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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
04 March 2021 (04.03.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/040703 A1

(51) International Patent Classification:
C07K 1/107 (2006.01) *C07K 16/00* (2006.01)
C07K 14/195 (2006.01) *C07K 16/18* (2006.01)

(21) International Application Number:
PCT/US2019/048508

(22) International Filing Date:
28 August 2019 (28.08.2019)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicants: **THE TRUSTEES OF PRINCETON UNIVERSITY** [US/US]; 87 Prospect Avenue, Princeton, New Jersey 08544 (US). **ALBERT EINSTEIN COLLEGE OF MEDICINE** [US/US]; 1300 Morris Park Avenue, Bronx, New York 10461 (US).

(72) Inventors: **MUIR, Tom W.**; c/o The Trustees of Princeton University, 87 Prospect Avenue, Princeton, New Jersey 08544 (US). **STEVENS, Adam**; c/o The Trustees of Princeton University, 87 Prospect Avenue, Princeton, New Jersey 08544 (US). **GRAMESPACHER, Josef**; c/o The Trustees of Princeton University, 87 Prospect Avenue, Princeton, New Jersey 08544 (US). **COWBURN, David**; c/o Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461 (US). **SEKAR, Giridhar**; c/o Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461 (US).

(74) Agent: **HOLTZ, William A.** et al.; Thompson Coburn LLP, One US Bank Plaza, St. Louis, Missouri 63101 (US).

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

(54) Title: ATYPICAL SPLIT INTEINS AND USES THEREOF

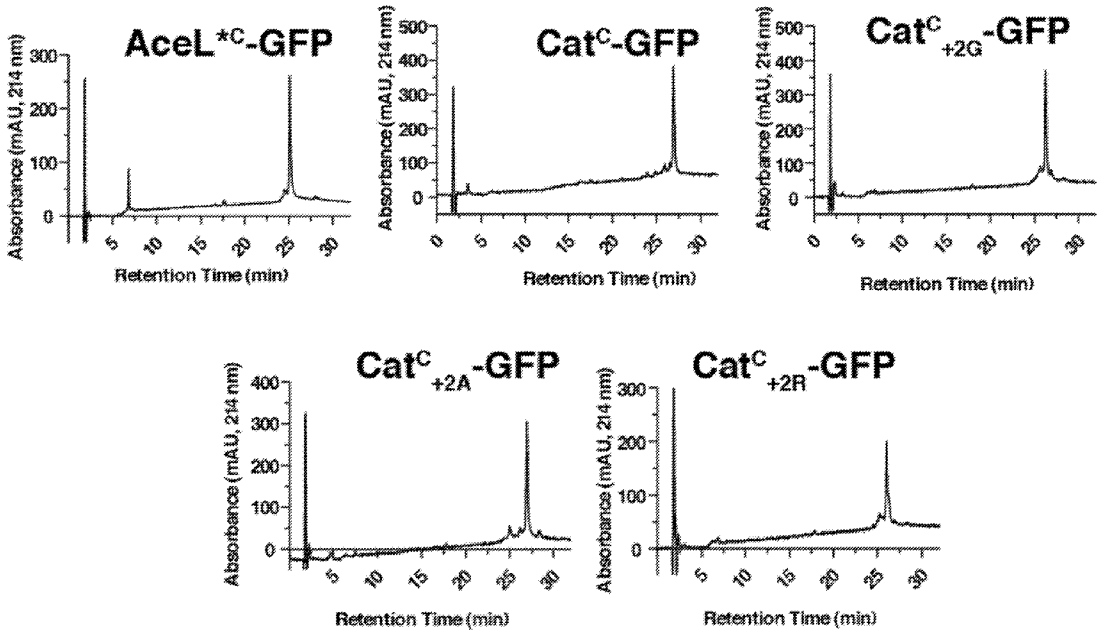


FIG. 1A

(57) Abstract: The present disclosure relates to atypical split N- and C-inteins and variants thereof. This disclosure also relates to complexes comprising the split N- or C-inteins of this disclosure and a compound of interest and compositions comprising said complexes. In addition, this disclosure relates to methods of using the atypical split N- and C- inteins.

WO 2021/040703 A1

[Continued on next page]

UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

— *with sequence listing part of description (Rule 5.2(a))*

ATYPICAL SPLIT INTEINS AND USES THEREOF

GOVERNMENT LICENSE RIGHTS

This invention was made with government support under Grant Nos. GM086868, OD016305, RR015495 and OD016432 awarded by the National Institutes of Health. The government has certain rights in the invention.

FIELD OF THE DISCLOSURE

The present disclosure is comprised within the field of biotechnology, it specifically relates to split inteins and their uses.

BACKGROUND

An intein is an intervening protein domain that undergoes a posttranslational auto-processing event called protein splicing in which it excises itself from a host protein while tracelessly ligating its flanking polypeptide sequences (exteins) to form a native peptide bond. Most inteins are found as contiguous domains embedded within a single gene and splice in cis. However, some exist naturally in split form, whereby each intein fragment is encoded on a separately expressed gene and must first associate prior to splicing in trans. These split inteins are commonly applied as tools in protein engineering, and are especially amenable to use in the cellular environment due to their highly specific recognition and unique activity.

Despite the growing use of inteins in chemical biology, their practical utility has been constrained by a number of common characteristics, namely (i) slow kinetics, (ii) context dependent efficiency with respect to the immediate flanking extein sequences, (iii) low expression levels of recombinant fusions to other proteins and (iv) suboptimal stability.

Thus, a need exists for more robust and more efficient split inteins for use in a variety of protein purification and protein modification applications.

SUMMARY

The authors of this disclosure provide herewith split inteins with atypical split sites which exhibit accelerated splicing rates and activity under adverse conditions, as it is shown in example 1 (figure 5, tables 5 and 6) of the present application. The disclosed inteins are useful in the N-terminal modification of expressed proteins and would complement other reported methods for protein N-terminal modification, such as

expressed protein ligation, transpeptidase-based ligation strategies, and various protein chemistry methods. In this regard, as the N-terminal intein fragments of these inteins are strikingly short, the isolated polypeptides are ideally suited for use in a range of protein modifications, since the complex protein of interest-split intein N-fragment can be easily obtained using solid-phase peptide synthesis.

Thus, an aspect of this disclosure relates to a split intein N-fragment comprising the amino acid sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1.

Another aspect of this disclosure relates to a complex comprising:

(i) a compound of interest,

(ii) the split intein N-fragment of this disclosure, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the complex optionally comprises a linker between (i) and (ii) and

wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

Another aspect of this disclosure relates to a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7.

Another aspect of this disclosure relates to a complex comprising:

(i) the split intein C-fragment of this disclosure or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120 and

(ii) a compound of interest

wherein the complex optionally comprises a linker between (i) and (ii) and

wherein

- the compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage.

In another aspect, this disclosure relates to a composition comprising the first complex and the second complex of this disclosure.

Another aspect of this disclosure relates to a complex comprising:

(i) the split intein C-fragment of this disclosure or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120

(ii) a compound of interest and

(iii) the split intein N-fragment of this disclosure, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

wherein the complex optionally comprises a linker between (i) and (ii) and/or between (ii) and (iii),

wherein

- the compound of interest is linked to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage and
- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

Another aspect of this disclosure relates to a conjugate comprising (a) the first complex of this disclosure and (b) a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, wherein the C-terminus of the split intein N-fragment is linked to the N-terminus of the split intein C-fragment by a peptide bond.

In another aspect, this disclosure relates to a polynucleotide encoding the split intein N-fragment of this disclosure, or the split intein C-fragment of this disclosure, or any one of the complexes of this disclosure wherein the compound of interest is a polypeptide or protein and the linker, if present, is a peptide linker.

In another aspect, this disclosure relates to a vector comprising the polynucleotide of this disclosure.

In another aspect, this disclosure relates to a host cell comprising the polynucleotide or the vector of this disclosure.

In another aspect, this disclosure relates to a composition comprising the first complex of this disclosure and the second complex of this disclosure.

In another aspect, this disclosure relates to a method to obtain a conjugate between a first compound of interest and a second compound of interest comprising

(i) contacting

(a) the first complex of this disclosure, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 with

(b) the second complex of this disclosure, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120

or

a complex comprising an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof and the second compound of interest, wherein the complex optionally comprises a linker between the split intein C-fragment and the second compound of interest and wherein

- the second compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the second compound of interest is bound to the linker by an amide linkage and/or the

linker is bound to the C-terminus of the split intein C-fragment by an amide linkage

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

In another aspect, this disclosure relates to a method to obtain a conjugate between a first compound of interest and a second compound of interest comprising

(i) contacting

(a) the first complex of this disclosure, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

or

a complex comprising the second compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) the second complex of this disclosure, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid

sequence selected from the group consisting of SEQ ID NO: 114-120

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

In another aspect, this disclosure relates to a method to obtain a conjugate of a compound of interest with a nucleophile comprising

(i) contacting

(a) the first complex of this disclosure, wherein the split intein N-fragment comprises the amino acid sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1 or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

or

a complex comprising a compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) a split intein C-fragment comprising an amino acid sequence selected from the group consisting of SEQ ID NO: 8, 9, 23-48 and 141-166, under appropriate conditions for binding between the split intein N-fragment and the split intein C-fragment to form an intein intermediate and

(ii) contacting the intein intermediate with an exogenous nucleophile.

In another aspect, this disclosure relates to a composition comprising:

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a variant thereof or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and
- a second polypeptide of interest

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and
- a second polypeptide of interest.

In another aspect, this disclosure relates to a method for expressing a gene of interest in a cell comprising:

(i) contacting the cell with

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:
- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:
- an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:
- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:
- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

(ii) allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and

(iii) allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.

In another aspect, this disclosure relates to a method for expressing a gene of interest comprising:

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

or

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant

thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

1. allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and secreted,
2. allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1. (A)-(E) *RP-HPLC analysis of inteins utilized in this study.* The masses corresponding to each RP-HPLC chromatogram are reported in Table 3.

Figure 2. (A)-(D) *Representative splicing gels of protein trans-splicing reactions.* (A) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat and AceL* at the indicated temperatures. Bands correspond to MBP-Int^{N} (N), Int^{C}-GFP (C) and the spliced product (SP) are indicated. (B) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat and AceL* at the indicated concentrations of urea. Bands corresponding to MBP-Int^{N} (N), Int^{C}-GFP (C) and the spliced product (SP) are indicated. (C) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat with the indicated -1 and -2 N-extein mutations (from the WT "FE" sequence). Bands corresponding to MBP-Cat^{N} (N), Cat^{C}-GFP (C) and the spliced product (SP) are indicated. C-terminal cleavage is observed for the -1A and -1P mutations and are indicated on the gel (GFP). (D) Representative SDS-PAGE gels of protein *trans*-splicing reactions for Cat with the indicated +2 and +3 C-extein mutations (from the WT

"EF"). Bands corresponding to MBP-Cat^{N} (N), Cat^{C}-GFP (C) and the spliced product (SP) are indicated.

Figure 3. (A)-(B) *Reaction progress curves*. (A) and (B) Reaction progress curves are presented for the splicing reactions carried out in this study. The best-fit lines for each reaction are shown.

Figure 4. (A)-(D) *Expression of Atypical Split Inteins*. Lanes correspond to (W) the whole cell lysate, (P) the inclusion body pellet, (S) the soluble fraction of the lysate, (FT) flow through of the soluble lysate batch bound to Ni-NTA affinity beads, (E) a 3 CV elution of 250 mM imidazole. (A) Purification of SUMO-GOS^{C}, SUMO-AceL*^{C}, and SUMO-Cat^{C} from *E. coli* expression (18 °C, 16 h). (B) Purification of SUMO-GOS^{C}, SUMO-AceL*^{C}-Sumo, and SUMO-Cat^{C} from *E. coli* expression (37 °C, 3 hours). (C) Purification of SUMO-GOS^{N}, SUMO-AceL*^{N}, and SUMO-Cat^{N} from *E. coli* expression (37 °C, 3 hours). (D) Purification of GOS^{C}-GFP, AceL*^{C}-GFP, and Cat^{C}-GFP from *E. coli* expression (18 °C, 16 hours).

Figure 5. (A)-(D) *Characterization of a consensus atypical (Cat) split intein*. (A) Pairwise sequence alignment of Cat and AceL* highlighting identical (black) and similar (gray) residues. (B) Reaction progress curve for Cat splicing at 30 °C. (C) Splicing rates for Cat and AceL* as a function of temperature (n = 3, error = SEM). AceL* is inactive at 50 °C. (D) Splicing rates for Cat and AceL* as a function of added Urea (n = 3, error = SEM). AceL* is not active in the presence of 2 M and 4 M Urea (NA).

Figure 6. (A)-(D) *Structural effects of Cat fragment association*. (A) ^{1}H-^{15}N HSQC spectra of 15N labeled Cat^{N} in free from (black) and in complex with unlabeled Cat^{C} (gray). (B) 1H-15N HSQC spectra of 15N labeled Cat^{C} in free form (black) and in complex with unlabeled Cat^{N} (gray). (C) Far UV circular dichroism spectra of Cat^{N} (black), Cat^{C} (dark gray) and the Cat^{N} + Cat^{C} complex (light gray). (D) Size exclusion chromatograms of Cat^{N} (black), Cat^{C} (dark gray), and the Cat^{N} + Cat^{C} complex (light gray).

Figure 7. (A)-(C) *Disorder to order transition of Cat^{N}* (A) (^{15}N-^{1}H) heteronuclear NOE of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). (B) Spin-spin relaxation rate of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). (C) Perturbation of Cα and Cβ chemical shifts of Cat^{N} in the presence of Cat^{C} (left) and in free form (right). Δδ(Cα,Cβ) = (δCβ- δCα)Observed- (δCβ-δCα)Random Coil.

Figure 8. (A)-(C) *Solution NMR structure of Cat.* (A) Backbone conformation of the 20 lowest energy conformers obtained in the structure calculation of the Cat^N (dark) – Cat^C (light) split intein complex. The Cat^C solubility tag is rendered in transparent gray. Structures are shown with a 180° rotation (top and bottom renderings). (B) Cartoon depiction of the lowest energy conformer. Structures are shown with a 180° rotation (top and bottom renderings). (C) Zoom view of the Cat active site with Ala_1, Ser_{75}, His_{78}, and His_{133} depicted as sticks. The distances between the carbonyl oxygen of Ala_1 and amide and hydroxyl protons of Ser_{75} are indicated.

Figure 9. (A)-(C) *Structure of Cat Complex*. (A) Average per residue Root Mean Square Deviation (RMSD) from average structure for 20 least energy conformers of Cat^N-Cat^C complex obtained in NMR structure calculation. (B) Average per residue RMSD plotted against residue number for Cat^N (gray) - Cat^C (black) complex. Extein regions are marked with a gray and the solubility tag used with Cat^C is shown as dashed lines. (C) Sequence logo of the Block B loop (left) Block F loop (middle) and C-terminal Block G (right) generated from an alignment of TerL intein homologues (Table 1).

Figure 10. (A)-(C) *Localization of Disorder in the Cat Fragments*. (A) RP-HPLC chromatogram stack from the limited proteolysis of Cat^N (left), Cat^C (middle) and a 1:1 Cat^N + Cat^C complex (right) with samples quenched after the indicated times. (B) Sequence of Cat with the disordered regions of Cat^C highlighted in dark gray and the protected center highlighted in light gray. (C) Model of Cat disorder mapped onto the NMR structure with the N-intein highlighted in light gray, disordered region of Cat^C highlighted in dark gray, and the protected center highlighted in medium gray. A zoom view of the active site is shown with the splicing residues rendered as sticks.

Figure 11. (A)-(B) *RP-HPLC analysis of limited Proteolysis of Cat fragments*. (A) RP-HPLC from the Cat^N (left) and Cat^C (right) proteolysis experiment (t = 30 min) with numbered samples corresponding to the ESI-MS data in Table 8. (B) Primary sequence of the Cat^N and Cat^C inteins used in the limited proteolysis experiment with the proteolysis fragments detected indicated below as brackets. The number of each bracket corresponds to the RP-HPLC peak in panel A.

Figure 12. (A)-(D) *Hydrophobic residues drive Cat association.* (A) Surface rendering of Cat^N with hydrophobic residues colored in grayscale based on the normalized consensus hydrophobicity scale. Cat^C is depicted as a cartoon. (B) Surface rendering of Cat^C with hydrophobic residues in grayscale. Cat^N is depicted as a cartoon. (C)

Equilibrium fluorescence anisotropy measurements of Fl-Cat^N (500 pM) in the presence of SUMO-Cat^C (indicated concentration) in low (100 mM NaClblack) and high (500 mM NaClgray dashed) salt buffers. (D) Concentration dependence of the observed rates of Fl-Cat^N+SUMO-Cat^C association in low (100mM NaClblack) and high (500 mM NaClgray dashed) salt buffers.

Figure 13. (A)-(C) *Electrostatic surface of Cat.* (A) Electrostatic surface potential of Cat^N with electronegative regions colored in smooth grayscale, electropositive regions colored in textured grayscale, and neutral regions colored in white. Cat^C is depicted as a cartoon. (B) Electrostatic surface potential of Cat^C with electronegative regions colored in smooth grayscale, electropositive regions colored in textured grayscale, and neutral regions colored in white. Cat^N is depicted as a cartoon. (C) Representative data and fits for kinetic binding experiments. Top: Single (left) and double (right) exponential models for the nonlinear least squares fitting of stopped flow anisotropy measurements of Fl-Cat^N upon mixing with SUMO-Cat^C. Bottom: Residual values obtained between experimental and predicted values are plotted for the single (left) and double (right) exponential fits.

Figure 14. (A)-(E) *Extein Dependence of Cat.* (A) Schematic of the assay used to investigate the impact of local extein sequences on Cat splicing. An N-extein maltose binding protein (MBP) is fused to Cat^N while a C-extein green fluorescent protein (GFP) is fused to Cat^C. The native extein sequences (Phe_{-2}, Glu_{-1}, Cys_{+1}, Glu_{+2}, Phe_{+3}) are shown within these fusion proteins. (B) Splicing rates for Cat in the presence of non-native C-extein residues (n = 3, error = SEM). Each indicated value corresponds to a single point mutation within the C-extein from the wild type (WT) sequence. (C) Splicing rates for Cat in the presence of non-native N-extein residues (n = 3, error = SEM). Each indicated value corresponds to a single point mutation within the N-extein from the wild type (WT) sequence. (D) Zoom view of the Cat active site with Cys_{+1}, Glu_{+2}, Asp_{115}, Asn_{123}, His_{133}, and Ala_{134} depicted as sticks. (E) Zoom view of Cat active site with Glu_{-1}, Ala_1, Ser_{75}, and His_{78} depicted as sticks.

DETAILED DESCRIPTION

The present disclosure relates to the provision of new atypical split inteins and its uses in biochemical engineering.

Split intein N-fragments

In a first aspect this disclosure relates to a split intein N-fragment comprising the amino acid sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1.

As used herein, the term "intein" means a naturally-occurring or artificially-constructed polypeptide sequence capable of catalyzing a protein splicing reaction that excises the intein sequence from a precursor protein and joins the flanking sequences (N- and C-exteins) with a peptide bond. They are typically 150-550 amino acids in size and may also contain a homing endonuclease domain. A list of known inteins is published on the world wide web at inteins.biocenter.helsinki.fi/.

The terms "polypeptide", "peptide" or “protein” are used interchangeably herein to refer to polymers of amino acids of any length.

The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Furthermore, the term "amino acid" includes both D- and L-amino acids (stereoisomers).

The term "natural amino acids" or “naturally occurring amino acid” comprises the 20 naturally occurring amino acids; those amino acids often modified post-translationally in vivo, including, for example, hydroxyproline, phosphoserine and phosphothreonine; and other unusual amino acids including, but not limited to, 2-aminoadipic acid, hydroxylysine, isodesmosine, nor-valine, nor-leucine and ornithine.

As used herein the term "non-natural amino acid" or “synthetic amino acid” refers to a carboxylic acid, or a derivative thereof, substituted with an amine group and being structurally related to a natural amino acid. Illustrative non- limiting examples of modified or uncommon amino acids include 2-aminoadipic acid, 3-aminoadipic acid, beta-alanine, 2-aminobutyric acid, 4-aminobutyric acid, 6-aminocaproic acid, 2-aminoheptanoic acid, 2-aminoisobutyric acid, 3-aminoisobutyric acid, 2-aminopimelic acid, 2,4-diaminobutyric acid, desmosine, 2,2'-diaminopimelic acid, 2,3-diaminopropionic acid, N-ethylglycine, N-ethylasparagine, hydroxy lysine, alio hydroxy lysine, 3-hydroxyproline, 4-hydroxyproline, isodesmosine, alloisoleucine, N-methylglycine, N-methyliso leucine, 6-N-methyl-lysine, N-methylvaline, norvaline, norleucine, ornithine, etc. This group also includes the D-isomers of the “natural amino acids”.

The term "split intein" as used herein refers to any intein in which the N-terminal and C-terminal amino acid sequences are not directly linked via a peptide bond, such that the N-terminal and C-terminal sequences become separate fragments that can non-covalently re-associate, or reconstitute, into an intein that is functional for trans-splicing reactions.

As used herein, the term “split intein N-fragment” or "N-terminal split intein" or "N-terminal intein fragment" or "N-terminal intein sequence" (abbreviated "Int N")" refers to any intein sequence that comprises an N-terminal amino acid sequence that is functional for trans-splicing reactions, that is, that is capable of associating with a functional split intein C- fragment to form a complete intein that is capable of excising itself from the host protein, catalyzing the ligation of the extein or flanking sequences with a peptide bond, or that upon association with a split intein C-fragment catalyzes the “N-terminal cleavage”, that is, the nucleophilic attack of the peptide bond between the extein and the N-terminus of the split intein N-fragment resulting in the breaking of said peptide bond.

It must be noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a split intein" includes a plurality of such split inteins and reference to "the polypeptide" includes reference to one or more polypeptides and equivalents thereof known to those skilled in the art, and so forth.

In certain embodiments, the split intein N-fragment comprises the amino acid sequence of SEQ ID NO: 1. The split intein N-fragment can comprise additional amino acid residues linked to the N- and/or C-terminus of the sequence of SEQ ID NO: 1. In certain embodiments, the split intein N-fragment comprises less than 10, less than 9, less than 8, less than 7, less than 6, less than 5, less than 4, less than 3, less than 2, or 1 additional amino acid residues linked to the N- and/or C-terminus of the sequence of SEQ ID NO: 1. In another embodiment, the split intein N-fragment consists on the amino acid sequence of SEQ ID NO: 1.

In certain embodiments, the split intein N-fragment comprises or consists of a variant of the amino acid sequence of SEQ ID NO: 1 having at least 90% sequence identity with SEQ ID NO: 1.

The term “variant” as used herein refers to a polypeptide molecule that is substantially similar to a particular polypeptide sequence. The variant may be similar in structure and biological activity to the polypeptide from which it derives. Thus, the variant may

refer to a mutant of a polypeptide sequence. The term "mutant" refers to a polypeptide molecule the sequence of which has one or more amino acids added, deleted, substituted or otherwise chemically modified in comparison to the polypeptide molecule from which it derives. The mutant may retain substantially the same properties as the polypeptide molecule from which it derives or lack the biological activity of the claimed sequences.

The variant of the split intein N-fragment of SEQ ID NO: 1 has at least 90% sequence identity with SEQ ID NO: 1. In certain embodiments, the variant of the split intein N-fragment of SEQ ID NO: 1 has at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% sequence identity with SEQ ID NO: 1.

In certain embodiments of this aspect of the present disclosure, the variant of the split intein N fragment of SEQ ID NO: 1 has a length of between 14 and 60 amino acids, for example, 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, 51, 52, 53, 54, 55, 56, 57, 58, 59 or 60 amino acids.

The terms "identity", "identical", "percent identity" or “sequence identity” in the context of two or more amino acid or nucleotide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues that are the same, when compared and aligned (introducing gaps, if necessary) for maximum correspondence, not considering any conservative amino acid substitutions as part of the sequence identity. The percent identity can be measured using sequence comparison software or algorithms or by visual inspection. Various algorithms and software are known in the art that can be used to obtain alignments of amino acid sequences. One such non-limiting example of a sequence alignment algorithm is the algorithm described in Karlin *et al.*, 1990, Proc. Natl. Acad. Sci., 87:2264-8, as modified in Karlin *et al.*, 1993, Proc. Natl. Acad. Sci., 90:5873-7, and incorporated into the N BLAST and XBLAST programs (Altschul *et al.*, 1991 , Nucleic Acids Res., 25:3389-402). In certain embodiments, Gapped BLAST can be used as described in Altschul *et al.*, 1997, Nucleic Acids Res. 25:3389-402. BLAST-2, WU-BLAST-2 (Altschul *et al.*, 1996, Methods in Enzymology, 266:460-80), ALIGN, ALIGN-2 (Genentech, South San Francisco, California) or Megalign (DNASTAR) are additional publicly available software programs that can be used to align sequences. In certain alternative embodiments, the GAP program in the GCG software package, which

incorporates the algorithm of Needleman and Wunsch (J. Mol. Biol. 48:444-53 (1970)) can be used to determine the percent identity between two amino acid sequences (e.g., using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5). Alternatively, in certain embodiments, the percent identity between amino acid sequences is determined using the algorithm of Myers and Miller (CABIOS, 4:1 1 -7 (1989)). For example, the percent identity can be determined using the ALIGN program (version 2.0) and using a PAM120 with residue table, a gap length penalty of 12 and a gap penalty of 4. Appropriate parameters for maximal alignment by particular alignment software can be determined by one skilled in the art. In certain embodiments, the default parameters of the alignment software are used. In certain embodiments, the percentage identity "X" of a first amino acid sequence to a second amino acid sequence is calculated as 100 x (Y/Z), where Y is the number of amino acid residues scored as identical matches in the alignment of the first and second sequences (as aligned by visual inspection or a particular sequence alignment program) and Z is the total number of residues in the second sequence. If the second sequence is longer than the first sequence, then the global alignment taken the entirety of both sequences into consideration is used, therefore all letters and null in each sequence must be aligned. In this case, the same formula as above can be used but using as Z value the length of the region wherein the first and second sequence overlaps, said region having a length which is substantially the same as the length of the first sequence.

As a non-limiting example, whether any particular polypeptide has a certain percentage sequence identity (e.g., is at least 80% identical, at least 85% identical, at least 90% identical, and in some embodiments, at least 95%, 96%, 97%, 98%, or 99% identical) to a reference sequence can, in certain embodiments, be determined using the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 5371 1). Bestfit uses the local homology algorithm of Smith and Waterman, Advances in Applied Mathematics 2:482-9 (1981), to find the best segment of homology between two sequences. When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present disclosure, the parameters are set such that the percentage of identity is calculated over the full length of the reference amino acid sequence and that gaps in homology of up to 5% of the total number of nucleotides in the reference sequence are allowed.

In certain embodiments, the variant of the split intein N-fragment of SEQ ID NO: 1 has at least 90% sequence identity with SEQ ID NO: 1 over the whole length of the sequence.

In certain embodiments, the variant of the split N-intein fragment of SEQ ID NO: 1 comprises or consists of an amino acid sequence selected from the group consisting of SEQ ID NO: 2- 6, and 125-127.

In another embodiment, the variant of the split N-intein fragment of SEQ ID NO: 1 is a functionally equivalent variant of SEQ ID NO: 1.

The term “functionally equivalent variant” as used herein is understood to mean all those proteins derived from a sequence by modification, insertion and/or deletion or one or more amino acids, whenever the function is substantially maintained, particularly in the case of a functionally equivalent variant of the split intein N-fragment refers to maintaining its activity.

In certain embodiments, the functionally equivalent variant of the split intein N-fragment of SEQ ID NO: 1 maintains or improves the activity from the split intein N-fragment of SEQ ID NO: 1.

The term “activity” as used herein referring to the split intein N-fragment, refers to the ability of the split intein N-fragment to bind to a split intein C-fragment and catalyze the “N-terminal cleavage”, that is, the nucleophilic attack of the peptide bond between the extein and the N-terminus of the split intein N-fragment, resulting in the breaking of said peptide bond. The activity of the split intein N-fragment can also refer to the “trans-splicing activity”, which is understood as the ability of said split intein N-fragment to bind to a functional split intein C-fragment excising the complete intein from the host protein, catalyzing the ligation of the extein or flanking sequences with a peptide bond. The activity is dependent on reaction conditions, including temperature, pH and the presence of chaotropic agents. The commonly used unit is $t_{1/2}$, which represents the time at which half of the catalyzed reaction has been completed. Additionally, intein activity is also measured by the rate constant (k) of the catalyzed reaction, that is, how many times per second does the reaction take place.

Suitable assays for determining whether a polypeptide is a functionally equivalent variant of a given split N-intein, in terms of its trans-splicing activity, include splicing assays, such as those described for example in the methods of the present application or disclosed in Shah NH et al (Shah NH et al., 2012, J Chem Soc, vol 134, 11338), as

long as in these assays the split intein N-fragment is combined with a functional split intein C-fragment, that is a split intein C-fragment which is capable of catalyzing "C-terminal cleavage". The assays described above allow to determine and characterize trans-splicing reactions in which functional N and C-intein fragments bind to each other and subsequently carry out a reaction by which they excise themselves out and form a new peptide bond between the N and C-exteins. Other assays have been developed, which rely on the use of functional N-intein and a C-intein mutant that prevents trans-splicing, so that the reaction is stopped after the cleavage of the N-extein from the N-intein. Such assays (Vila-Perelló et al. J Am Cem Soc. 2013, 135(1): 286-292) allow to characterize the ability of an N-intein to perform the N-terminal cleavage reaction. Additionally, other assays exist to measure the affinity between N and C-terminal inteins (Shah et al. Angew Chem Int Ed Engl. 2011, 50(29): 6511-5).

According to the present disclosure, the activity of the split N-intein of this disclosure is substantially maintained if the functionally equivalent has at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or at least 100% of its activity. Furthermore, the activity of the split N-intein of this disclosure is substantially improved if the functionally equivalent variant has at least 1%, at least 2%, 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 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%, or at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 100%, at least 150%, at least 200%, at least 300%, at least 400%, at least 500%, at least 1000%, or more of its activity.

As mentioned above, the activity of the split N-intein of this disclosure depends on a number of reaction parameters, including temperature, chaotropic environment and pH. Thus, in one embodiment, the functionally equivalent variant of the split intein N-fragment of this disclosure maintains or improve its activity at a temperature of at least 0°C, at least 5°C, at least 10°C, at least 15°C, at least 20°C, at least 25°C, at least 30°C, at least 35°C, at least 37°C, at least 40°C, at least 45°C, at least 50°C, at least 55°C, at least 60°C, at least 65°C, at least 70°C or higher; in certain embodiments at a temperature of 50ºC. Likewise, in another embodiment the functionally equivalent variant of the split N-intein of this disclosure maintains or improves its activity at least at pH 2.0, or at least at pH 2.5, or at least at pH 3.0, or at least at pH 3.5, or at least at pH 4.0, or at least at pH 4.5, or at least at pH 5.0, or at least at pH 5.5, or at least at pH

6.0, or at least at pH 6.5, or at least at pH 7.0, or at least at pH 7.2, or at least at pH 7.5, or at least at pH 8.0, or at least at pH 8.5, or at least at pH 9.0, or at least at pH 9.5, or at least at pH 10.0, or at least at pH 10.5, or at least at pH 11.0, or at least at pH 11.5, or at least at pH 12.0, or at least at pH 12.5, or at least at pH 13.0, or at least at pH 13.5, or at least at pH 14; in certain embodiments at pH 7.2. In another embodiment, the functionally equivalent variant of the split N-intein of this disclosure maintains or improves its activity at urea 1 M, or at least at urea 1.5 M, or at urea least 2 M, or at least urea 3 M, or at least urea 3.5 M, or at least urea 4 M, or at least urea 4.5 M, or at least urea 5 M; in certain embodiments at urea 2 M or at urea 4 M. In certain embodiments, the functionally equivalent variant of the split N-intein of this disclosure maintains or improves its activity at urea 2 M or urea 4 M. In certain embodiments, the functionally equivalent variant of the split N-intein of this disclosure maintains or improves its at a temperature of 50ºC, at pH 7.2 and at urea 2 M or urea 4 M. All possible combinations of temperatures, urea concentration, other denaturants and pH are also contemplated by this disclosure.

