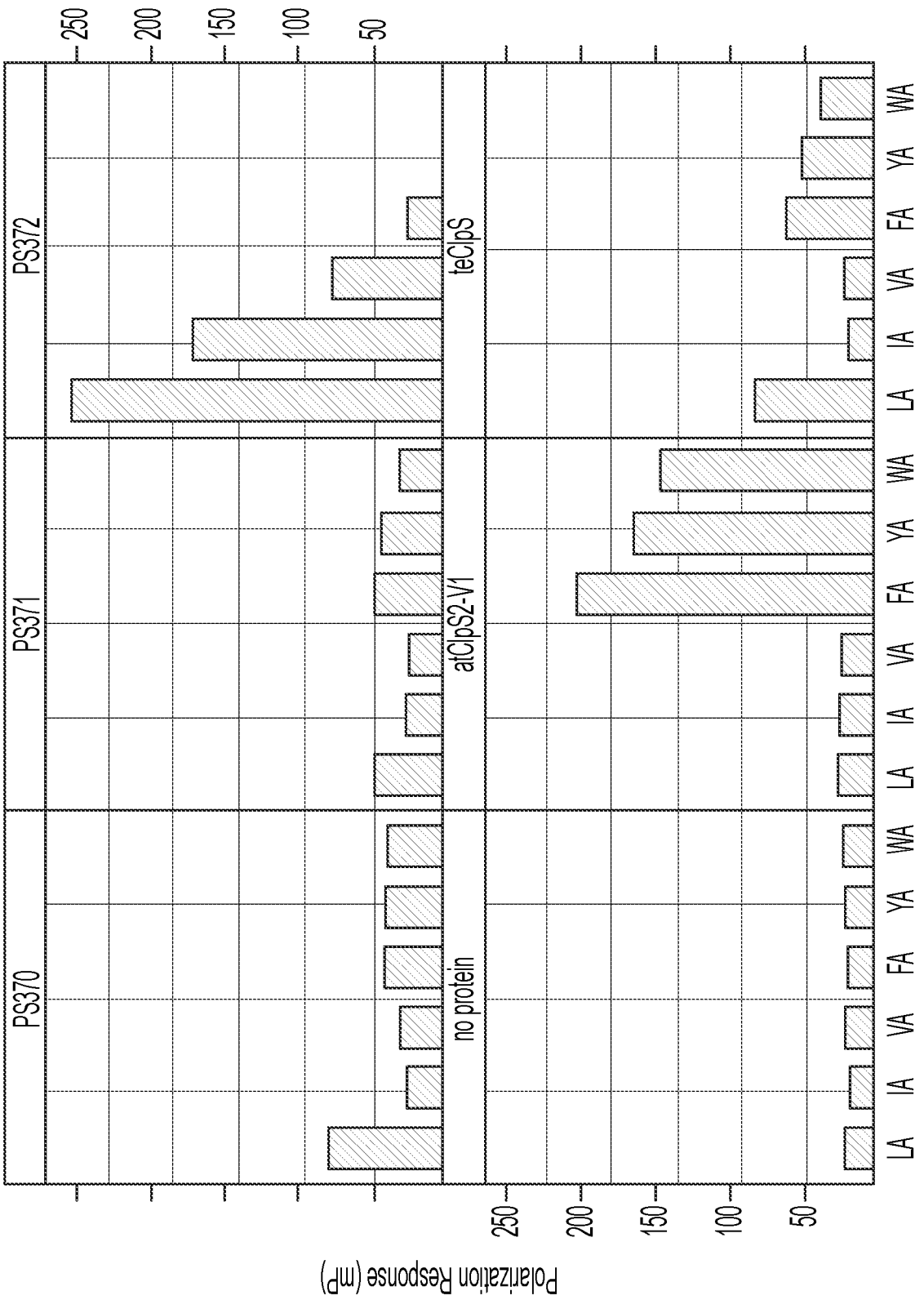

PS370
PS371
PS372
no protein
atClpS2-V1
teClpS
LA IA VA FA YA WA
Polarization Response (mP)
250
200
150
100
50

FIG. 29E

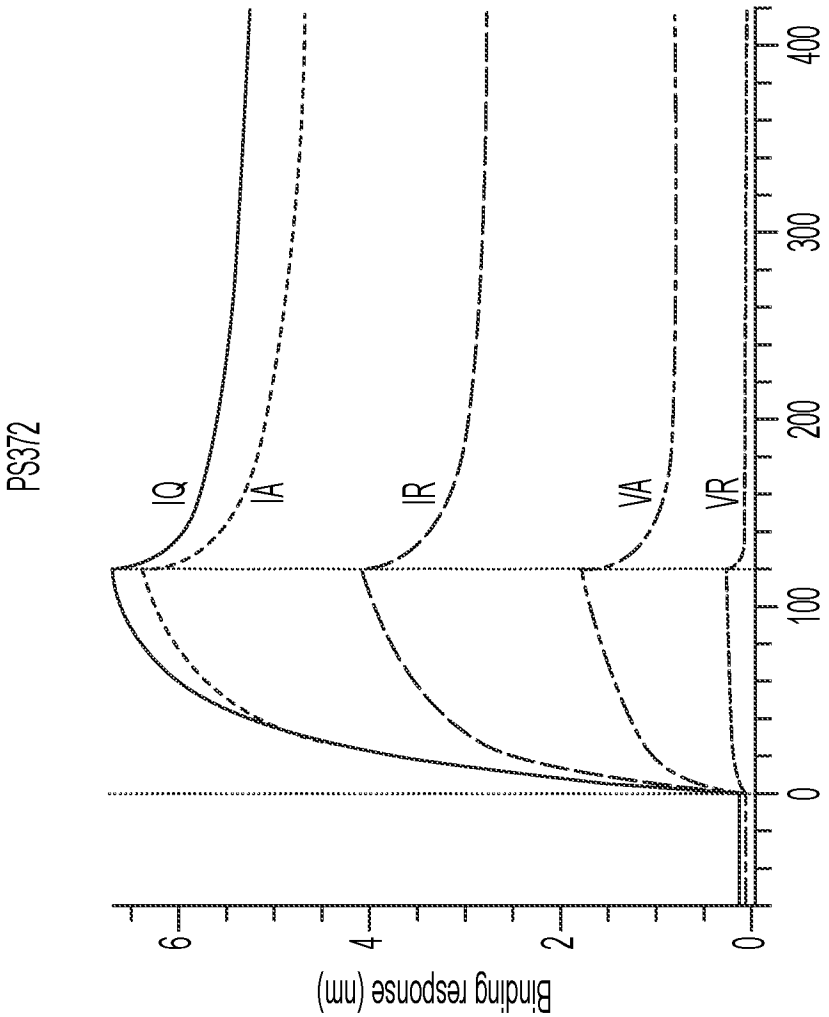

PS372
IQ
IA
IR
VA
VR
Binding response (nm)
0
2
4
6
0
100
200
300
400
Time (s)

FIG. 29F

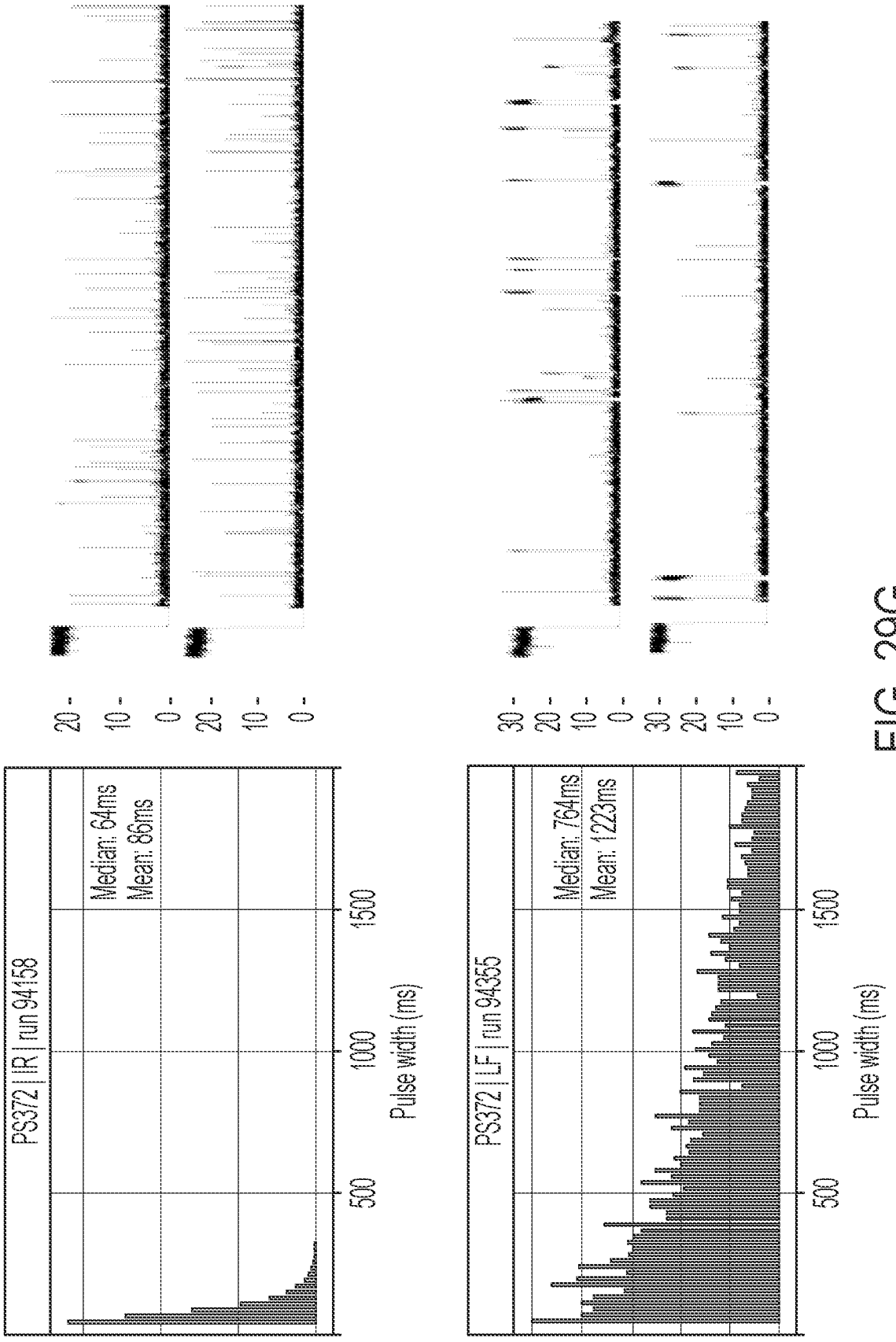

PS372 | IR | run 94158
Median: 64ms
Mean: 86ms
Pulse width (ms)
500
1000
1500
20
10
0
PS372 | LF | run 94355
Median: 764ms
Mean: 1223ms
Pulse width (ms)
500
1000
1500
30
20
10
0

FIG. 29G

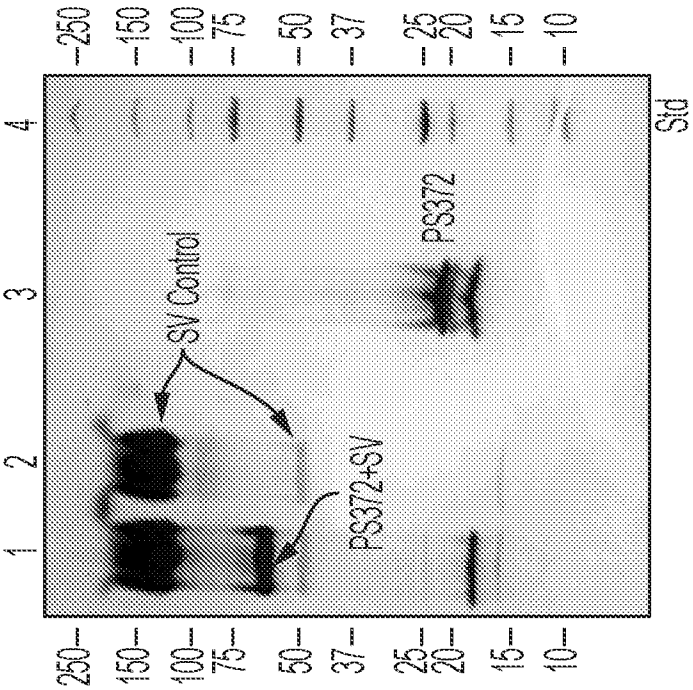


FIG. 30A

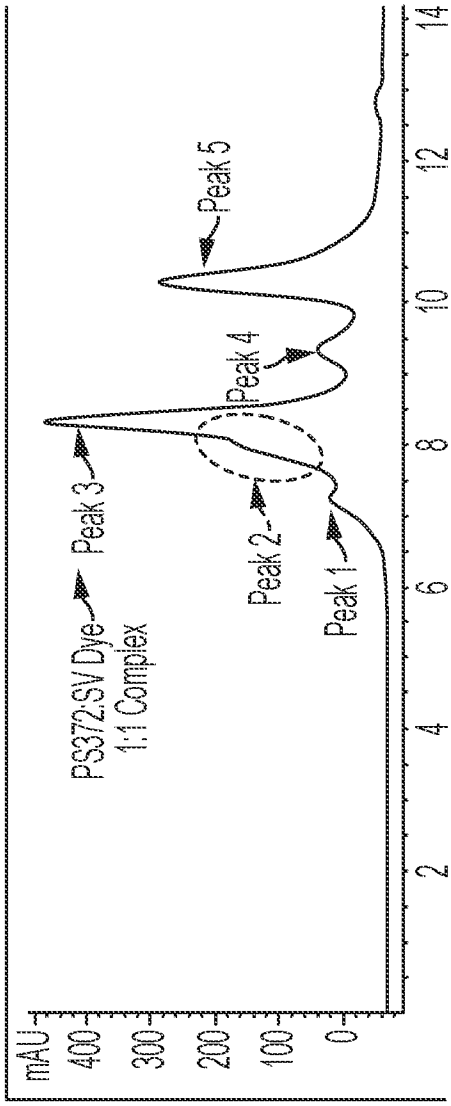


FIG. 30B

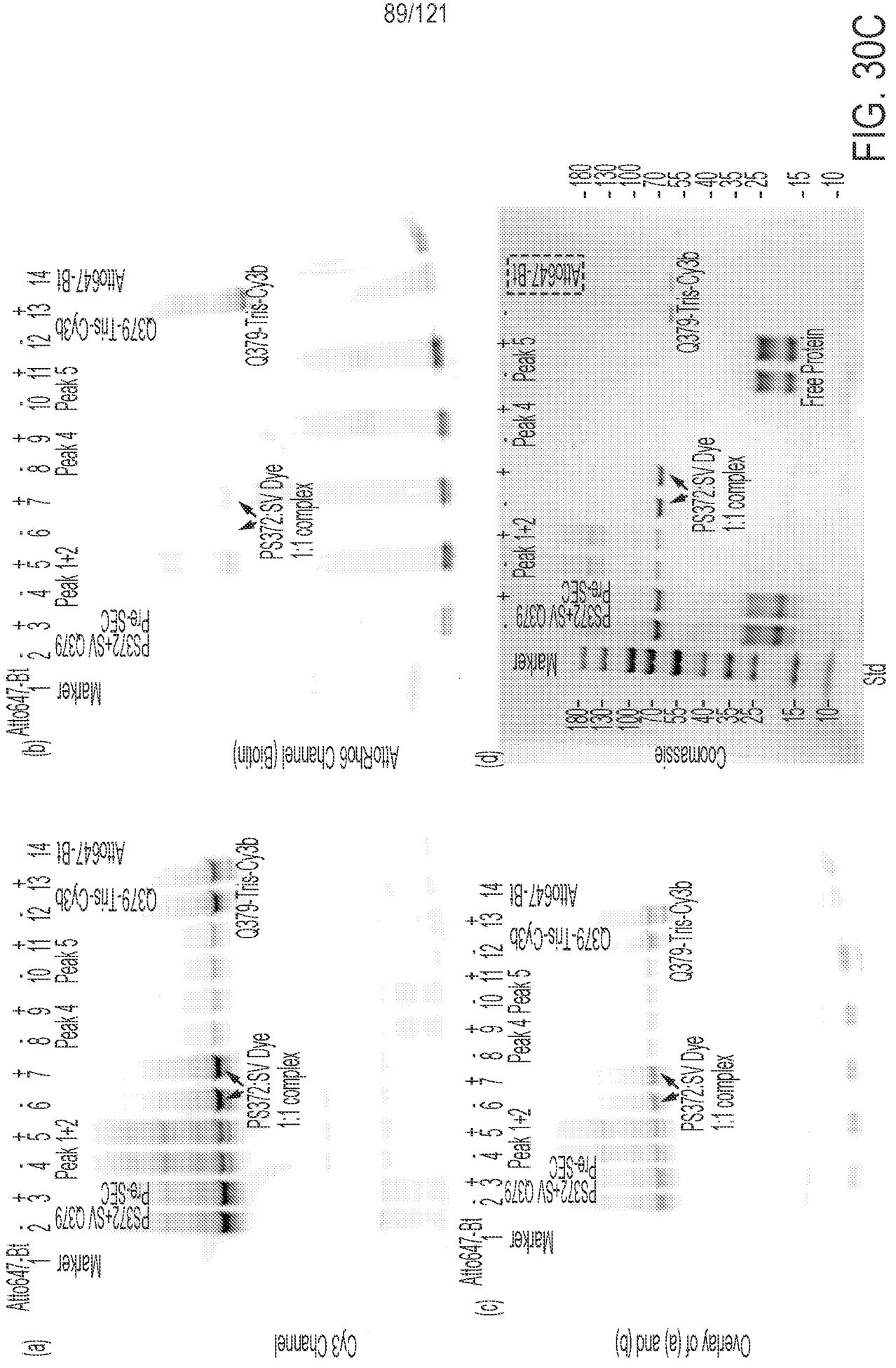

(a)
Atto647-Bt
Marker
PS372+SV Q379 Pre-SEC
Peak 1+2
Peak 4
Peak 5
Q379-Tris-Cy3b
Atto647-Bt
PS372:SV Dye 1:1 complex
Q379-Tris-Cy3b
Cy3 Channel
(b)
Atto647-Bt
Marker
PS372+SV Q379 Pre-SEC
Peak 1+2
Peak 4
Peak 5
Q379-Tris-Cy3b
Atto647-Bt
PS372:SV Dye 1:1 complex
Q379-Tris-Cy3b
AttoRhob Channel (Biotin)
(c)
Atto647-Bt
Marker
PS372+SV Q379 Pre-SEC
Peak 1+2
Peak 4 Peak 5
Q379-Tris-Cy3b
Atto647-Bt
PS372:SV Dye 1:1 complex
Q379-Tris-Cy3b
Overlay of (a) and (b)
(d)
Coomassie
Marker
PS372+SV Q379 Pre-SEC
Peak 1+2
Peak 4
Peak 5
Atto647-Bt
PS372:SV Dye 1:1 complex
Q379-Tris-Cy3b
Free Protein
Std
180
130
100
70
55
40
35
25
15
10