In certain embodiments, the functionally equivalent variant of the split intein N-fragment of this disclosure that maintains or improves its activity has at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98% or at least 99% sequence identity with SEQ ID NO: 1.

In another embodiment, the functionally equivalent variant of the split intein N-fragment of SEQ ID NO: 1 comprises or consist of the amino acid sequence of SEQ ID NO: 4 or SEQ ID NO: 125.

Complex comprising a split intein N-fragment

In another aspect, this disclosure relates to a complex, hereinafter first complex of this disclosure, comprising:

(i) a compound of interest,

(ii) the split intein N-fragment of this disclosure, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the complex optionally comprises a linker between (i) and (ii) and

wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or

- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

As used herein, the term “compound of interest” include any synthetic or naturally occurring molecule, including a protein or peptide, a single or doubled stranded oligonucleotide, small molecule a drug or a cytotoxic molecule. The term therefore encompasses those compounds traditionally regarded as drugs, vaccines, and biopharmaceuticals including molecules such as proteins, peptides, and the like. Examples of therapeutic agents are described in well-known literature references such as the Merck Index (14th edition), the Physicians' Desk Reference (64th edition), and The Pharmacological Basis of Therapeutics (1st edition), and they include, without limitation, medicaments; substances used for the treatment, prevention, diagnosis, cure or mitigation of a disease or illness; substances that affect the structure or function of the body, or pro-drugs, which become biologically active or more active after they have been placed in a physiological environment. In addition, the “compound of interest” may include any non-protein molecule having a carboxylic group able to bind the amino-terminus end of the N-intein.

Optionally, the compound of interest and the split intein N-fragment may be joined through a linker, so the linker is located in between the compound of interest and the N-intein. The nature of the linker will depend on the nature of the compound of interest. In certain embodiments, the linker is a peptide. In certain embodiments, the linker is a peptide having a length of 1, 2, 3, 4, 5, 10, 20, 50, 100 or more amino acid residues; specifically, it may be 1 to 3 amino acid residues. If the compound of interest is a peptide or protein, the N-terminus of the linker is linked to the C-terminus of the compound of interest and the C-terminus of the linker is linked to the N-terminus of the N-intein through peptide bonds.

In certain embodiments, the linker is a non-peptide linker. Non-peptide linkers are for example, alkyl linkers such as -HN-(CH_2)s —CO—, wherein s=2-20 can be used. These alkyl linkers may further be substituted by any non-sterically hindering group such as lower alkyl (e.g., Ci -Ce), halogen (e.g., Cl, Br), CN, NH2, phenyl, etc.

Another type of non-peptide linker is a polyethylene glycol group, such as: —HN-(CH2)2-(0-CH2-CH2)n-0-CH2-CO, wherein n is such that the overall molecular weight of the linker ranges from approximately 101 to 5000; in certain embodiments 101 to 500.

In another embodiment, the non-peptide linker comprises a basic nucleotide, polyether, polyamine, polyamide, carbohydrate, lipid, polyhydrocarbon, or other polymeric compounds.

In certain embodiments, the complex does not comprise a linker between the compound of interest and the split intein N-fragment. In this embodiment, the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage.

In certain embodiments, the complex comprises a linker between the compound of interest and the split intein N-fragment. In this embodiment, the compound of interest may be bound to the linker by any suitable means, depending on the chemical nature of the compound of interest and of the linker. In this embodiment, the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage. In a another embodiment, the compound of interest is bound to the linker by an amide linkage, in which case the linker may be found to the N-terminus of the split intein N-fragment by any suitable means. In another embodiment, the compound of interest is bound to the linker by a amide linkage and the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

In another embodiment, the compound of interest is a protein having the C-terminal amino acid residues of the extein capable of being spliced by an intein comprising the N-intein of SEQ ID NO: 1. In another embodiment, the compound of interest is a protein having the sequence Glu-Phe-Glu in its C-terminus. In another embodiment, the compound of interest is a protein having the sequence Phe-Glu in its C-terminus. In another embodiment, the compound of interest is a protein having the residue Glu in its C-terminus.

In another embodiment, when the compound of interest is not a protein, the N-intein comprises or consists on the polypeptide of SEQ ID NO: 4-6, 125-127 or 168-170. In another embodiment, when the compound of interest is not a protein, the compound of interest and the N-intein are joined through a linker in which case, the linker is a peptide having the C-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split intein N-fragment of sequence SEQ ID NO: 1; in certain embodiments, the linker is a peptide having the sequence Glu-Phe-Glu, Phe-Glu or Glu in its C-terminus.

In another embodiment, the compound of interest is a protein that does not have the C-terminal amino acid residues of the extein capable of being spliced by an intein

comprising the split intein N-fragment of SEQ ID NO: 1, in which case (i) the N-intein comprises or consists on the polypeptide of sequence SEQ ID NO: 4-6, 125-127 or 168-170 or (ii) the compound of interest and the N-intein are joined through a linker in which case, the linker is a peptide having the C-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split intein N-fragment of SEQ ID NO: 1; in certain embodiments, the linker is a peptide having the sequence Glu-Phe-Glu, Phe-Glu or Glu in its C-terminus.

The phrase "peptide bond" refers to a covalent chemical bond —CO—NH— formed between two molecules when the carboxy part of one molecule, referred to as a carboxy component, reacts with the amino part of another molecule, referred to as an amino component, causing the release of a molecule. For example, proteinogenic L-amino acids can form the peptide bond upon joining with the release of a molecule of water. Therefore, proteins and peptides can be regarded as chains of amino acid residues held together by peptide bonds. A peptide bond is an "amide bond" or "amide linkage".

In certain embodiments, the compound of interest is a protein or polypeptide.

In another embodiment, the compound of interest is a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa.

In certain embodiments, the protein is Cas9, or a fragment of Cas9. The term "Cas9" or "CRISPR-associated endonuclease Cas9", as used herein, refers to a protein, which is the hallmark protein of the type II CRISPR-Cas system, and is a large monomeric DNA nuclease guided to a DNA target sequence adjacent to the PAM (protospacer adjacent motif) sequence motif by a complex of two noncoding RNAs: CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA). The Cas9 protein contains two nuclease domains homologous to RuvC and HNH nucleases. The HNH nuclease domain cleaves the complementary DNA strand whereas the RuvC-like domain cleaves the non-complementary strand and, as a result, a blunt cut is introduced in the target DNA. Heterologous expression of Cas9 together with a sgRNA can introduce site-specific double strand breaks (DSBs) into genomic DNA of live cells from various organisms. The Cas9 can be of any origin, including for example, *Streptocccus thermophilus*, *Streptococcus pyogenes*, *Staphylococcus aeureus*, *Francisella tularensis*, *Actinomyces naeslundii*, *Neiserria meningitides*, *Listeria innocua*, among others. In certain embodiments, the term "Cas9" refers to any one of the proteins defined by the UniProtKB/Swiss-Prot accession numbers G3ECR1 (entry version 31 of 10 April 2019,

sequence version 2 of 13 June 2012), Q99ZW2 (entry version 112 of 31 July 2019, sequence version 1 of 1 June 2001), J7RUA5 (entry version 33 of 8 May 2019, sequence version 1 of 31 October 2012), A0Q5Y3 (entry version 62 of 16 January 2019, sequence version 1 of 9 January 2007), J3F2B0 (entry version 33 of 8 May 2019, sequence version 1 of 3 October 2012), Q03JI6 (entry version 70 of 8 May 2019, sequence version 1 of 14 November 2006), C9X1G5 (entry version 47 of 31 July 2019, sequence version 1 of 24 November 2009), Q927P4 (entry version 94 of 8 May 2019, sequence version 1 of 1 December 2001).

In certain embodiments, the compound of interest of the complex is a polypeptide or protein, and if the complex comprises a linker, the linker is a peptide linker. In this embodiment, the complex is a fusion protein.

The term "fusion protein" is well known in the art, referring to a single polypeptide chain artificially designed which comprises two or more sequences from different origins, natural and/or artificial. The fusion protein, per definition, is never found in nature as such.

The term "single polypeptide chain", as used herein means that the polypeptide components of the fusion protein can be conjugated end-to-end but also may include one or more optional peptide or polypeptide "linkers" or "spacers" intercalated between them, linked by a covalent bond.

In another embodiment, the polypeptide of interest is an antibody of a fragment of an antibody.

As used herein, the term "antibody" relates to a monomeric or multimeric protein which comprises at least one polypeptide having the capacity for binding to a determined antigen, or epitope within the antigen, and comprising all or part of the light or heavy

The term antibody also includes any type of known antibody, such as, for example, polyclonal antibodies, monoclonal antibodies and genetically engineered antibodies, such as chimeric antibodies, humanized antibodies, primatized antibodies, human antibodies, camelid antibodies and bispecific antibodies (including diabodies), multispecific antibodies (e.g. bispecific antibodies), and antibody fragments so long as they exhibit the desired biological activity.

The term "antibody fragment" includes antibody fragments such as Fab, F(ab')2, Fab', single chain Fv fragments (scFv), diabodies and nanobodies.

An illustrative non-limitative example of antibody is an antibody against the DEC-205 receptor. The term “DEC-205 receptor”, or “lymphocyte antigen 75”, or “C-type lectin domain family 13 member B”, as used herein, refers to a protein which acts as an endocytic receptor to direct captured antigens from the extracellular space to a specialized antigen-processing compartment and is found mainly on dendritic cells. In certain embodiments, the DEC-205 is the human protein defined by the UniProtKB/Swiss-Prot accession number O60449 (entry version 170 of 31 July 2019, sequence version 3 of 11 January 2011). In certain embodiments, the anti-DEC205 antibody is a monoclonal antibody. The anti-DEC-205 antibody can be of any origin, for example, from mouse, rabbit, human, or can be a humanized antibody. In certain embodiments, the compound of interest is a chain of the anti-DEC-205 antibody; in certain embodiments, the heavy chain. In another embodiment, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In another embodiment, the compound of interest is a fragment of a protein; in certain embodiments, a fragment of a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa.

In another embodiment, the compound of interest is an N-terminal fragment of a protein; in certain embodiments, a fragment of a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa. The term “N-terminal fragment of a protein”, as used herein, refers to a fragment of variable length that includes the N-terminus of the protein. In certain embodiments, the N-terminal fragment is a fragment comprising less than 100%, less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than 5% of the length of the whole protein.

In certain embodiments, the complex comprises a split intein N-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 111, 112 and 113.

In certain embodiments, the sequences of SEQ ID NO: 112 and 113 have higher thermal stability than the sequence of SEQ ID NO: 1.

In certain embodiments, the complex comprises a split intein N-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 49-68 or variant thereof. In certain embodiments, the variant is a functionally equivalent variant.

The terms “variant” and “functionally equivalent variant” have been previously defined. In certain embodiments, the functionally equivalent variants of the split intein N-fragments of SEQ ID NO: 49-68 have at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95% or at least 99% sequence identity with the sequence from which they derive.

In certain embodiments, the functionally equivalent variants of the split intein N-fragments of SEQ ID NO: 49-68 maintain or improve the activity from the sequence from which they derive. The term “activity” as well as methods to measure this activity have been previously defined in connection with the functionally equivalent variants of the split intein N-fragment of SEQ ID NO: 1. The embodiments regarding the activity of the variants of the split intein N-fragment of SEQ ID NO: 1 fully applies to the activity of the variants of the split intein N-fragments of SEQ ID NO: 49-68.

Split intein C-fragment

In another aspect, this disclosure relates to a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7.

As interchangeably used herein, the terms “split intein C-fragment”, "C-terminal split intein", "C-terminal intein fragment" and "C-terminal intein sequence" (abbreviated "Int^C") refer to any intein sequence that comprises a C-terminal amino acid sequence that is functional for trans-splicing reactions, that is, that is capable of associating with a functional split intein N-fragment to form a complete intein that is capable of excising itself from the host protein, catalyzing the ligation of the extein or flanking sequences with a peptide bond, or that upon association with a split N-intein catalyzes the “C-terminal cleavage”, that is, the nucleophilic attack of the peptide bond between the extein and the C-terminus of the split intein C-fragment resulting in the breaking of said peptide bond. An Int^C thus also comprises a sequence that is spliced out when trans-splicing occurs. An Int^C can comprise a sequence that is a modification of the C-terminal portion of a naturally occurring intein sequence. For example, it can comprise additional amino acid residues and/or mutated residues so long as the inclusion of such additional and/or mutated residues does not render the Int^C non-functional in trans-splicing. In certain embodiments, the inclusion of the additional and/or mutated residues improves or enhances the trans-splicing activity of the Int^C.

In certain embodiments, the split intein C-fragment comprises the amino acid sequence of SEQ ID NO: 7. The split intein C-fragment can comprise additional amino acid

residues linked to the N- and/or C-terminus of the sequence of SEQ ID NO: 7. In certain embodiments, the split intein C-fragment comprises less than 10, less than 9, less than 8, less than 7, less than 6, less than 5, less than 4, less than 3, less than 2, or 1 additional amino acid residues linked to the N- and/or C-terminus of the sequence of SEQ ID NO: 7. In another embodiment, the split intein N-fragment consists on the amino acid sequence of SEQ ID NO: 7.

In certain embodiments, the split intein C-fragment comprises or consists on a variant of the amino acid sequence of SEQ ID NO: 7 having at least 88% sequence identity with SEQ ID NO: 7.

The terms “amino acid” and “variant” have been already described within the context of the N-inteins and equally apply to the present case.

The variant of the split intein C-fragment of SEQ ID NO: 7 has at least 88% sequence identity with SEQ ID NO: 7. In certain embodiments, the variant of the split intein C-fragment of SEQ ID NO: 7 has at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% sequence identity with SEQ ID NO: 7.

In certain embodiments, the variant of the split intein C-fragment of SEQ ID NO: 7 has a length of between 50 and 160 amino acids; and in certain embodiments, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155 or 160 amino acids.

In certain embodiments, the variant of the split intein C-fragment of SEQ ID NO: 7 has at least 88% sequence identity with SEQ ID NO: 7 over the whole length of the sequence.

In certain embodiments, the variant of the split intein C-fragment of sequence SEQ ID NO: 7 comprises or consist on an amino acid sequence selected from the group consisting of SEQ ID NO: 848 and 128-166.

In another embodiment, the variant of the split C-intein of SEQ ID NO: 7 is a functionally equivalent variant of SEQ ID NO: 7.

The term “functionally equivalent variant” has been previously defined for the split intein C-fragment. In the case of the functionally equivalent variant of the split intein C-fragment of SEQ ID NO: 7, the activity of the split intein C-fragment refers to its ability to bind to a split intein N-fragment and catalyze the “C-terminal cleavage”, that is, the nucleophilic attack of the peptide bond between the extein and the C-terminus of the

split intein C-fragment, resulting in the breaking of said peptide bond. The activity of the split intein C-fragment can also refer to the "trans-splicing activity", which is understood as the ability of said split intein C-fragment to bind to a functional split intein N-fragment excising the complete intein from the host protein, catalyzing the ligation of the extein or flanking sequences with a peptide bond. Suitable assays for determining whether a polypeptide is a functionally equivalent variant of a given split C-intein, in terms of its trans-splicing activity, include splicing assays, such as those describe in example the methods of the present application or disclosed in Shah NH et al (Shah NT et al., 2012, J Chem Soc, vol 134, 11338), as long as in these assays the split intein C-fragment is combined with a functional split intein N-fragment, that is a split intein N-fragment which is capable of catalyzing the N-terminal cleavage. Other more specific assays have also been described which allow characterizing each of the steps of the protein splicing, and particularly the last step involving the cleavage of the peptide bond between the C-intein and the C-extein, herein referred as "C-terminal cleavage" (Shah et al. JACS 2013).

According to the present disclosure, the activity of an C-intein is substantially maintained if the functionally equivalent has at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or at least 100% of the activity of the intein of the claimed sequences. Furthermore, the activity of the C-intein is substantially improved if the functionally equivalent variant has at least 1%, at least 2%, 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 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%, or at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 100%, at least 150%, at least 200%, at least 300%, at least 400%, at least 500%, at least 1000%, or more of the activity of the C-inteins of this disclosure.

As mentioned above, the activity of the split intein C-fragment of this disclosure depend on a number of reaction parameters, including temperature, chaotropic environment and pH. Thus, in one embodiment, the functionally equivalent variant of the split intein C-fragment of this disclosure maintains or improve its activity at a temperature of at least 0°C, at least 5°C, at least l0°C, at least l5°C, at least 20°C, at least 25°C, at least 30°C, at least 35°C, at least 37°C, at least 40°C, at least 45°C, at least 50°C, at least 55°C, at least 60°C, at least 65°C, at least 70°C or higher. In certain embodiments, the functionally equivalent variant of the split intein C-fragment of this disclosure maintains

or improve its activity at a temperature of 50ºC. Likewise, in another embodiment the functionally equivalent variant of the split intein C-fragment of this disclosure maintains or improves its activity at least at pH 0.1, or at least at pH 0.5, or at least at pH 1.0, or at least at pH 1.5, or at least at pH 2.0, or at least at pH 2.5, or at least at pH 3.0, or at least at pH 3.5, or at least at pH 4.0, or at least at pH 4.5, or at least at pH 5.0, or at least at pH 5.5, or at least at pH 6.0, or at least at pH 6.5, or at least at pH 7.0, or at least at pH 7.2, or at least at pH 7.5, or at least at pH 8.0, or at least at pH 8.5, or at least at pH 9.0, or at least at pH 9.5, or at least at pH 10.0, or at least at pH 10.5, or at least at pH 1 1.0, or at least at pH 11.5, or at least at pH 12.0, or at least at pH 12.5, or at least at pH 13.0, or at least at pH 13.5, or at pH 14. In certain embodiments, the functionally equivalent variant of the split intein C-fragment of this disclosure maintains or improves its activity at pH 7.2. In another embodiment, the functionally equivalent variant of the split intein C-fragment of this disclosure maintains or improves its activity at urea 1 M, or at least at urea 1.5 M, or at least urea 2 M , or at least urea 3 M, or at least urea 3.5 M, or at least urea 4 M, or at least urea 4.5 M, or at least urea 5 M. In certain embodiments, the functionally equivalent variant of the split C-intein of this disclosure maintains or improves its activity at urea 2 M or urea 5 M. In certain embodiments, the functionally equivalent variant of the split C-intein of this disclosure maintains or improves its activity at a temperature of 50ºC, at pH 7.2 and at urea 2 M or urea 4 M. All possible combinations of temperatures, urea concentration and pH are also contemplated by this disclosure.

In certain embodiments, the functionally equivalent variant of the split intein C-fragment of this disclosure that maintains or improves its activity has at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98% or at least 99% sequence identity with SEQ ID NO: 7.

In another embodiment, the functionally equivalent variant of the split intein C-fragment comprises or consist on an amino acid sequence selected from the group consisting of SEQ ID NO: 10-22 and 128-140.

Complex comprising a split intein C-fragment

In another aspect, this disclosure relates to a complex, hereinafter second complex of this disclosure, comprising:

(i) the split intein C-fragment of SEQ ID NO: 7 or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120 and

(ii) a compound of interest

wherein the complex optionally comprises a linker between (i) and (ii) and wherein

- the compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest if bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by and amide linkage.

The terms “compound of interest” and “linker” have been previously defined in connection with the first complex of this disclosure. All the embodiments of the compound of interest and linker of the first complex of this disclosure fully apply to the second complex of this disclosure.

In certain embodiments, the complex does not comprise a linker between the compound of interest and the split intein C-fragment. In this embodiment, the compound of interest is linked to the C-terminus of the split intein C-fragment by an amide linkage.

In certain embodiments, the complex comprises a linker between the compound of interest and the split intein C-fragment. In this embodiment, the compound of interest may be bound to the linker by any suitable means, depending on the chemical nature of the compound of interest and of the linker. In this embodiment, the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage. In another embodiment, the compound of interest is bound to the linker by an amide linkage, in which case the linker may be bound to the C-terminus of the split intein C-fragment by any suitable means. In another embodiment, the compound of interest is bound to the linker by an amide linkage and the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage.

In another embodiment, the compound of interest is a protein having the N-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split intein C-fragment of sequence SEQ ID NO: 7. In another embodiment, the compound of interest is a protein having the sequence Cys-Xaa_1-Xaa_2 or Cys-Xaa_1-Xaa_2-Leu in its N-terminus, where:

- Xaa_1 and Xaa_2 are any amino acid;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is any amino acid;
- Xaa_1 is any amino acid and Xaa_2 is Gly, Glu, Ala or Arg;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is Gly, Glu, Ala or Arg.

In another embodiment, the compound of interest is a protein having a sequence selected from Cys-Glu-Phe, Cys-Ala-Phe; Cys-Gly-Phe; Cys-Arg-Phe, Cys-Phe-Phe, Cys-Glu-Gly, Cys-Glu-Glu, Cys-Glu-Ala, Cys-Glu-Phe-Leu, Cys-Ala-Phe-Leu; Cys-Gly-Phe-Leu; Cys-Arg-Phe-Leu, Cys-Phe-Phe-Leu, Cys-Glu-Gly-Leu, Cys-Glu-Glu-Leu and Cys-Glu-Ala-Leu in its N-terminus.

In another embodiment, when the compound of interest is not a protein, the C-intein comprises or consists on a polypeptide selected from the group consisting of SEQ ID NO: 10-48 or SEQ ID NO: 128-166. In another embodiment, when the compound of interest is not a protein, the compound of interest and the C-intein are joined through a linker in which case, the linker is a peptide having the N-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split intein C-fragment of sequence SEQ ID NO: 7; in certain embodiments, the linker is a peptide having the sequence Cys-Xaa_1-Xaa_2 or Cys-Xaa_1-Xaa_2-Leu in its N-terminus, where:

- Xaa_1 and Xaa_2 are any amino acid;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is any amino acid;
- Xaa_1 is any amino acid and Xaa_2 is Gly, Glu, Ala or Arg;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is Gly, Glu, Ala or Arg;

or the linker is a peptide having a sequence selected from Cys-Glu-Phe, Cys-Ala-Phe, Cys-Gly-Phe, Cys-Arg-Phe, Cys-Phe-Phe, Cys-Glu-Gly, Cys-Glu-Glu, Cys-Glu-Ala, Cys-Glu-Phe-Leu, Cys-Ala-Phe-Leu, Cys-Gly-Phe-Leu, Cys-Arg-Phe-Leu, Cys-Phe-Phe-Leu, Cys-Glu-Gly-Leu, Cys-Glu-Glu-Leu and Cys-Glu-Ala-Leu in its N-terminus.

In another embodiment, the compound of interest is a protein that does not have the N-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split C-intein of SEQ ID NO: 7, in which case (i) the C-intein comprises or consists on the polypeptide of sequence SEQ ID NO: 10-44 or 128-166 or (ii) the compound of interest and the C-intein are joined through a linker in which case, the linker is a peptide having the C-terminal amino acid residues of the extein capable of being spliced by an intein comprising the split intein C-fragment of SEQ ID NO: 7; in certain embodiments, the linker is a peptide having the sequence Cys-Xaa_1-Xaa2 or Cys-Xaa_1-Xaa_2-Leu in its N-terminus, where:

- Xaa_1 and Xaa_2 are any amino acid;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is any amino acid;
- Xaa_1 is any amino acid and Xaa_2 is Gly, Glu, Ala or Arg;
- Xaa_1 is Ala, Gly, Art or Phe and Xaa_2 is Gly, Glu, Ala or Arg;

or the linker is a peptide having a sequence selected from Cys-Glu-Phe, Cys-Ala-Phe, Cys-Gly-Phe, Cys-Arg-Phe, Cys-Phe-Phe, Cys-Glu-Gly, Cys-Glu-Glu, Cys-Glu-Ala, Cys-Glu-Phe-Leu, Cys-Ala-Phe-Leu; Cys-Gly-Phe-Leu; Cys-Arg-Phe-Leu, Cys-Phe-Phe-Leu, Cys-Glu-Gly-Leu, Cys-Glu-Glu-Leu and Cys-Glu-Ala-Leu in its N-terminus.

In certain embodiments, the compound of interest is a protein or polypeptide.

In another embodiment, the compound of interest is a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa.

In certain embodiments, the protein is Cas9 or a fragment of Cas9.In certain embodiments, the compound of interest is a polypeptide or protein, and if the complex comprises a linker, the linker is a peptide linker. In this embodiment, the complex is a fusion protein.

In another embodiment, the polypeptide of interest is an antibody or a fragment of an antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDec205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In another embodiment, the compound of interest is a fragment of a protein; in certain embodiments, a fragment of a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa. In another embodiment, the compound of interest is a C-terminal fragment of a protein. The term “C-terminal fragment of a protein”, as used herein, refers to a fragment of variable length that includes the C-terminus of the protein. In certain embodiments, the C-terminal fragment is a fragment comprising less than 100%, less than 90%, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than 5% of the length of the whole protein.

In another embodiment, the compound of interest is an antibody. The term antibody has been described within the context of the N-inteins and equally apply to the present case.

In certain embodiments, the complex comprises a split intein C-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120.

In certain embodiments, the sequences of SEQ ID NO: 123 and 124 have higher thermal stability than the sequence of SEQ ID NO: 7.

In certain embodiments, the complex comprises a split intein C-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 69-87 or a variant thereof. In certain embodiments, the variant is a functionally equivalent variant.

The terms “variant” and “functionally equivalent variant” have been previously defined. In certain embodiments, the functionally equivalent variants of the split intein C-fragments of SEQ ID NO: 69-87 have at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95% or at least 99% sequence identity with the sequence from which they derive.

In certain embodiments, the functionally equivalent variants of the split intein C-fragments of SEQ ID NO: 69-87 maintain or improve the activity from the sequence from which they derive. The term “activity” as well as methods to measure this activity have been previously defined in connection with the functionally equivalent variants of the split intein N-fragment of SEQ ID NO: 7. The embodiments regarding the activity of the variants of the split intein C-fragment of SEQ ID NO: 7 fully applies to the activity of the variants of the split intein C-fragments of SEQ ID NO: 69-87.

<u>Complex comprising a split intein N-fragment and a split intein C-fragment</u>

In another aspect, this disclosure relates to a complex, hereinafter third complex of this disclosure, comprising:

(iv) the split intein C-fragment of this disclosure or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120

(v) a compound of interest and

(vi) the split intein N-fragment of this disclosure, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

wherein the complex optionally comprises a linker between (i) and (ii) and/or between (ii) and (iii),

wherein

- the compound of interest is linked to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage and
- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

The terms “compound of interest” and “linker” have been previously defined in connection with the first complex of this disclosure. All the embodiments of the compound of interest and linker of the first complex of this disclosure fully apply to the second complex of this disclosure.

In certain embodiments, the compound of interest is a protein or polypeptide.

In another embodiment, the compound of interest is a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa. In certain embodiments, the compound of interest is a polypeptide or protein, and if the complex comprises a linker, the linker is a peptide linker. In this embodiment, the complex is a fusion protein.

In certain embodiments, the polypeptide of interest is an antibody of a fragment of an antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In certain embodiments, the complex comprises a split intein C-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120.

In certain embodiments, the sequences of SEQ ID NO: 123 and 124 have higher thermal stability than the sequence of SEQ ID NO: 7.

In certain embodiments, the complex comprises a split intein C-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID

NO: 69-87 or a variant thereof. In certain embodiments, the variant is a functionally equivalent variant.

In certain embodiments, the complex comprises a split intein N-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 111, 112 and 113.

In certain embodiments, the sequences of SEQ ID NO: 112 and 113 have higher thermal stability than the sequence of SEQ ID NO: 1.

In certain embodiments, the complex comprises a split intein N-fragment comprising or consisting of an amino acid sequence selected from the group consisting of SEQ ID NO: 49-68 or a variant thereof. In another embodiment, the variant is a functionally equivalent variant.

The terms “variant” and “functionally equivalent variant” have been previously defined. The embodiments regarding these terms fully applies to the third complex of this disclosure.

Composition comprising the complexes of this disclosure

In another aspect, this disclosure relates to a composition, hereinafter first composition of this disclosure, comprising the first and the second complex of this disclosure.

The term “composition” is intended to encompass a product containing the specified components, as well as any product that results, directly or indirectly, from a combination of the specified components in the specified amounts. The components of the composition may be packed together in a single formulation or separately in different formulations. Thus in an embodiment the first complex of this disclosure is packed together with the second complex of this disclosure in a single formulation. In another embodiment, the first complex of this disclosure and of the second complex of this disclosure are separately packed.

In one embodiment, the first and the second complex comprise the N-terminal fragment and the C-terminal fragment of the same protein respectively, in such a way that when both complexes are combined according to the methods of this disclosure, the N-terminal fragment of the protein is linked to the C-terminal fragment of the protein generating the whole protein.

Conjugates of this disclosure

In another aspect, this disclosure relates to a conjugate, hereinafter first conjugate of this disclosure, comprising the first complex of this disclosure and the second complex of this disclosure, wherein the C-terminus of the split intein N-fragment is linked to the N-terminus of the split intein C-fragment by a peptide bond.

In another aspect, this disclosure relates to a conjugate, hereinafter second conjugate of this disclosure, comprising (a) the first complex of this disclosure and (b) a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, wherein the C-terminus of the split intein N-fragment is linked to the N-terminus of the split intein C-fragment by a peptide bond.

In certain embodiments, the conjugate comprises a split intein C-fragment comprising or consisting of a sequence selected from SEQ ID NO: 121-124.

In certain embodiments, the conjugate comprises a split intein C-fragment comprising or consisting of a sequence selected from SEQ ID NO: 69-87 or a variant thereof. In certain embodiments, the variant is a functionally equivalent variant. The functionally equivalent variants of the split intein C-fragment of SEQ ID NO: 69-87 have been previously defined.

In certain embodiments, the compound of interest is a protein or polypeptide.

In another embodiment, the compound of interest is a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa.

In certain embodiments, the protein is Cas9 or a fragment of Cas9.

In certain embodiments, the compound of interest is a polypeptide or protein, and if the complex comprises a linker, the linker is a peptide linker.

In certain embodiments, the polypeptide of interest is an antibody or a fragment of an antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

Polynucleotides, vectors and host cells of this disclosure

In another aspect, this disclosure relates to a polynucleotide encoding:

- the split intein N-fragment of this disclosure, or
- the split intein C-fragment of this disclosure, or
- the first, second or third complex of this disclosure, wherein the compound of interest is a polypeptide or protein and the linker, if present, is a peptide linker, or
- the conjugate of this disclosure.

As used herein, the term "polynucleotide" refers to a polymer composed of a multiplicity of nucleotide units (deoxyribonucleotides or ribonucleotides, or related structural variants or synthetic analogues thereof) linked via phosphodiester bonds (or related structural variants on synthetic analogues thereof). The term polynucleotide includes double or single stranded genomic and cDNA, RNA, any synthetic and genetically manipulated polynucleotide, and both sense and anti-sense polynucleotide (although only sense stands are being disclosed in the present disclosure). This includes single- and double-stranded molecules, i.e., DNA-DNA, DNA-RNA and RNA-RNA hybrids.

The polynucleotide of this disclosure can be found isolated as such or forming part of vectors allowing the propagation of said polynucleotides in suitable host cells. Therefore, in another aspect, this disclosure relates to a vector comprising the polynucleotide of this disclosure as described above.

Vectors suitable for the insertion of said polynucleotide are vectors derived from expression vectors in prokaryotes such as pUC18, pUC19, Bluescript and the derivatives thereof, mpl8, mpl9, pBR322, pMB9, ColEl, pCRl, RP4, phages and "shuttle" vectors such as pSA3 and pAT28; expression vectors in yeasts such as vectors of the type of 2 micron plasmids, integration plasmids, YEP vectors, centromere plasmids and the like; expression vectors in insect cells such as vectors of the pAC series and of the pVL; expression vectors in plants such as plBl, pEarleyGate, pAVA, pCAMBIA, pGSA, pGWB, pMDC, pMY, pORE series and the like; and expression vectors in eukaryotic cells, including baculovirus suitable for transfecting insect cells using any commercially available baculovirus system. The vectors for eukaryotic cells include viral vectors (adenoviruses, adeno associated viruses (AAV),retroviruses and lentiviruses) as well as non-viral vectors such as pSilencer 4.1-CMV (Ambion), pcDNA3, pcDNA3.1/hyg, pHMCV/Zeo, pCR3.1, pEFI/His, pIND/GS, pRc/HCMV2, pSV40/Zeo2, pTRACER-HCMV, pUB6/V5-His, pVAXl, pZeoSV2, pCl, pSVL and PKSV-10, pBPV-1, pML2d and pTDTl .

The vectors may also comprise a reporter or marker gene which allows identifying those cells that have incorporated the vector after having been put in contact with it.

Useful reporter genes in the context of the present disclosure include lacZ, luciferase, thymidine kinase, GFP and on the like. Useful marker genes in the context of this disclosure include, for example, the neomycin resistance gene, conferring resistance to the aminoglycoside G418; the hygromycin phosphotransferase gene, conferring resistance to hygromycin; the ODC gene, conferring resistance to the inhibitor of the ornithine decarboxylase (2-(difluoromethyl)-DL-ornithine (DFMO); the dihydrofolatereductase gene, conferring resistance to methotrexate; the puromycin-N-acetyl transferase gene, conferring resistance to puromycin; the ble gene, conferring resistance to zeocin; the adenosine deaminase gene, conferring resistance to 9-beta-D-xylofuranose adenine; the cytosine deaminase gene, allowing the cells to grow in the presence of N-(phosphonacetyl)-L-aspartate; thymidine kinase, allowing the cells to grow in the presence of aminopterin; the xanthine-guanine phosphoribosyltransferase gene, allowing the cells to grow in the presence of xanthine and the absence of guanine; the trpB gene of *E. coli*, allowing the cells to grow in the presence of indol instead of tryptophan; the hisD gene of *E. coli*, allowing the cells to use histidinol instead of histidine. The selection gene is incorporated into a plasmid that can additionally include a promoter suitable for the expression of said gene in eukaryotic cells (for example, the CMV or SV40 promoters), an optimized translation initiation site (for example, a site following the so-called Kozak's rules or an IRES), a polyadenylation site such as, for example, the SV40 polyadenylation or phosphoglycerate kinase site, introns such as, for example, the beta-globulin gene intron. Alternatively, it is possible to use a combination of both the reporter gene and the marker gene simultaneously in the same vector.

On the other hand, as the skilled person in the art knows, the choice of the vector will depend on the host cell in which it will subsequently be introduced. By way of example, the vector in which said polynucleotide is introduced can also be a yeast artificial chromosome (YAC), a bacterial artificial chromosome (BAC) or a PI -derived artificial chromosome (PAC). The characteristics of the YAC, BAC and PAC are known by the person skilled in the art. Detailed information on said types of vectors has been provided, for example, by Giraldo and Montoliu (Giraldo, P. & Montoliu L., 2001 Size matters: use of YACs, BACs and PACs in transgenic animals, Transgenic Research 10(2): 83-110). The vector of this disclosure can be obtained by conventional methods

known by persons skilled in the art (Sambrook J. *et al.*, 2000 "Molecular cloning, a Laboratory Manual", 3rd ed., Cold Spring Harbor Laboratory Press, N.Y. Vol 1-3).

The polynucleotide of this disclosure can be introduced into the host cell in vivo as naked DNA plasmids, but also using vectors by methods known in the art, including but not limited to transfection, electroporation (e.g. transcutaneous electroporation), microinjection, transduction, cell fusion, DEAE dextran, calcium phosphate precipitation, use of a gene gun, or use of a DNA vector transporter. Methods for formulating and administering naked DNA to mammalian muscle tissue are also known. See Feigner P, *et al.*, US 5,580,859, and US 5,589,466. Other molecules are also useful for facilitating transfection of a nucleic acid in vivo, such as cationic oligopeptides, peptides derived from DNA binding proteins, or cationic polymers. See Bazile D, *et al.*, WO 1995021931, and Byk G, *et al.*, WO 1996025508.

Another well-known method that can be used to introduce polynucleotides into host cells is particle bombardment (aka biolistic transformation). Biolistic transformation is commonly accomplished in one of several ways. One common method involves propelling inert or biologically active particles at cells. See Sanford J, *et al.*, US 4,945,050, US 5,036,006, and US 5,100,792.

Alternatively, the vector can be introduced *in vivo* by lipofection. The use of cationic lipids can promote encapsulation of negatively charged nucleic acids, and also promote fusion with negatively charged cell membranes. See Feigner P, Ringold G, Science 1989; 337:387-388. Useful lipid compounds and compositions for transfer of nucleic acids have been described. See Feigner P, *et al.*, US 5,459,127, Behr J, *et al.*, WO1995018863, and Byk G, WO1996017823.

Thus, in another aspect, this disclosure relates to a host cell comprising the polynucleotide or the vector of this disclosure. The cells can be obtained by conventional methods known by persons skilled in the art (see e.g. Sambrook *et al.*, cited ad supra).

The term "host cell", as used herein, refers to a cell into which a nucleic acid of this disclosure, such as a polynucleotide or a vector according to this disclosure, has been introduced and is capable of expressing the split intein N-fragment of this disclosure or the fusion protein comprising said split intein N-fragment. The terms "host cell" and "recombinant host cell" are used interchangeably herein. It should be understood that such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding

generations due to either mutation or environmental influences, such progeny may not, in fact be identical to the parent cell, but are still included within the scope of the term as used herein. The term includes any cultivatable cell that can be modified by the introduction of heterologous DNA. In certain embodiments, a host cell is one in which the polynucleotide of this disclosure can be stably expressed, post-translationally modified, localized to the appropriate subcellular compartment, and made to engage the appropriate transcription machinery. The choice of an appropriate host cell will also be influenced by the choice of detection signal. For example, reporter constructs, as described above, can provide a selectable or screenable trait upon activation or inhibition of gene transcription in response to a transcriptional regulatory protein; in order to achieve optimal selection or screening, the host cell phenotype will be considered. A host cell of the present disclosure includes prokaryotic cells and eukaryotic cells. Prokaryotes include gram negative or gram positive organisms, for example, *E. coli* or *Bacilli*. It is to be understood that in certain embodiments prokaryotic cells will be used for the propagation of the transcription control sequence comprising polynucleotides or the vector of the present disclosure. Suitable prokaryotic host cells for transformation include, for example, *E. coli*, *Bacillus subtilis*, *Salmonella typhimurium*, and various other species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*. Eukaryotic cells include, but are not limited to, yeast cells, plant cells, fungal cells, insect cells (e.g., baculovirus), mammalian cells, and the cells of parasitic organisms, e.g., trypanosomes. As used herein, yeast includes not only yeast in a strict taxonomic sense, i.e., unicellular organisms, but also yeast-like multicellular fungi of filamentous fungi. Exemplary species include *Kluyverei lactis*, *Schizosaccharomyces pombe*, and *Ustilaqo maydis*, and *Saccharomyces cerevisiae*. Other yeasts which can be used in practicing the present disclosure are *Neurospora crassa*, *Aspergillus niger*, *Aspergillus nidulans*, *Pichia pastoris*, *Candida tropicalis*, and *Hansenula polymorpha*. Mammalian host cell culture systems include established cell lines such as COS cells, L cells, 3T3 cells, Chinese hamster ovary (CHO) cells, embryonic stem cells, BHK, HeK, or HeLa cells. In certain embodiments, eukaryotic cells are used for recombinant gene expression.

Methods to conjugate two compounds of interest

In another aspect, this disclosure relates to a method to obtain a conjugate between a first compound of interest and a second compound of interest comprising:

(i) contacting

(a) the first complex of this disclosure, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 with

(b) the second complex of this disclosure, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120

or

a complex comprising an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof and the second compound of interest, wherein the complex optionally comprises a linker between the split intein C-fragment and the second compound of interest and wherein

- the second compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the second compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

In another aspect, this disclosure relates to a method to obtain a conjugate between a first compound of interest and a second compound of interest comprising

(i) contacting

(a) the first complex of this disclosure, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

or

a complex comprising complex comprising a compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) the complex of any one of claims 17 to 21, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

The term “AceL-TerL intein”, as used herein, refers to a family of non-canonical split inteins identified in the Antarctic permanently stratified saline lake, Ace Lake. This family of inteins was described by Thiel et al., Angew. Chem. Int. Ed 2014, 53: 1306-1310. In certain embodiments, the AceL-TerL split intein N-fragment comprises or consists on the sequence of SEQ ID NO: 101 or 102. In certain embodiments, the AceL-TerL split intein C-fragment comprises or consists on the sequence of SEQ ID NO: 99 or 100.

The terms “compound of interest” and “functionally equivalent variant” have been previously defined. In some embodiments, the first compound and/or the second compound is or includes a peptide or a polypeptide. In some embodiments the first compound and/or the second compound is or includes an antibody, antibody chain, or antibody heavy chain. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest

is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

. In some embodiments, the first compound and/or the second compound is or includes a peptide, oligonucleotide, drug, or cytotoxic molecule.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 111-113.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 121-124.In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof and the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

The appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate can be easily determined by the skilled person. In certain embodiments, these conditions involve contacting the first and second complex at temperature between 0ºC and 70ºC, for example, between 5ºC and 65ºC, between 10ºC and 60ºC, between 15ºC and 55ºC, between 20ºC and 50ºC, between 25ºC and 45ºC, between 30ºC and 40ºC, between 25ºC and 35ºC, between 45ºC and 55ºC; in certain embodiments at 30ºC or 50ºC. In another embodiment the conditions involve contacting the first and second complex at a pH between 0.1 and 14, for example between 0.5 and 13.5, between 1.0 and 13.0, between 1.5 and 12.5, between 2.0 and 12.0, between 2.5 and 11.5, between 3.0 and 11.0, between 3.5 and 10.5, between 4.0 and 10.0, between 4.5 and 9.5, between 5.0 and 9.0, between 5.5 and 8.5, between 6.0 and 8.0, between 6.5 and 7.5; in certain embodiments at pH 7.2. In another embodiment, these conditions involve contacting the first and second complex

in the absence of urea, or in the presence of urea at a concentration between 1 M and 5 M, for example between 1.5 M and 4.5 M, between 2 M and 4.0 M, between 2.5 M and 3.5 M; in certain embodiments at urea 2 M or at urea 4 M. In certain embodiments. In certain embodiments, these conditions involve contacting the first and second complex at a temperature of 50ºC, at pH 7.2 and in the presence of urea 2 M or urea 4 M. All possible combinations of temperatures, urea concentration and pH are also contemplated by this disclosure.

Method to obtain a conjugate of a compound of interest and a nucleophile

In another aspect this disclosure relates to a method to obtain a conjugate of a compound of interest with a nucleophile comprising

(i) contacting

(a) the first complex of this disclosure, wherein the split intein N-fragment comprises the amino acid sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1 or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

or

a complex comprising a compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) a split intein C-fragment comprising an amino acid sequence selected from the group consisting of SEQ ID NO: 8, 9, 23-48 and 141-166, under appropriate conditions for binding between the split intein N-fragment and the split intein C-fragment to form an intein intermediate and

(ii) contacting the intein intermediate with an exogenous nucleophile.

The terms “AceL-TerL split intein N-fragment”, “compound of interest” and “functionally equivalent variant” have been previously defined. In certain embodiments, the AceL-

TerL split intein N-fragment comprises or consist on the sequence of SEQ ID NO: 101 or 102. In some embodiments, the first compound and/or the second compound is or includes a peptide or a polypeptide. In some embodiments the first compound and/or the second compound is or includes an antibody, antibody chain, or antibody heavy chain. In certain embodiments, the polypeptide of interest is an antibody or a fragment of an antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In some embodiments, the first compound and/or the second compound is or includes a peptide, oligonucleotide, drug, or cytotoxic molecule.

The term “nucleophile,” as used herein, refers to any chemical species that donates an electron pair to an electrophile to form a chemical bond in relation to a reaction. All molecules or ions with a free pair of electrons or at least one pi bond can act as nucleophiles. Because nucleophiles donate electrons, they are by definition Lewis bases. In one embodiment of the present disclosure, a nucleophile may be either a sulfur nucleophile or a nitrogen nucleophile.

The term “sulfur nucleophile,” as used herein, refers to a nucleophile comprising at least one sulfur atom. The example of sulfur nucleophile may include hydrogen sulfide and its salts, thiols (RSH), thiolate anions (RS −), anions of thiolcarboxylic acids (RC(O)—S −), and anions of dithiocarbonates (RO—C(S)—S −) and dithiocarbamates (R 2N—C(S)—S −). In one embodiment of the present disclosure, the sulfur nucleophile is MESNA or DTT.

The term “nitrogen nucleophile,” as used herein, refers to a nucleophile comprising at least one nitrogen atom. Nitrogen nucleophiles include ammonia, azide, amines, hydrazines, and nitrites. In one embodiment of the present disclosure, the nitrogen nucleophile is hydrazine.

The term “exogenous nucleophile”, as used herein, means that the nucleophile does not form part of the complex of this disclosure or of the split intein C-fragment.

Thus, in the present method, wherein the compound of interest is a protein or a polypeptide, the intein intermediate is reacted with a nucleophile to release the polypeptide of interest from the bound intein N- and C-fragments thereby obtaining a

protein or polypeptide having a C-terminus modified by the nucleophile. The type of modification will depend on the type of nucleophile. For example, when the nucleophile is a thiol, the modified polypeptide of interest is an α-thioester, which in turn can be further modified, e.g., with a different nucleophile (e.g., a drug, a polymer, another polypeptide, a oligonucleotide), or any other moiety using the well-known α -thioester chemistry for protein modification at the C-terminus. One advantage of this chemistry is that only the C-terminus is modified with a thioester for further modification, thus allowing for selective modification only at the C terminus and not at any other acidic residue in the polypeptide. In the case wherein the compound of interest is not a protein or a polypeptide the compound of interest will carry a moiety able to react with the nucleophile, that is, an electrophile. Suitable electrophiles capable to react with a nucleophile are commonly known in the field.

In certain embodiments, the nucleophile is added to the reaction after contacting the first complex of this disclosure and the split intein C-fragment. In another embodiment, the first complex of this disclosure, the split intein C-fragment and the nucleophile are contacted simultaneously.

In certain embodiments, the method further comprises contacting the conjugate of the compound of interest and the nucleophile with a second exogenous nucleophile.

The nucleophile that is used in the methods disclosed herein either with the intein intermediate or as a subsequent or second nucleophile reacting with, e.g., an α-thioester, can be any compound or material having a suitable nucleophilic moiety. For example, to form a thioester, a thiol moiety is contemplated as the nucleophile. In some cases, the thiol is a 1,2 aminothiol, or a 1,2-aminoselenol. An α-selenothioester can be formed by using a selenothiol (R-SeH). Alternative nucleophiles contemplated include amines (i.e. aminolysis to give amides directly), hydrazines (to give hydrazides), amino-oxy groups (to give hydroxamic acids). Additionally, the nucleophile can be a functional group within a compound of interest for conjugation to the polypeptide of interest (e.g., a drug to form a protein-drug conjugate) or could alternatively bear an additional functional group for subsequent known bioorthogonal reactions such as an azide or an alkyne (for a click chemistry reaction between the two function groups to form a triazole), a tetrazole, an a-ketoacid, an aldehyde or ketone, or a cyanobenzothiazole.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 111-113.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

Composition comprising polynucleotides

In another aspect, this disclosure relates to a composition, hereinafter second composition of this disclosure, comprising:

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a variant thereof or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and
- a second polypeptide of interest

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and

- a second polypeptide of interest.

In certain embodiments, the variants are functionally equivalent variants.

The term “composition” has been previously defined. In certain embodiments, the first polynucleotide is packed together with the second polynucloetide in a single formulation. In another embodiment, the first polynucleotide and of the second polynucleotide are separately packed.

The term “AceL-TerL intein” has been previously defined. In certain embodiments, the AceL-TerL split intein N-fragment comprises or consists on the sequence of SEQ ID NO: 101 or 102. In certain embodiments, the AceL-TerL split intein C-fragment comprises or consists on the sequence of SEQ ID NO: 99 or 100.

In certain embodiments, the first polypeptide of interest is the N-terminal fragment of a protein and the second polypeptide of interest is the C-terminal fragment of said protein; in certain embodiments a protein of more than 25 KDa, more than 50 KDa or more than 100 KDa, such that upon covalently linking the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest the whole protein is obtained.

In some embodiments the first compound and second compound is or includes an antibody, antibody chain, or antibody heavy chain. In certain embodiments, the polypeptide of interest is an antibody or a fragment of an antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 111-113.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 121-124.

In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof and the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

The second composition of this disclosure can be used for expressing a gene of interest in a cell using the method of this disclosure.

Methods for expressing a gene of interest

In another aspect, this disclosure relates to a method for expressing a gene of interest in a cell, hereinafter fist method for expressing a gene of interest, comprising:

(i) contacting the cell with

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a

functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

(ii) allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and

(iii) allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.

In another aspect, this disclosure relates to a method for expressing a gene of interest, hereinafter second method for expressing a gene of interest of this disclosure, comprising:

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:
  - a first polypeptide of interest and
  - a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:
  - an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a

- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

or

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

(iii) allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and secreted,

(iv) allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.

The term “AceL-TerL intein” has been previously defined. In certain embodiments, the AceL-TerL split intein N-fragment comprises or consists on the sequence of SEQ ID NO: 101 or 102. In certain embodiments, the AceL-TerL split intein C-fragment comprises or consists on the sequence of SEQ ID NO: 99 or 100.

In certain embodiments, the first polypeptide of interest is the N-terminal fragment of a protein and the second polypeptide of interest is the C-terminal fragment of said protein; in certain embodiments a protein of more than 25 KDa, more than 50 KDa or

more than 100 KDa, so that upon covalently linking the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest the whole protein is obtained.

In certain embodiments, the first or second polypeptide of interest is Cas9 or a fragment of Cas9. In certain embodiments, the first polypeptide of interest is an N-terminal fragment of Cas9, and the second polypeptide of interest is a C-terminal fragment of Cas9. In another embodiment, when the first polypeptide of interest is an N-terminal fragment of Cas9 and the second polypeptide of interest is a C-terminal fragment of Cas9, upon covalently linking the C-terminus of the N-terminal fragment of Cas9 to the N-terminus of the C-terminal fragment of Cas9, the whole Cas9 protein is obtained

In some embodiments the first compound and/or the second compound is or includes an antibody, an antibody fragment, an antibody chain, or antibody heavy chain. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 antibody. In certain embodiments, the polypeptide of interest is the heavy chain of an anti-DEC-205 monoclonal antibody. In certain embodiments, the compound of interest is the heavy chain of the mouse αDEC-205 monoclonal antibody, as described by Stevens et al., JACS 2016, 138: 2162-5.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 111-113.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 121-124.

In certain embodiments, the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

In certain embodiments, the split intein N-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof and the split intein C-fragment comprises or consists of a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

The contacting of the cell with the first and/or second polynucleotide can be made by any suitable means for allowing introducing a polynucleotide of interest into a cell, for example, transfection, electroporation, microinjection, transduction, lipofection, cell fusion, DEAE dextran, calcium phosphate precipitation, use of a gene gun, or use of a DNA vector transporter. In the first method for expressing a gene of interest of this disclosure, it is contemplated that the cell is contacted simultaneously with the first and second polynucleotide, or sequentially with the first and second polynucleotide in any order, that is, the cell can be contacted firstly with the first polynucleotide and secondly with the second polynucleotide or firstly with the second polynucleotide and secondly with the first polynucleotide.

Any cell previously defined as a host cell can be used in these methods.

The term "signal peptide" or "secretory signal peptide", as used herein, refers to a peptide of a relatively short length, generally between 5 and 30 amino acid residues, directing proteins synthesized in the cell towards the secretory pathway. The signal peptide usually contains a series of hydrophobic amino acids adopting a secondary alpha helix structure. Additionally, many peptides include a series of positively-charged amino acids that can contribute to the protein adopting the suitable topology for its translocation. The signal peptide tends to have at its carboxyl end a motif for recognition by a peptidase, which is capable of hydrolyzing the signal peptide giving rise to a free signal peptide and a mature protein. The signal peptide can be cleaved once the protein of interest has reached the appropriate location. Any secretory signal peptide may be used in the present disclosure.

In certain embodiments, the signal peptide is linked to the N-terminus of the first polypeptide of interest in the first fusion protein.

In certain embodiments, the signal peptide is linked to the N-terminus of the split intein C-fragment in the second fusion protein.

***

The invention will be described by way of the following examples which are to be considered as merely illustrative and not limitative of the scope of this disclosure.

EXAMPLES

Materials and Methods

Materials

Oligonucleotides and synthetic genes were purchased from Integrated DNA Technologies (Coralville, IA). Pfu Ultra II Hotsart fusion polymerase for cloning was purchased from Agilent (La Jolla, CA). All restriction enzymes and 2x Gibson Assembly Master Mix were purchased from New England Biolabs (Ipswich, MA). High-competency cells used for cloning and protein expression were generated from One Shot Bl21 (DE3) chemically competent *E. coli* and sub-cloning efficiency DH5α competent cells purchased from Invitrogen (Carlsbad, CA). DNA purification kits were purchased from Qiagen (Valencia, CA). All plasmids were sequenced by GENEWIZ (South Plainfield, NJ). Luria Bertani (LB) media, and all buffering salts were purchased from Fisher Scientific (Pittsburgh, PA). Dimethylformamide (DMF), dichloromethane (DCM), Coomassie brilliant blue, triisopropylsilane (TIS), β-mercaptoethanol (BME), DL-dithiothreitol (DTT), sodium 2-mercaptoethanesulfonate (MESNa), 5(6)-carboxyfluorescein, and thermolysin were purchased from Sigma-Aldrich (Milwaukee, WI). Tris (2-carboxyethyl) phosphine hydrochloride (TCEP) and isopropyl-β-D-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). Roche Complete Protease Inhibitors were used for protein purification (Roche, Branchburg, NJ). Nickel-nitrilotriacetic acid (Ni-NTA) resin was purchased from Thermo scientific (Rockford, IL). Fmoc amino acids were purchased from Novabiochem (Darmstadt, Germany) or Bachem (Torrance, CA). *O*-(Benzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate (HBTU) was purchased from Genscript (Piscataway, NJ). Trifluoroacetic acid (TFA) was purchased from Halocarbon (North Augusta, SC). MES-SDS running buffer was purchased from Boston Bioproducts (Ashland, MA).

Equipment

Analytical reverse phase high performance liquid chromatography (RP-HPLC) was carried out on Hewlett-Packard 1100 and 1200 series instruments equipped with a C18 Vydac column (5 µm, 4.6 x 150 mm). All HPLC runs used the following solvents at a flow rate of 1 mL/min: 0.1 % TFA (trifluoroacetic acid) in water (solvent A) and 90 % acetonitrile in water with 0.1 % TFA (solvent B). All peptides and proteins were analyzed using the gradient: 0% B for 2 min followed by 0-73% B for 30 min. Electrospray ionization mass spectrometric analysis (ESI-MS) was carried out on a Bruker Daltonics MicroTOF-Q II mass spectrometer. Size-exclusion chromatography (SEC) was performed on an AKTA FPLC system (GE Healthcare) with a Superdex S75 16/60 column (125 mL column volume) for preparative runs and a Superdex S75 10/300 column for analytical runs. Gels were imaged with a LI-COR Odyssey Infrared

Imager. Circular dichroism experiments were carried out on a Chirascan Circular Dichroism spectrometer (Applied Photophysics). Cell lysis was carried out using a S-450D Branson Digital Sonifier. NMR experiments were carried out on a Bruker 900, 800, 600 and 500 MHz spectrometers with 5 mm TCI triple resonance cryoprobes. Steady state fluorescence measurements were performed on a Horiba Flourmax 4 fluorimeter. Stopped flow anisotropy measurements were performed on an Applied Photophysics SX20 stopped-flow spectrometer.

Consensus Protein Design

Homologues of AceL TerL were identified through a BLAST search of metagenomic data in the NCBI (nucleotide collection) and JGI databases using the TerL DNA sequence. This led to the identification of TerL N- and C-inteins with high sequence identity to AceL (Table 1). Because the cognate N- and C- inteins could not been matched, the split inteins were treated as two distinct datasets and analyzed separately. MSAs of these split inteins were then generated in Jalview[4], and the consensus sequence was determined. At some positions in the N-intein, additional residues from the alignment corresponding to loops not present in AceL were included in the consensus sequence.

Table 1. Identified TerL Inteins

C-intein	Source	Project #	Sequence	SEQ ID NO
AceL	Thiel et al[1]		MFRTNTNNIKILSPNGFSNFNGIQKVERNLY QHIIFDDDTEIKTSINHPFGKDKILARDVKV GDYLNSKKVLYNELVNENIFLYDPINVEKES LYITNGVVSHN	99
AceL*	Thiel et al[1]		MFRTNTNNIKILGPNGFSNFIGIQKVERDQY QHIIFDDDTEIKTSINHPFGKDKILARDVKV GDYLNSKKVLYNELVNENIFLYDPINVEKES LYITNGVVSHN	100
Cep	NCBI	CEPX01183120	MFKTNTNNIKILSPDGFSNFNGIQKVKRKLY QHIIFEGGTEIKTSINHSFGKDKILARDIKV GDYLNNKKVLYNELVNEKIFLYDPINVEKEN LYITNDVVSHN	69
Mdt	NCBI	MDTC01246584	MYKVNNNIKVKTPTGFQSFSGIQKVFKPFYH WIIFDDGSEIKCSDNHSFGSEKIKASSLKLD DIIQGKKVLYNEIVEEGIYLYDLLDVGEENL YITNKIISHN	70
Meh	NCBI	MEHZ010888690	MSKTYEVLSPSGFVKFSGIQKVSRSKYRHFI FDDGAELKCSLNHRFGKDEILASSLWPSSDL QGKNILYAEDVEEDIDLYDLLNVGGGNLYYT NGLVSHN	71
Aac	NCBI	AACY020064168	MFKINKNIKVKTPDGFKDFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGKEKIKASTIKVD DILQEKKVLYNEIVEEGIYLYDLLDVGEDNL YYSNNIVSHN	72
Aac2	NCBI	AACY023445674	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGKEKIKASTIRVD DFLQGKKVVYNEIVEEGIYLYDLLDVGENNL YYSNNIISHN	73
Cen	NCBI	CENI01048299	MYKLNSSIKVKTPRGFKKFAGIQKVRKPVYQ WIIFGDDSEIKCSLDHSFGEEQVKAHTIKTG	74

			DLLQHKEVVYSEIVEEPIDLYDLLEVEDGNL YNTNGVVSHN	
Cep2	NCBI	CEPQ01016765	MYEVLSPSGFVKFSGVQKVSRSKYRHFIFDD GTEIKCSLDHRFGGLWDEDEILASSLNRGEY LQGKKILYVEDVEEQIDLYDLMNVDGGNLYY TNGLVSHN	75
Cep3	NCBI	CEPZ01087314	MSKTYEVLSPSGFVKFSGIQKVSHSKYRHFI FDDGTELKCSFNHRFGKDEILASSLCRGSDL QGKKILYAEDVEEDIDLYDLLNVGGGNLYYT NGLVSHN	76
Cep4	NCBI	CEPZ01013308	MYIRYQKTTSKTYEVLSPSGFVNFSGIQTVP HSKYRHFIFDDGTELKCSLNHRFDKDEILAS SLWRGAELQGKQILYAEDIEEDIDLYDLLNV GGGNLYYTNGLVSHN	77
Cep5	NCBI	CEPS01165861	MFTKYKILTPNGYESFDGVNRIKRDMYSHLI FSSGIEIRCSLNHPLYISKGDIIKSYELKIG DKVLSKNGWEIVTYNEIIEEPIYLYDIINSG KDHNYYTNDILSHN	78
Cep6	NCBI	CEPZ01055800	MISKNFTKYKILTPNGYESFDGVNRIKRDMY SHLIFSSGIEIRCSLNHPLYISKGDIIKSYE LKIGDKVLSKNGWEIVTYNEIIEEPIYLYDI INSGKDHNYYTNDILSHN	79
Lak	JGI	Ga0169931	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGKEKIKASTIRVD DFLQGKKVVYNEIVEEGIYLYDLLDVGENNL YYSNDIISHN	80
Kab	JGI	Ga0172376	MFKLNKNIRVKTPSGFKSFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGEEQIKASSIKVD DFLQGKKVVYNEIVEEGIYLYDLLDVGEDNL YYSNDVVSHN	81
Chb	JGI	Ga0129336	MFKLNKNIKVKTPRGFKFFSGIQKVYKPYYH WIIFDDGSEIKCSDNHSFGKEKTKASTIKVD DFLQGKKVVYNEIVEEGIYLYDLLDVGEDNL YYSNEIISHN	82
Del	JGI	Ga0075462	MFKLNKNIKVKTPSGFKYFSGIQKVYKPFYH WIIFDDGTEIKCSDNHSFGKEQIKASMIKVD DFFQGKKVVYNEIVEEEIYLYDLLDVGEDNL YFSNGIISHN	83
Del2	JGI	Ga0075478	MFKLNKNIEVKTPDGFKSFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGKEKIKASTIKVD DLLQGKKVVYNEIVEEGIYLYDLLDVGEDNL YYSNNLVSHN	84
Del3	JGI	Ga0075478	MFKLNKNITVKTPSGFKSFSGIQKVYKPFYH WIIFDDGSEIKCSDNHSFGEEQIKASMIKVD DFLQGKKVVYNEVVEEGVYLYDLLDVGEDNL YYSNNIISHN	85
AceL2	JGI	Ga0075117	MFRTNTDNIKILSPSGFSNFNGIQKVERDLY QHIIFDDKSEIKTSINHPFGKDKILARNIKV GDYLNSKKVLYNELVAEKITLYDPINVEKEN LYITNGVISHN	86
AceL3	JGI	Ga0075117	MFRTNTDNIKILSPSGFSNFNGIQKVERDLY QHIIFDDKSEIKTSINHPFGKDKILARNIKV GDYLNSKKVLYNELVNEKITLYDPINVEKEN LYITNGVISHN	87
N-intein	**Source**	**Project #**	**Sequence**	**SEQ ID NO**
AceL	Thiel et al[1]		CVYGDTMVETEDGKIKIEDLYKRLA	101
AceL*	Thiel et al[1]		CVSGDTMVETEDGKIKIEDLYKRLA	102
AAC	NCBI	AACY023445674	CLGGDTIIEIQDDDGITQKISMEDLYERL	49
AAC2	NCBI	AACY020064168	CLGGDTEIEILDDNGIVQKTSMENLYERL	50
FUW	NCBI	FUWD010114546	CLGGETLIEIQDDNENISKVSMEDLYDRM	51
FUW2	NCBI	FUWD010387041	CVDGDTIVEIYDKKTKEEYCVKIKDLYDLI	52
FUW3	NCBI	FUWD012964875	CLSGDTQIEIKNVNDKIESVSMEELYERM	53
AAC3	NCBI	AACY020820060	CLSGDTMIEILDENGIPQKISMEDLYQR	54

MDT	NCBI	MDTB01192700	CVSGDTNIEIECEDGVETTTIKDLYDRM	55
CEP	NCBI	CEPX01183120	CVDGDTMVETEDGKIKIEDLYKKL	56
MEH	NCBI	MEHZ011579446	CVYGDTMVETEDGKIKIEDLYKKL	57
MDT2	NCBI	MDTC01246584	CVRGDTLVEVEKDDVISEMRIEDLYNRM	58
ABL	NCBI	ABLX01341501	CVGGNTLVEVEKDDIISEMRIEDLYNTM	59
SSF	JGI	Ga0102963	CLSGDTTIEILDVDGIPQKISMEDLYQRL	60
Del	JGI	Ga0075478_10047284	CLSGDTMIEILDESGIPQKISMKELYQRM	61
Del2	JGI	Ga0075478_10000264	CLDGNTSIEILDENNTIQKISMENLYKRL	62
Del3	JGI	Ga0070746	CLGGDTIIEIQDDDGITQKISMEDLYQRL	63
Del4	JGI	Ga0070752	CLDGGTSIEILDTNNITQKISLENLYERL	64
Del5	JGI	Ga0070749	CLSGDTSIEILDENNTIQKISMEDLYERL	65
Del6	JGI	Ga0070754	CLSGDTLIEIIDDDGNTQKISMEDLY	66
Del7	JGI	Ga0070751	CLSGDTLIEIIDDDGNTQKISMEDLYQ	67
Kab	JGI	Ga0172375	CLGGDTIIEIKDDDGITQKISMEDLYQRL	68

[1] Angewandte Chemie International Edition 2014, 53(5): 1306-1310

Cloning of Recombinant DNA

Synthetic genes were purchased and introduced into pET-30 expression vectors using Gibson assembly. Targeted mutations were introduced using inverse PCR with Pfu Ultra II HF Polymerase. The identity of all recombinant plasmids was confirmed through sequencing and the corresponding protein sequences are reported in Table 2.