FIG. 30C

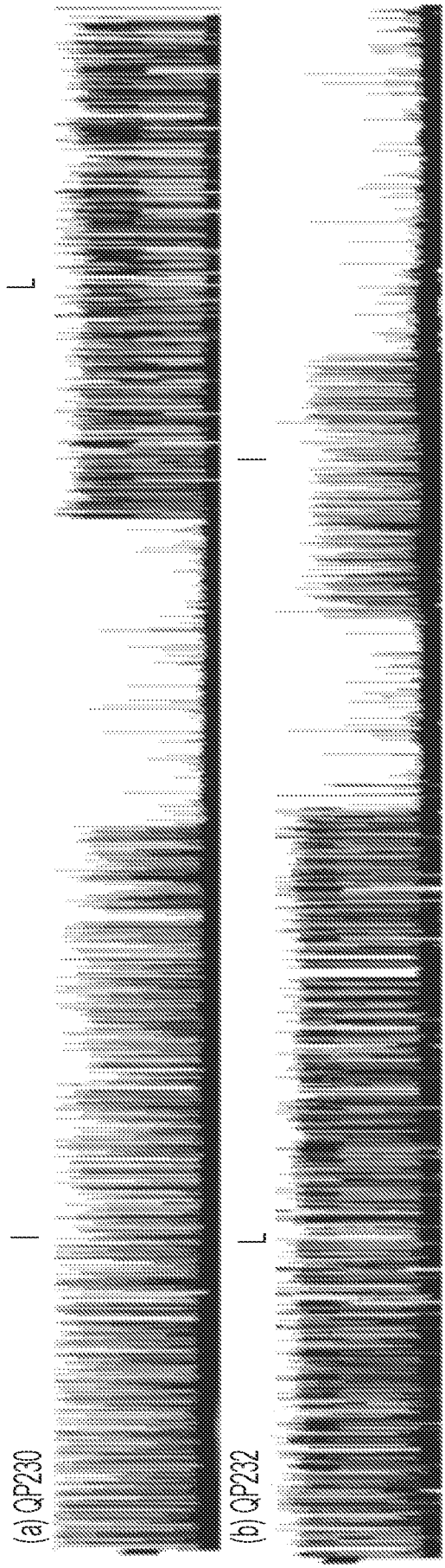


FIG. 30D

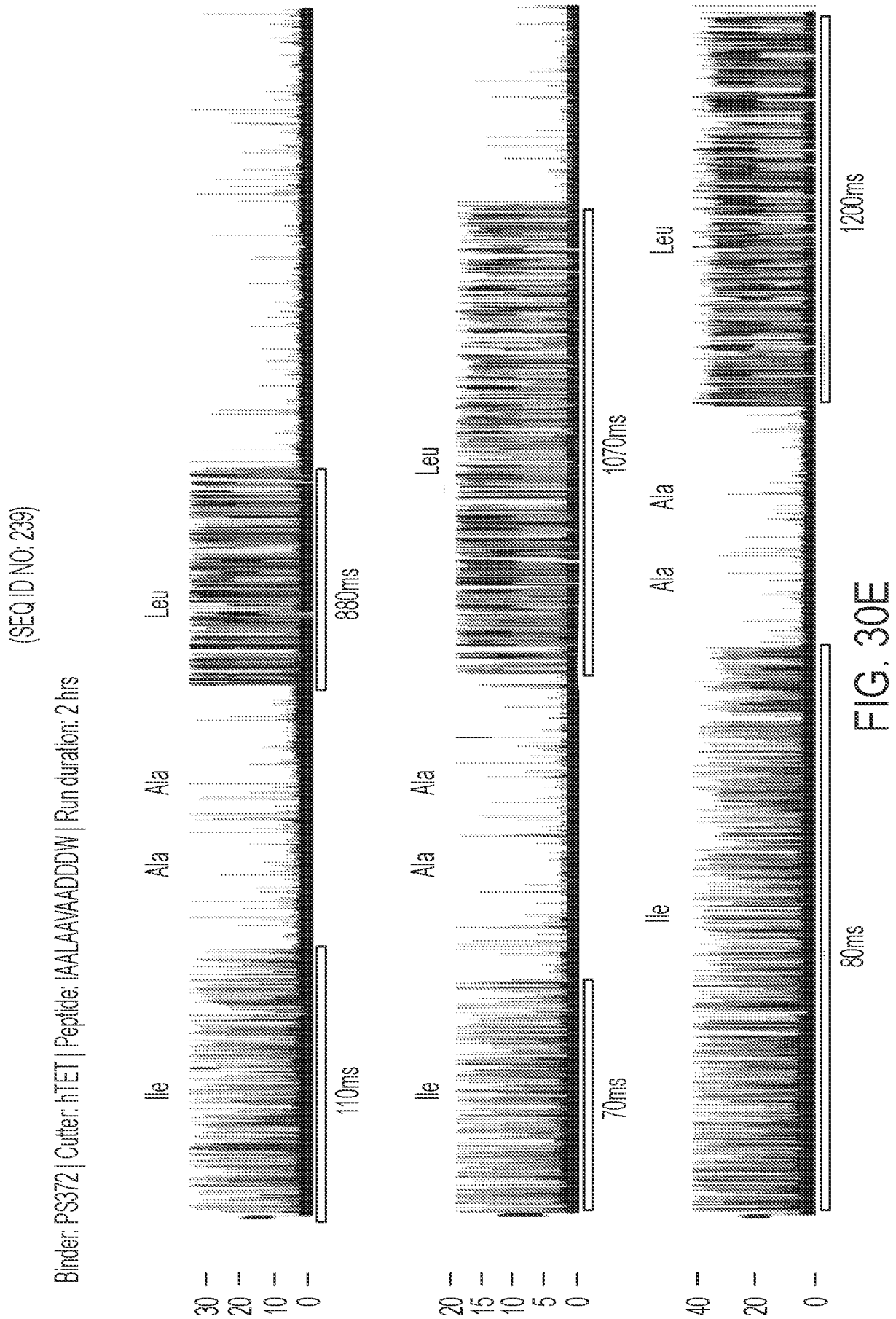

Binder: PS372 | Cutter: hTET | Peptide: IAALAAVAADDDW | Run duration: 2 hrs
(SEQ ID NO: 239)
Ile
Ala
Ala
Leu
110ms
880ms
30
20
10
0
Ile
Ala
Ala
Leu
70ms
1070ms
20
15
10
5
0
Ile
Ala
Ala
Leu
80ms
1200ms
40
20
0

FIG. 30E

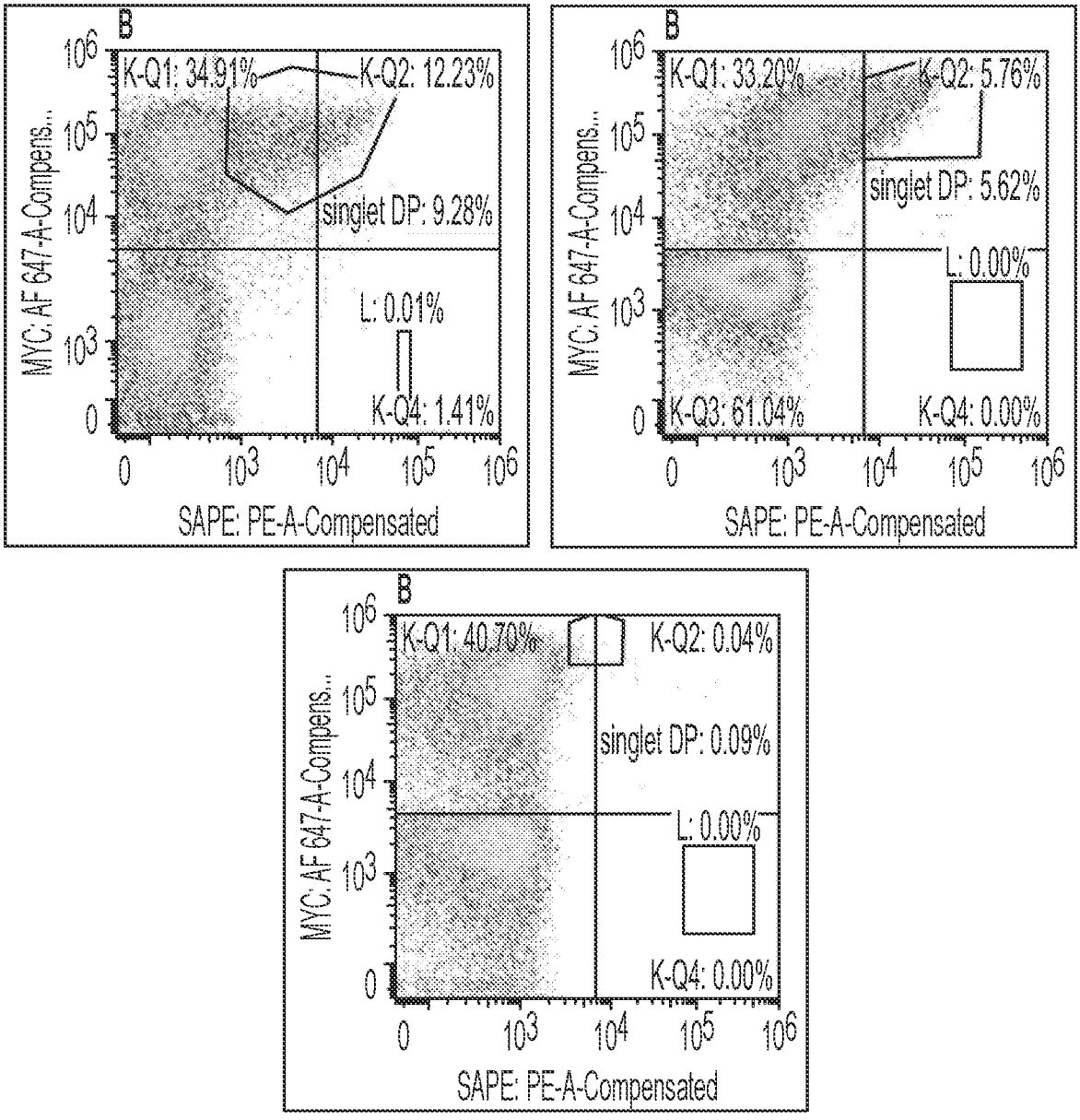

B
K-Q1: 34.91%
K-Q2: 12.23%
singlet DP: 9.28%
L: 0.01%
K-Q4: 1.41%
MYC: AF 647-A-Compens...
SAPE: PE-A-Compensated
B
K-Q1: 33.20%
K-Q2: 5.76%
singlet DP: 5.62%
L: 0.00%
K-Q3: 61.04%
K-Q4: 0.00%
MYC: AF 647-A-Compens...
SAPE: PE-A-Compensated
B
K-Q1: 40.70%
K-Q2: 0.04%
singlet DP: 0.09%
L: 0.00%
K-Q4: 0.00%
MYC: AF 647-A-Compens...
SAPE: PE-A-Compensated