Table 2. Sequence of proteins utilized in the present application.

Construct	[a]Sequence	SEQ ID NO
SUMO-Cat^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**EFEALSGDTMIEILDDDGIIQKISM EDLYQRLA**	88
SUMO-AceL*^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**EFEAVSGDTMVETEDGKIKIEDLYK RLA**	89
SUMO-GOS^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**EFEAISQESYINIEVNGKVETIKIG DLYKKLSFNERKFNE**	90
SUMO-Cat^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**MFKLNTKNIKVLTPSGFKSFSGIQK VYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDYLQGKKVLY NEIVEEGIYLYDLLNVGEDNLYYTNGIVSHACEFL**	91
[b]SUMO-Flag-Cat^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**DYKDDDDKMFKLNTKNIKVLTPSGF KSFSGIQKVYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDY LQGKKVLYNEIVEEGIYLYDLLNVGEDNLYYTNGIVSHACESRGK**	92
SUMO-AceL*^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD	93

	QTPEDLDMEDNDIIEAHREQIGG**MFRTNTNNIKILGPNGFSNFIGIQK VERDQYQHIIFDDDTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLY NELVNENIFLYDPINVEKESLYITNGVVSHACEFL**	
SUMO-GOS^{C}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**MKLPESVVKNNINLKIETPYGFENF YGVNKIKKDKYIHLEFTNGEKLKCSLDHPLSTIDGIVKAKDLDKYTEV YTKFGGCFLKKSKVINESIELYDIVNSGLKHLYYSNNIISHACEFL**	94
AceL*^{C}-GFP	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**MFRTNTNNIKILGPNGFSNFIGIQK VERDQYQHIIFDDDTEIKTSINHPFGKDKILARDVKVGDYLNSKKVLY NELVNENIFLYDPINVEKESLYITNGVVSHNCEFLMVSKGEELFTGVV PILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL VTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKT RAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADK QKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQ SALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKDYKDDDDK**	95
^{c}Cat^{C}-GFP	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**MFKLNTKNIKVLTPSGFKSFSGIQK VYKPFYHHIIFDDGSEIKCSDNHSFGKDKIKASTIKVGDYLQGKKVLY NEIVEEGIYLYDLLNVGEDNLYYTNGIVSHNCEFLMVSKGEELFTGVV PILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL VTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKT RAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADK QKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQ SALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKDYKDDDDK**	96
MBP-AceL*^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**KIEEGKLVIWINGDKGYNGLAEVGK KFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQS GLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAF KYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAF NKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAG INAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAK DPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEAP KDAQTNEFECVSGDTMVETEDGKIKIEDLYKRLA**	97
^{c}MBP-Cat^{N}	MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLK VSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQAD QTPEDLDMEDNDIIEAHREQIGG**KIEEGKLVIWINGDKGYNGLAEVGK KFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQS GLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAF KYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAF NKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAG INAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAK DPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEAP KDAQTNEFECLSGDTMIEILDDDGIIQKISMEDLYQRLA**	98
Fl-Cat^{N}	**Fl-GEFEALSGDTMIEILDDDGIIQKISMEDLYQRLA**	167

[a]The sequences shown correspond to the complete protein expressed by the pET-30 expression vector. The sequence corresponding to the protein cleaved from the SUMO expression tag is shown in bold. [b]The optimized Cat^{C} intein construct with appended charged residues utilized for the structural studies [c]The WT intein sequences are shown for both MBP-Cat^{N} and Cat^{C}-GFP. The underlined residues correspond to the positions of mutation for the extein activity screen.

Expression and Purification of Inteins for Splicing Assay

Expression and purification of the inteins was carried out as previously described. The expressed N-intein constructs contained the following architecture: His_6-SUMO-MBP-EFE-Int^N, where "His_6" is a 6x polyhistidine affinity tag, "SUMO" is the ubiquitin-like protein SMT3, "MBP" is maltose binding protein, "EFE" is the wild type -1, -2, and -3 N-extein sequence of TerL inteins, and Int^N is the N-intein. The expressed C-intein constructs contained the following architecture: His_6-SUMO-Int^C-CEFL-GFP, where "Int^C" is the C-intein, "CEFL" is the +1, +2, +3, and +4 C-extein residues of TerL inteins, and "GFP" is green fluorescent protein. For the screen of extein dependence, constructs corresponding to each indicated point mutation in the "EFE" or "CEFL" extein sequences were utilized.

E. coli BL21(DE3) cells were transformed with an MBP-Int^N or Int^C-GFP intein plasmid and grown at 37 °C in 1 L of LB containing 50 μg/mL of kanamycin. Once the culture reached an OD_{600}=0.6, 0.5 mM IPTG was added to induce expression (0.5 mM final concentration, 18 h at 18 °C). For test expression of the SUMO-Cat^C constructs, expression tests were also carried out at 37 °C for 3 hours upon addition of IPTG. Following expression, the cells were pelleted via centrifugation (5,000 rcf, 30 min) and stored at -80 °C.

The cell pellet was then resuspended in 30 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, pH 8.0) containing a protease inhibitor cocktail. The cells were lysed by sonication (35% amplitude, 8 x 20 s pulses on / 30 s off) and then pelleted by centrifugation (35,000 rcf, 30 min). The supernatant was incubated with 4 mL of Ni-NTA resin for 30 min at 4 °C to bind the His-tagged inteins. The slurry was then loaded onto a fritted column, the flow through was collected, and the column was washed with 20 mL of lysis buffer. The protein was then eluted from the column with 20 mL of elution buffer (lysis buffer + 250 mM imidazole).

The eluted protein was dialyzed into lysis buffer while being treated with 10 mM TCEP and Ulp1 protease overnight at 4 °C to cleave the His_6-SUMO expression tag. The dialyzed protein was then incubated with 4 mL Ni-NTA resin for 30 min at 4 °C, after which it was applied to a fritted column with the flow through collected together with a 10 mL wash of lysis buffer. The protein was then treated with 10 mM TCEP, concentrated to 2 mL, and purified over an S75 16/60 gel filtration column using degassed splicing buffer (100 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, pH 7.2) as the mobile phase. Fractions were analyzed by analytical RP-HPLC and ESI-MS

(FIG 1, Table 3), and either immediately utilized in the splicing assay or stored long term in glycerol (20% v/v) after being flash-frozen in liquid N_2.

Table 3. Masses of purified proteins.

eGFP Proteins	Expected Mass (Da)	Observed Mass (Da)
AceL*^{C}-GFP	40424.6	40424.3
Cat^{C}-GFP	40277.7	40275.6
Cat$^{C}_{+2G}$-GFP	40205.5	40205.8
Cat$^{C}_{+2A}$-GFP	40219.5	40218.7
Cat$^{C}_{+2R}$-GFP	40304.6	40305.8
Cat$^{C}_{+2F}$-GFP	40295.6	40294.9
Cat$^{C}_{+3R}$-GFP	40286.6	40286.4
Cat$^{C}_{+3A}$-GFP	40201.5	40201.0
Cat$^{C}_{+3G}$-GFP	40187.5	40187.3
Cat$^{C}_{+3E}$-GFP	40259.5	40258.4
MBP Proteins	**Expected Mass (Da)**	**Observed Mass (Da)**
MBP-Cat^{N}	44094.0	44093.4
MBP-AceL*^{N}	43508.3	43508.1
MBP-Cat$^{N}_{-1R}$	44121.1	44120.2
MBP-Cat$^{N}_{-1A}$	43036.0	44034.5
MBP-Cat$^{N}_{-1G}$	44022.0	44022.4
MBP-Cat$^{N}_{-1P}$	44062.0	44062.8
MBP-Cat$^{N}_{-2R}$	44103.0	44101.9
MBP-Cat$^{N}_{-2A}$	44017.9	44017.2
MBP-Cat$^{N}_{-2E}$	44076.0	44076.4
MBP-Cat$^{N}_{-2G}$	44003.9	44002.8
Proteins for NMR	**Expected Mass (Da)**	**Observed Mass (Da)**
FLAG-Cat^{C}	13499.2	13499.9
Cat^{N}	3773.2	3773.0
^{15}N, ^{13}C FLAG-Cat^{C}	14257.6	14247.5
^{15}N, ^{13}C Cat^{N}	3974.7	3972.7
Proteins for Binding	**Expected Mass (Da)**	**Observed Mass (Da)**
Fl-Cat^{N}	4187.6	4186.7
SUMO-Flag-Cat^{C}	26766.0	26764.7

Splicing Assays

Splicing assays were carried out as adapted from a previously described protocol.[8] Briefly, N- and C-inteins (4 μM Int^{N}, 4μM Int^{C}) were individually preincubated in splicing buffer (100 mM sodium phosphates, 150 mM NaCl, 1 mM EDTA, pH 7.2) with 2 mM TCEP for 15 min. Splicing reactions were carried out at indicated temperatures and concentrations of urea. For the extein characterization, the Cat^{C}-GFP and MBP-Cat^{N} proteins containing the indicating extein mutations were spliced with their cognate wild type N- or C- intein at 30 °C. Splicing of Cat and AceL* in the presence of urea was carried out at 30 °C. Splicing was initiated by mixing equal volumes of N- and C- inteins

with aliquots removed at the indicated times and quenched by the 1:1 addition of 4X loading dye (160 mM Tris, 40% glycerol, 4% SDS, 0.08% Bromophenol Blue, 8 % BME). Samples were analyzed by SDS-PAGE gel electrophoresis (12 % bis-tris, 60 min, 150 v) and quantified by densitometry (FIG. 2 and 3).

Kinetic analysis of *trans*-splicing reactions

To determine the splicing rates of *trans*-splicing reactions, the data was fit to the first order rate equation using GraphPad Prism software.

$$[P](t) = [P]_{max} \cdot \left(1 - e^{-kt}\right)$$

Where [*P*] is the normalized intensity of product, $[P]_{max}$ is the reaction plateau, and *k* is the rate constant (s^{-1}). The mean and standard error for each value are reported (n = 3).

Expression of Inteins for Structural Studies

Construct optimization was required in order to isolate Cat^C with minimal extein sequence for structural characterization. Compared to $AceL^{*C}$ and GOS^C, SUMO-Cat^C had increased yields during recombinant expression in *E. coli* (18 °C, 16 h or 37 °C for 3 h) (FIG 4). However, removal of the SUMO expression tag resulted in Cat^C aggregating upon cleavage (possibly due its neutral charge at physiological pH, pI = 7.2). Charged residues were therefore appended immediately flanking Cat^C to improve the solubility of the protein in solution, specifically an N-terminal FLAG epitope tag and "CESRGK" C-extein sequence (SUMO-Flag-Cat^C). The Cat^N construct utilized in these structural studies was expressed as a SUMO fusion (SUMO-Cat^N) and contains the minimal "EFE" N-extein following SUMO cleavage. In addition, inactivating C1A and N134A mutations were included in the constructs to prevent splicing during structural analysis of the associated complex. Expression and purification of these Cat^N and Cat^C constructs for structural study were carried out as described above for the proteins utilized for splicing.

For use in NMR spectroscopy, expression of the isotopically enriched Cat proteins was carried out as previously described. The intein plasmids were used to transform BL-21 (DE3) cells, and the cells were grown overnight in 5 mL LB starter cultures (37 °C, 18 h). The starter cultures were then spun down (4,000 rcf, 5 min). The supernatant was discarded, and the cells were then resuspended and grown in 1L of M9 medium supplemented with ^{13}C-glucose and $^{15}NH_4Cl$ as the sole carbon and nitrogen sources (50 µg/mL kanamycin, 37 °C). Once the cells reached OD_{600} = 0.6, expression was

induced with the addition of IPTG (0.5 mM, 18 h, 18 °C). Following expression, the cells were spun down by centrifugation (5,000 rcf, 30 min) and stored at -80 °C. Purification was carried out with the general method described above for intein constructs. The masses of the purified proteins correspond to an isotopic labeling efficiency of 99% for both the Cat^N and Cat^C proteins.

NMR Spectroscopy

NMR experiments were performed using Cat^N and Cat^C in free form and in complex. NMR samples were prepared by buffer exchanging purified protein to 20 mM sodium phosphate 150 mM NaCl, 2 mM TCEP (pH 6.8, 37 °C). The uniformly labeled ^{15}N, ^{13}C, ^{1}H proteins were concentrated to final concentrations of ~300-600 µM. For the HSQC experiments of the complex reported in figures 3A, 3B, the isotopically labeled intein fragments were mixed with the complementary unlabeled intein solution in a ratio of 1:1.5 and concentrated to a final concentration similar to the free protein and measured directly. For structure determination isotopically labeled intein fragments were mixed at a Cat^N:Cat^C ratio of 1.5:1. The complex was further purified by size exclusion chromatography to remove the free forms.

Experiments were performed at field strengths of 600, 700, 800 or 900 MHz and Non-Uniform Sampling (NUS) acquisition was employed as appropriate. NMR spectra were processed using Bruker Topspin 3.0 or NMR Pipe software and NUS spectra were reconstructed by compressed sensing using qMDD.

Chemical shift assignment

Backbone chemical shifts were assigned using HNCO, HN(CA)CO, HNCACB, CBCA(CO)NH triple resonance experiments. Side chain assignments were obtained from H(CC)(CO)NH, (H)CC(CO)NH, H(C)CH-TOCY and (H)CCH-TOCSY experiments. Aromatic assignments were obtained from CT-^{13}C-resolved [^{1}H,^{1}H]-NOESY (mixing time = 100 ms), (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE experiments. CcpNmr Analysis software was used for manual chemical shift assignment and other data analysis. Chemical shift values have been validated and deposited to the Biological Magnetic Resonance Bank (BMRB No : 30480). Random coil chemical shifts were calculated using CcpNmr analysis.

Spin relaxation measurements

Spin-spin relaxation (R_2) rates of ^{15}N spins (mixing times of 0, 17, 34, 51, 85, 119, 170, 255, 340, 510, 680 ms) and [^{15}N-^{1}H] NOE experiments were measured at a field strength of 600 MHz.

Structure determination

Dihedral angle restrains were calculated from chemical shifts using TALOS software.[13] NOE cross peaks were picked from ^{15}N-resolved [^{1}H,^{1}H]-NOESY (mixing time = 80 ms), ^{13}C-resolved-[^{1}H,^{1}H]-NOESY (mixing time = 80 ms), CT-^{13}C-resolved aromatic [^{1}H,^{1}H]-NOESY experiments (mixing time = 100 ms) and assigned automatically using ARIA and CNS softwares. Assignment and structure calculation was done in 8 cycles, calculating 20 structures in each step. The assigned NOEs were verified manually and violation analysis was done. The verified NOE peak lists were used to generate distance restraints. 3,283 unambiguous restraints, 206 ambiguous restraints and 180 dihedral angle restraints were used to finally calculate 256 structures. 20 least energy structures were selected and water refinement was performed. Structures have been validated and deposited to the Protein Data Bank (PDB ID : 6DSL).

Circular Dichroism (CD)

Cat^N, Cat^C, and 1:1 complex of Cat^N and Cat^C were dialyzed into CD buffer (25 mM sodium phosphate, 50 mM NaF, 1 mM DTT, pH 7.2). CD spectra were measured at 25 °C in a 1 mm pathlength cuvette (10 µM sample concentration).

Analytical Size Exclusion Chromatography (SEC)

Analytical SEC experiments were run on an S75 10/300 column at 4 °C in splicing buffer (25 mM sodium phosphate, 150 mM NaCl, 1 mM DTT, pH 7.2. For all runs, UV absorbance was monitored at 214 nm. Samples were injected with a sample volume of 500 µL (25 µM) and eluted with a flow rate of 0.5 mL/min.

Limited Proteolysis

EFE-Cat^N, Flag-Cat^C, and 1:1 complex of EFE-Cat^N and Flag-Cat^C were dialyzed into thermolysin buffer (50 mM Tris HCl, 100 mM NaCl, 2 mM MgSO4, 2 mM CaCl2, 1 mM DTT, pH 7.4) and diluted to a concentration of 10 µM. Thermolysin powder (Sigma) dissolved to 0.4 mg/mL in thermolysin buffer was then prepared and added to each solution (1:50 v/v). At the indicated times, aliquots were removed and quenched with the 1:3 addition of 8 M Guanidine HCL 4% TFA. The samples were then analyzed by RP-HPLC and ESI-MS. Masses from each peak were compared to predicted cleavage products of the inteins from ProteinProspector (UCSF).

Production of Inteins for Binding Experiments

The fluorescein labeled Cat^N (Fl-Cat^N) peptide was synthesized by standard 9-fluorenylmethyl-oxycarbonyl (Fmoc) solid phase peptide synthesis (SPPS). After coupling the last amino acid in the peptide, the N-terminus was capped with 5(6)-carboxyfluorescein. The synthesized Fl-Cat^N peptide was purified by preparative RP-HPLC and characterized by analytical RP-HPLC and ESI-MS. The C-intein expressed for the binding experiments was SUMO-Flag-Cat^C construct detailed above. Instead of carrying out an Ulp1 digestion, the expressed SUMO-Flag-Cat^C protein was purified directly over the S75 16/60 gel filtration column following Ni-NTA enrichment.

Steady State Fluorescence Anisotropy

Equilibrium measurements were performed using 500 pM Fl- Cat^N with given concentrations of SUMO-Flag-Cat^C (0 pM - 2,500 pM) in low salt (50 mM sodium phosphate, 100 mM NaCl, 1mM DTT, 1mM EDTA, pH 7.0) and high salt (50 mM sodium phosphate, 500 mM NaCl, 1mM DTT, 1mM EDTA, pH 7.0) buffers. Proteins were diluted from stock solutions to desired concentrations and incubated at 25 °C for 30 min. Samples were transferred to a cuvette of 1 cm path-length and the fluorescence anisotropy was measured immediately. Constants in the one site binding equation were obtained using non-linear least squares curve fitting method in MATLAB. For both the high and low salt conditions, the constants obtained from these fits (Table 4) fall below the concentration of Cat^N used for the measurements. We therefore report the K_d as < 500 pM, as we were unable to measure fluorescence anisotropy at lower concentrations of Cat^N.

Table 4. Kinetic binding constants.

	100 mM NaCl		**500 mM NaCl**	
[Cat^C] (nM)	**k_{obs1} (s^{-1})**	**k_{obs2} (s^{-1})**	**k_{obs1} (s^{-1})**	**k_{obs2} (s^{-1})**
200	0.60 ± .07	0.08 ± 0.09	0.44 ± 0.04	0.05 ± 0.009
325	0.78 ± .08	0.07 ± 0.06	0.92 ± 0.12	0.09 ± 0.009
500	1.10 ± .12	0.10 ± 0.015	0.90 ± 0.08	0.10 ± 0.011
750	2.08 ± .16	0.15 ± 0.014	1.87 ± 0.22	0.16 ± 0.013
1000	2.74 ± .18	0.20 ± 0.010	2.32 ± 0.22	0.20 ± 0.020
	k_{on1} (M^{-1} s^{-1})	**k_{on2} (M^{-1} s^{-1})**	**k_{on1} (M^{-1} s^{-1})**	**k_{on2} (M^{-1} s^{-1})**
Fit	(2.80 ± .28) × 10^6	(0.16 ± .019) × 10^6	(2.34 ± 0.30) × 10^6	(0.18 ± 0.016) × 10^6

Stopped flow fluorescence anisotropy

The stopped flow syringes were loaded with Fl-Cat^N and SUMO-Flag-Cat^C protein solutions so as to obtain final concentrations of 100 nM Cat^N and reported concentrations of Cat^C (200, 325, 500, 750, 1000 nM). Change in anisotropy values were measured in low salt and high salt buffers for a duration of 50 s. The change in anisotropy over time was fit to a double exponential kinetic model previously reported using non-linear least squares curve fitting method in MATLAB to obtain kinetic constants of binding (k_{obs1} and k_{obs2}) for each concentration.[16] The k_{obs1} and k_{obs2} values were then plotted as a function of Cat^C concentration, fit to a line, and the slope of the line was interpreted as k_{on}.

Results

1. **Design of a consensus atypical split intein with enhanced stability and activity**

In order to determine the mechanism of fragment association, an atypically split intein with minimal extein residues was isolated. Both naturally occurring atypical split inteins whose splicing rates have been characterized in vitro were identified within the T4-bacteriophage-type DNA-packaging terminase large subunit (TerL) from metagenomic sequencing data. The first, from the saline meromictic Ace Lake in Antarctica (AceL), exhibits an optimal splicing rate at 8 °C (t1/2 = 7 min). In addition, directed evolution found stabilizing mutations within AceL (AceL*) that increase activity at 37 °C (t1/2 = 6 min). The second characterized atypical split intein was sequenced in a sample collected from Punta Cormorant in the global ocean sampling project (GOS) and splices at an optimal temperature of 30 °C (t1/2 = 3 min). Purification of soluble GOS^N (i.e. the N-terminal GOS intein fragment), GOS^C, or AceL*^{C} from expression in *E. coli* was performed by means of large stabilizing extein proteins (FIG 4). The extraction of atypically split inteins lacking solubilizing exteins from the insoluble inclusion body fraction with chaotropic agents was unsuccessful due to aggregation issues while refolding.

Consensus design is a protein engineering strategy that utilizes evolutionary information from homologous protein sequences to predict stabilizing mutations and has previously been applied to generate a highly active and thermostable naturally split DnaE intein (Cfa). Seeking to engineer an atypically split intein amenable to in vitro structural characterization, a consensus atypical (Cat) TerL intein from multiple

sequence alignments (MSA) of $TerL^N$ and $TerL^C$ inteins discovered from BLAST searches of metagenomic sequencing information in the JGI and NCBI databases was designed (Table 1). Both Cat^N (60%) and Cat^C (64%) contain high sequence similarity to $AceL^{*N}$ and $AceL^{*C}$ respectively, with the nonidentical residues spread throughout the primary sequence (FIG 5). The Cat intein pair was isolated fused to model exteins to measure its in vitro trans-splicing activity (Table 5). Cat exhibits ultrafast splicing activity ($t_{1/2}$ = 59 s at 30 °C) and consistently outperforms AceL* across an array of temperatures (FIG 5). Moreover, Cat remains active at 50 °C, a temperature at which AceL* fails to splice. PTS was also measured in the presence of chaotropic agents, which are often utilized to solubilize aggregation-prone extein fragments.1 Cat displays enhanced chaotropic stability and can splice in both 2 M and 4 M urea (FIG 5, Table 6), while AceL* is inactive under both of these conditions. The accelerated splicing rates and activity under adverse conditions establish Cat as the fastest and most robust atypical split intein reported to date, and it should therefore serve as a tool for the synthetic N-terminal modification of proteins.

Table 5. Protein Splicing at Indicated Temperatures.

Intein	Temp (°C)	k_{splice} (s^{-1})	$t_{1/2}$ (s)
AceL*	4	(3.70 ± .26) × 10^{-4}	1873 ± 132
AceL*	15	(9.17 ± 1.2) × 10^{-4}	756 ± 100
AceL*	30	(7.68 ± .58) × 10^{-4}	902 ± 68
AceL*	37	(3.03 ± .30) × 10^{-4}	2287 ± 228
Cat	4	(7.54 ± .48) × 10^{-4}	919 ± 58
Cat	15	(4.81 ± .48) × 10^{-3}	144 ± 14
Cat	30	(1.17 ± .10) × 10^{-2}	59 ± 5
Cat	37	(1.32 ± .06) × 10^{-2}	52 ± 2
Cat	50	(1.58 ± .12) × 10^{-2}	44 ± 3

Table 6. Protein Splicing in Chaotropic Agents.

Intein	[Urea] (M)	k_{splice} (s^{-1})	$t_{1/2}$ (s)
Cat	2	(1.86 ± .18) × 10^{-3}	373 ± 35
Cat	4	(1.73 ± 1.7) × 10^{-4}	3826 ± 368

2. Fragment assembly drives a disorder to order structural transition

To investigate the association process of atypical split inteins, Cat^N and Cat^C bearing minimal exteins were expressed in isotopically enriched media (^{15}N, ^{13}C), purified, and analyzed by nuclear magnetic resonance (NMR) spectroscopy. Note, these constructs also included inactivating C1A and N134A mutations to prevent splicing during structural analysis of the complex. The ^{1}H-^{15}N HSQC spectrum of Cat^N in isolation

displays minimal dispersion along the ^{1}H dimension, a common phenomenon among disordered proteins and previously observed for Ssp^{C} and Npu^{C} (FIG 6). A stark transition occurs upon addition of unlabeled Cat^{C}, resulting in a well dispersed ^{1}H-^{15}N HSQC spectrum, which is consistent with Cat^{N} folding (FIG 6). Furthermore, measurements of ^{1}H-^{15}N heteronuclear NOEs, spin-spin relaxation rates, and Cα-Cβ chemical shift perturbation in Cat^{N} provide additional evidence for a disorder to order transition in Cat^{N} upon binding Cat^{C} (FIG 7). The ^{1}H-^{15}N HSQC of Cat^{C} in isolation exhibits far fewer crosspeaks than expected from the number of residues in the protein, a feature present in dynamic proteins that are undergoing chemical exchange and previously observed in both SspN and NpuN (FIG 6). Addition of unlabeled Cat^{N} leads to the appearance of new crosspeaks, which indicates a transition to a more ordered complex (FIG 6). Although the spectral quality of Cat^{C} in free form precluded our ability to assign the protein, some crosspeaks overlap those observed in the bound form, which suggests that Cat^{C} in free and bound form share a partial structural identity.

In line with the NMR studies, analysis by circular dichroism spectroscopy indicates that unbound Cat^{N} is largely unstructured with some propensity to sample secondary structure, and that both Cat^{N} and Cat^{C} inteins undergo a structural transition upon association (FIG 6). Further evidence for folding upon binding was observed by size exclusion chromatography (SEC), as Cat^{C} elutes at an earlier time than the bound complex despite having a lower molecular weight (FIG 6). The SEC elution profile is consistent with a compaction of Cat^{C} upon binding its cognate intein.

3. Solution structure of an atypical split intein complex

The isotopically enriched Cat^{N} and Cat^{C} proteins were assembled into a complex, and its structure was calculated from distance restraints and dihedral angle constraints obtained from NMR spectroscopy. The twenty lowest energy conformers obtained from the structure calculation are shown (FIG 8A, PDB ID: 6DSL). The structure ensemble is precise in all regions of the protein (with the exception of a short solubility tag in Cat^{C} and the exteins) with a mean backbone RMSD of 1.19 Å to the average structure (Table 7). Residue wise backbone RMSD values of < 0.5 Å were obtained across the structured regions of the protein (FIG 9A and 9B). The structure of Cat is predominantly β-sheet, with the last 8 residues present in the C-terminus of Cat^{N} being the only α-helix (FIG 8). It has a horseshoe-like shaped structure that is typical for proteins containing the HINT domain. The structure of Cat is similar to that of DnaE inteins, such as Npu (PDB ID: 2KEQ, RMSD 1.45 Å over 92 aligned Cα atoms) and Ssp (PDB

ID: 1ZDE, RMSD 1.34 Å over 90 aligned Cα atoms) with the notable exception that Npu and Ssp have an additional helix, which is absent in Cat.

In the Cat active site, a serine residue (Ser_{75}) replaces the threonine located in the canonical TXXH B-block motif (FIG 9C). The carbonyl oxygen of C1A is proximal to the amide proton (2.4 Å) and the hydroxyl proton (3.7 Å) of Ser75 (FIG 8C). The threonine residue in DnaE inteins adopts a similar conformation, suggesting that Ser75 supplants the role of threonine in assisting the cleavage of the N-terminal scissile peptide bond. Another notable feature in the structure is the lack of an F-block histidine (FIG 9C), and therefore resolution of the branched intermediate is likely mediated by the penultimate G-block histidine (His133).

Table 7: Statistics from NMR structure determination calculations of Cat complex in solution.

Parameter	**Value**
Restraints	
Distance restraints	3489
Unambiguous restraints	3283
Intra-residue	1667
Sequential	642
Short range	266
Long range	708
Ambiguous restraints	206
Dihedral angle restraints	180
Structure statistics	
NOE Violations > 0.5 Å	12 (+/- 4)
Dihedral violations > 5	0
Total Energy (kcal/mol)	-5074 (+/- 163)
RMSD from mean structure	
Backbone (all residues)	1.99 Å(+/- 0.4)
Heavy atoms (all residues)	2.52 Å(+/- 0.4)
Backbone (structured*)	1.19 Å(+/- 0.3)
Heavy atoms (structured*)	2.04 Å(+/- 0.3)
Ramachandran plot analysis	
Most favoured regions	85.7%
Additional allowed regions	13.5%
Generously allowed	0.8%

regions	
Disallowed regions	0.0%
*excluding exteins and solubility tag	

4. Mapping disorder localization in Cat

Limited proteolysis by thermolysin digestion was applied to investigate the distribution of local structure in Cat (FIG 10A). In isolation, Cat^N undergoes rapid degradation, while Cat^C displays slightly greater resistance to proteolysis. The intein complex, however, remains intact after 30 minutes. The variation in protease susceptibility observed is consistent with a largely disordered Cat^N, partially disordered Cat^C, and formation of a globular fold upon binding. We next examined cleavage products (t = 30 min) using electrospray ionization mass spectrometry (ESI-MS) to determine the regions protected from proteolysis, which should correspond to localized structural elements (FIG 11, Table 8). For Cat^N, cut sites appeared to be evenly spread throughout the primary sequence. Conversely, a large portion of Cat^C is resistant to proteolysis. Numerous peaks corresponding to intact fragments centered on residues 57 through 112 were observed, which points to this area as a structured region flanked by disordered N- and C-terminal peptides (FIG 10B). Mapping this model onto the structure of Cat indicates that the disordered N- and C-terminal ends of Cat^C directly interact with Cat^N (FIG 10C). Moreover, key catalytic residues for succinimide formation (Asp115, His133, and Asn_{134}) are present within the disordered region of Cat^C.