FIG. 31A

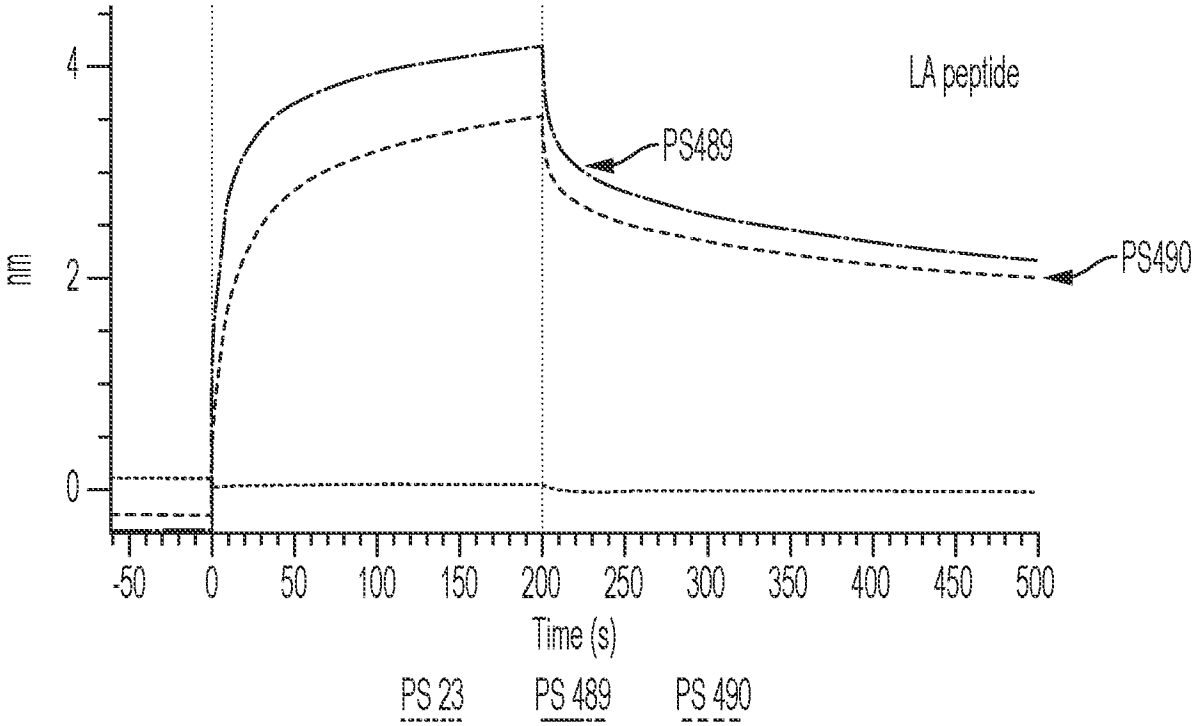


FIG. 31B

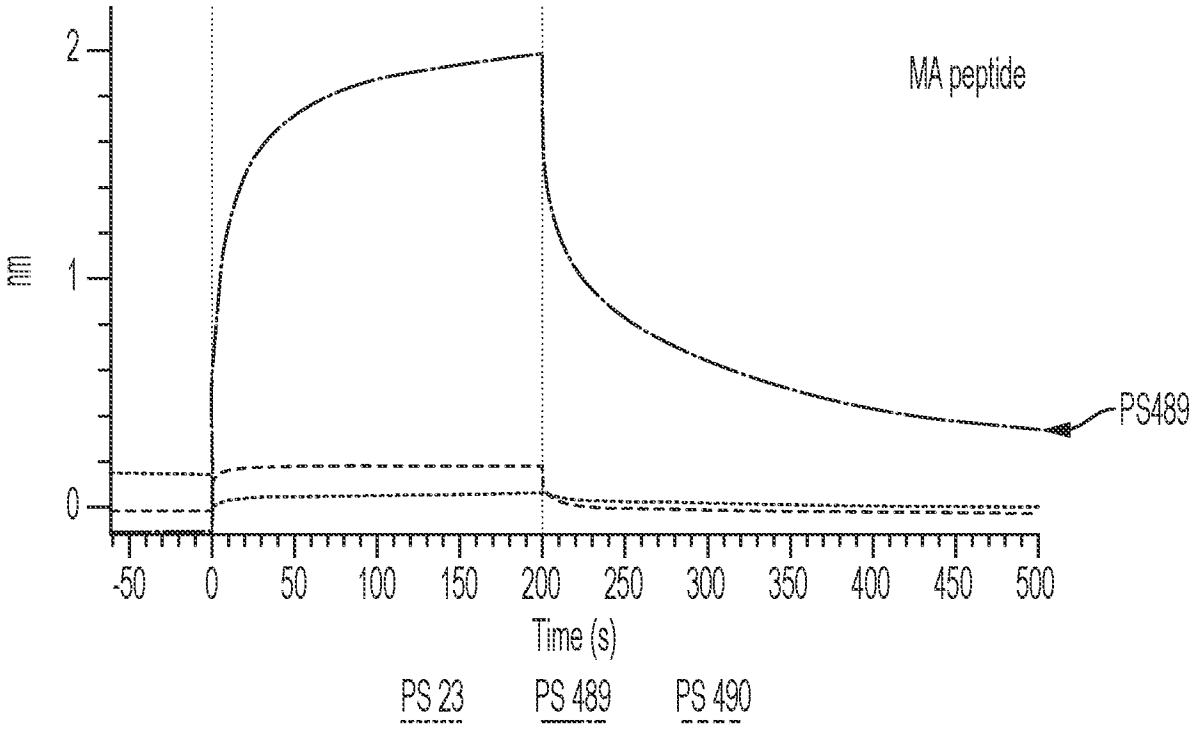


FIG. 31C

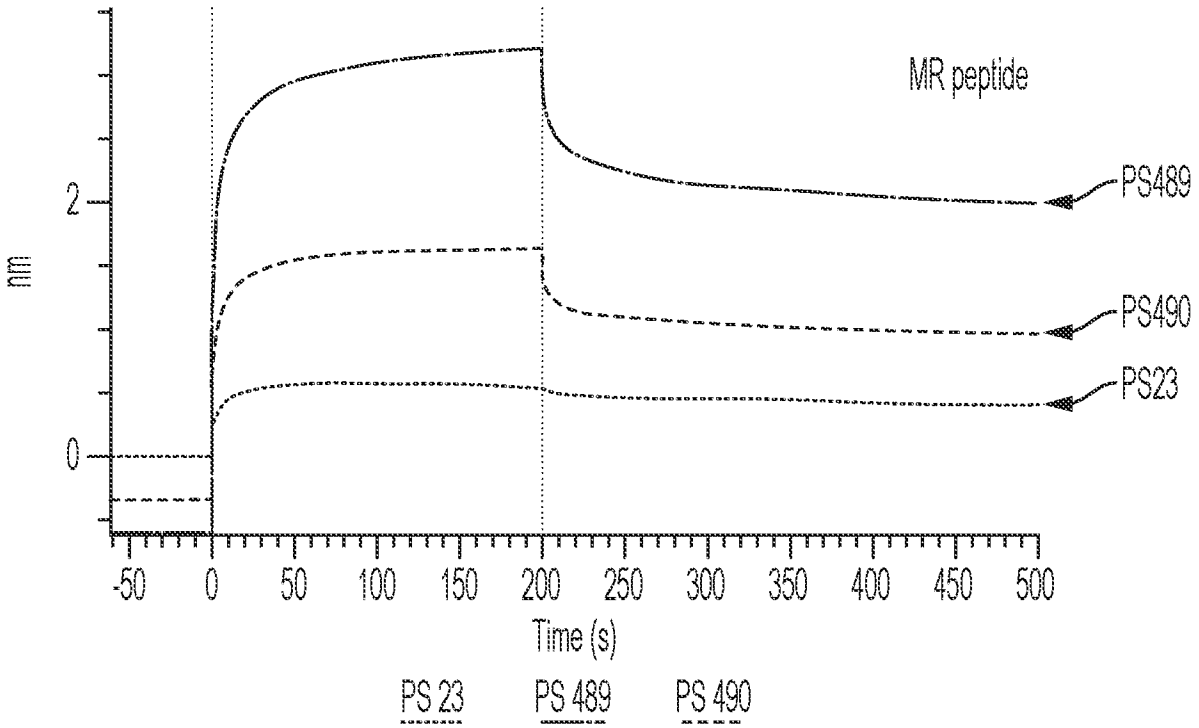


FIG. 31D

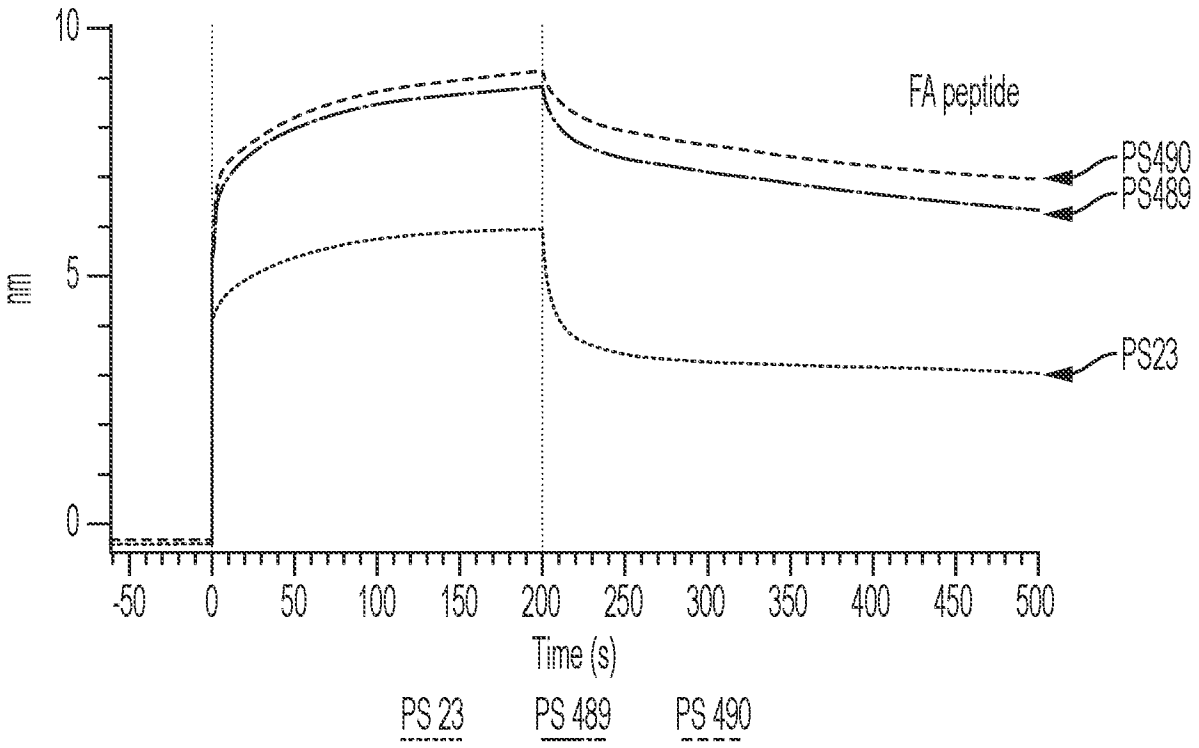


FIG. 31E

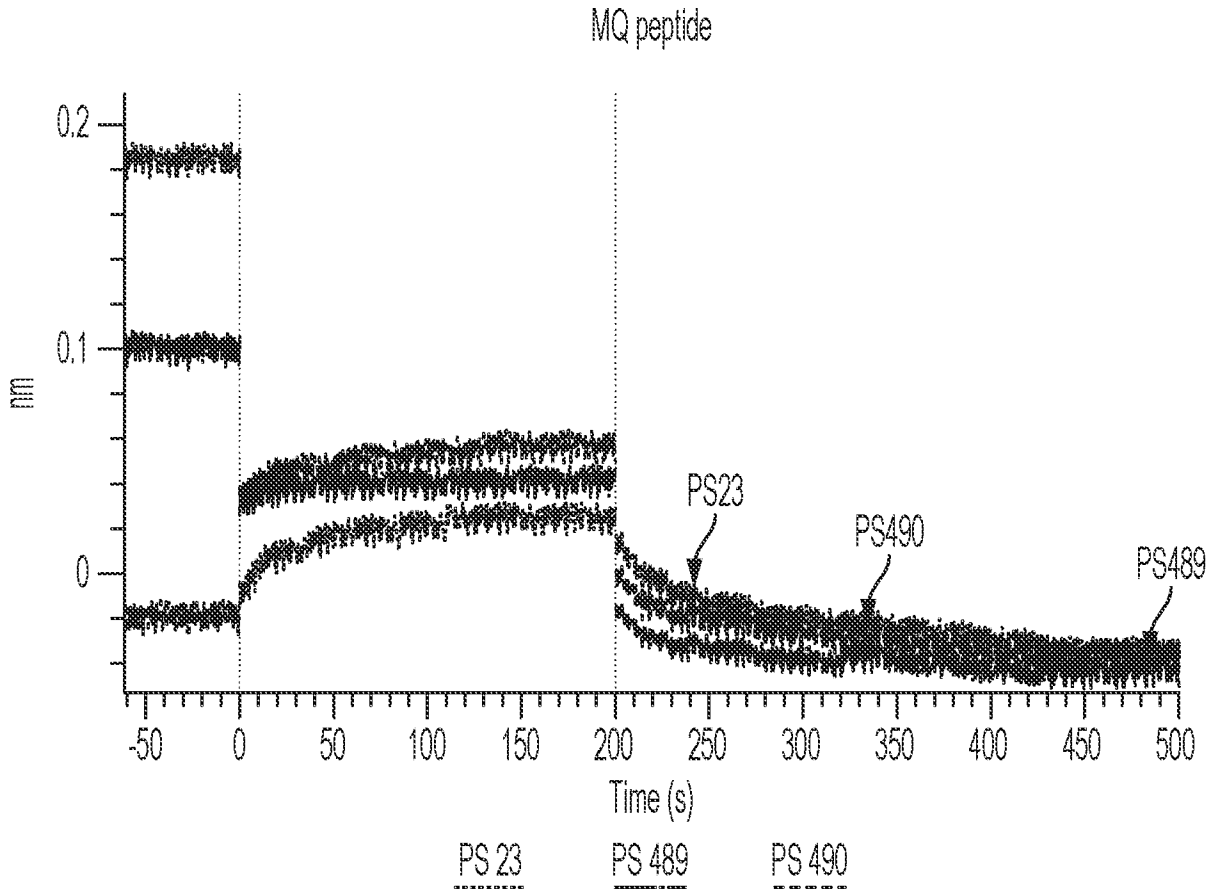

MQ peptide
0.2
0.1
0
nm
PS23
PS490
PS489
-50
0
50
100
150
200
250
300
350
400
450
500
Time (s)
PS 23
PS 489
PS 490

FIG. 31F

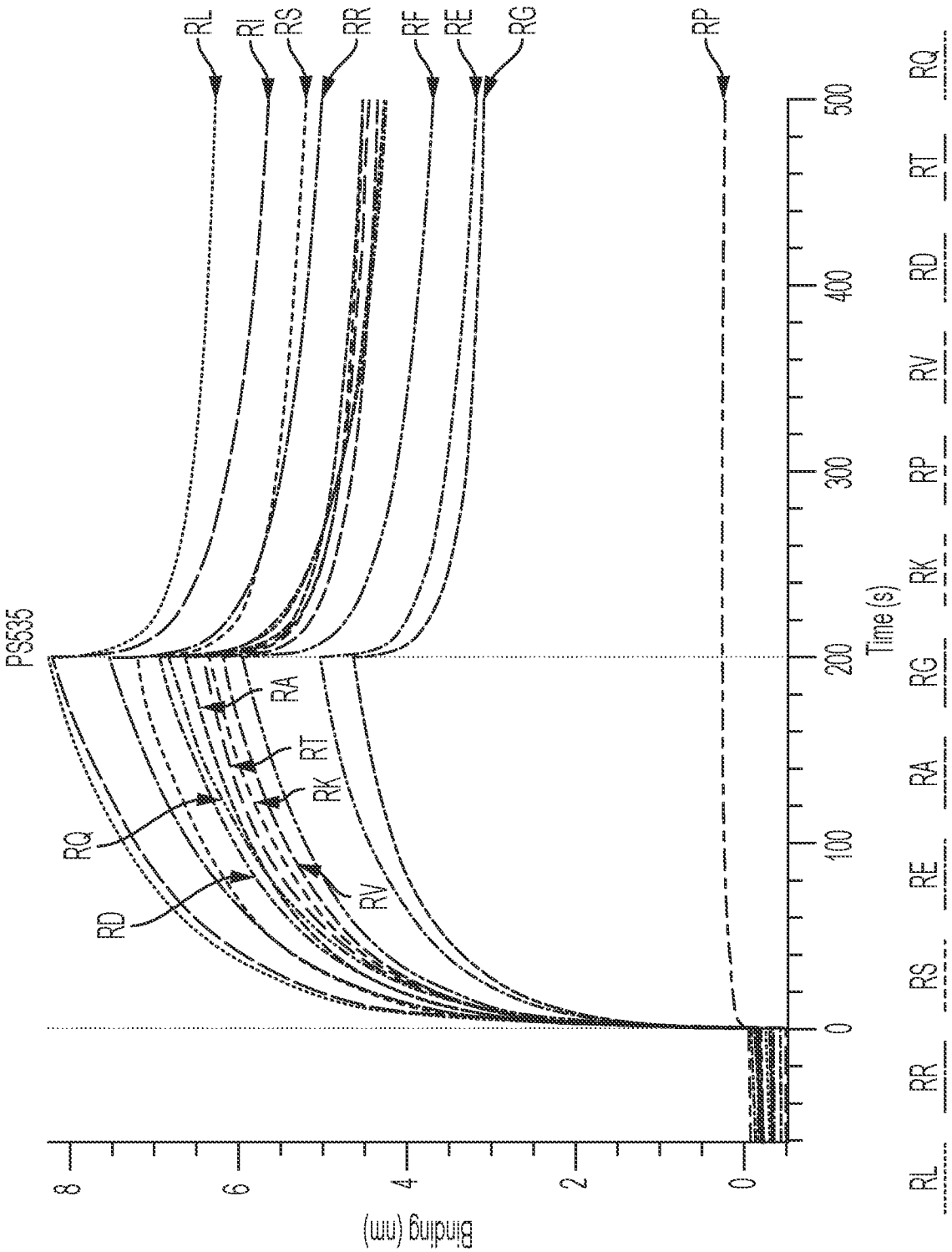

RL
RI
RS
RR
RF
RE
RG
RP
PS535
RA
RT
RK
RQ
RV
RD
Binding (mm)
8
6
4
2
0
Time (s)
0
100
200
300
400
500
RL RR RS RE RA RG RK RP RV RD RT RQ
RI
RF

FIG. 32A

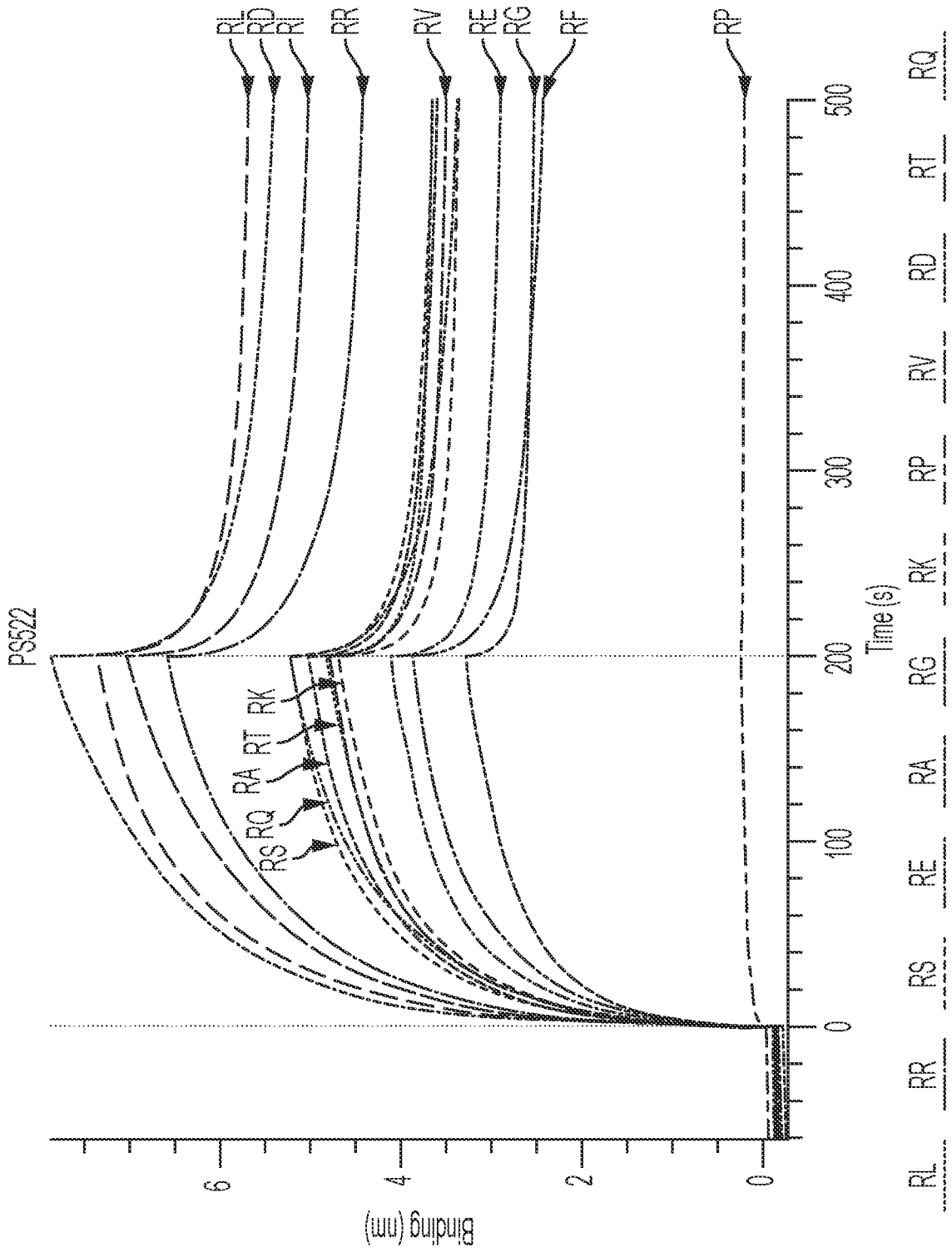

PS522
RL
RD
RI
RR
RV
RE
RG
RF
RP
RS
RQ
RA
RT
RK
Binding (nm)
6
4
2
0
Time (s)
0
100
200
300
400
500
RL
RR
RS
RE
RA
RG
RK
RP
RV
RD
RT
RQ
RI
RF

FIG. 32B

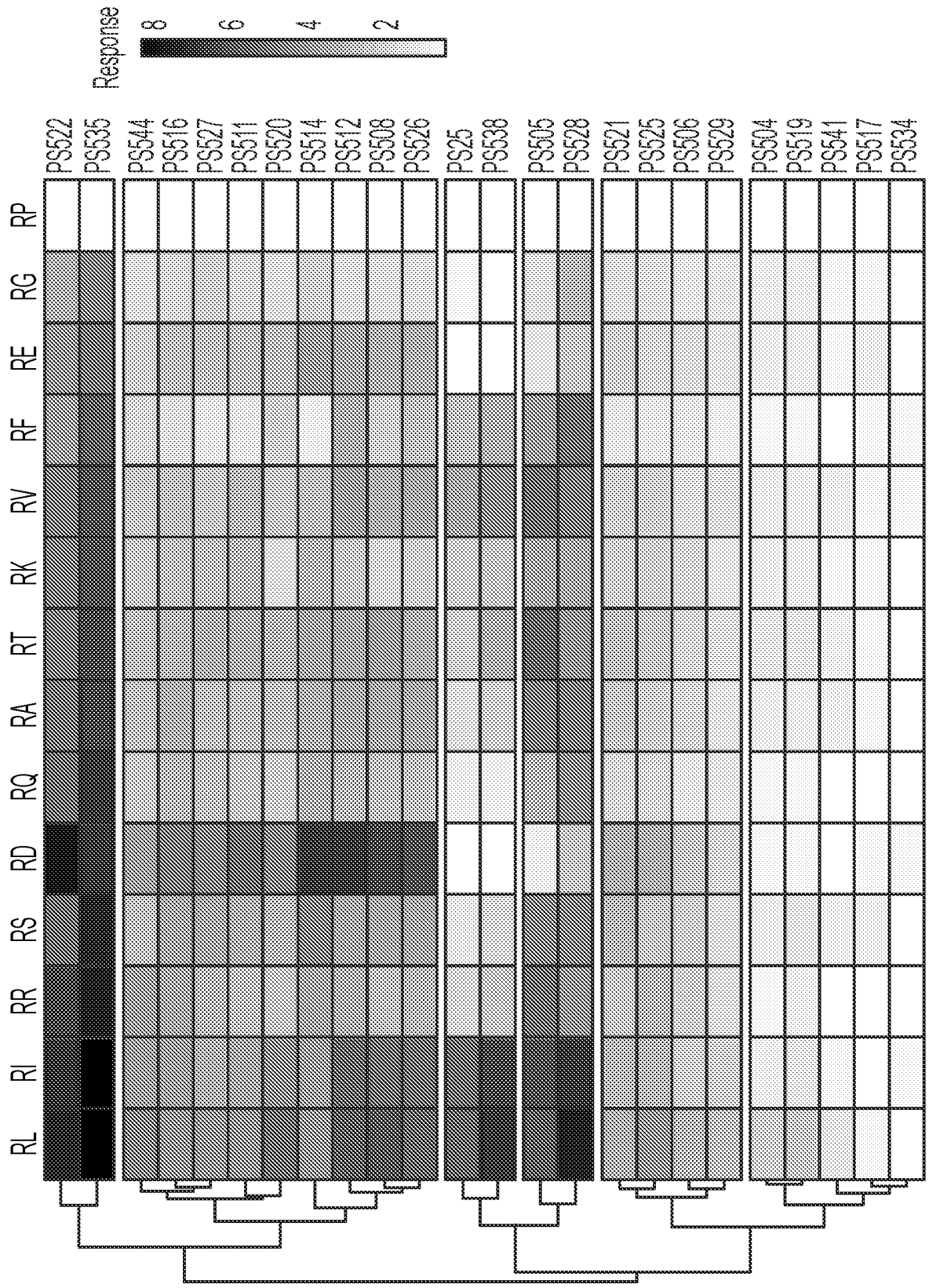

Response
8
6
4
2
PS522
PS535
PS544
PS516
PS527
PS511
PS520
PS514
PS512
PS508
PS526
PS25
PS538
PS505
PS528
PS521
PS525
PS506
PS529
PS504
PS519
PS541
PS517
PS534
RL
RI
RR
RS
RD
RQ
RA
RT
RK
RV
RF
RE
RG
RP

FIG. 32C

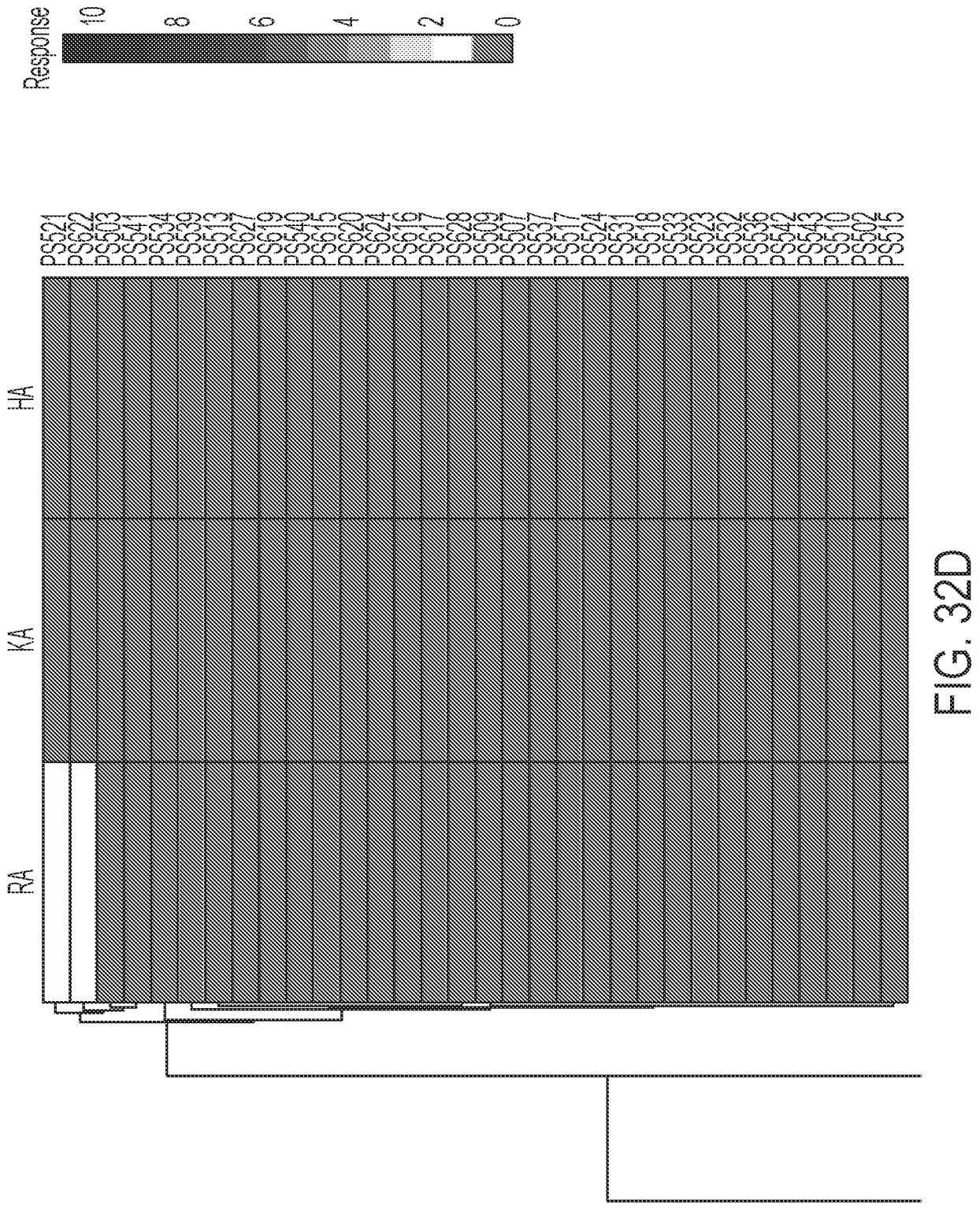

Response
10
8
6
4
2
0
PS521
PS622
PS503
PS541
PS534
PS539
PS513
PS527
PS519
PS540
PS515
PS520
PS624
PS516
PS617
PS528
PS509
PS507
PS537
PS517
PS524
PS531
PS518
PS533
PS523
PS532
PS536
PS542
PS543
PS510
PS502
PS515
HA
KA
RA

FIG. 32D

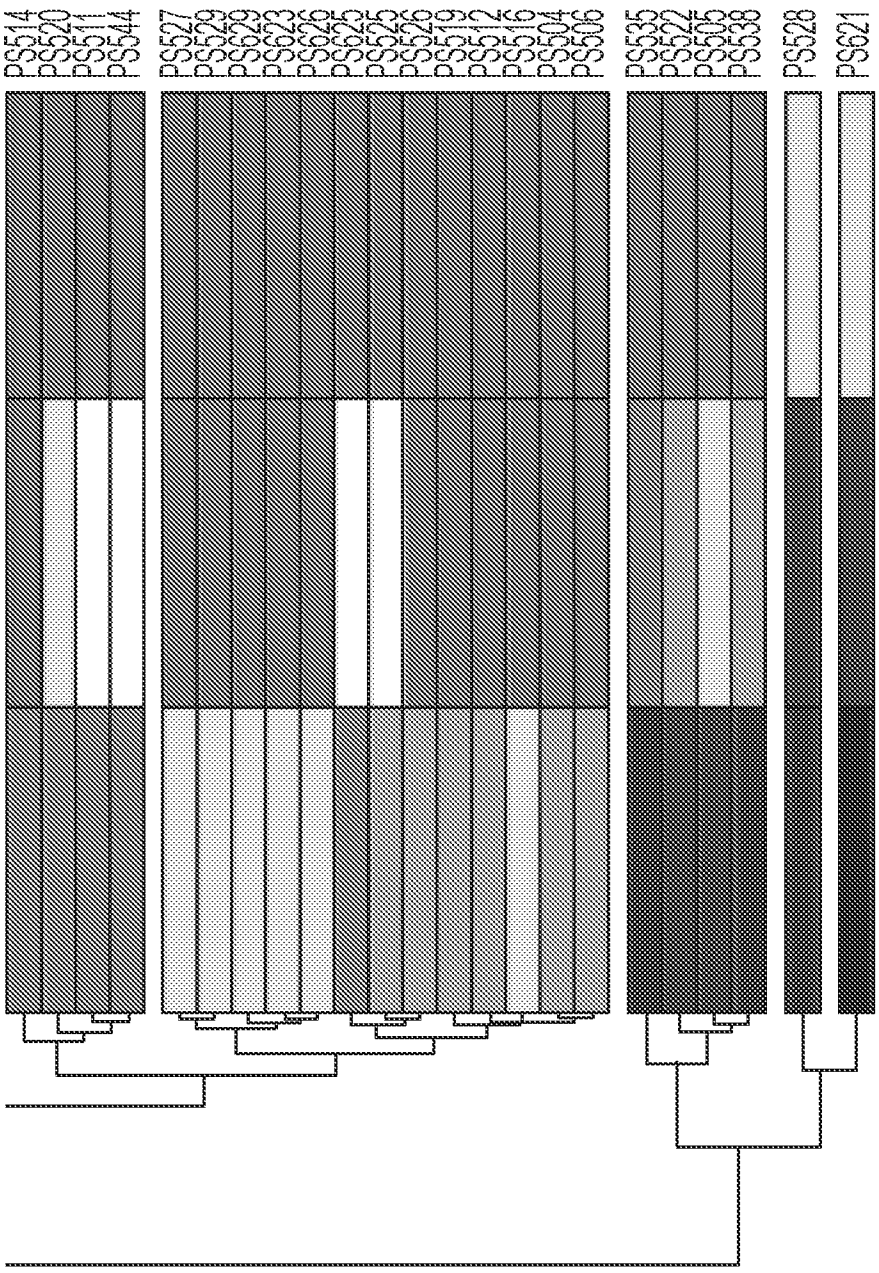


FIG. 32D
CONTINUED

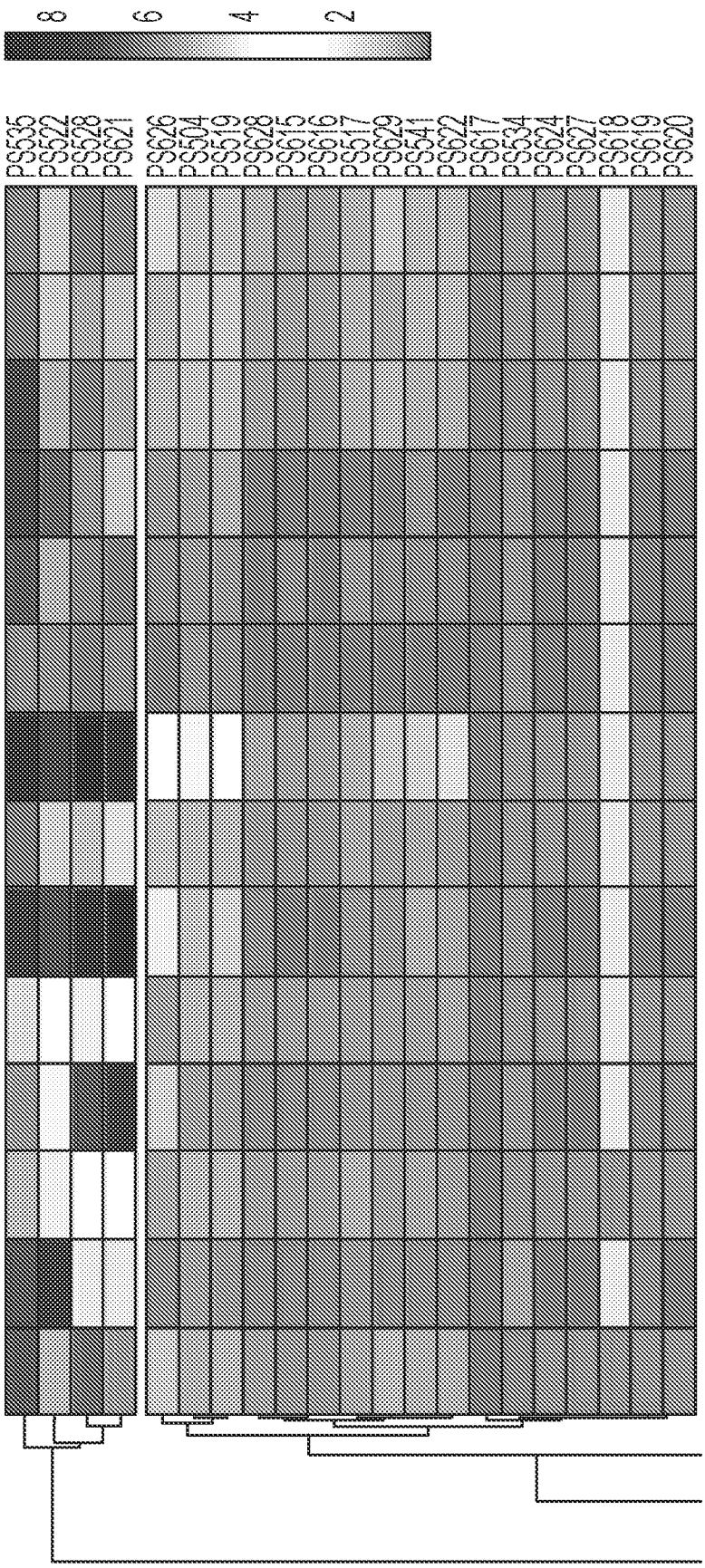

8
6
4
2
PS535
PS522
PS528
PS621
PS626
PS504
PS519
PS628
PS615
PS616
PS517
PS629
PS541
PS622
PS617
PS534
PS624
PS627
PS618
PS619
PS620

FIG. 32E

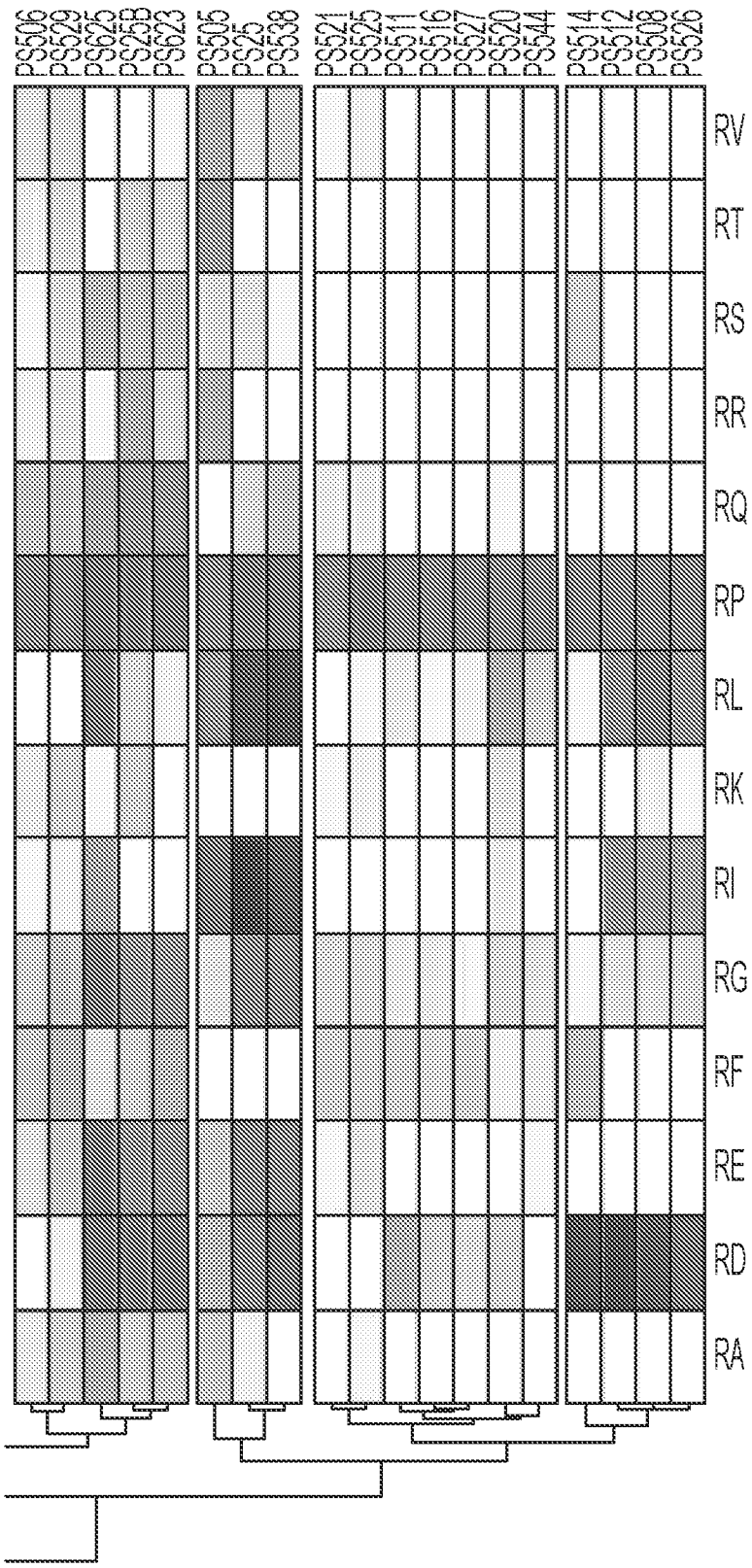


FIG. 32E
CONTINUED

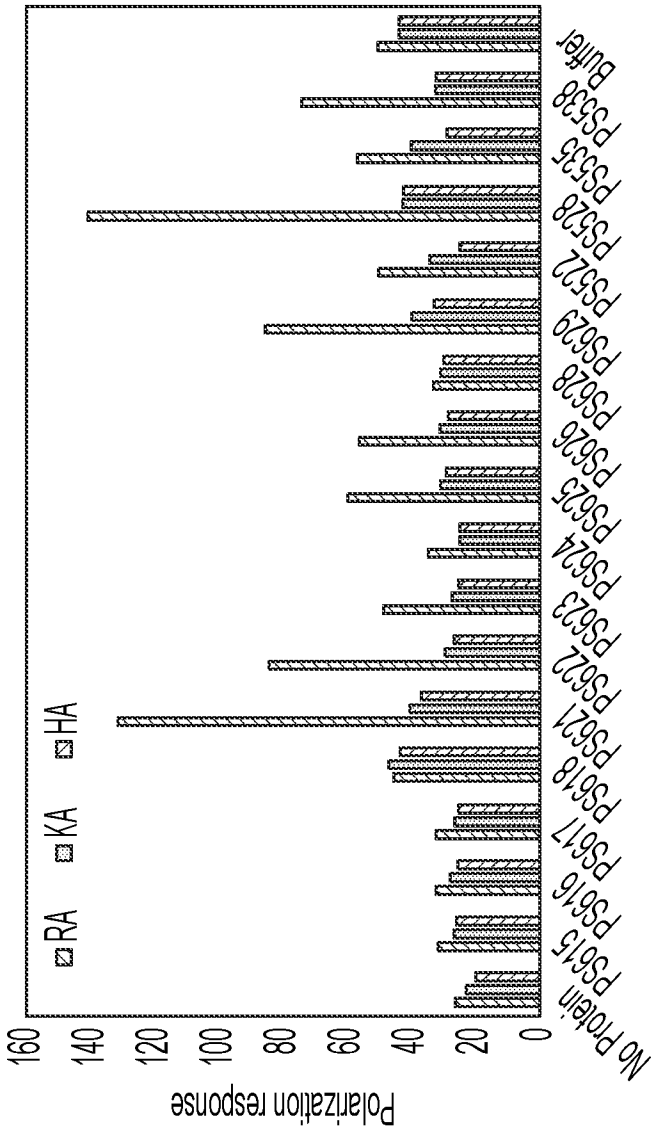

Polarization response
160
140
120
100
80
60
40
20
0
RA
KA
HA
No Protein
PS615
PS616
PS617
PS618
PS621
PS622
PS623
PS624
PS625
PS626
PS628
PS629
PS522
PS528
PS535
PS538
Buffer