Table 8. Masses from limited proteolysis.

Cat^N			
Peak[a]	**$Mass_{obs}$ (Da)**	**$Mass_{exp}$ (Da)**	**Position**
1	623.27	623.27	2 to 7
2	495.21	495.21	-3 to 1
3	1154.56	1154.55	20 to 28
4	1371.75	1371.74	8 to 19
5	1976.00	1975.99	2 to 19
Cat^C			
Peak[a]	**$Mass_{obs}$ (Da)**	**$Mass_{exp}$ (Da)**	**Position**
1	1186.62	1186.6	130 to +6
2	410.2	410.19	113 to 115
3	760.35	760.35	117 to 123
4	1144.47	1144.45	-8 to 31
5	1094.65	1094.62	80 to 89

6	875.56	875.53	53 to 59
6	730.34	730.4	124 to 129
7	864.515	864.49	32 to 38
7	836.45	836.45	42 to 49
8	686.35	686.34	112 to 116
8	1471.67	1471.68	117 to 129
9	1779.8	1779.79	64 to 79
9	1347.8	1347.82	90 to 101
10	1176.71	1176.7	39 to 49
11	1584.76	1584.75	117 to 130
12	1313.66	1313.76	50 to 60
13	6093.1	6093.8	60 to 112
13	6580.3	6581.4	56 to 112
13	6950.6	6950.9	53 to 112
14	6483.2	6482.3	57 to 112
14	6972.5	6972.9	56 to 115
14	7341.7	7342.3	53 to 115
15	7632.9	7633.6	50 to 115
16	7455.8	7455.5	53 to 116
17	8197.1	8197.2	53 to 123
18	8908.4	8909	53 to 129
18	9199.6	9200.3	50 to 129
18	8538.2	8539.6	56 to 129

[a]The indicated peak number corresponds to the RP-HPLC traces in Figure 11

5. Assembly is largely driven by hydrophobic interactions

After examining the structural properties of the Cat fragments in split form, identification the molecular components that drive association were sought. Although the primary sequences of Cat^N and Cat^C exhibit separation of charge, the binding surface of Cat^N-Cat^C is rich in hydrophobic residues (FIG 12A and B). In the complex, the charged residues of both Cat^N and Cat^C are excluded towards the exterior of the protein while hydrophobic residues are clustered within the binding interface (FIG 13A and B). To validate that these hydrophobic interactions drive complex formation, the effect of buffer ionic strength on fragment association was evaluated using a fluorescence anisotropy-based binding assay. Cat^N containing an N-terminal fluorescein (Fl-Cat^N) was synthesized by solid phase peptide synthesis, and an increase in fluorescence anisotropy was observed upon association with a SUMO-Cat^C fusion protein (FIG 12C. This increased anisotropy is consistent with an expected increase in rotational correlation time for the Cat complex compared to unbound Cat^N, and was used as a

measure of Cat complex formation. Like other split inteins, Cat^N and Cat^C exhibit high binding affinity in vitro, with Kd values below 500 pM, which was the limit of detection of the assay (**Table 9**). Importantly, the binding isotherm for Cat complex formation is minimally perturbed by a change in ionic strength of the buffer, consistent with an association process driven by hydrophobic interactions.

Kinetics of binding between Fl-Cat^N and SUMO-Cat^C were next monitored by stopped-flow fluorescence, and the data was found to be best fit to a double exponential model (FIG 13C). Both determined rate constants (kobs1 and kobs2) exhibit concentration dependence leading to a calculated kon1 of (2.80 ± 0.28) × 106 M-1 s-1 and kon2 of (0.16 ± 0.019) × 106 M-1 s-1 under low salt conditions and kon1 of (2.34 ± 0.30) × 106 M-1 s-1 and kon2 of (0.18 ± 0.016) × 106 M-1 s-1 under high salt conditions (FIG 12D, Table 4). This model suggests that parallel association events may proceed from distinct conformers of the intein, with subsets of conformers being kinetically distinguishable. Moreover, the observation that both kobs1 and kobs2 are unperturbed by buffer ionic strength across all measured Cat^C concentrations further suggests that association is largely driven by hydrophobic interactions.

Table 9. Steady state Binding Constants.

	100 mM NaCl	**500 mM NaCl**
[Cat^C] (pM)	**Anisotropy**	**Anisotropy**
0	0.064 ± 0.001	0.084 ± 0.006
100	0.087 ± 0.009	0.103 ± 0.010
200	0.011 ± 0.010	0.120 ± 0.009
312.5	0.136 ± 0.013	0.142 ± 0.017
500	0.169 ± 0.004	0.168 ± 0.009
625	0.189 ± 0.008	0.185 ± 0.013
750	0.191 ± 0.006	0.185 ± 0.002
1000	0.195 ± 0.003	0.194 ± 0.005
1250	0.196 ± 0.008	0.199 ± 0.002
1875	0.203 ± 0.002	0.205 ± 0.005
2500	0.199 ± 0.003	0.203 ± 0.005
Fit	**K_d (pM)**	**K_d (pM)**
	33.87 ± 8.69	80.41 ± 15.41

6. The Extein Dependence of Cat

To date, all characterized inteins exhibit splicing rates dependent on their flanking extein residues. Deviation from the native extein sequence often decelerates splicing

and consequently may limit applications of PTS. The extein dependence of TerL inteins has yet to be thoroughly characterized, and we therefore sought to identify the sequence preferences of Cat by introducing substitutions that vary charge and steric bulk from the native residues (FIG 14A). Substitutions from the native C-extein, which is Cys+1, Glu+2, Phe+3, were introduced at the +2 and +3 positions and assayed in vitro (FIG 14B, Table 10). Cat demonstrates remarkable C-extein promiscuity, splicing with half-lives ranging from 1 to 3 minutes. This broad tolerance to C-extein substitutions is superior even to an engineered version of Npu previously designed to possess promiscuous activity. Unlike the tolerance to C-extein substitution, Cat exhibits a stark dependence on the identity of the -1 residue: decreased activity results from inserting alanine (t1/2 = 54 min), glycine (t1/2 = 146 min), or proline (t1/2 = 158 min) at this position (FIG 14C, Table 10). The measured in vitro extein dependence is likely explained by interactions observed in the solution structure of the Cat complex. Both Glu+2 and Phe+3 appear to have minimal contact with active site-catalytic residues, agreeing with the experimentally observed C-extein promiscuity (FIG 14D). Interestingly, Glu+2 does contact Asn123, which is present in place of an F-block histidine. Conversely, Glu-1 directly interacts with Ser75 and His78, two conserved residues with implications in thioester formation (FIG 14E). N-extein substitutions may therefore directly interfere with the capability of Ser75 and His78 to catalyze protein splicing.

Table 10. Protein splicing of Cat in varying Extein Contexts.

[a]N-extein$_{-2,\ -1}$	[a]C-Extein$_{+1,+2,+3}$	k_{splice} (s^{-1})	$t_{1/2}$ (s)
F<u>A</u>	CEF	(2.14 ± .08) × 10^{-4}	3244 ± 116
F<u>P</u>	CEF	(7.33 ± .44) × 10^{-5}	9451 ± 575
F<u>G</u>	CEF	(7.92 ± .33) × 10^{-5}	8749 ± 364
F<u>R</u>	CEF	(1.38 ± .06) × 10^{-3}	504 ± 22
<u>R</u>E	CEF	(4.53 ± .33) × 10^{-3}	153 ± 11
<u>A</u>E	CEF	(3.16 ± .36) × 10^{-3}	220 ± 25
<u>G</u>E	CEF	(1.30 ± .06) × 10^{-3}	532 ± 23
<u>E</u>E	CEF	(1.76 ± .06) × 10^{-3}	394 ± 14
FE	C<u>A</u>F	(9.75 ± .60) × 10^{-3}	71 ± 4
FE	C<u>G</u>F	(8.57 ± .91) × 10^{-3}	80 ± 9
FE	C<u>R</u>F	(5.16 ± .42) × 10^{-3}	134 ± 11
FE	C<u>F</u>F	(7.08 ± .51) × 10^{-3}	98 ± 7
FE	CE<u>G</u>	(6.47 ± .40) × 10^{-3}	107 ± 7
FE	CE<u>E</u>	(4.23 ± .23) × 10^{-3}	164 ± 9
FE	CE<u>A</u>	(9.20 ± 1.42) × 10^{-3}	75 ± 12
FE	CE<u>R</u>	(5.65 ± .25) × 10^{-3}	123 ± 5

[a]The position of mutation from the wild type extein sequence is underlined.

CLAIMS

What is claimed is:

1. A split intein N-fragment comprising the amino acid sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1.

2. The split intein N-fragment of claim 1, wherein the variant comprises an amino acid sequence selected from the group consisting of SEQ ID NO: 2-6, 125-127 and 168-170.

3. The split intein N-fragment of claim 2, wherein the variant is a functionally equivalent variant of SEQ ID NO: 1.

4. The split intein N-fragment of claim 3, wherein the functionally equivalent variant comprises the amino acid sequence of SEQ ID NO: 4 or SEQ ID NO: 125.

5. A complex comprising:

(i) a compound of interest,

(ii) the split intein N-fragment of any one of claims 1 to 4, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the complex optionally comprises a linker between (i) and (ii) and

wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

6. The complex of claim 5, wherein the split intein N-fragment comprises an amino acid sequence selected from the group consisting of SEQ ID NO: 49-68 or a variant thereof.

7. The complex of any one of claims 5 or 6, wherein the compound of interest is a polypeptide or protein, and wherein if the complex comprises a linker, the linker is a peptide linker.

8. The complex of claim 7, wherein the polypeptide of interest is an antibody or a fragment of a protein.

9. The complex of claim 8, wherein the compound of interest is an N-terminal fragment of a protein.

10. A polynucleotide encoding the split intein N-fragment of any one of claims 1 to 5 or the complex of claim 7.

11. A vector comprising the polynucleotide of claim 10.

12. A host cell comprising the polynucleotide of claim 10 or the vector of claim 11.

13. A split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7.

14. The split intein C-fragment of claim 13, wherein the variant comprises an amino acid sequence selected from the group consisting of SEQ ID NO: 8-48 and 128-166.

15. The split intein C-fragment of claim 14, wherein the variant is a functionally equivalent variant.

16. The split intein C-fragment of claim 15, wherein the functionally equivalent variant comprises an amino acid sequence selected from the group consisting of SEQ ID NO: 10-22 and 128-140.

17. A complex comprising:

(i) the split intein C-fragment of any one of claims 13 to 16 or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120 and

(ii) a compound of interest

wherein the complex optionally comprises a linker between (i) and (ii) and

wherein

- the compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage.

18. The complex of claim 17, wherein the split intein C-fragment comprises a sequence selected from SEQ ID NO: 69-87 or a variant thereof.

19. The complex of any one of claims 17 or 18, wherein the compound of interest is a polypeptide or protein, and wherein if the complex comprises a linker, the linker is a peptide linker.

20. The complex of claim 19, wherein the compound of interest is an antibody or a fragment of a protein.

21. The complex of claim 20, wherein the compound of interest is the C-terminal fragment of a protein.

22. A complex comprising:

(i) the split intein C-fragment of any one of claims 13 to 16 or a split intein C-fragment comprising a sequence selected from the group consisting of SEQ ID NO: 114-120

(ii) a compound of interest and

(iii) the split intein N-fragment of any one of claims 1 to 4, or a split intein N-fragment comprising the amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

wherein the complex optionally comprises a linker between (i) and (ii) and/or between (ii) and (iii),

wherein

- the compound of interest is linked to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the C-terminus of the split intein C-fragment by an amide linkage

and

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

23. A polynucleotide encoding the split intein C-fragment of any one of claims 13 to 16 or the complex of claim 19 or the complex of claim 22 wherein the conjugate of interest is a protein, and wherein if the complex comprises a linker, the linker is a peptide linker.

24. A vector comprising the polynucleotide of claim 23.

25. A host cell comprising the polynucleotide of claim 23 or the vector of claim 24.

26. A composition comprising the complex of any one of claims 5 to 9 and the complex of any one of claims 17 to 21.

27. A conjugate comprising the complex of any one of claims 5 to 9 and the complex of any one of claims 17 to 21, wherein the C-terminus of the split intein N-fragment is linked to the N-terminus of the split intein C-fragment by a peptide bond.

28. A conjugate comprising (a) the complex of claim 7 and (b) a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a variant thereof

having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, wherein the C-terminus of the split intein N-fragment is linked to the N-terminus of the split intein C-fragment by a peptide bond.

29. A polynucleotide encoding the conjugate of claim 28 or a vector comprising said polynucleotide.

30. A host cell comprising the polynucleotide or the vector of claim 29.

31. A method to obtain a conjugate between a first compound of interest and a second compound of interest comprising

(i) contacting

(a) the complex of any one of claims 5 to 9, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 with

(b) the complex of any one of claims 17 to 21, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120

or

a complex comprising an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof and the second compound of interest, wherein the complex optionally comprises a linker between the split intein C-fragment and the second compound of interest and wherein

- the second compound of interest is bound to the C-terminus of the split intein C-fragment by an amide linkage or
- if the complex comprises a linker, the second compound of interest is bound to the linker by an amide linkage and/or the

linker is bound to the C-terminus of the split intein C-fragment by an amide linkage

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

32. A method to obtain a conjugate between a first compound of interest and a second compound of interest comprising

(i) contacting

(a) the complex of any one of claims 5 to 9, wherein the complex comprises the first compound of interest and a split intein N-fragment comprising the amino sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

or

a complex comprising the second compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) the complex of any one of claims 17 to 21, wherein the complex comprises the second compound of interest and a split intein C-fragment comprising the amino acid sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid

sequence selected from the group consisting of SEQ ID NO: 114-120

under appropriate conditions for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate and

(ii) allowing the intein intermediate to react to form a conjugate between the first and the second compound of interest.

33. A method to obtain a conjugate of a compound of interest with a nucleophile comprising

(i) contacting

(a) the complex of any one of claims 5 to 9, wherein the split intein N-fragment comprises the amino acid sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% sequence identity with SEQ ID NO: 1 or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110

or

a complex comprising a compound of interest and an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, wherein the complex optionally comprises a linker between the compound of interest and the split intein N-fragment, and wherein

- the compound of interest is linked to the N-terminus of the split intein N-fragment by an amide linkage or
- if the complex comprises a linker, the compound of interest is bound to the linker by an amide linkage and/or the linker is bound to the N-terminus of the split intein N-fragment by an amide linkage.

with

(b) a split intein C-fragment comprising an amino acid sequence selected from the group consisting of SEQ ID NO: 8, 9, 23-48 and 141-166, under appropriate conditions for binding between the split intein N-fragment and the split intein C-fragment to form an intein intermediate and

(ii) contacting the intein intermediate with an exogenous nucleophile.

34. The method of claim 33, further comprising contacting the conjugate of the compound of interest and the nucleophile with a second exogenous nucleophile.

35. The method of claim 34, wherein the nucleophile is a thiol.

36. The method of any one of claims 31 to 35, wherein the split intein N-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

37. The method of any one of claims 31 to 35, wherein the split intein C-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

38. A composition comprising:

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a variant thereof or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7 or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and
- a second polypeptide of interest

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a variant thereof having at least 90% sequence identity with SEQ ID NO: 1, or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110 and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7

or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120 and

- a second polypeptide of interest.

39. The composition of claim 38, wherein the first polypeptide of interest is the N-terminal fragment of a protein and the second polypeptide of interest is the C-terminal fragment of said protein, and wherein upon covalently linking the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest the whole protein is obtained.

40. The composition of any one of claims 38 or 39, wherein the split intein N-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

41. The composition of any one of claims 38 or 39, wherein the split intein C-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

42. A method for expressing a gene of interest in a cell comprising:

(i) contacting the cell with

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

or

(a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110, and

(b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest,

(ii) allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and

(iii) allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.

43. A method for expressing a gene of interest comprising:

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

or

(i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus:

- a first polypeptide of interest and
- an AceL-TerL split intein N-fragment or a functionally equivalent variant thereof, or a split intein N-fragment comprising the sequence of SEQ ID NO: 1 or a functionally equivalent variant thereof having at least 90% or an amino acid sequence selected from the group consisting of SEQ ID NO: 103-110,

wherein the first fusion protein comprises a signal peptide, and

(ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus:

- a split intein C-fragment comprising the sequence of SEQ ID NO: 7 or a functionally equivalent variant thereof having at least 88% sequence identity with SEQ ID NO: 7, or an amino acid sequence selected from the group consisting of SEQ ID NO: 114-120, and a
- second polypeptide of interest

wherein the second fusion protein comprises a signal peptide,

(iii) allowing the expression of the first and the second polynucleotides so that the first and the second fusion proteins are produced and secreted,

(iv) allowing the contact between the first and second fusion proteins so that the split intein N-fragment binds to the split intein C-fragment to form a intein intermediate and the intein intermediate reacts to covalently link the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest.The method of any one of claims 42 or 43, wherein

the first polypeptide of interest is the N-terminal fragment of a protein and the second polypeptide of interest is the C-terminal fragment of said protein, and wherein upon covalently linking the C-terminus of the first polypeptide of interest to the N-terminus of the second polypeptide of interest the whole protein is obtained.

44. The method of any one of claims 42 to 44, wherein the split intein N-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 49-68 or a functionally equivalent variant thereof.

45. The method of any one of claims 44 to 44, wherein the split intein C-fragment comprises a sequence selected from the group consisting of SEQ ID NO: 69-87 or a functionally equivalent variant thereof.

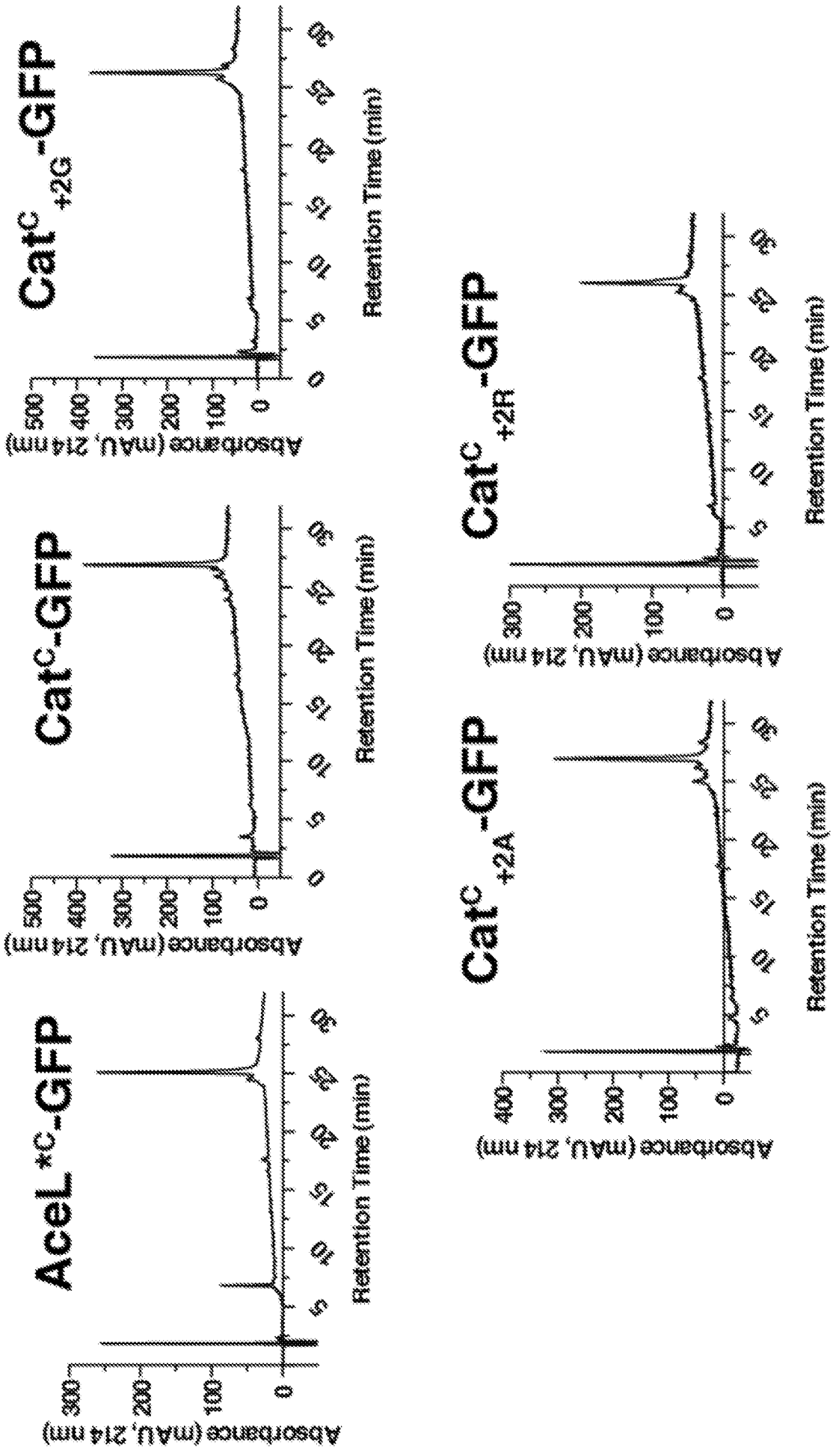

AceL*C-GFP
CatC-GFP
CatC+2G-GFP
CatC+2A-GFP
CatC+2R-GFP
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 1A

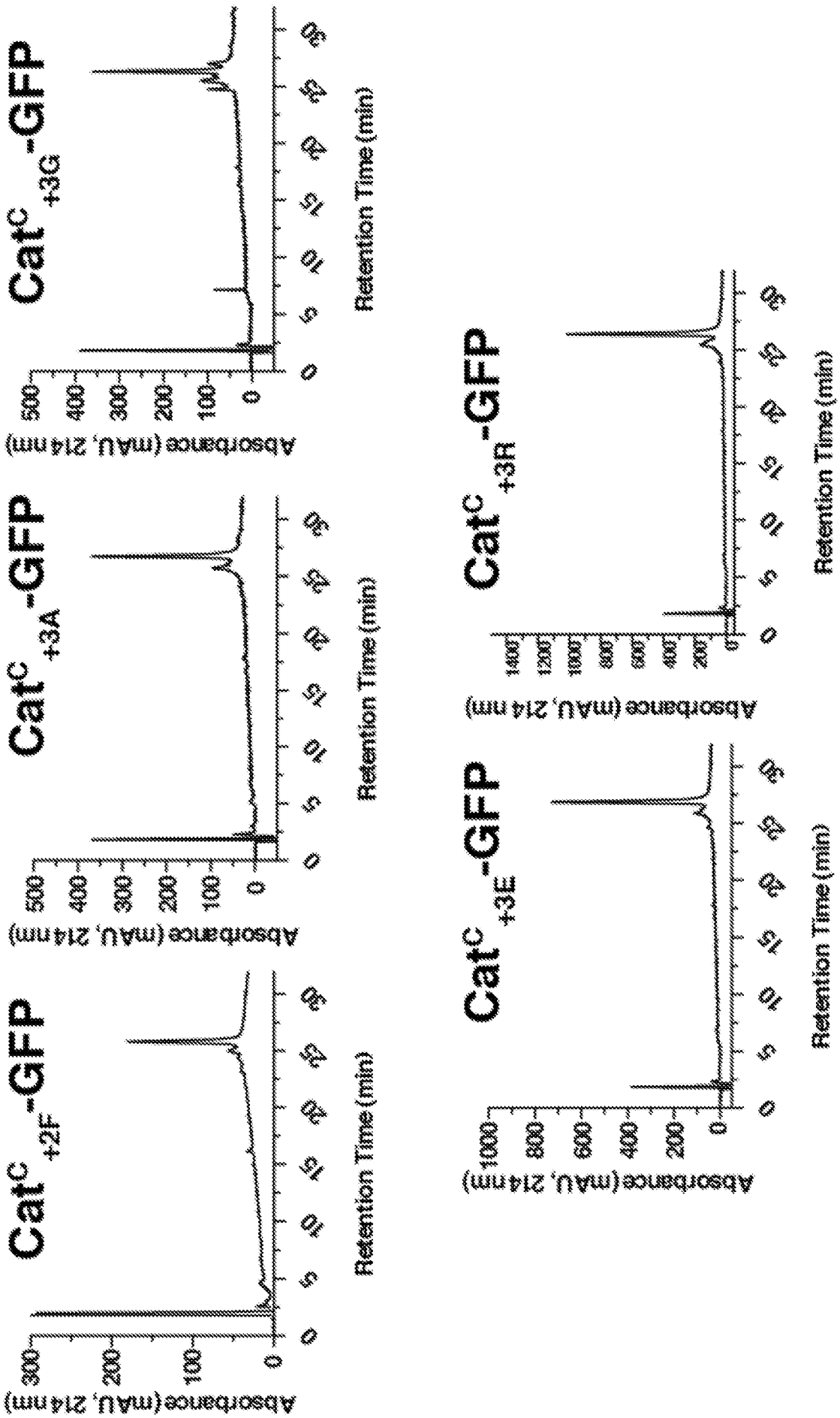

Cat^C_+2F-GFP
Cat^C_+3A-GFP
Cat^C_+3G-GFP
Cat^C_+3E-GFP
Cat^C_+3R-GFP
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 1B

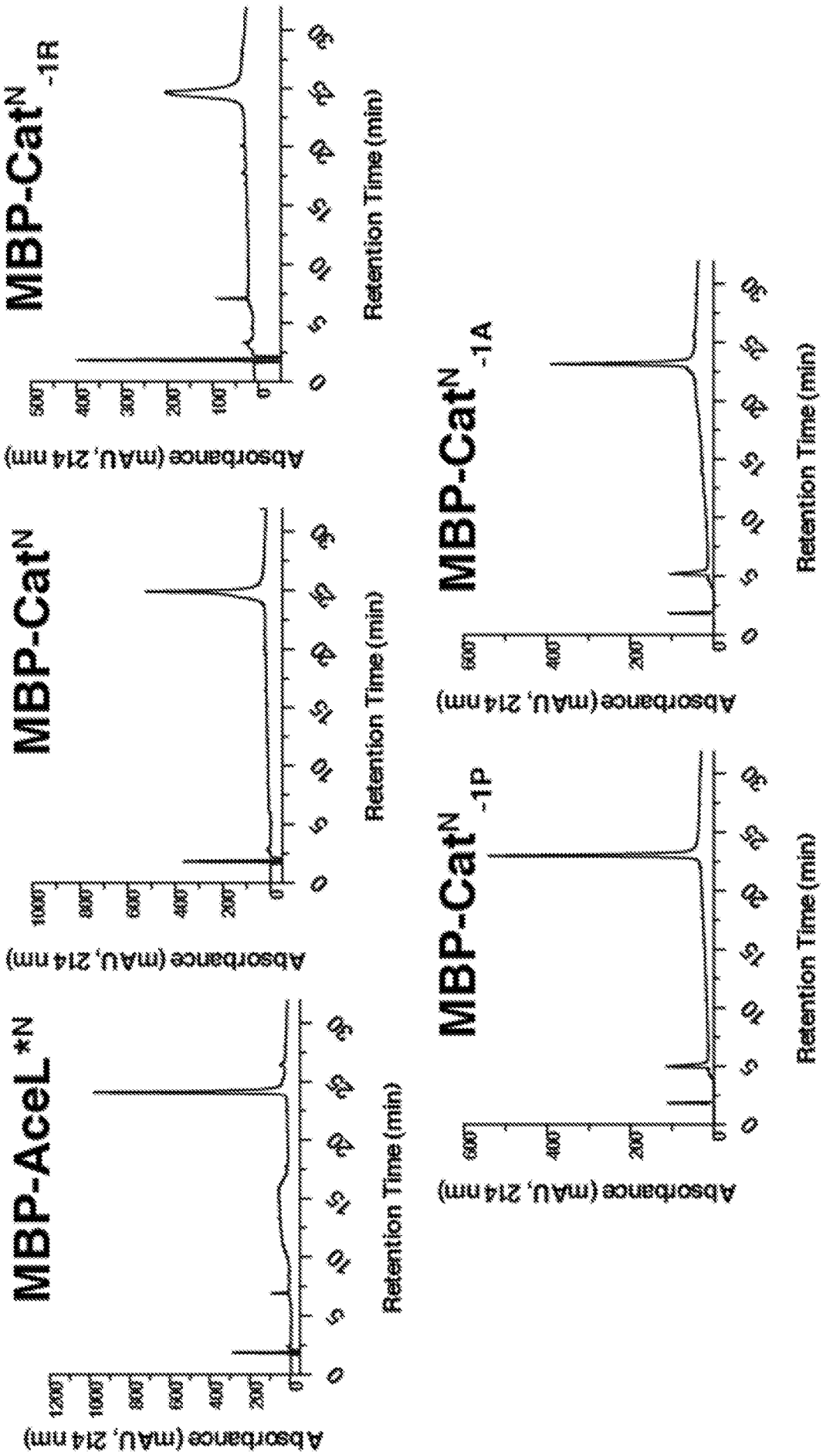

MBP-AceL*N
MBP-CatN
MBP-CatN-1P
MBP-CatN-1A
MBP-CatN-1R
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 1C