FIG. 32F

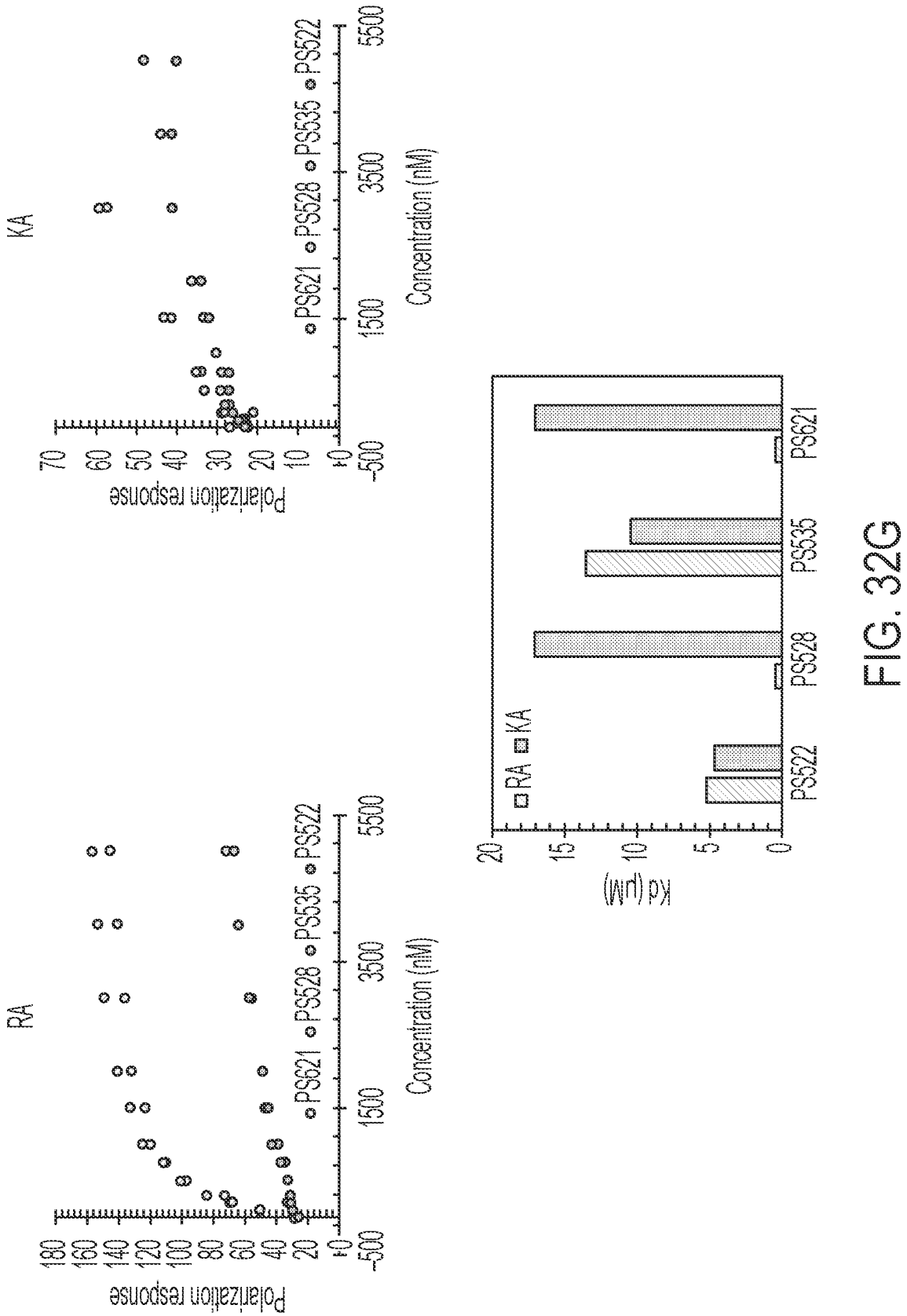

KA
Polarization response
70
60
50
40
30
20
10
0
-500
1500
3500
5500
Concentration (nM)
PS621
PS528
PS535
PS522
RA
Polarization response
180
160
140
120
100
80
60
40
20
0
-500
1500
3500
5500
Concentration (nM)
PS621
PS528
PS535
PS522
RA
KA
Kd (μM)
20
15
10
5
0
PS522
PS528
PS535
PS621

FIG. 32G

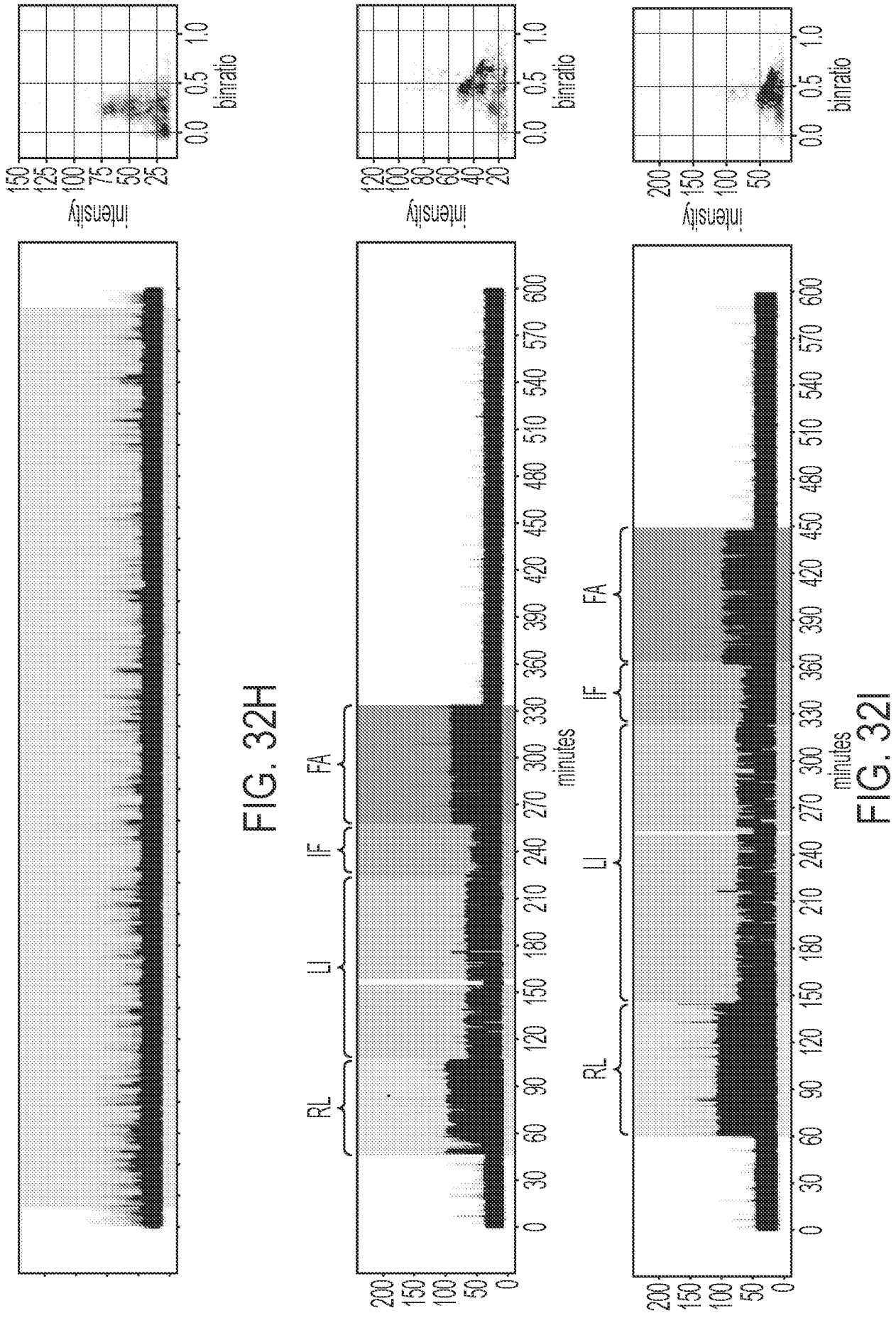


FIG. 32H

FIG. 32I

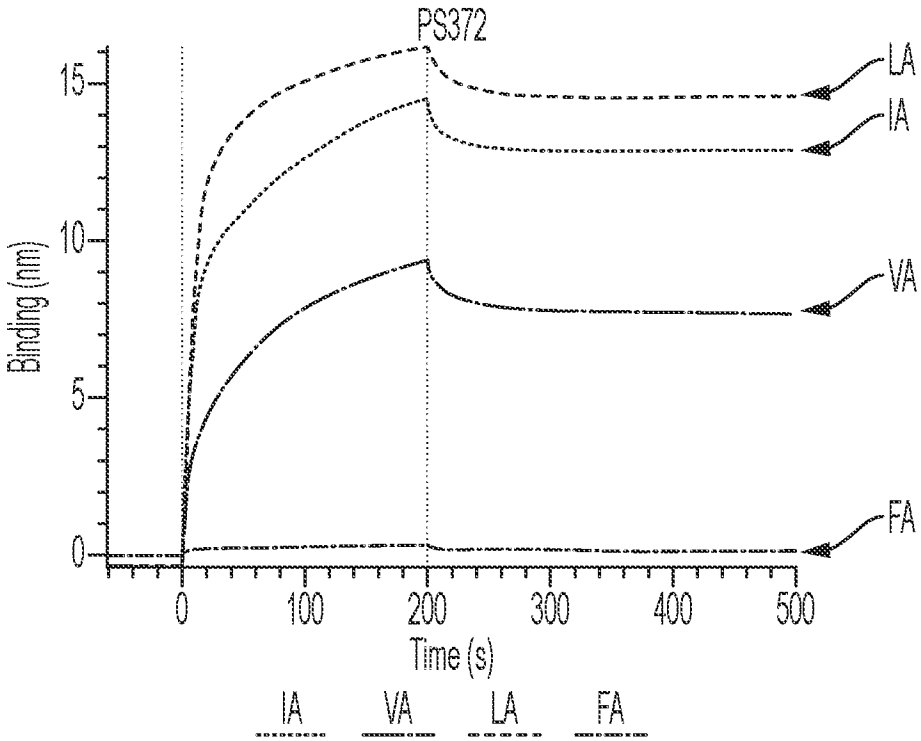

PS372
LA
IA
VA
FA
Binding (nm)
15
10
5
0
0
100
200
300
400
500
Time (s)
IA
VA
LA
FA