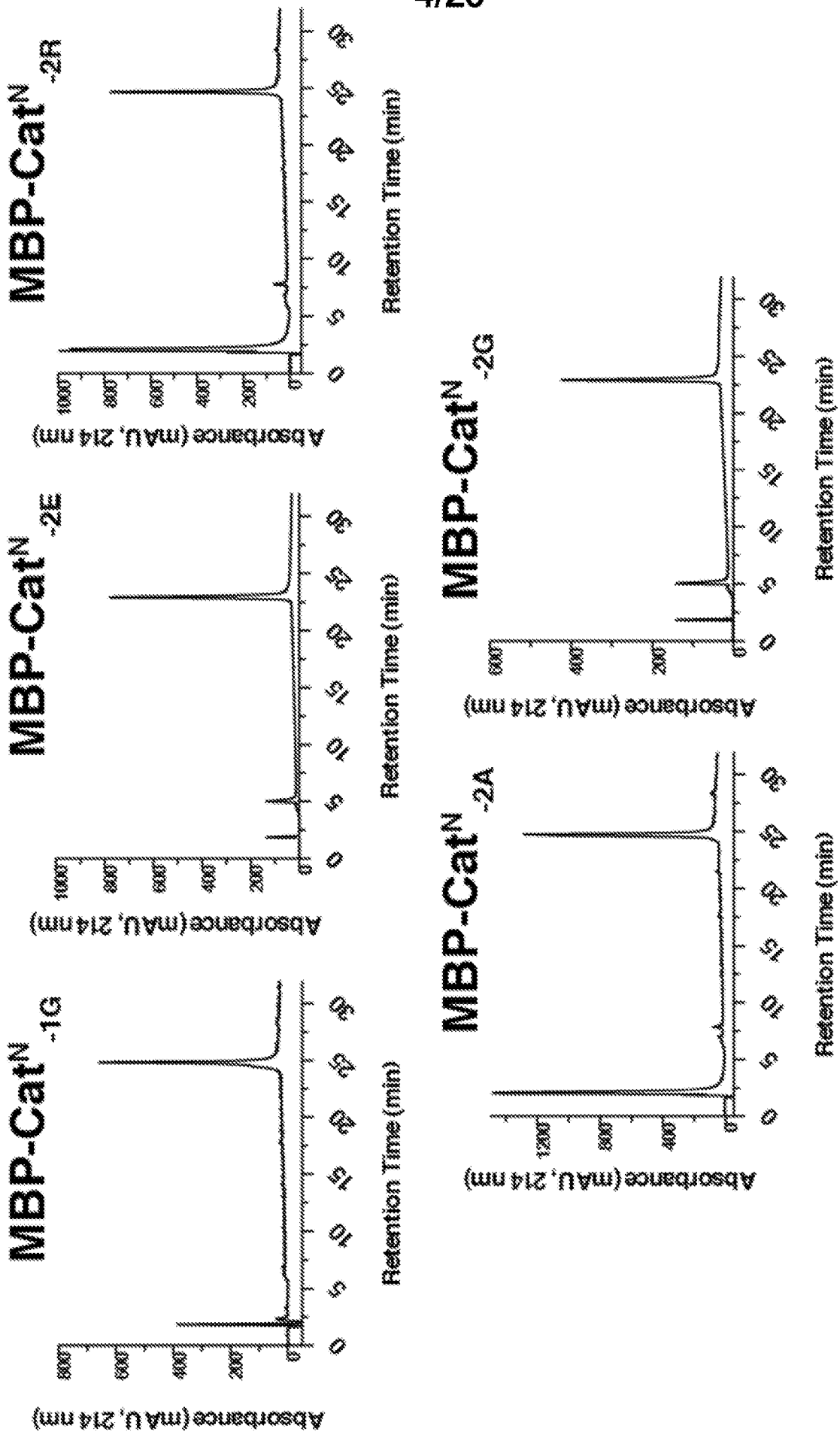

MBP-Cat^{N}-1G
MBP-Cat^{N}-2E
MBP-Cat^{N}-2R
MBP-Cat^{N}-2A
MBP-Cat^{N}-2G
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 1D

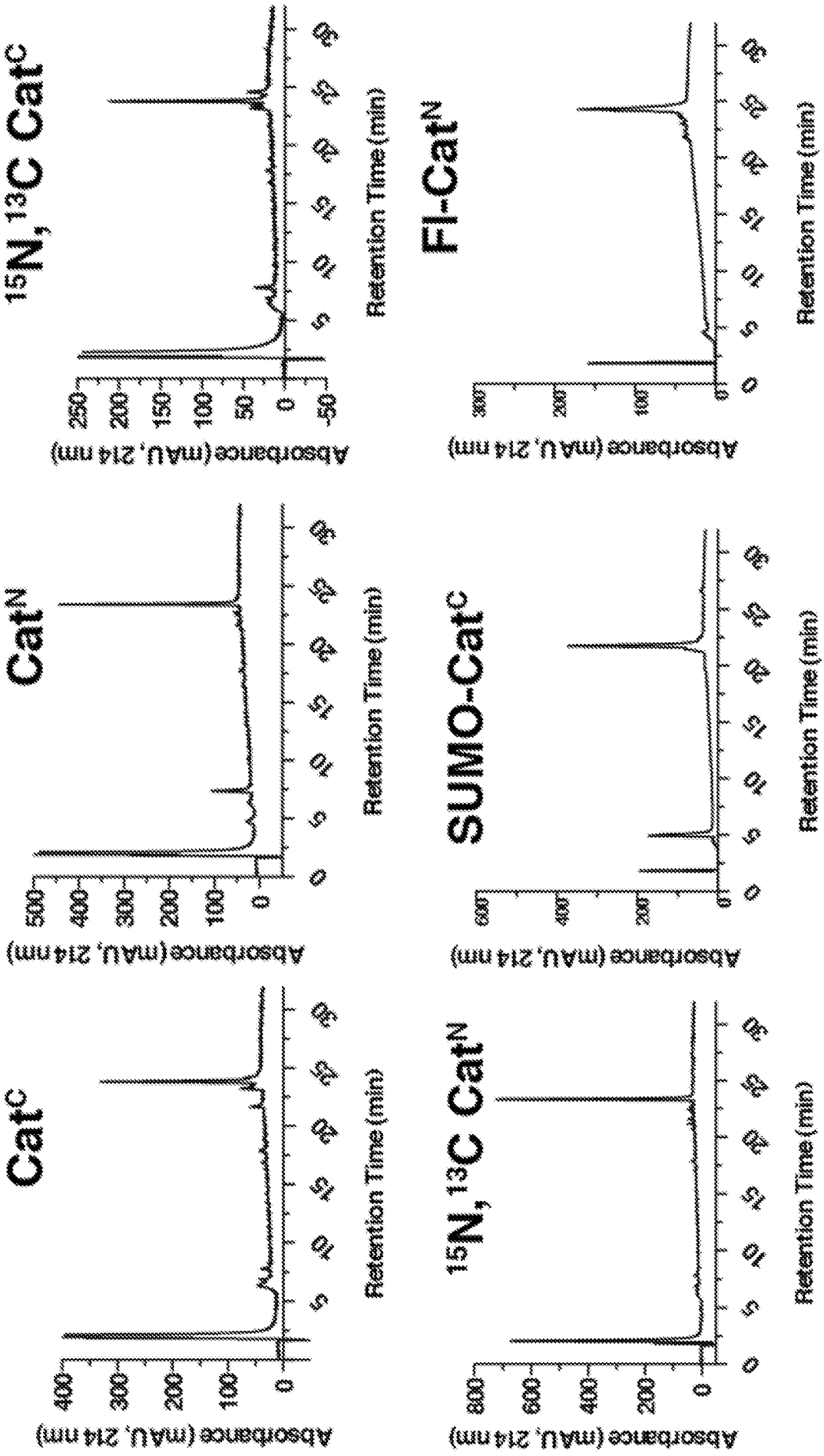

$^{15}N, ^{13}C$ Cat^{C}
Absorbance (mAU, 214 nm)
Retention Time (min)
Fl-Cat^{N}
Absorbance (mAU, 214 nm)
Retention Time (min)
Cat^{N}
Absorbance (mAU, 214 nm)
Retention Time (min)
SUMO-Cat^{C}
Absorbance (mAU, 214 nm)
Retention Time (min)
Cat^{C}
Absorbance (mAU, 214 nm)
Retention Time (min)
$^{15}N, ^{13}C$ Cat^{N}
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 1E

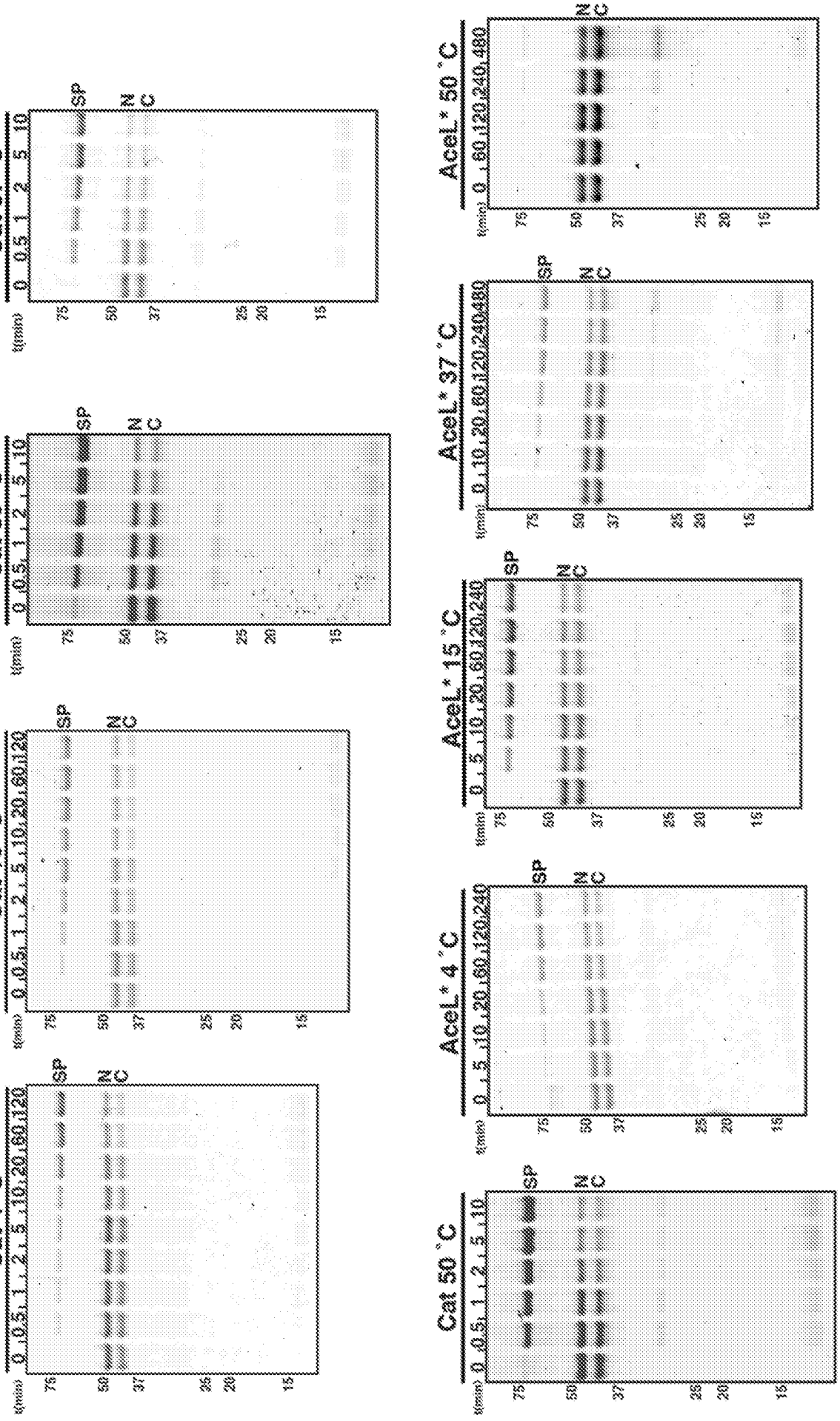

Cat 4 ˚C
Cat 15 ˚C
Cat 30 ˚C
Cat 37 ˚C
AceL* 4 ˚C
AceL* 15 ˚C
AceL* 37 ˚C
AceL* 50 ˚C
Cat 50 ˚C
t(min)
SP
N
C

FIG. 2A

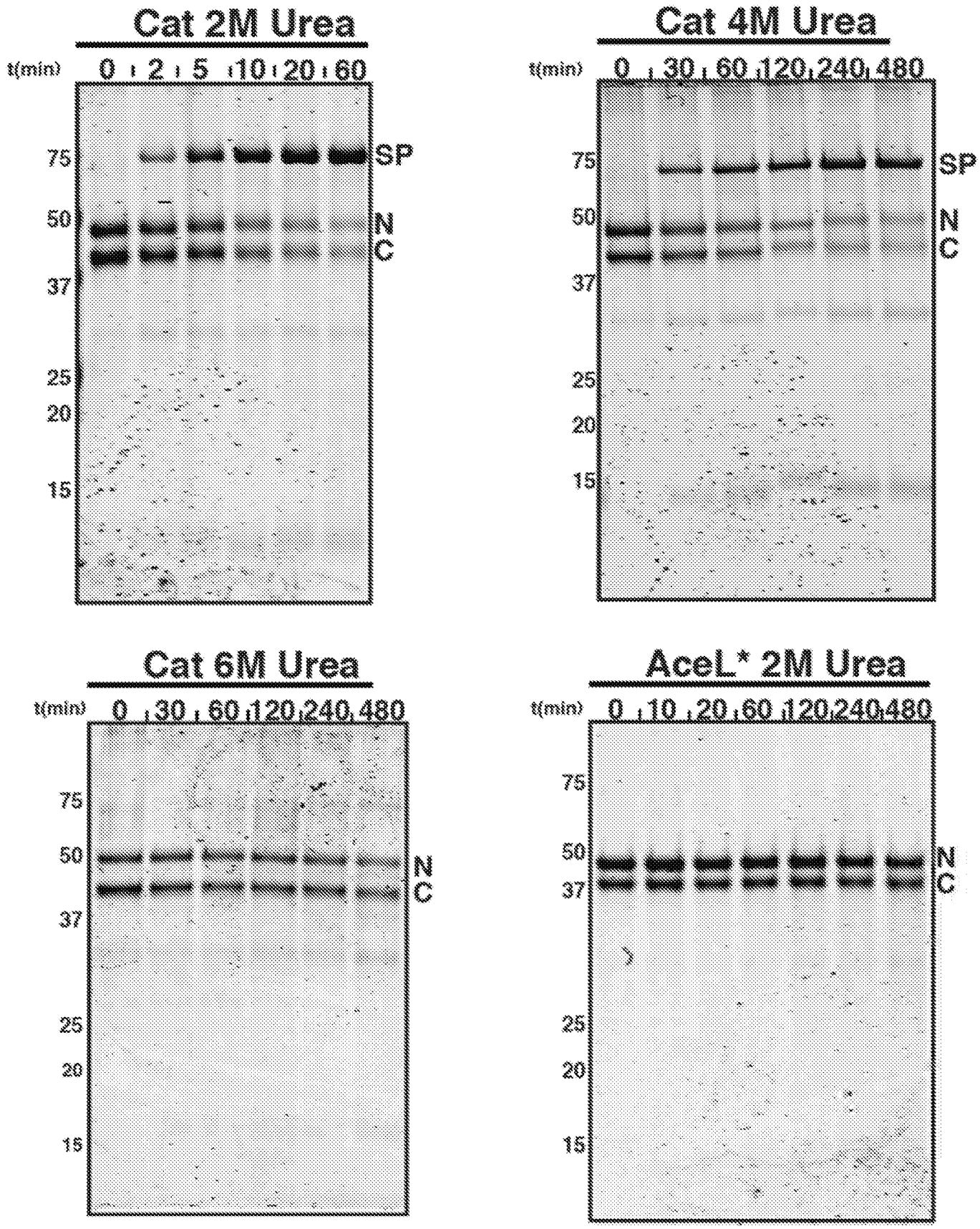

Cat 2M Urea
t(min) 0 2 5 10 20 60
75
50
37
25
20
15
SP
N
C
Cat 4M Urea
t(min) 0 30 60 120 240 480
75
50
37
25
20
15
SP
N
C
Cat 6M Urea
t(min) 0 30 60 120 240 480
75
50
37
25
20
15
N
C
AceL* 2M Urea
t(min) 0 10 20 60 120 240 480
75
50
37
25
20
15
N
C

FIG. 2B

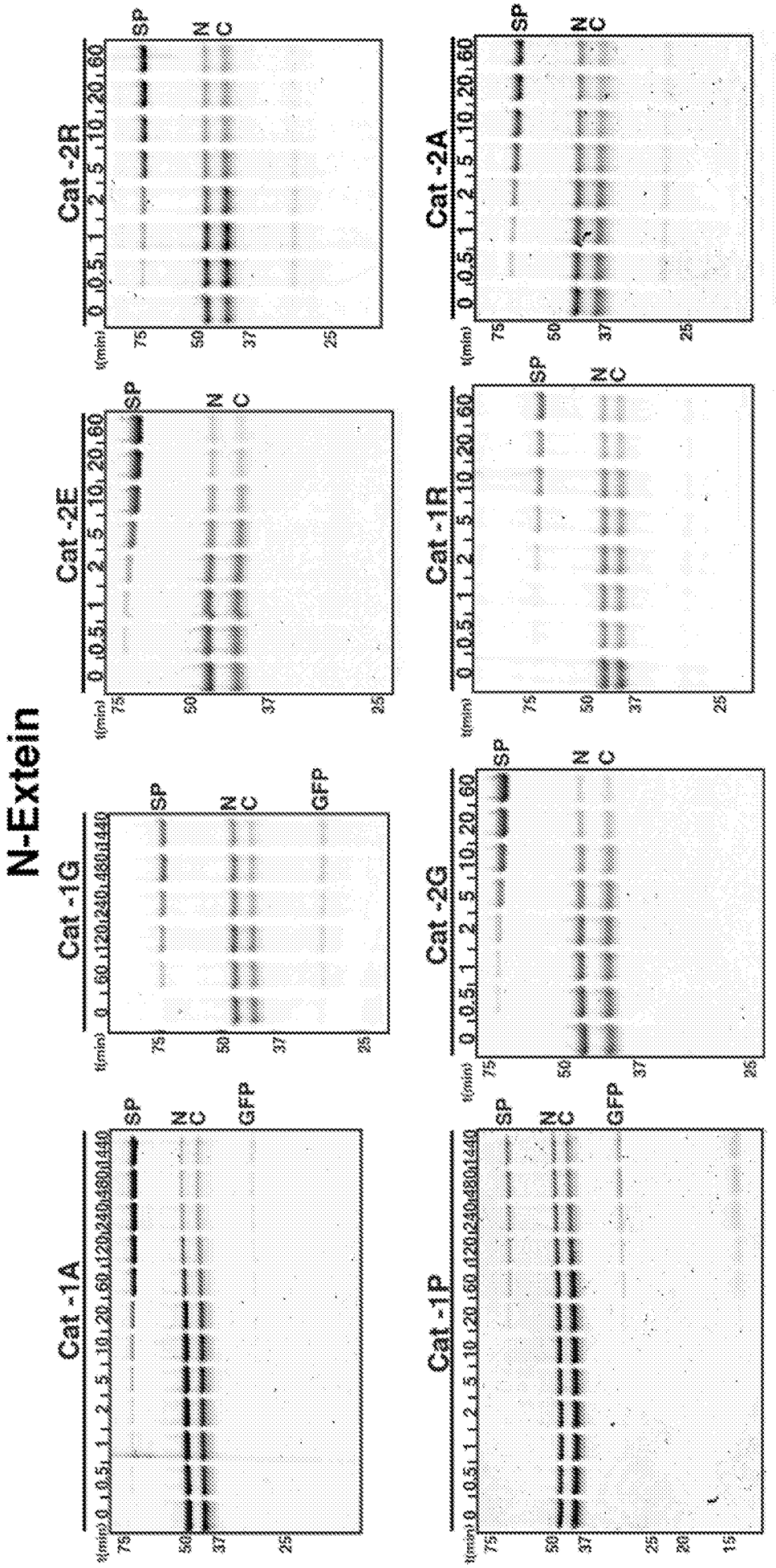

N-Extein
Cat -1A
Cat -1G
Cat -2G
Cat -1P
Cat -2E
Cat -1R
Cat -2A
Cat -2R
SP
N
C
GFP
t(min)

FIG. 2C

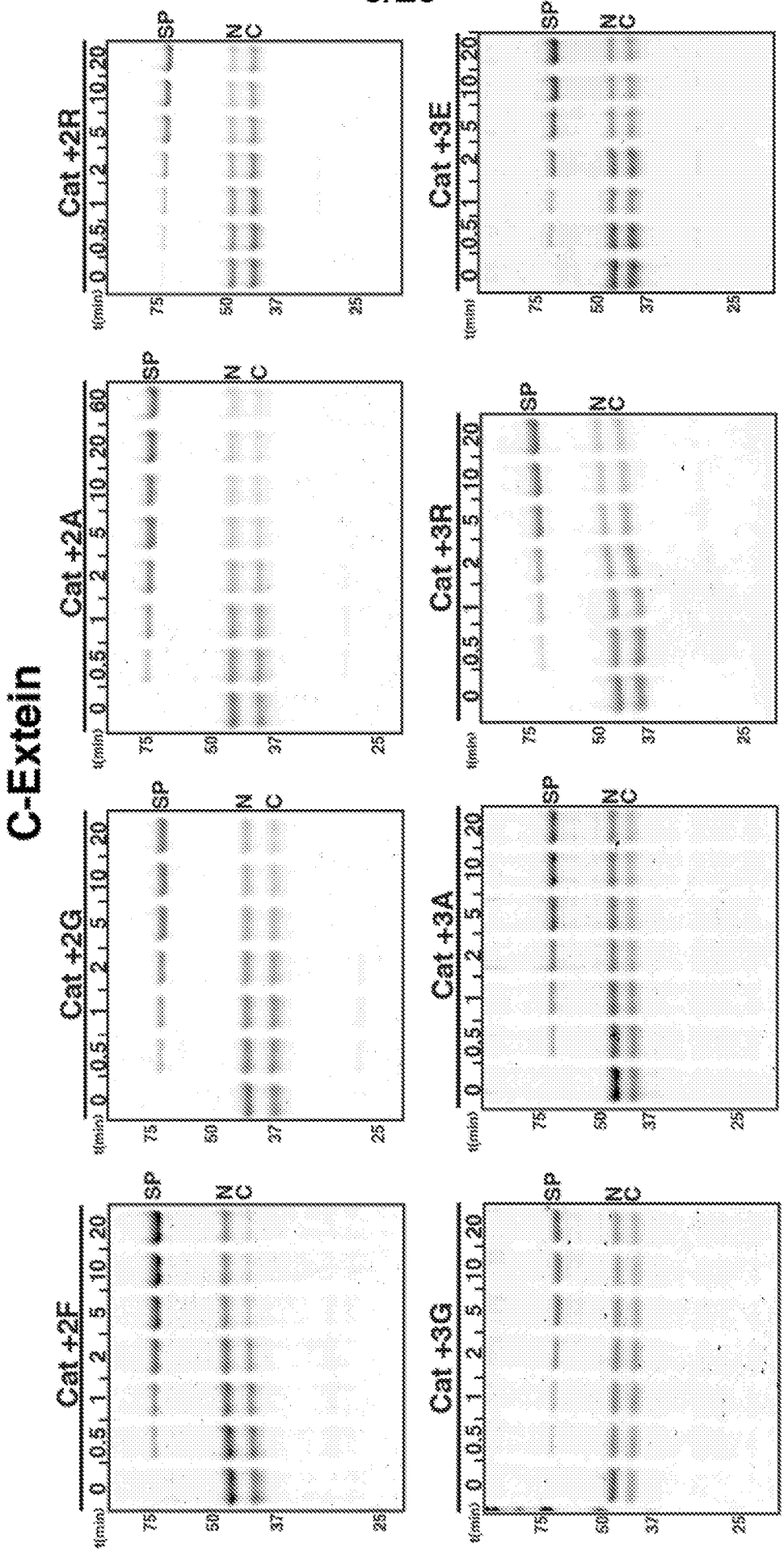

C-Extein
Cat +2R
Cat +2A
Cat +2G
Cat +2F
Cat +3E
Cat +3R
Cat +3A
Cat +3G
t(min)
0 0.5 1 2 5 10 20
0 0.5 1 2 5 10 20 60
SP
N
C
75
50
37
25

FIG. 2D

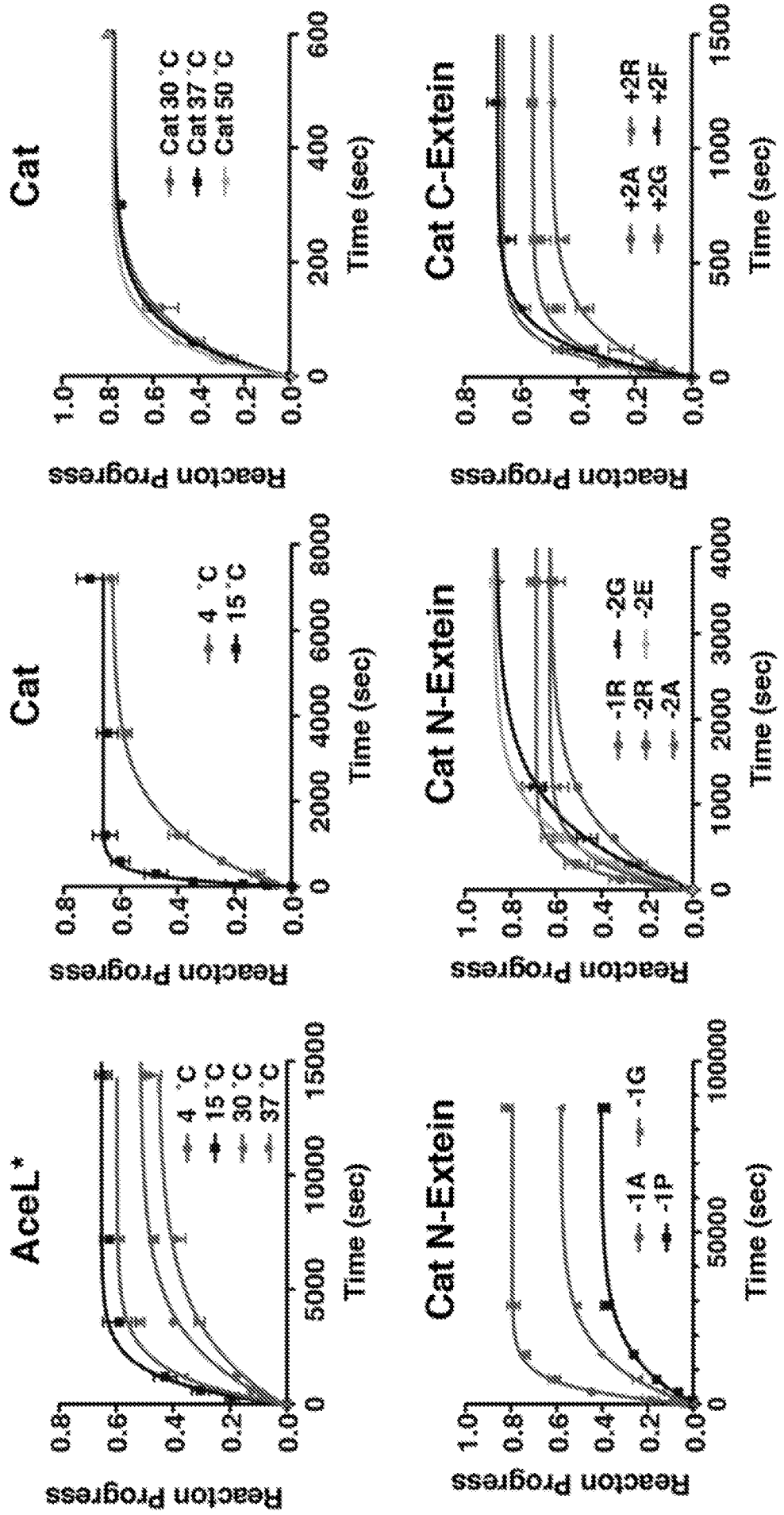

Cat
Cat 30 ˚C
Cat 37 ˚C
Cat 50 ˚C
Reaction Progress
Time (sec)
Cat C-Extein
+2A
+2R
+2G
+2F
Cat
4 ˚C
15 ˚C
Cat N-Extein
-1R
-2G
-2R
-2E
-2A
AceL*
4 ˚C
15 ˚C
30 ˚C
37 ˚C
Cat N-Extein
-1A
-1G
-1P

FIG. 3A

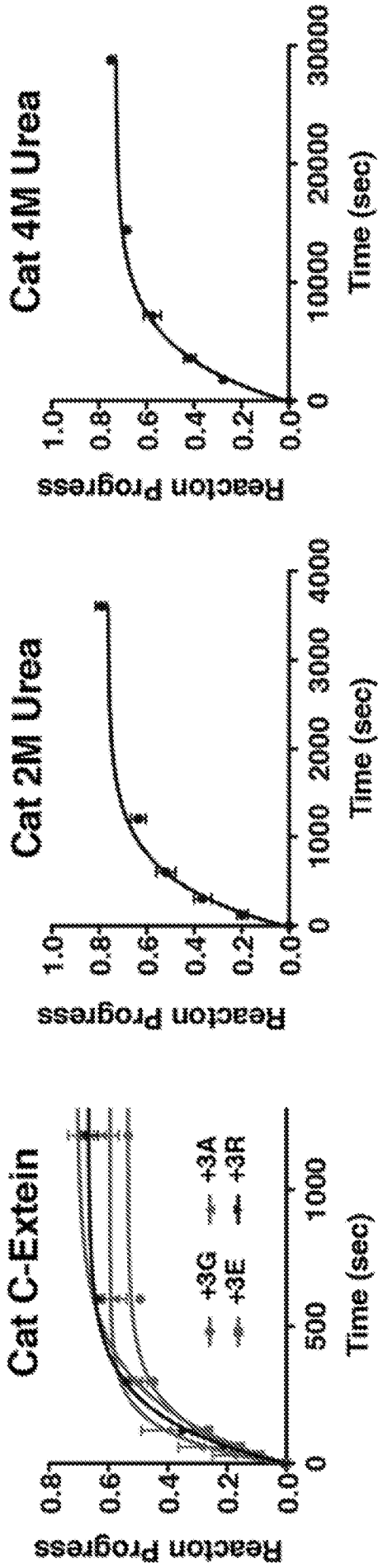

Cat 4M Urea
Reaction Progress
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
Time (sec)
Cat 2M Urea
Reaction Progress
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
Time (sec)
Cat C-Extein
Reaction Progress
0.8
0.6
0.4
0.2
0.0
0
500
1000
Time (sec)
+3G
+3E
+3A
+3R