FIG. 33A

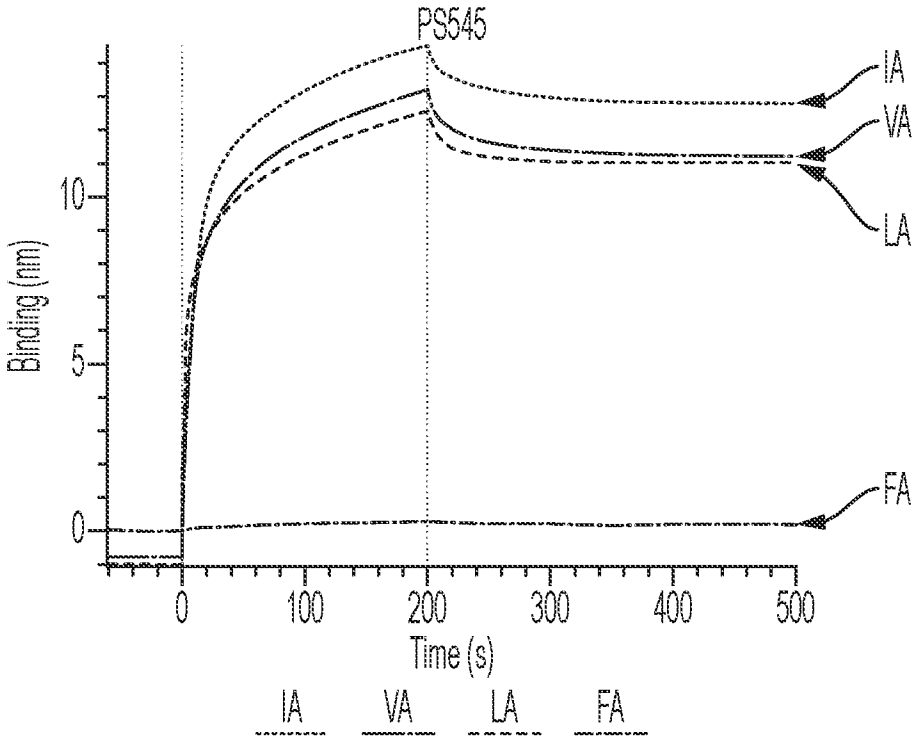


FIG. 33B

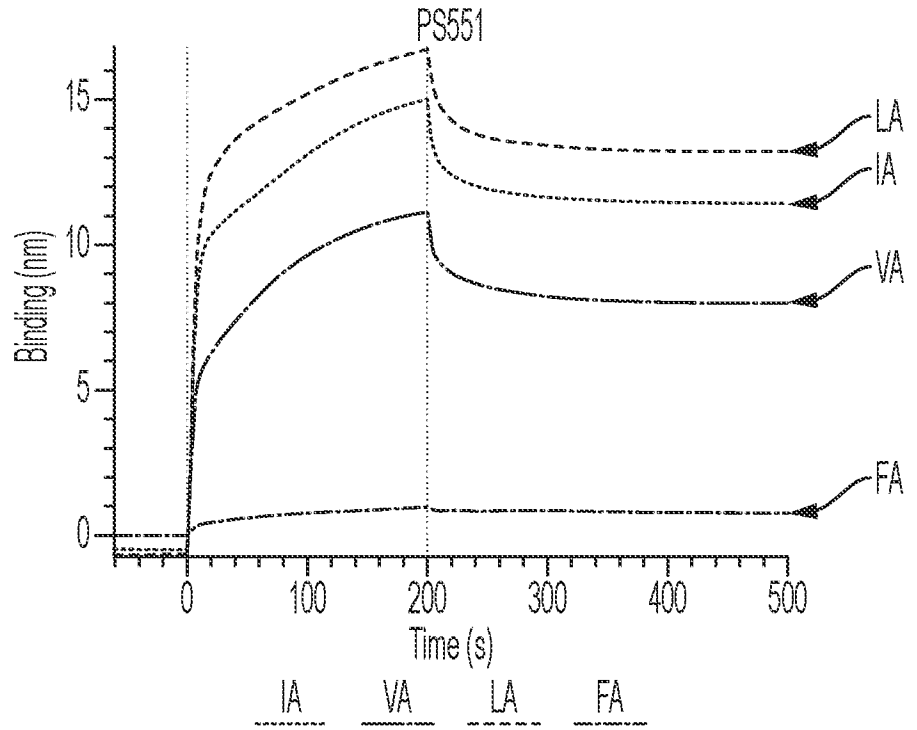


FIG. 33C

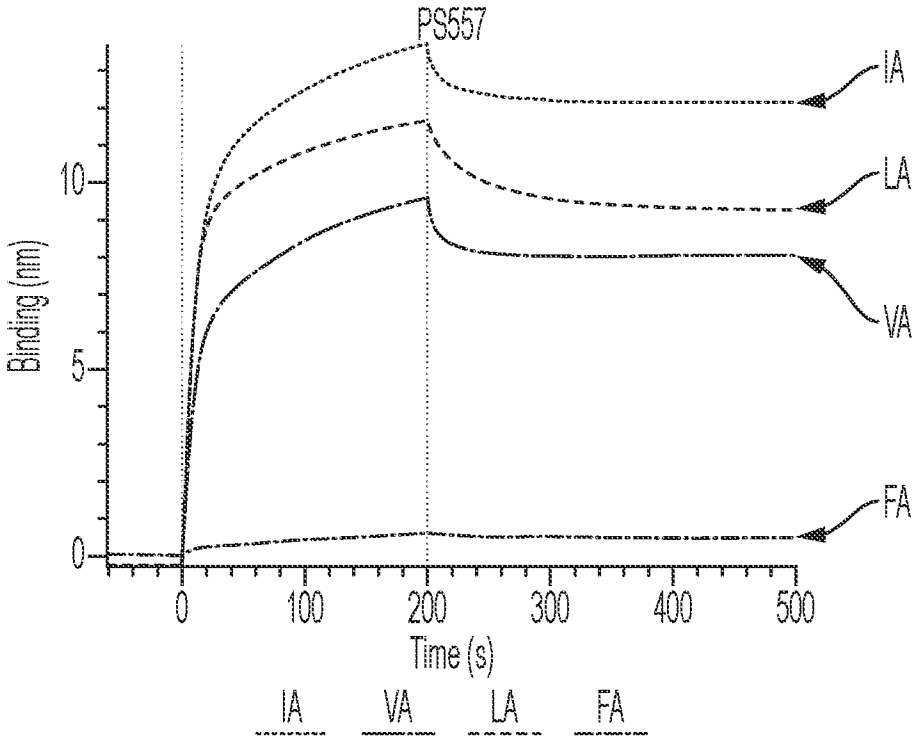


FIG. 33D

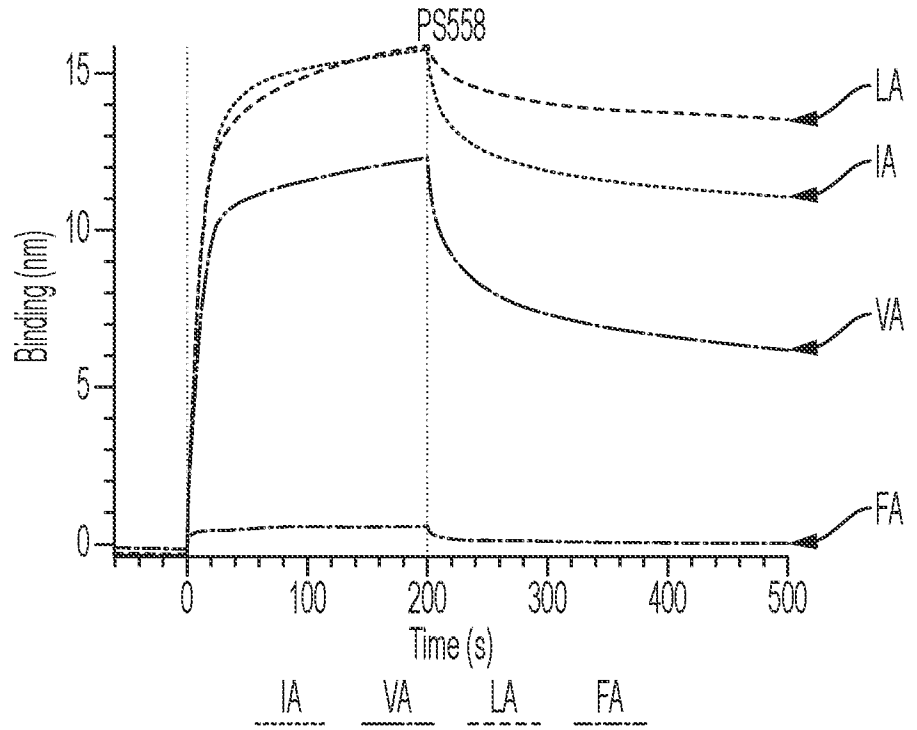


FIG. 33E

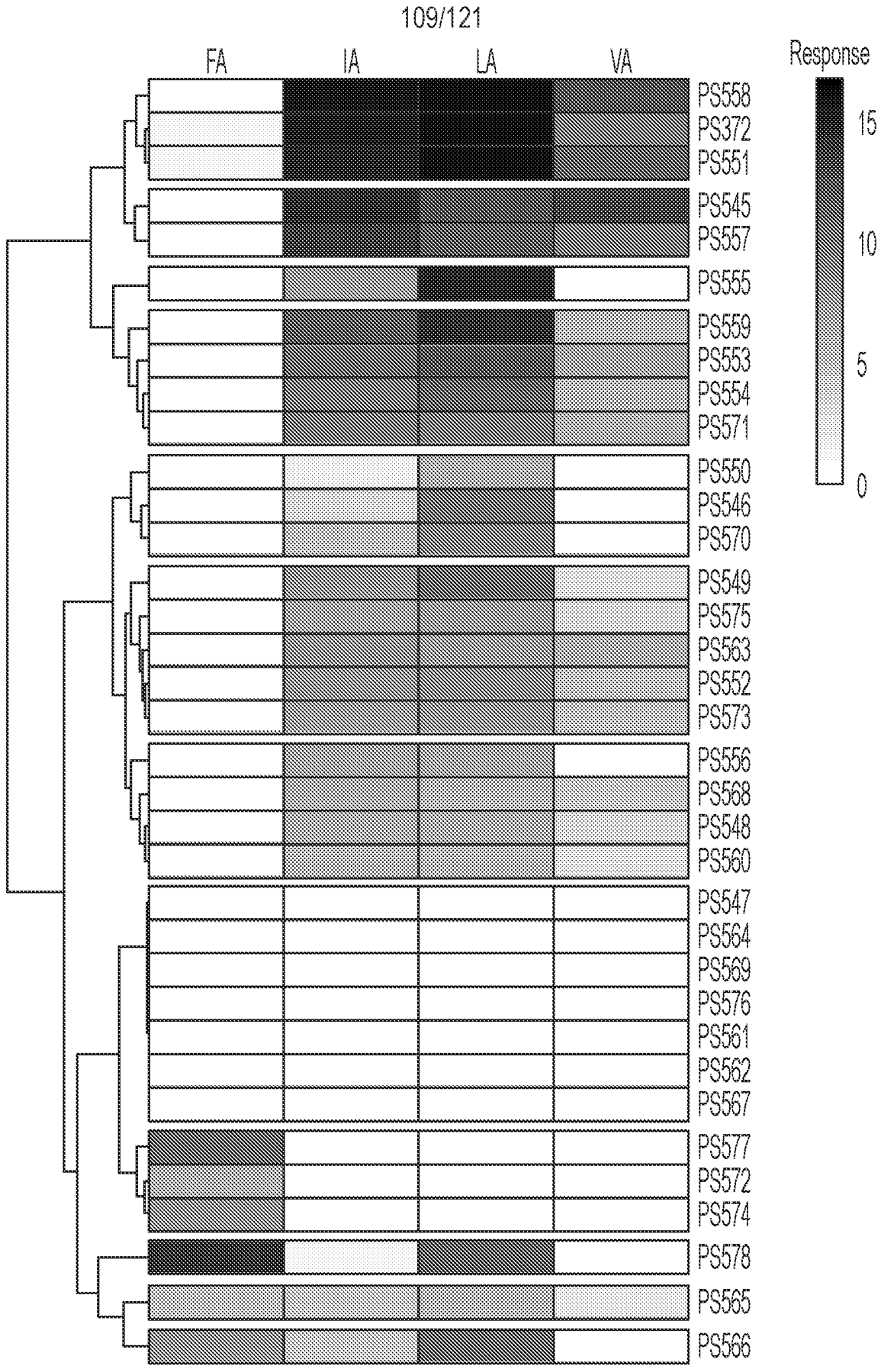

109/121
FA
IA
LA
VA
Response
15
10
5
0
PS558
PS372
PS551
PS545
PS557
PS555
PS559
PS553
PS554
PS571
PS550
PS546
PS570
PS549
PS575
PS563
PS552
PS573
PS556
PS568
PS548
PS560
PS547
PS564
PS569
PS576
PS561
PS562
PS567
PS577
PS572
PS574
PS578
PS565
PS566

FIG. 33F

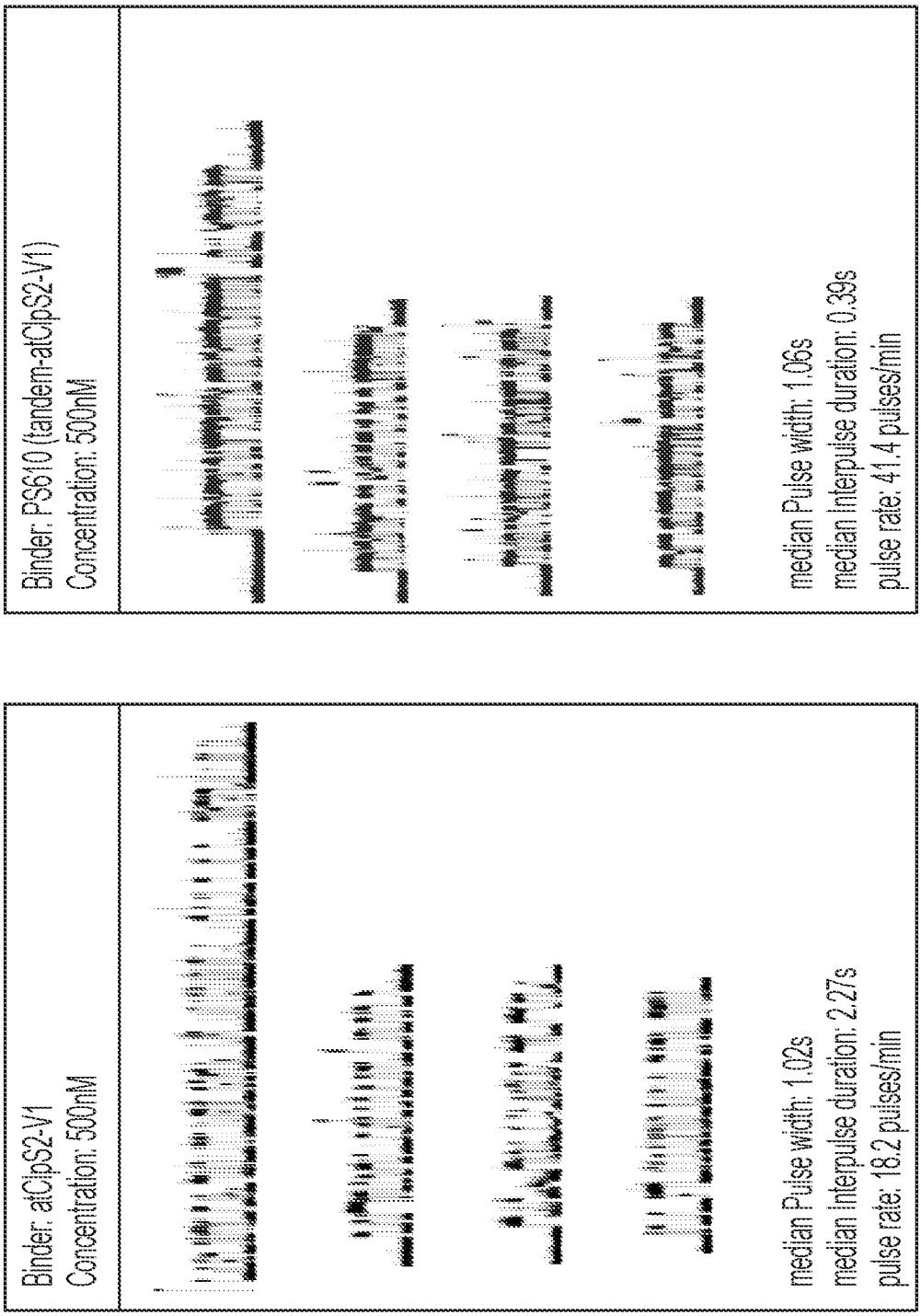

Binder: atClpS2-V1
Concentration: 500nM
median Pulse width: 1.02s
median Interpulse duration: 2.27s
pulse rate: 18.2 pulses/min
Binder: PS610 (tandem-atClpS2-V1)
Concentration: 500nM
median Pulse width: 1.06s
median Interpulse duration: 0.39s
pulse rate: 41.4 pulses/min

FIG. 34A

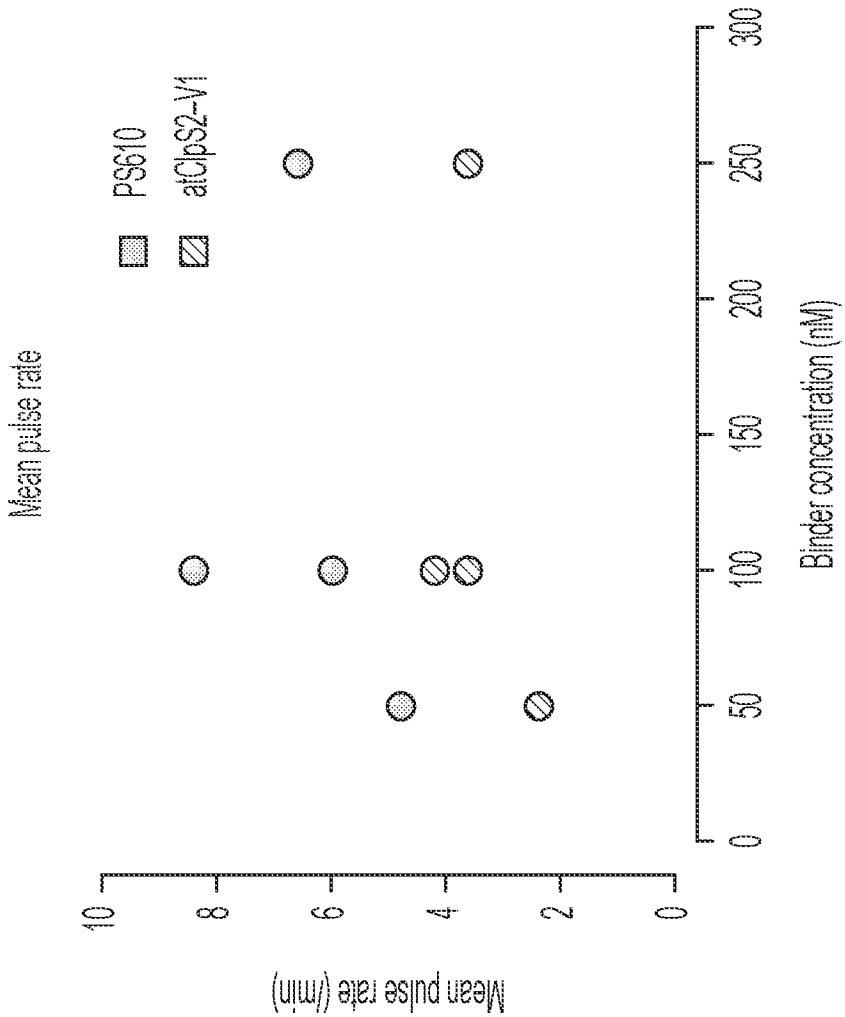
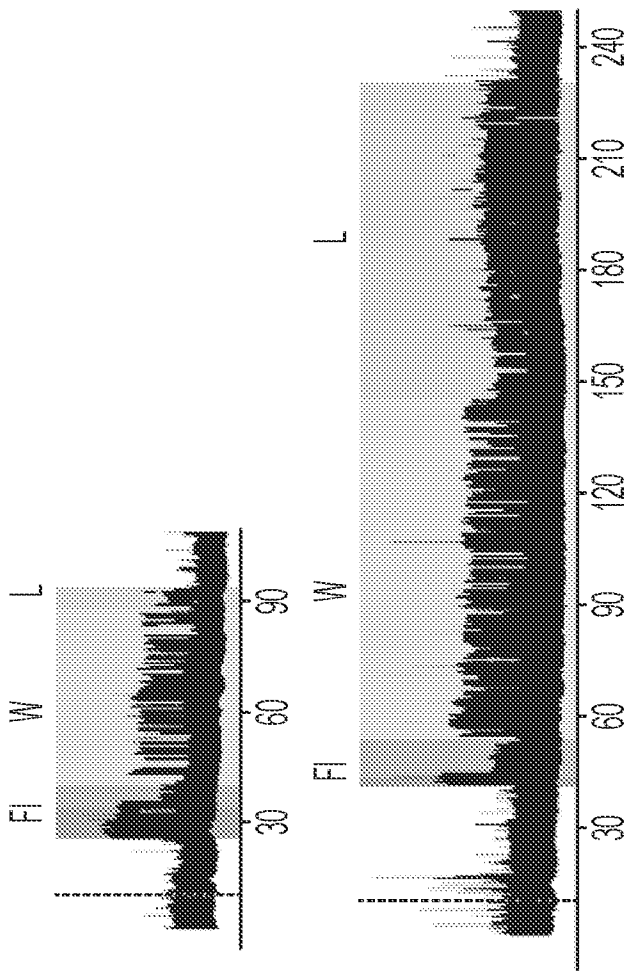
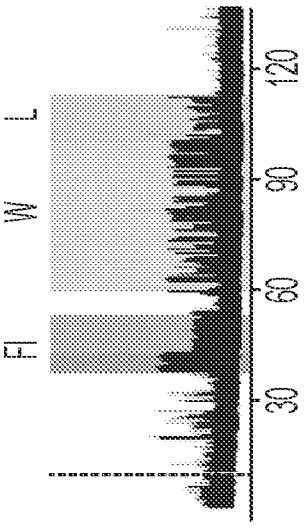

Mean pulse rate
PS610
atClpS2-V1
Mean pulse rate (/min)
10
8
6
4
2
0
0
50
100
150
200
250
300
Binder concentration (nM)

FIG. 34B





Peptide: GLP1 (EFIAWLVK) (SEQ ID NO: 248)
F/Y/W Recognizer: PS610 (tandem atChipS2-V1)
L/I Recognizer: PS372
Cutters: hTET, pfuTET

FIG. 34C

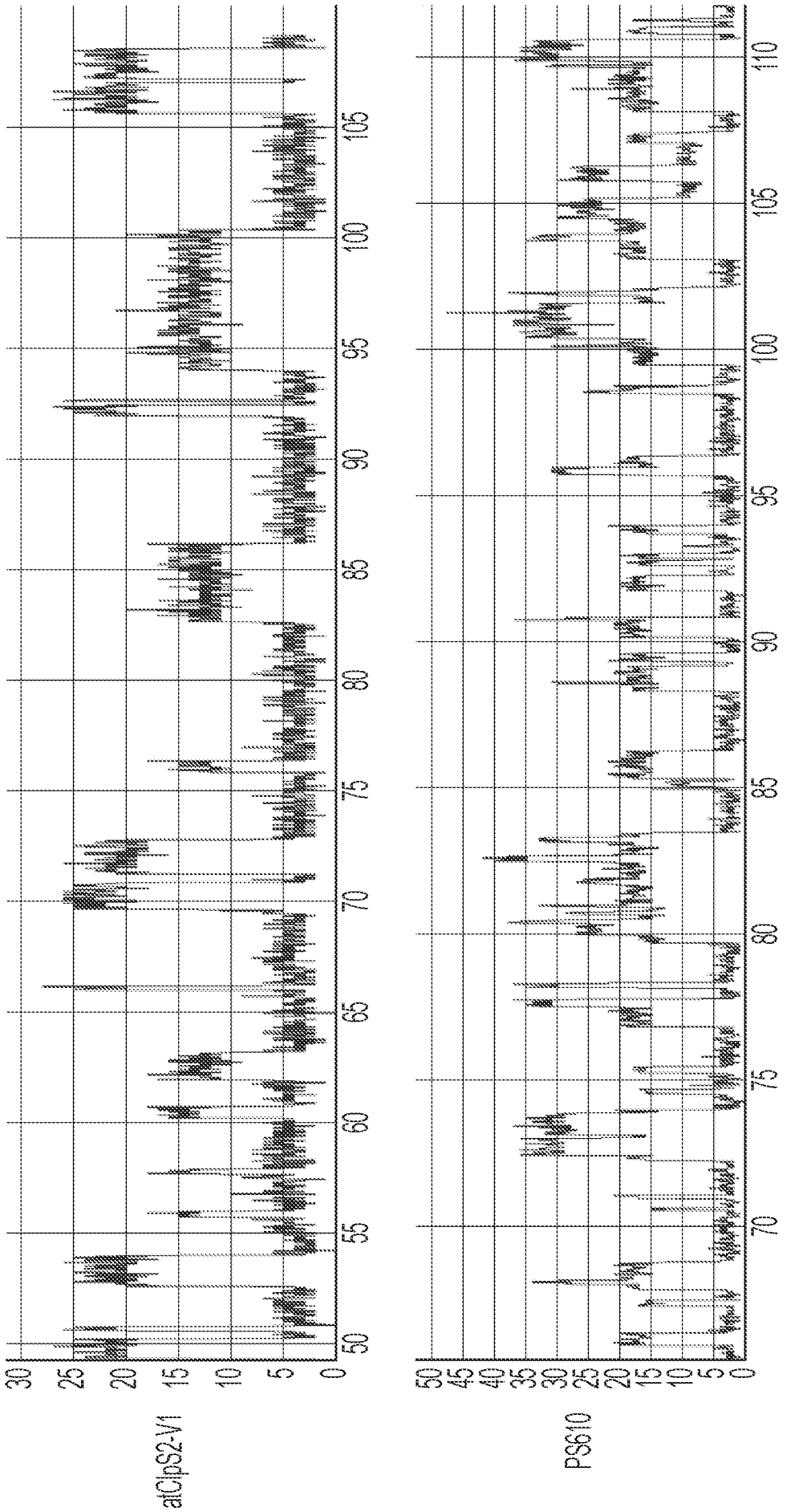


FIG. 34D

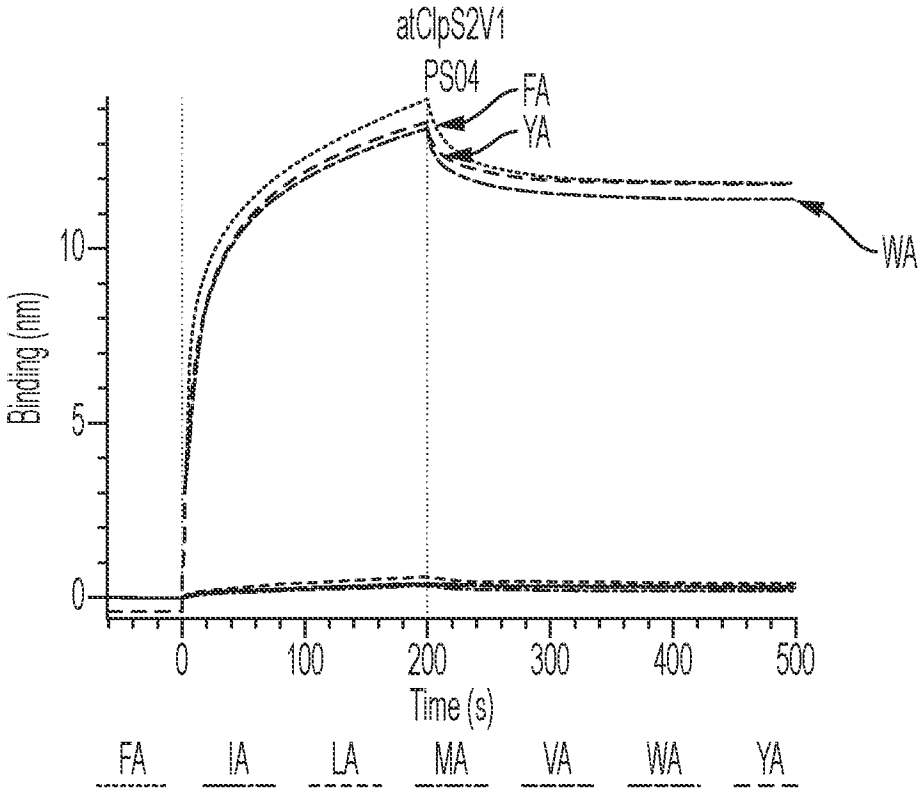


FIG. 35A

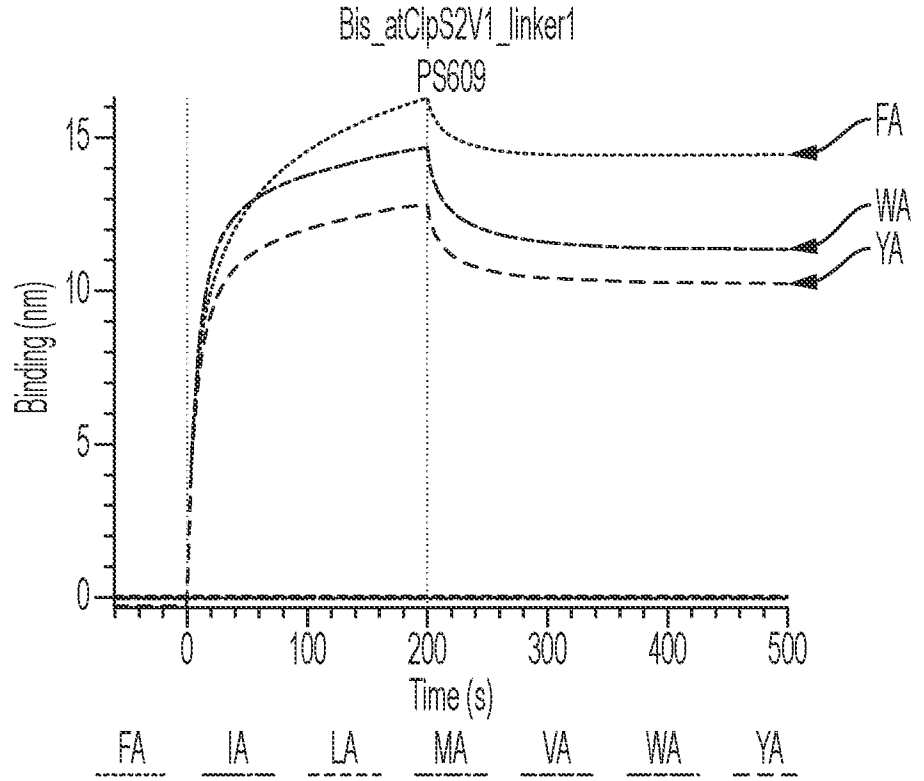


FIG. 35B

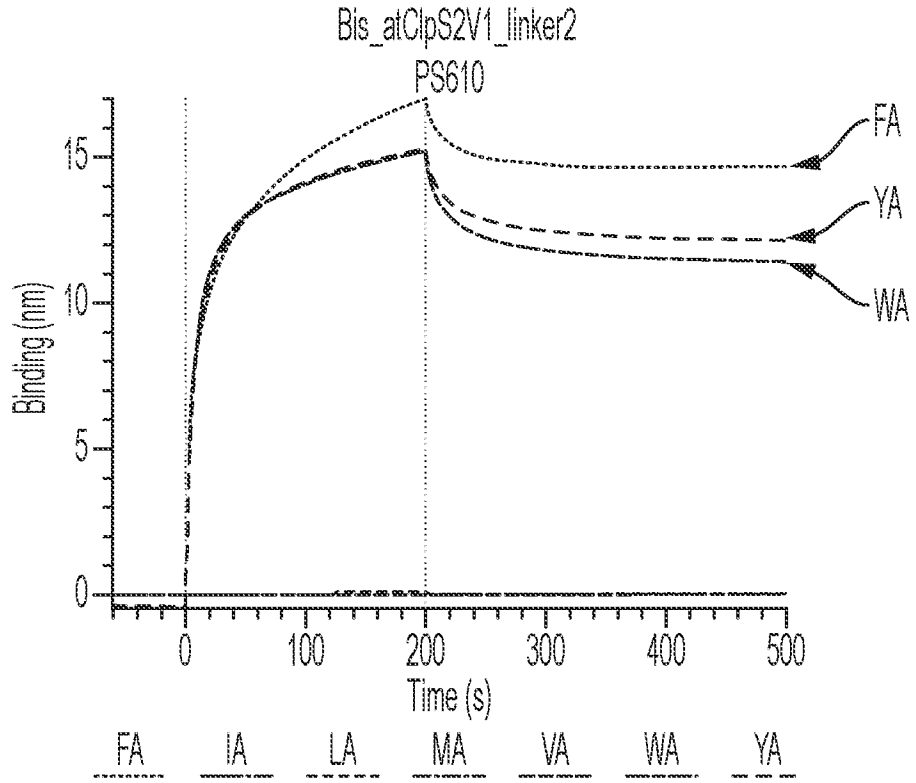


FIG. 35C

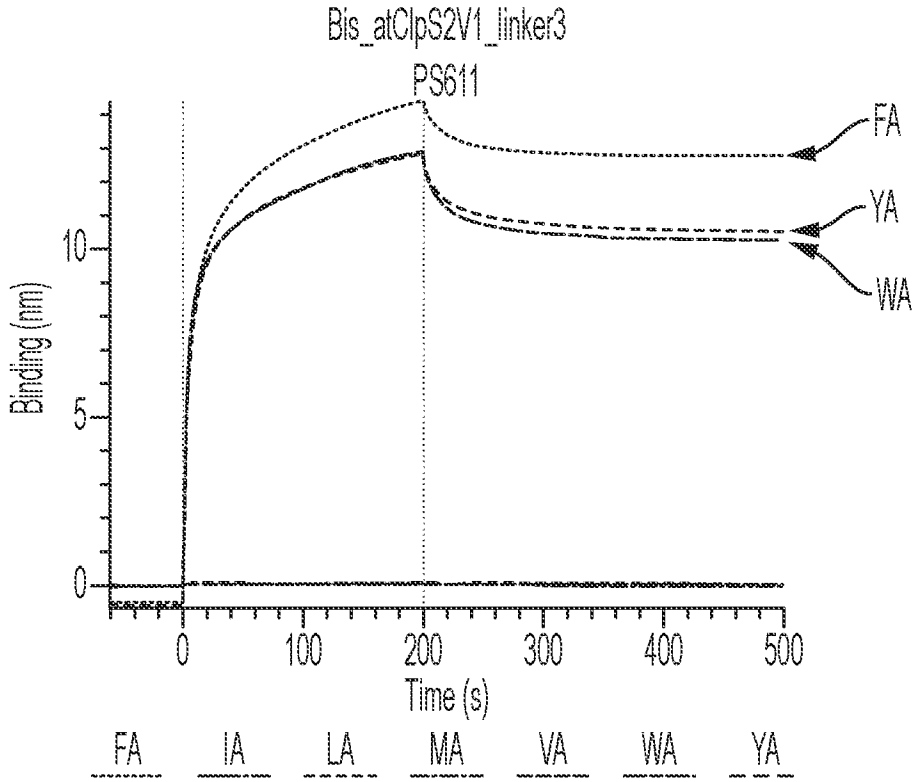


FIG. 35D

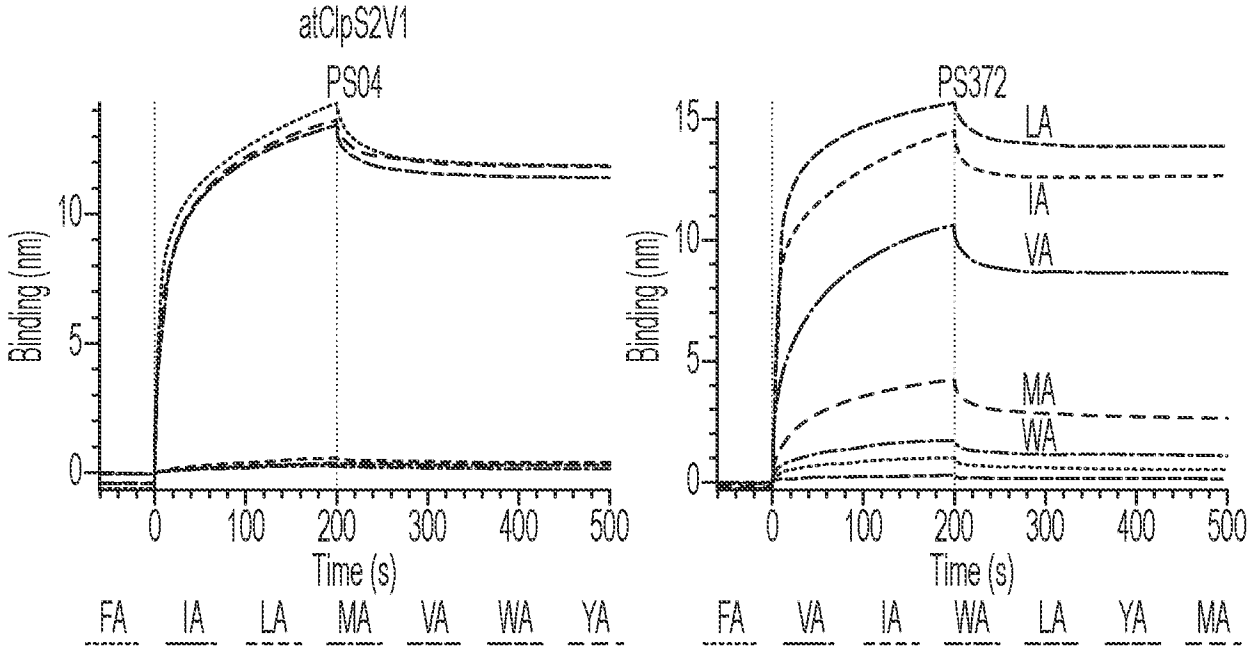


FIG. 36A

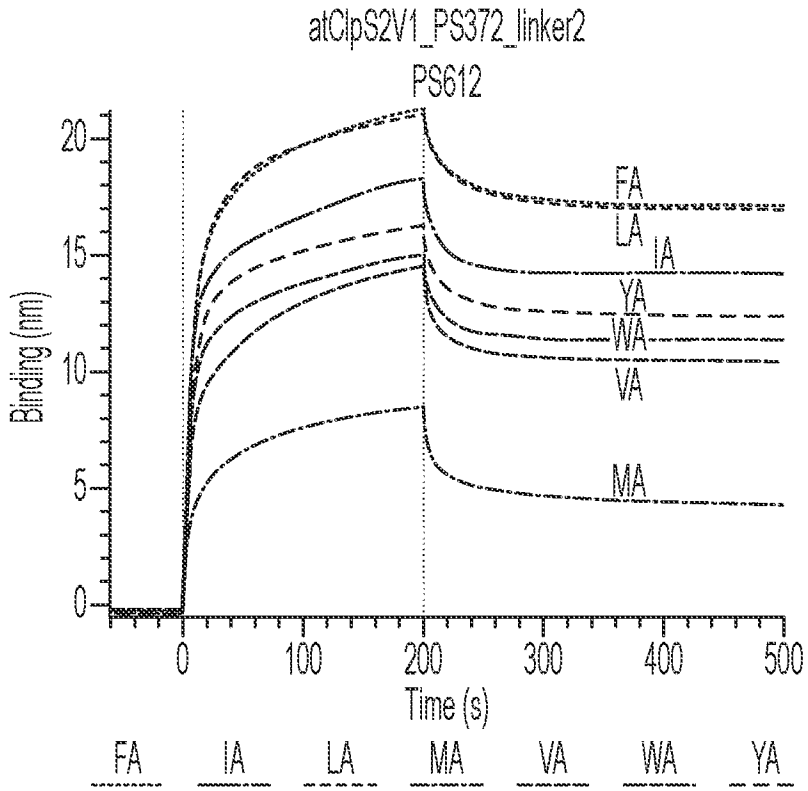


FIG. 36B

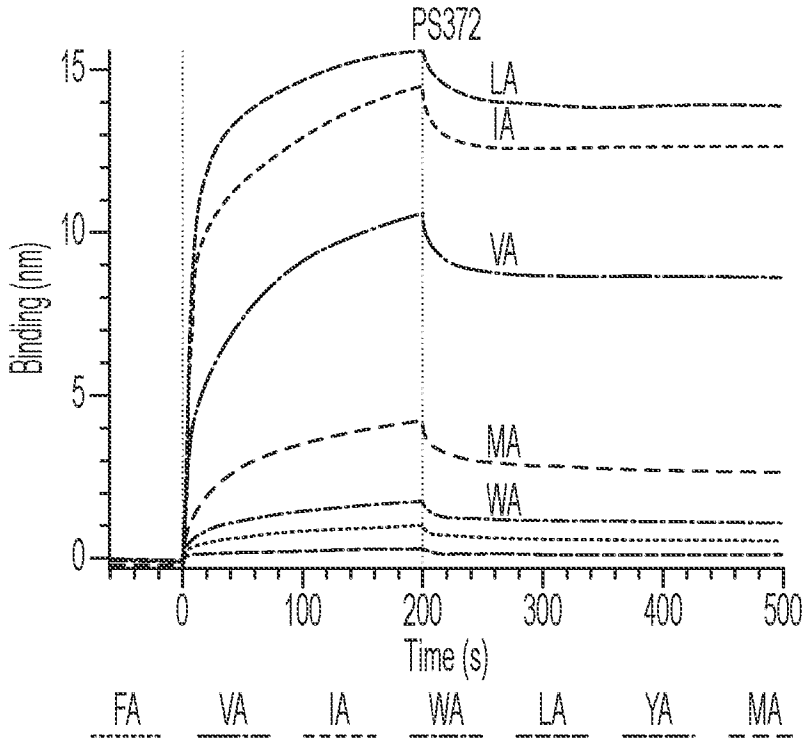


FIG. 36C

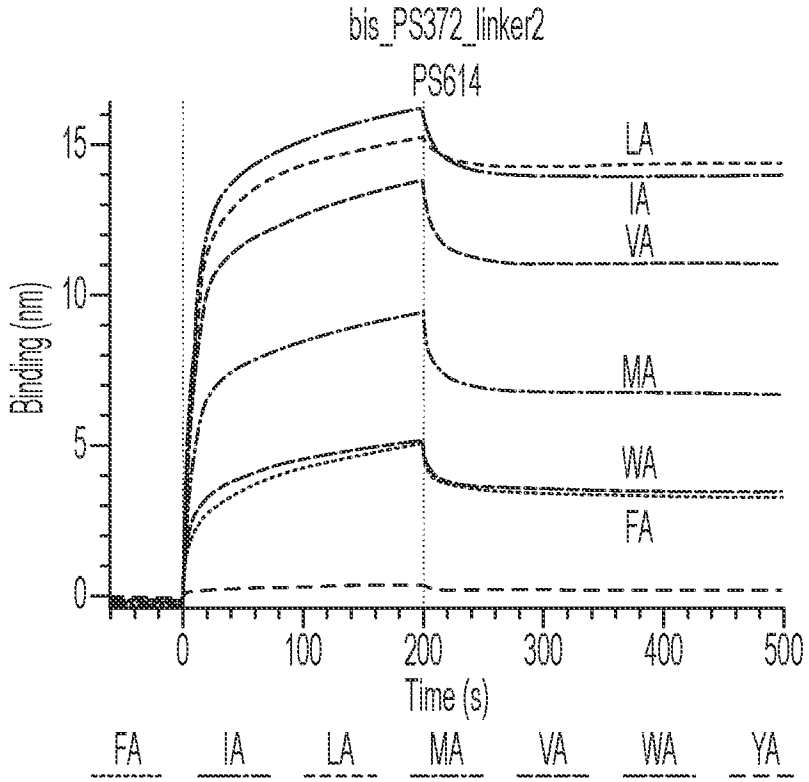


FIG. 36D

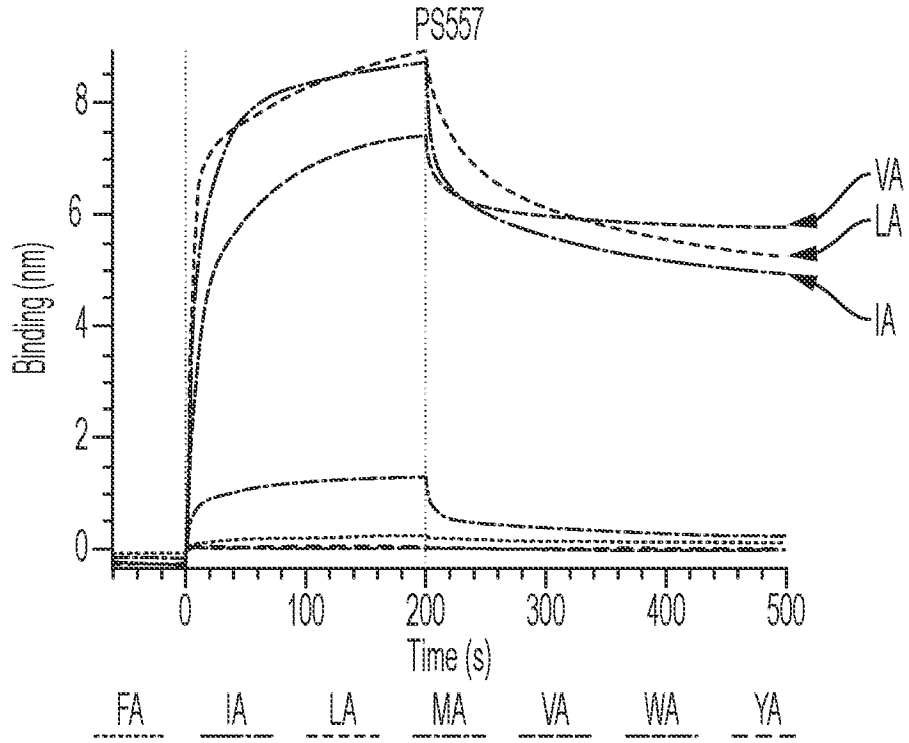