FIG. 3B

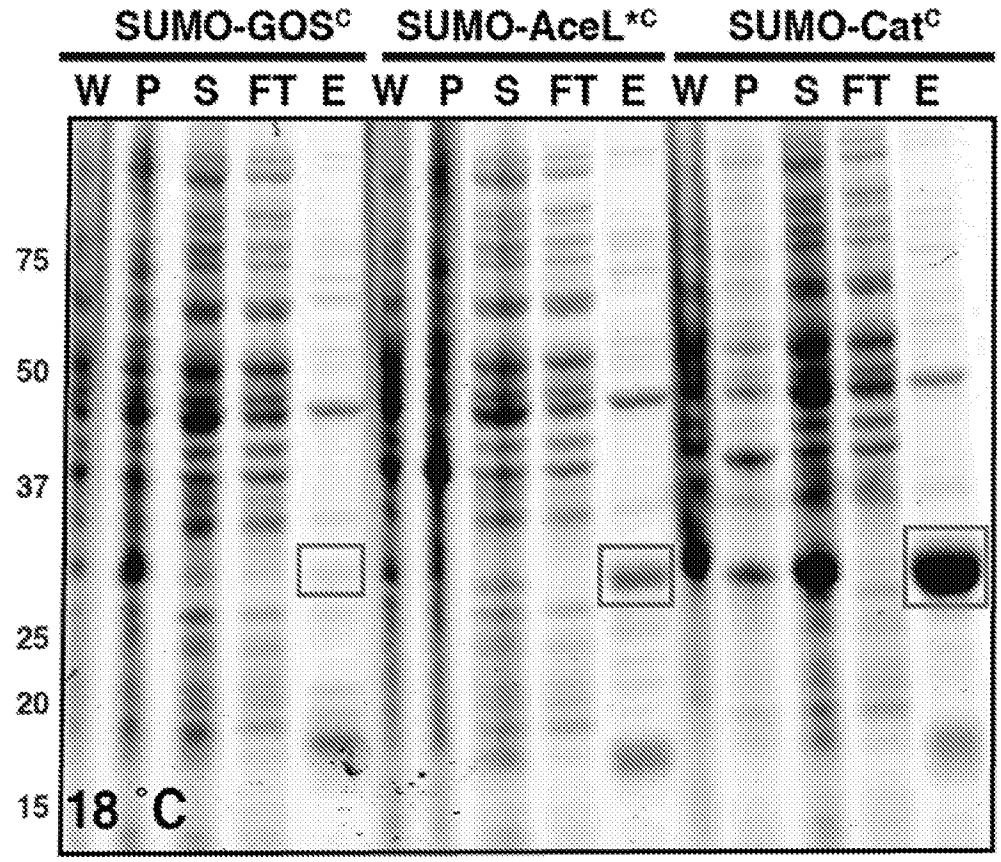

SUMO-GOS^{C}
SUMO-AceL*^{C}
SUMO-Cat^{C}
W P S FT E W P S FT E W P S FT E
75
50
37
25
20
15
18 °C

FIG. 4A

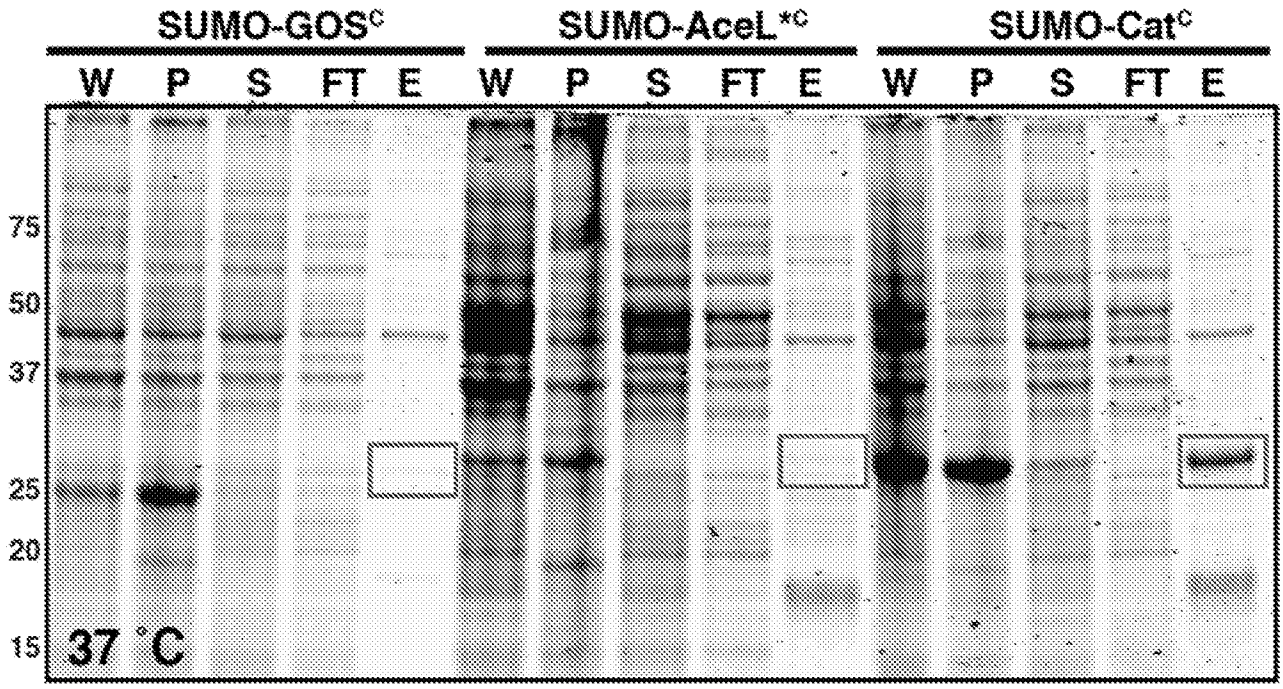

SUMO-GOS^{C}
SUMO-AceL*^{C}
SUMO-Cat^{C}
W P S FT E W P S FT E W P S FT E
75
50
37
25
20
15
37 °C

FIG. 4B

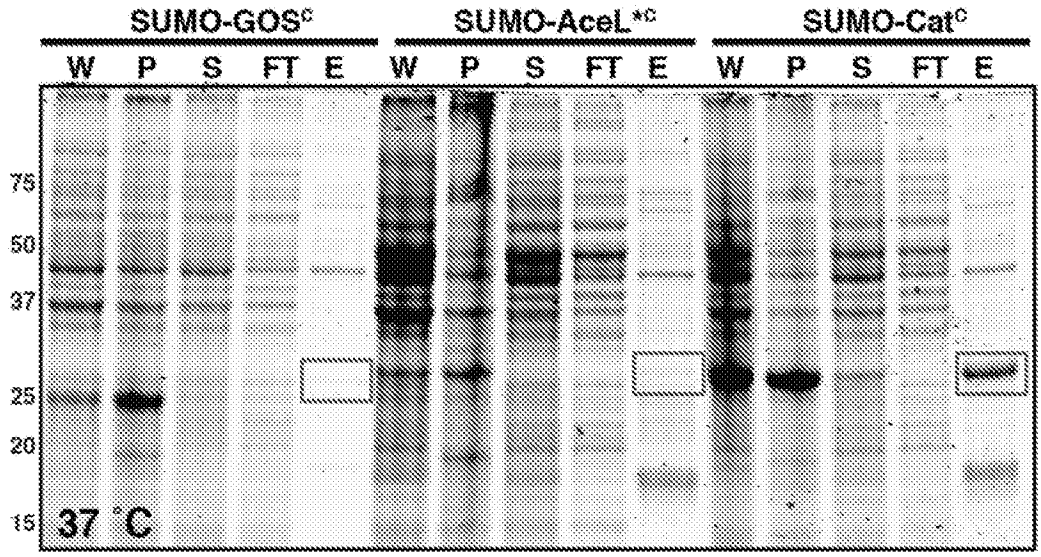

SUMO-GOS^{C}
SUMO-AceL^{*C}
SUMO-Cat^{C}
W P S FT E W P S FT E W P S FT E
75
50
37
25
20
15
37 °C

FIG. 4C

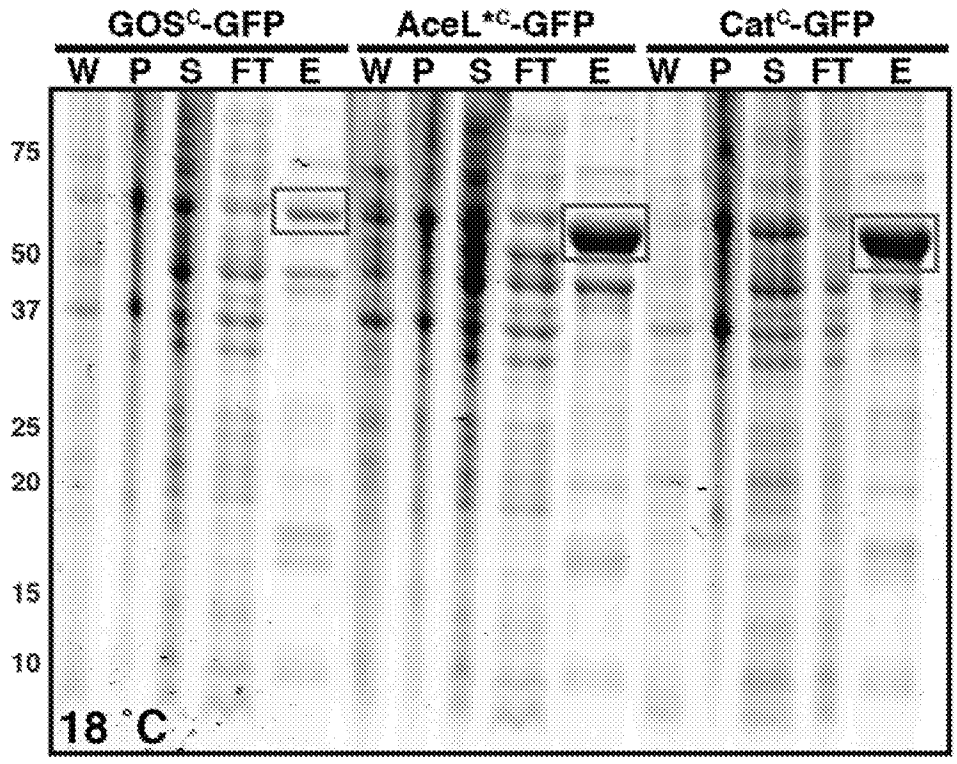

GOS^{C}-GFP
AceL^{*C}-GFP
Cat^{C}-GFP
W P S FT E W P S FT E W P S FT E
75
50
37
25
20
15
10
18 °C

FIG. 4D

```
Cat^N       1  CLSGDTMIEILDDDGIIQKISMEDLYQRLA
AceL*^N     1  CVSGDTMVET--EDG---KIKIEDLYKRLA

Cat^C      31  MFKLNTKNIKVLTPSGFKSFSGIQKVYKPFYHHIIFDDGSEIKCSDNHSF
AceL*^C    26  MFRTNTNNIKILGPNGFSNFIGIQKVERDQYQHIIFDDDTEIKTSINHPF

Cat^C      81  GKDKIKASTIKVGDYLQGKKVLYNEIVEEGIYLYDLLNVGEDNLYYTNGI
AceL*^C    76  GKDKILARDVKVGDYLNSKKVLYNELVNENIFLYDPINVEKESLYITNGV

Cat^C     131  VSHN
AceL*^C   126  VSHN
```

FIG. 5A

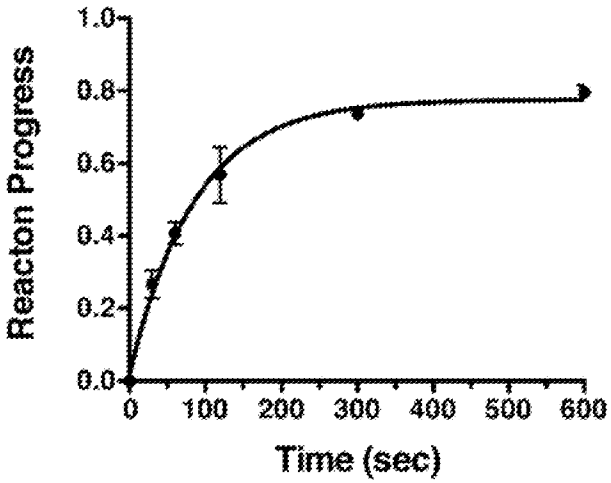


FIG. 5B

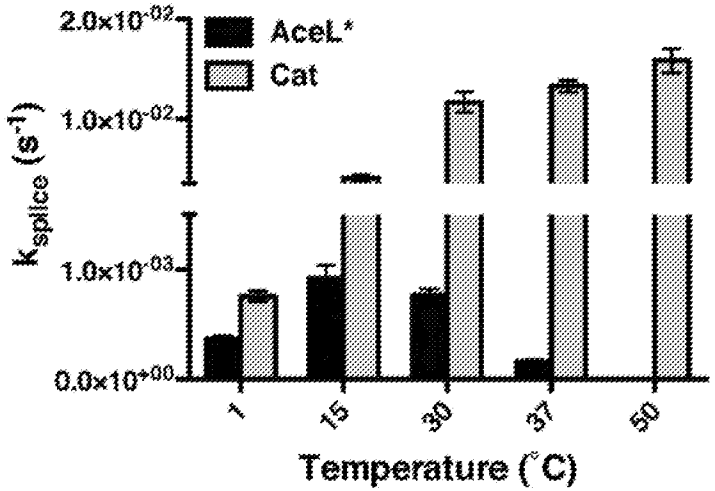


FIG. 5C

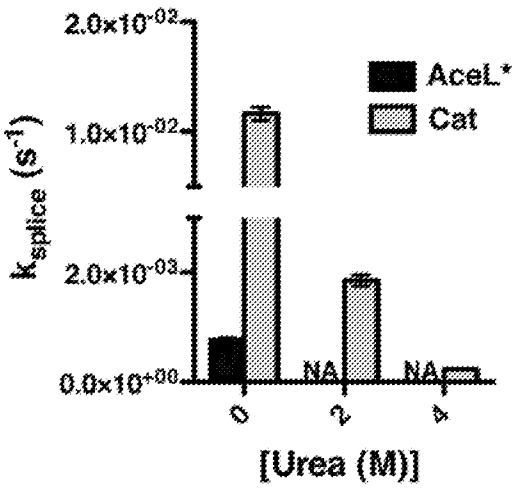


FIG. 5D

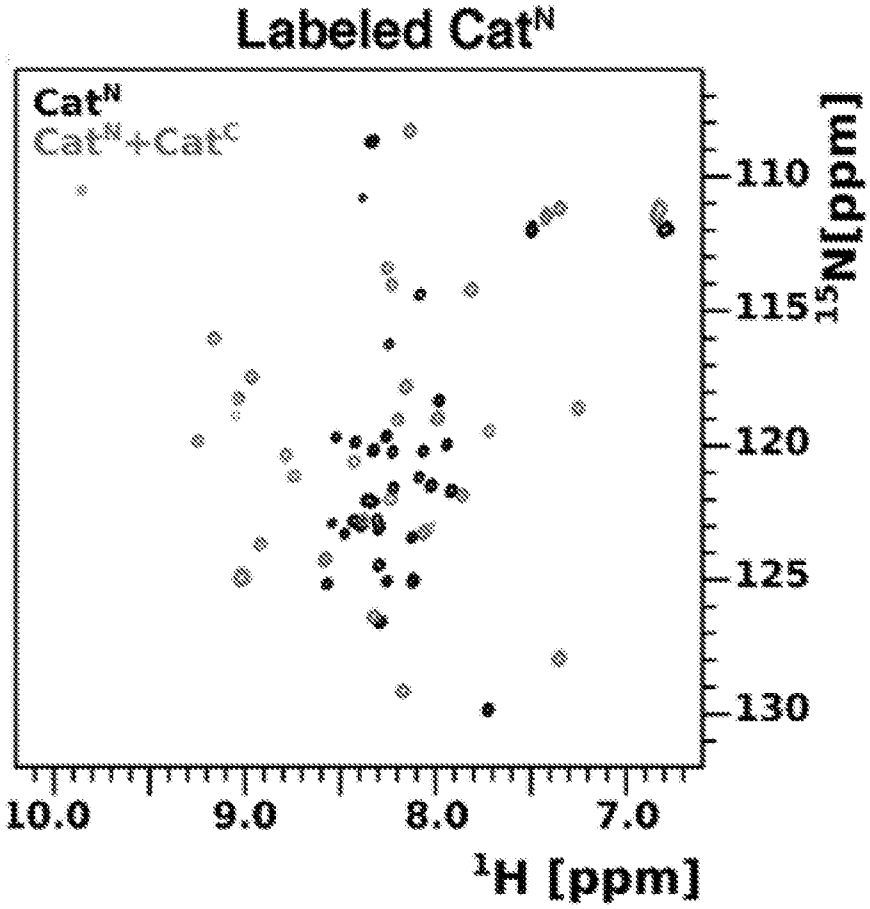

Labeled Cat^N
Cat^N
Cat^N+Cat^C
110
115
120
125
130
15N[ppm]
10.0
9.0
8.0
7.0
1H [ppm]

FIG. 6A

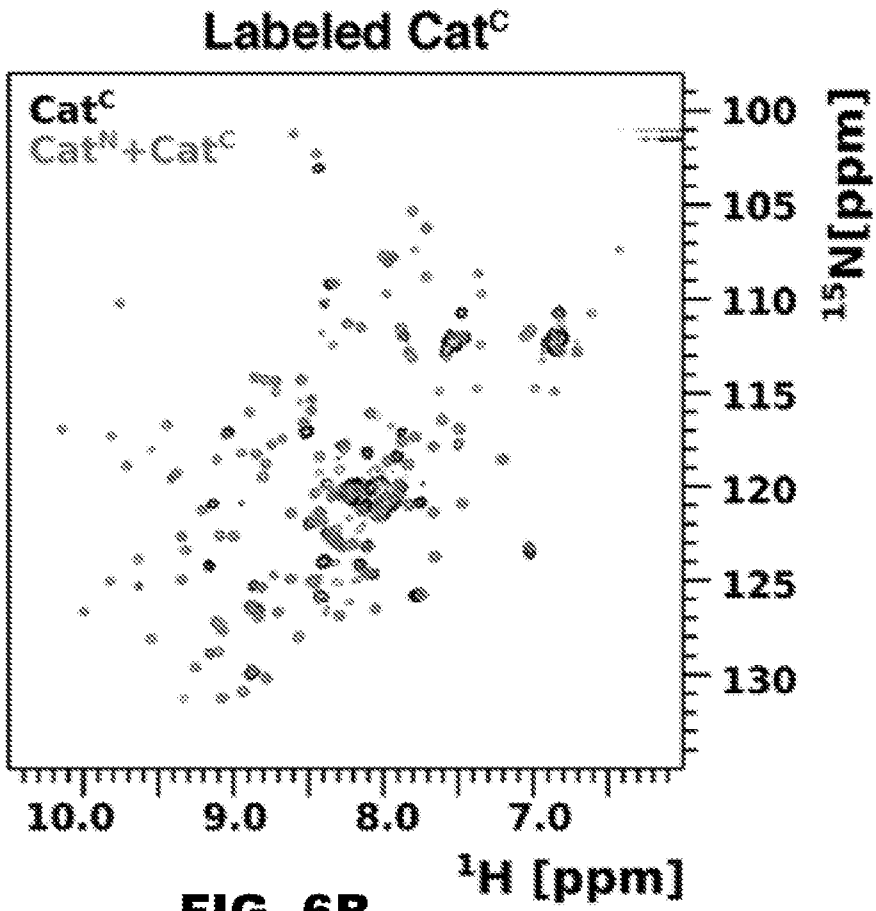

Labeled Cat^C
Cat^C
Cat^N+Cat^C
100
105
110
115
120
125
130
15N[ppm]
10.0
9.0
8.0
7.0
1H [ppm]

FIG. 6B

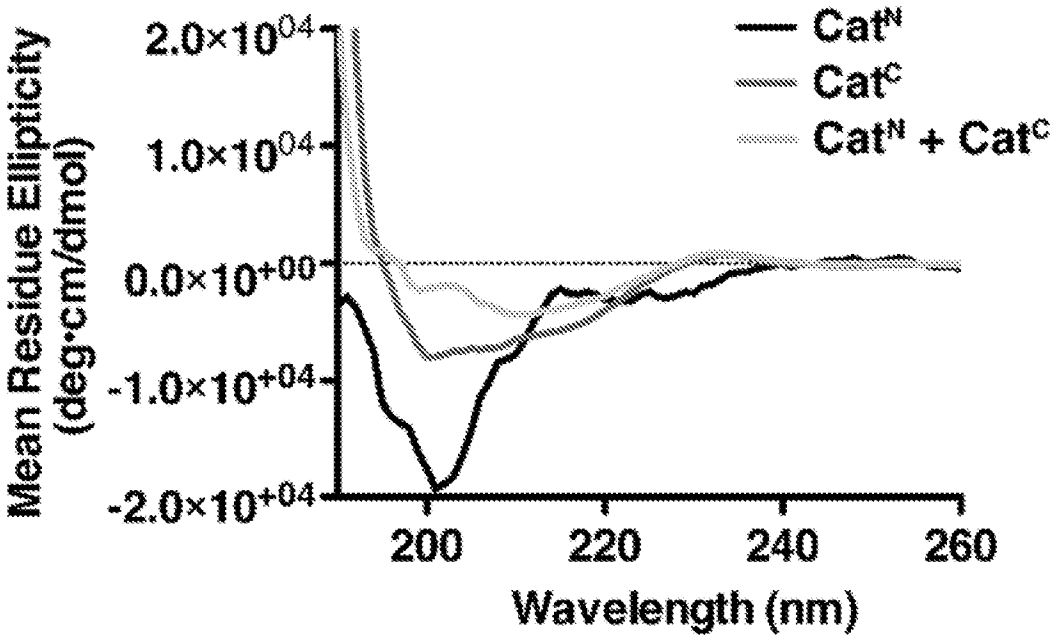

Mean Residue Ellipticity (deg·cm/dmol)
2.0×10^04
1.0×10^04
0.0×10^+00
-1.0×10^+04
-2.0×10^+04
200
220
240
260
Wavelength (nm)
Cat^N
Cat^C
Cat^N + Cat^C

FIG. 6C

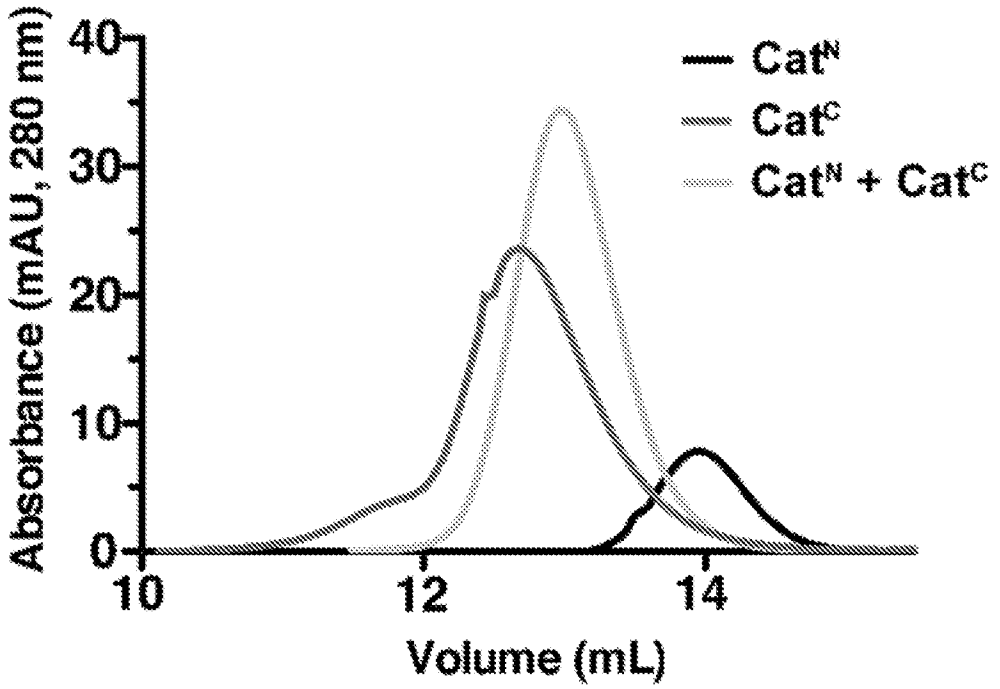

Absorbance (mAU, 280 nm)
40
30
20
10
0
10
12
14
Volume (mL)
Cat^N
Cat^C
Cat^N + Cat^C

FIG. 6D

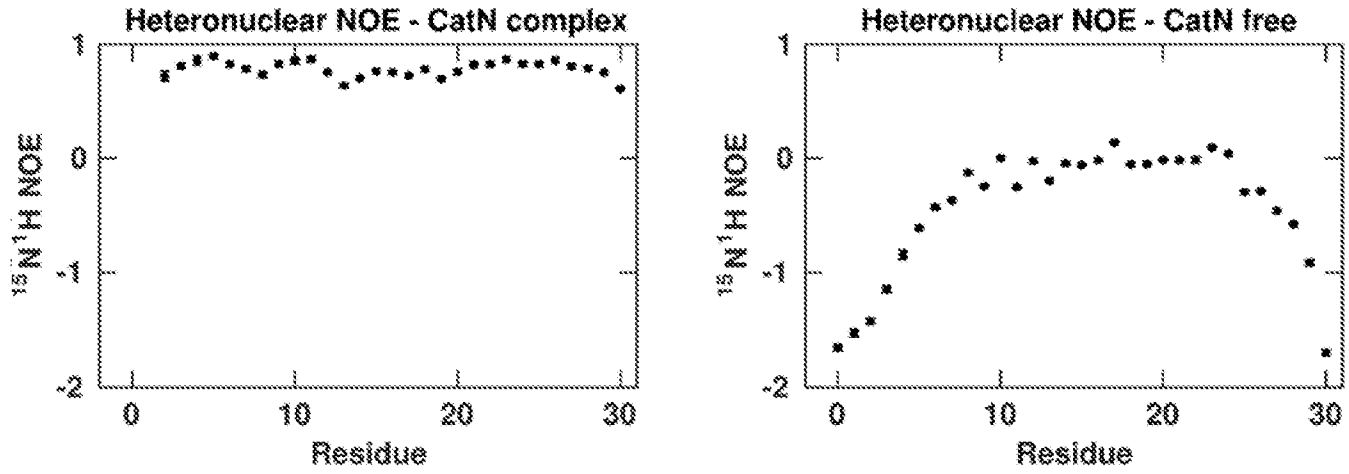


FIG. 7A

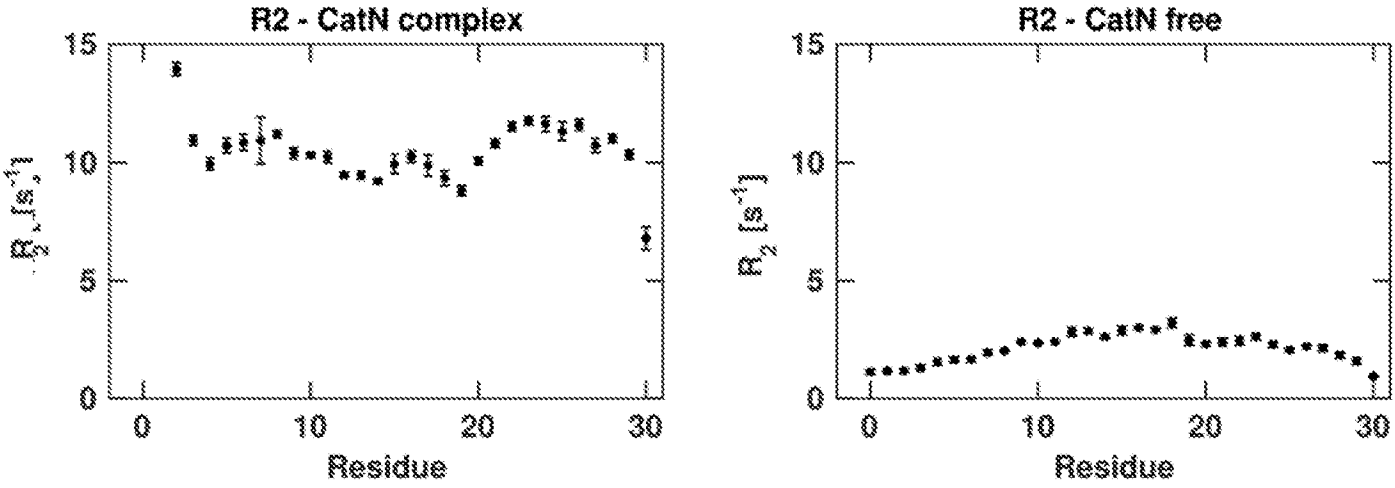


FIG. 7B

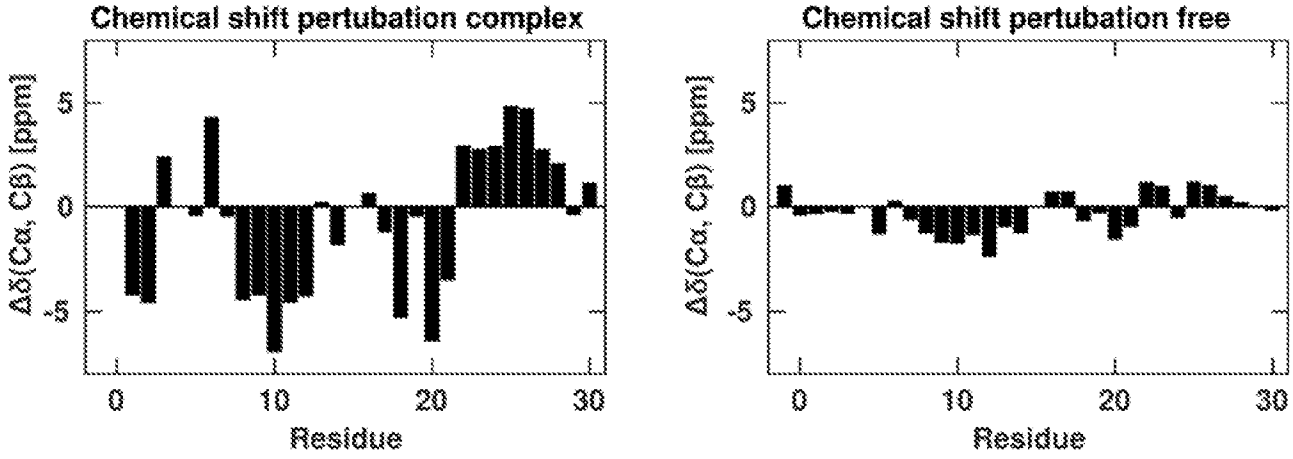


FIG. 7C

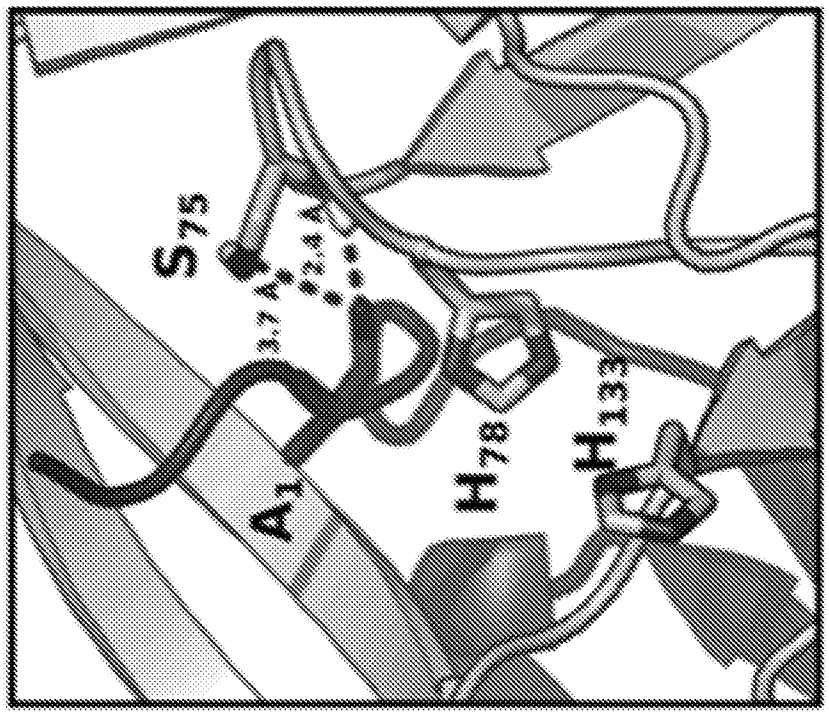


FIG. 8C



FIG. 8B

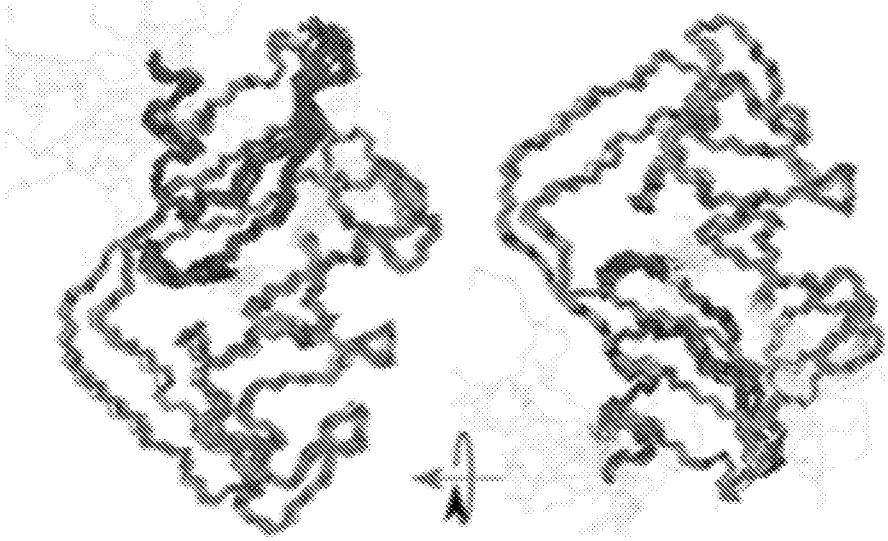


FIG. 8A

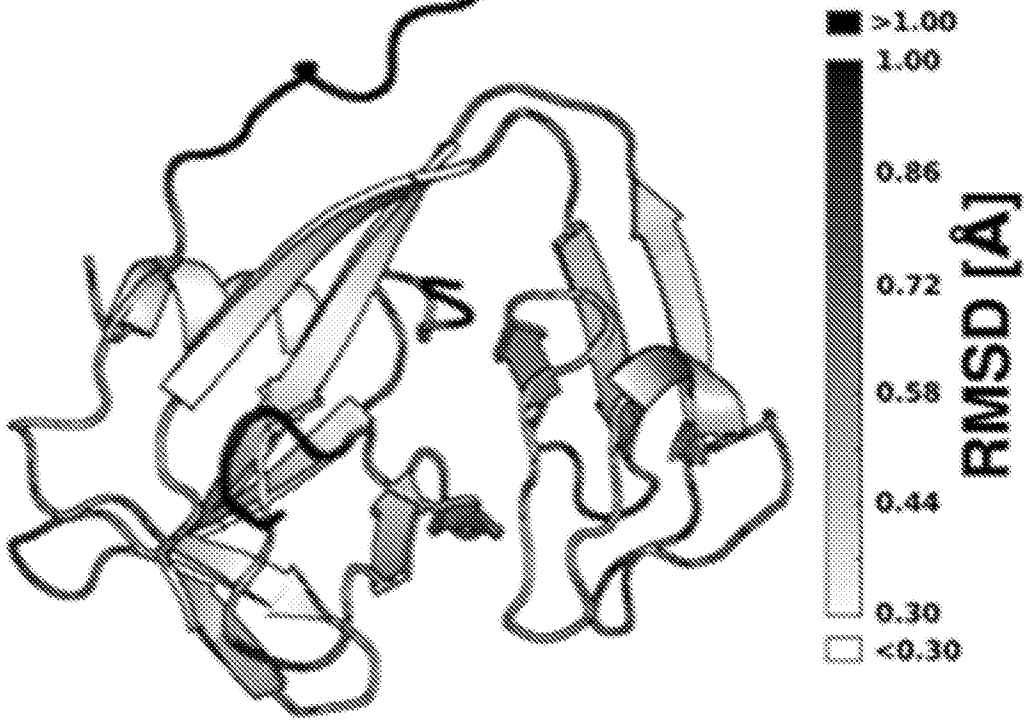

>1.00
1.00
0.86
0.72
0.58
0.44
0.30
<0.30
RMSD [Å]