FIG. 36E

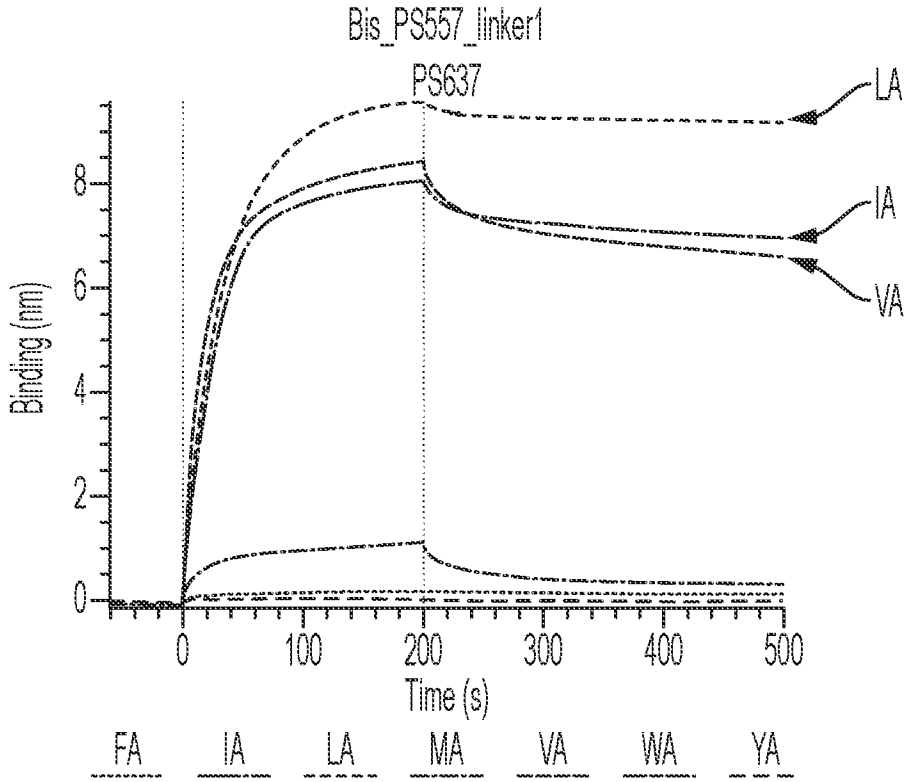


FIG. 36F

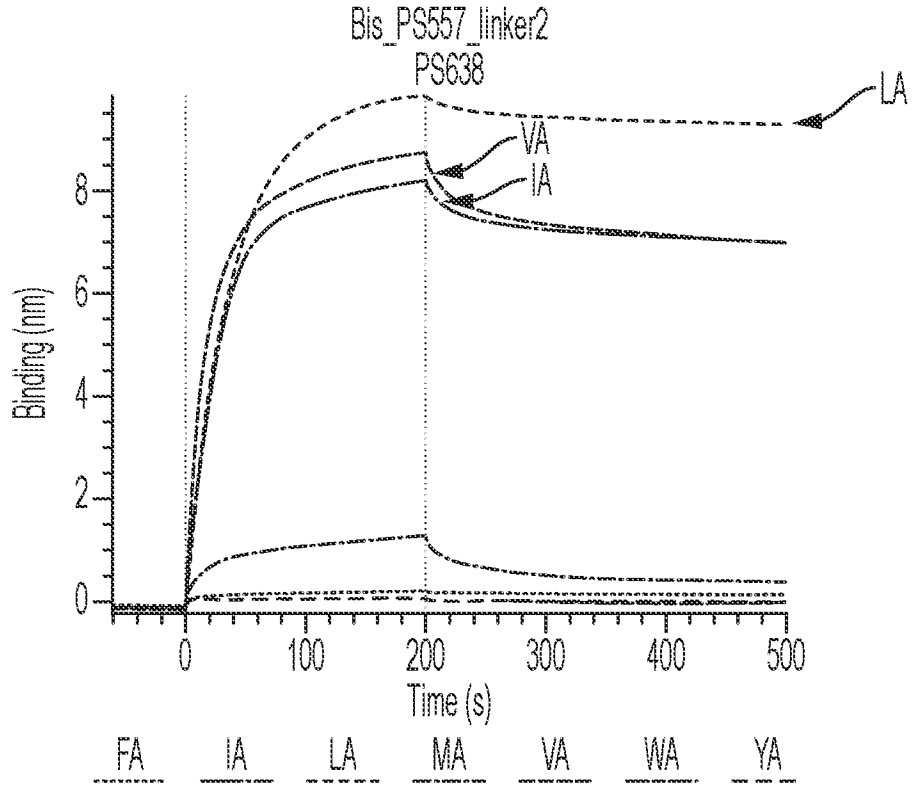


FIG. 36G

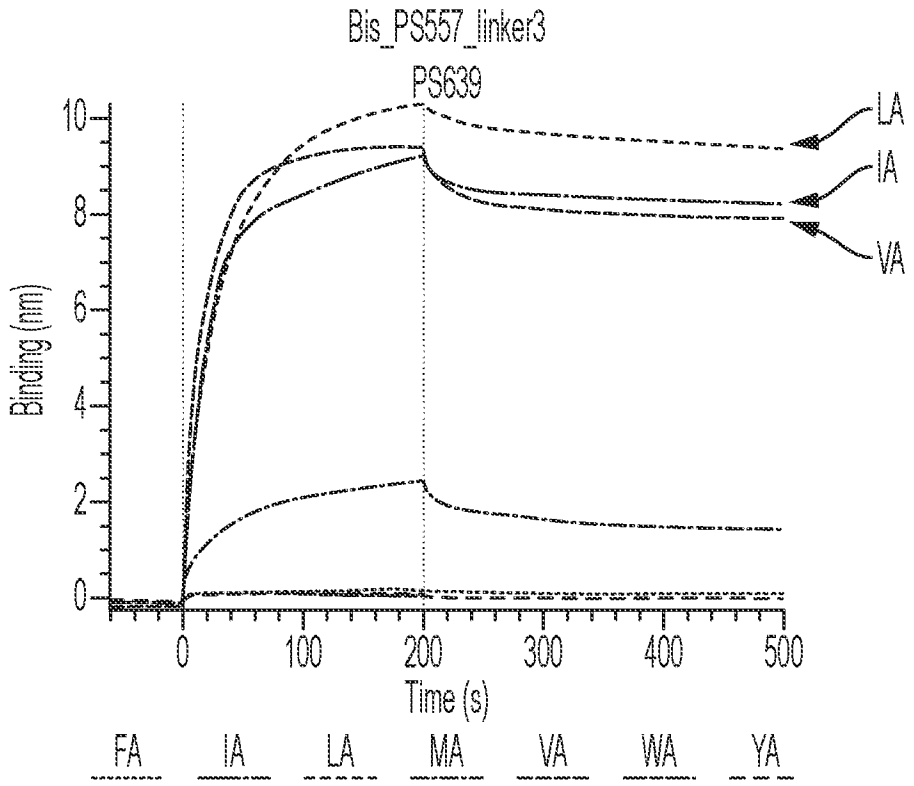
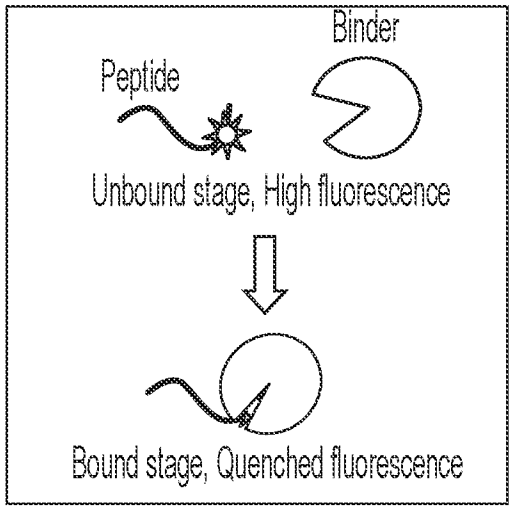
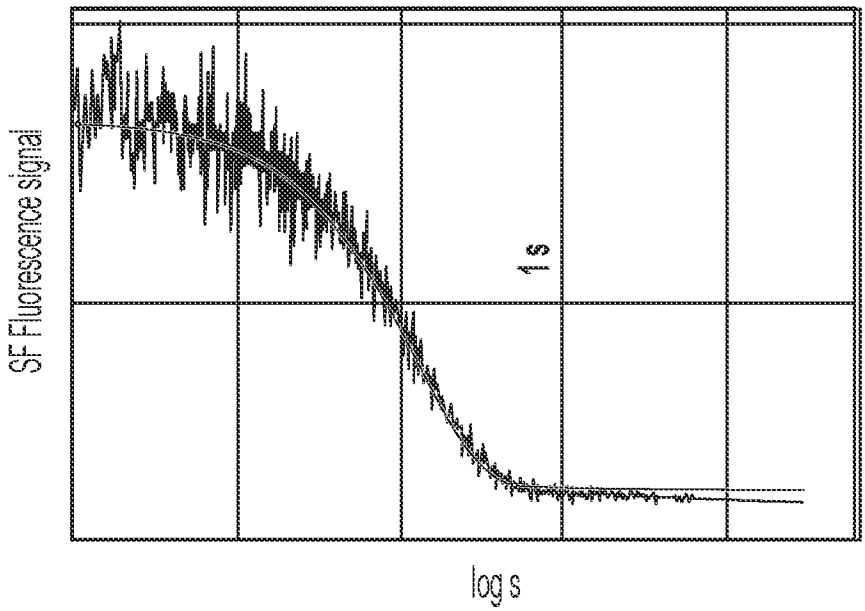


FIG. 36H

120/121





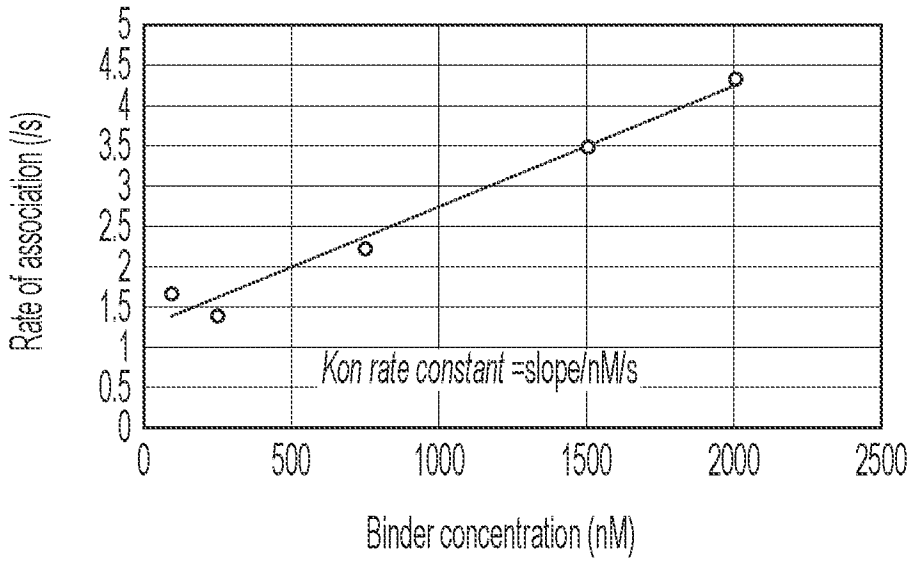


FIG. 37A

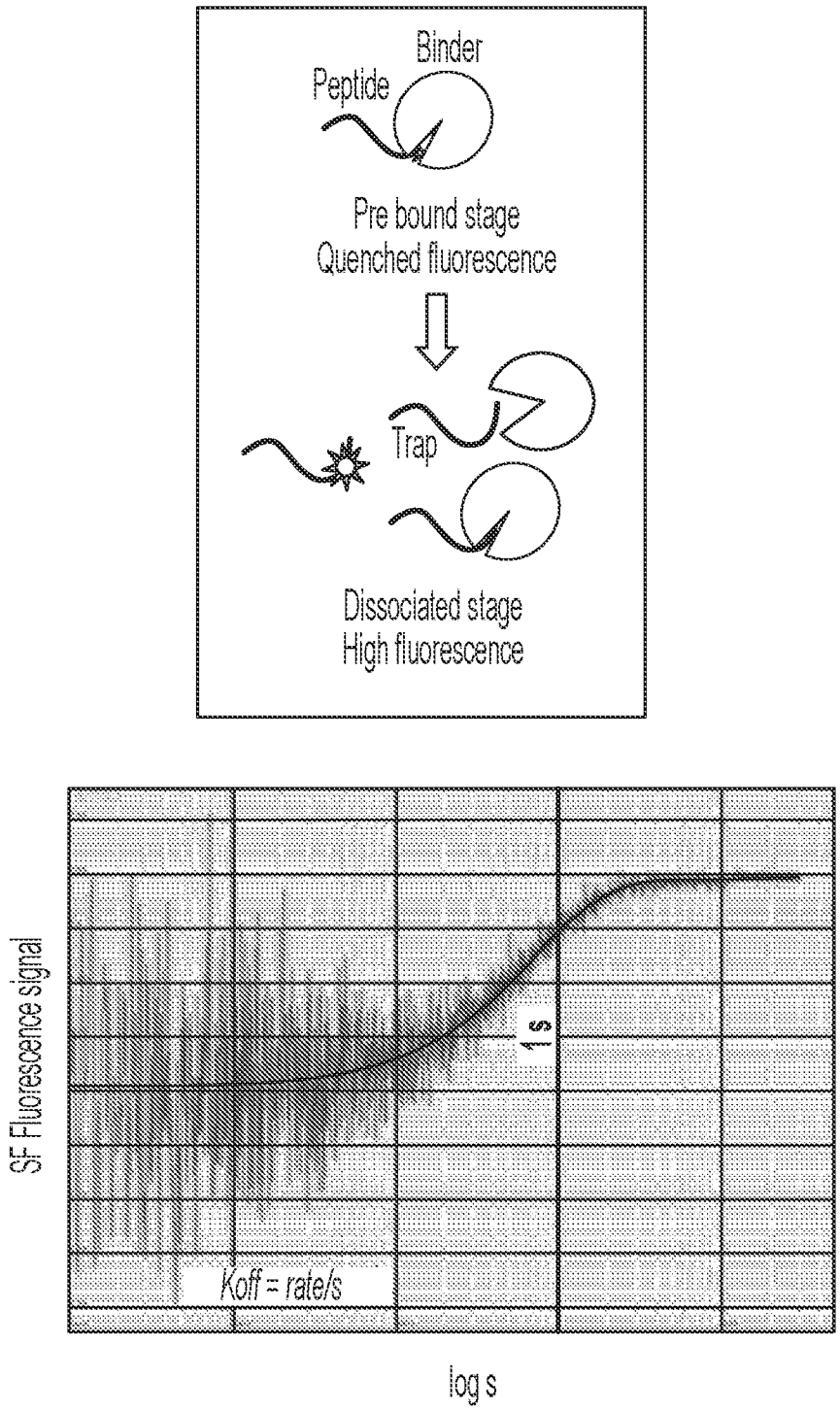

Binder
Peptide
Pre bound stage
Quenched fluorescence
Trap
Dissociated stage
High fluorescence
SF Fluorescence signal
1 s
Koff = rate/s
log s

FIG. 37B