FIG. 9A

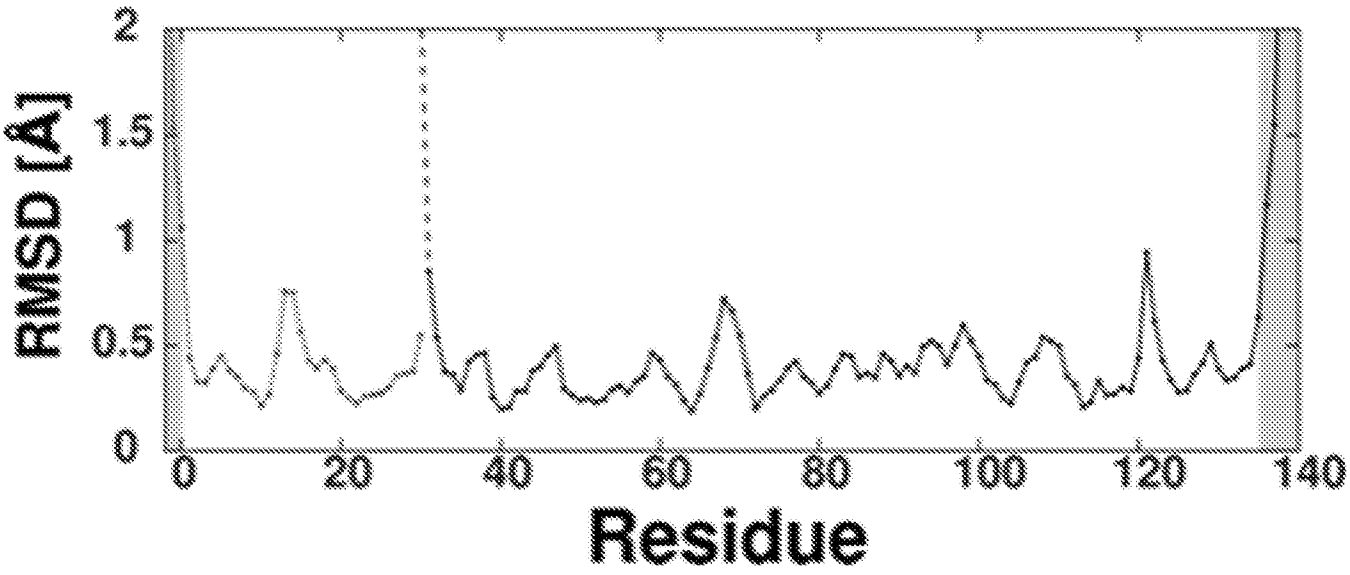

RMSD [Å]
2
1.5
1
0.5
0
0
20
40
60
80
100
120
140
Residue

FIG. 9B

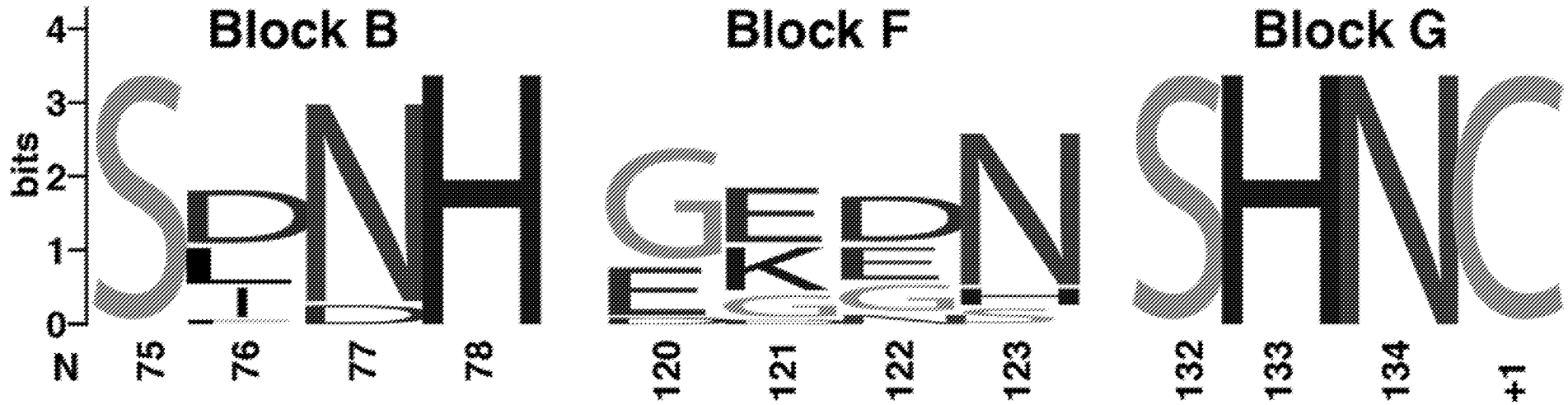


FIG. 9C

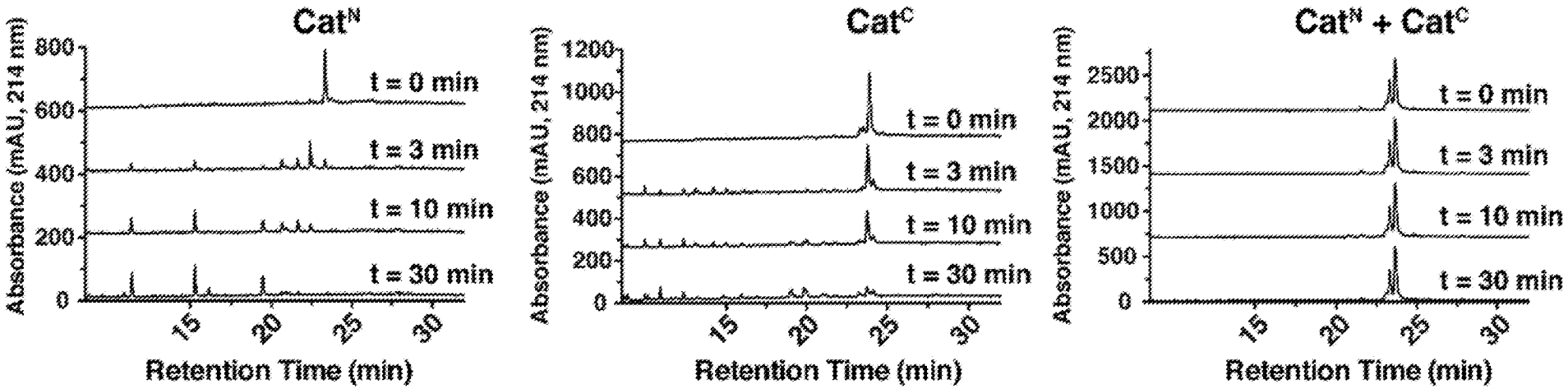


FIG. 10A

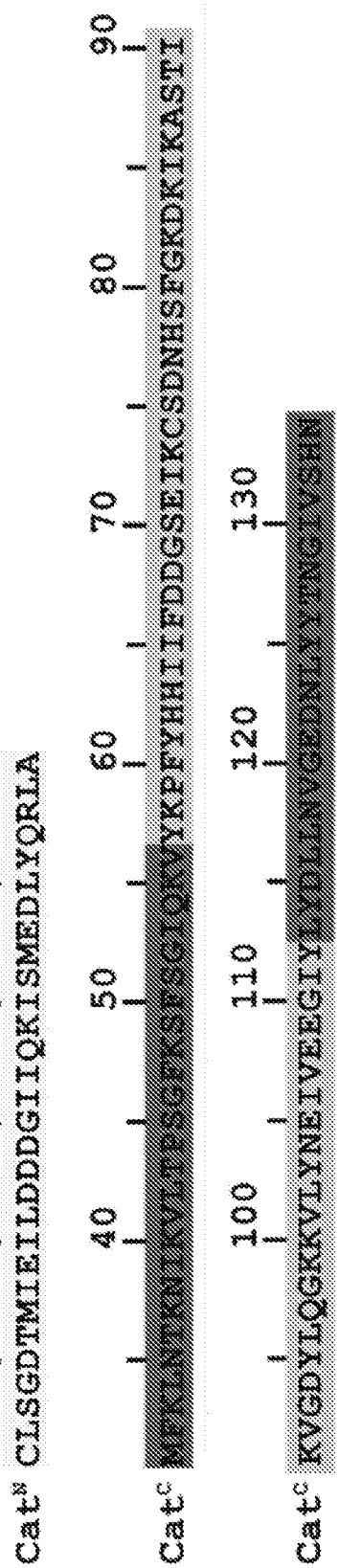


FIG. 10B

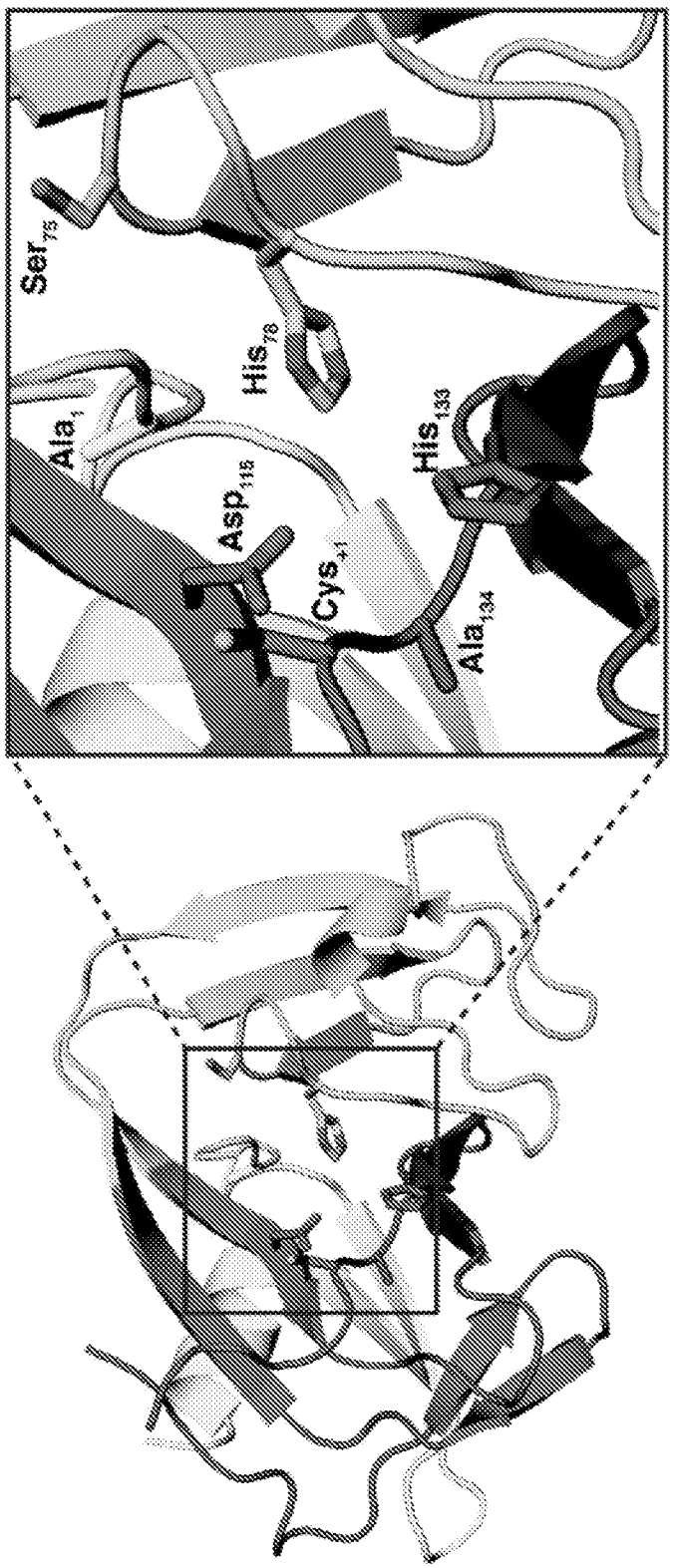


FIG. 10C

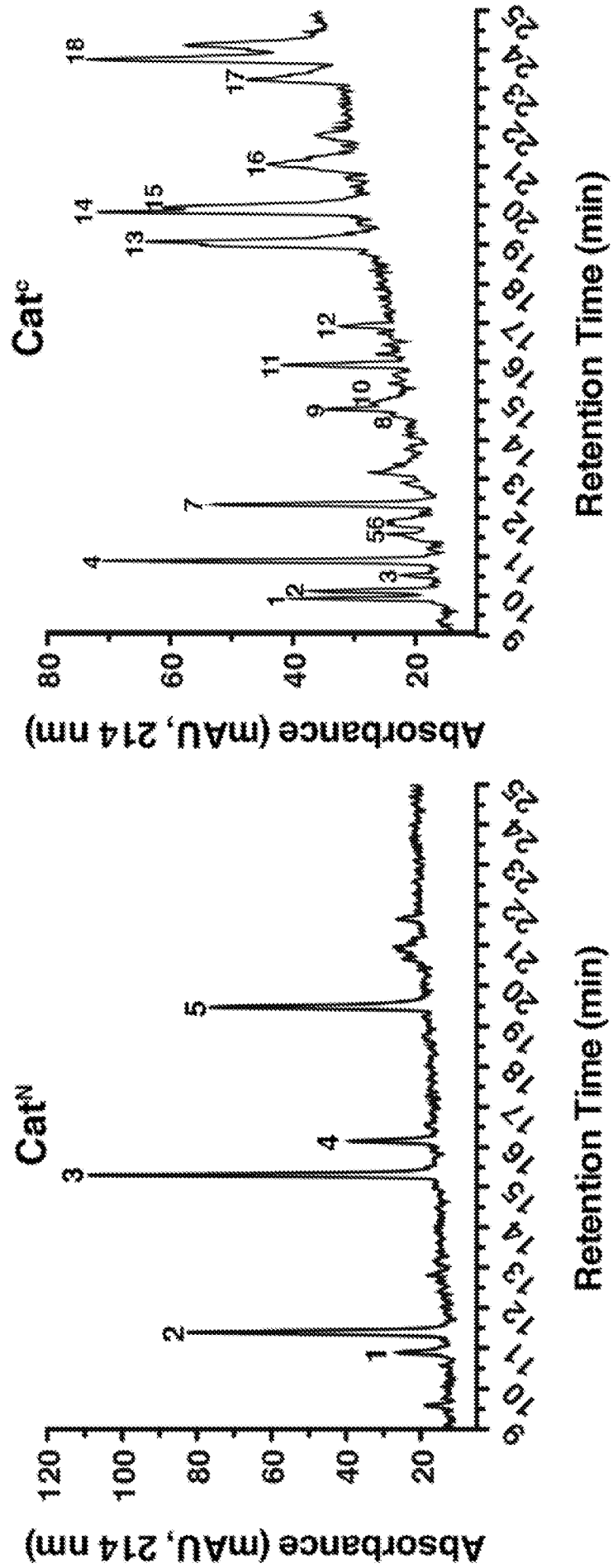

Cat°
Absorbance (mAU, 214 nm)
Retention Time (min)
Cat^N
Absorbance (mAU, 214 nm)
Retention Time (min)

FIG. 11A

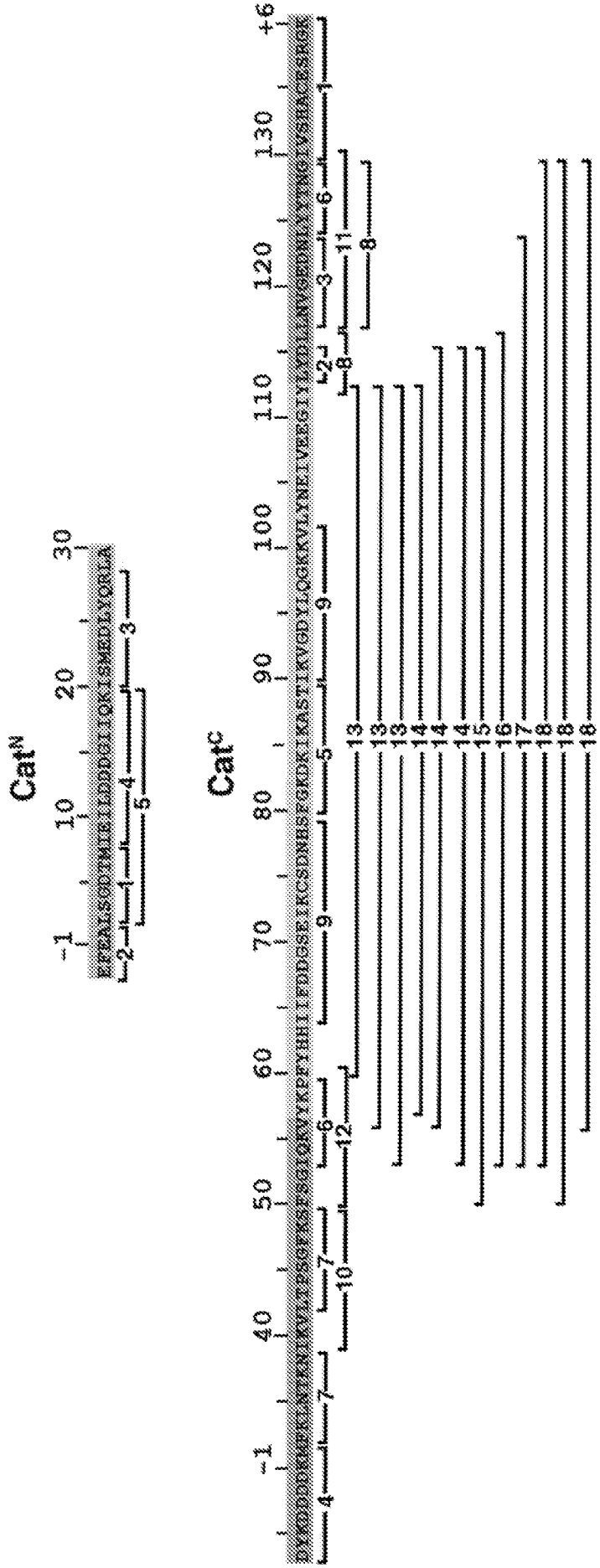


FIG. 11B

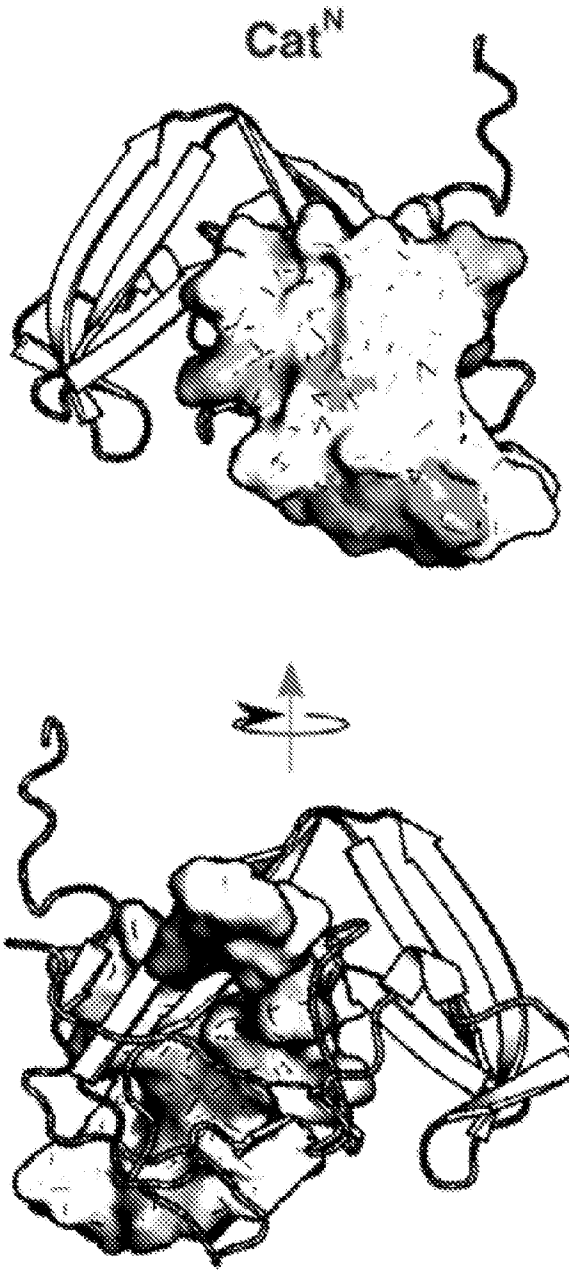


FIG. 12A

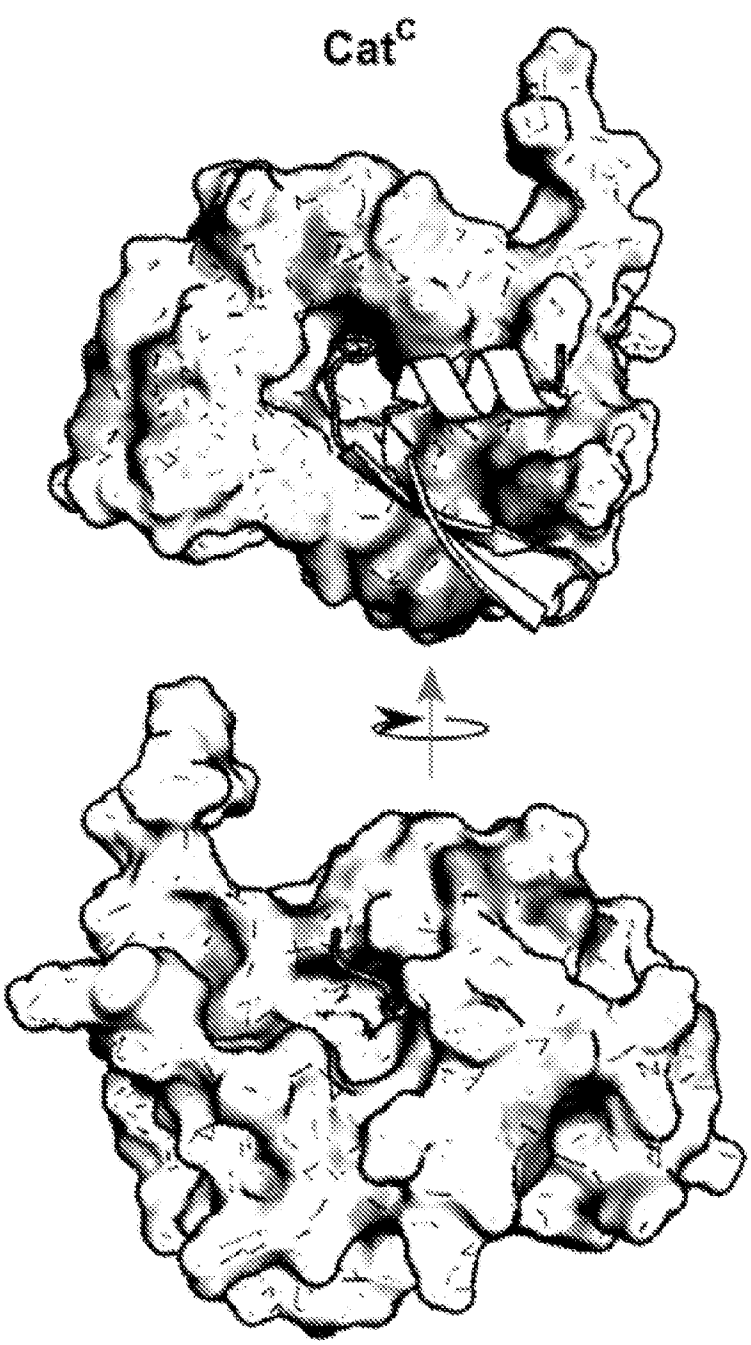


FIG. 12B

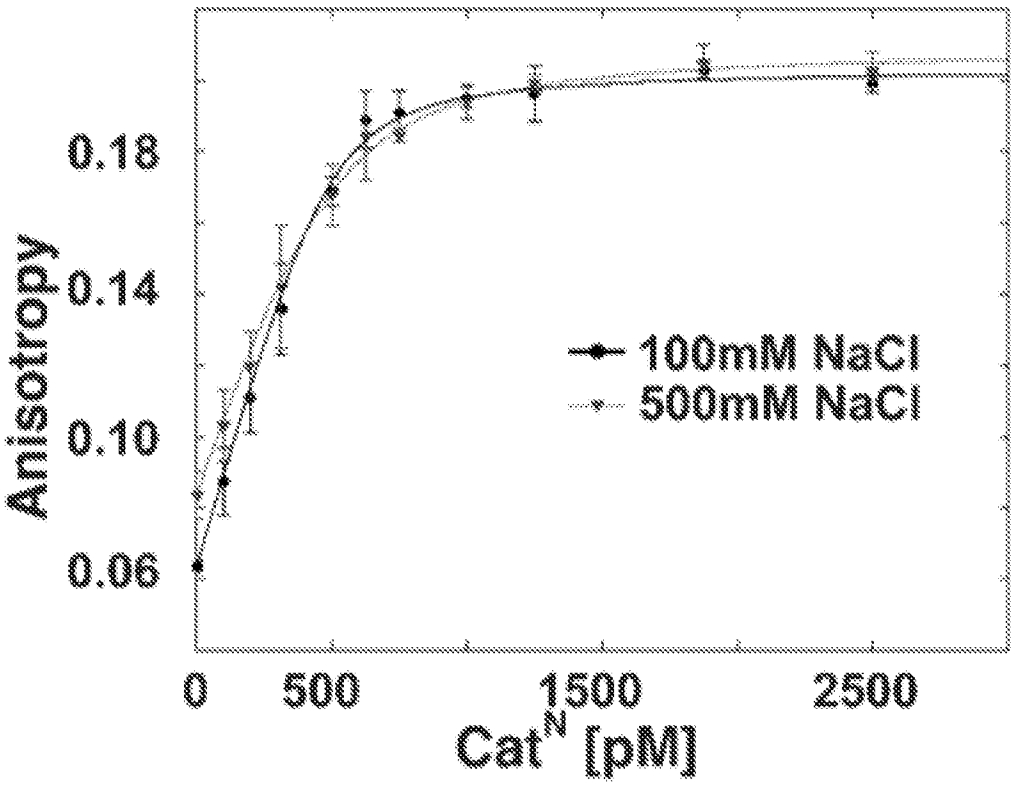


FIG. 12C

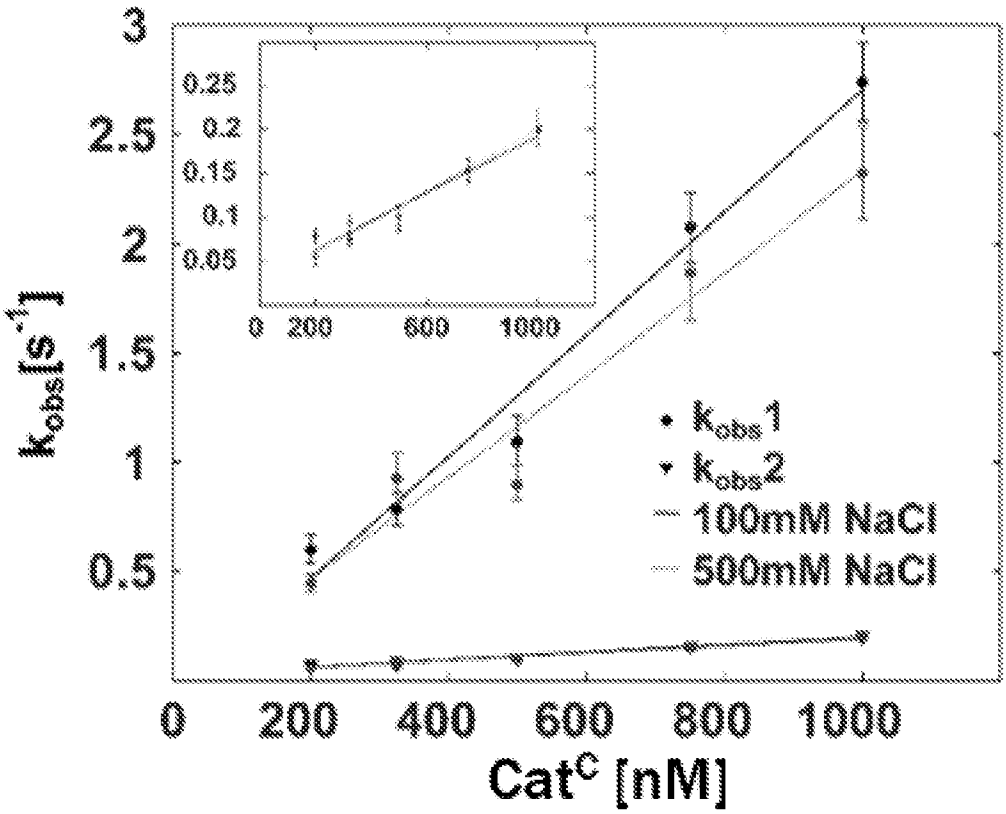


FIG. 12D

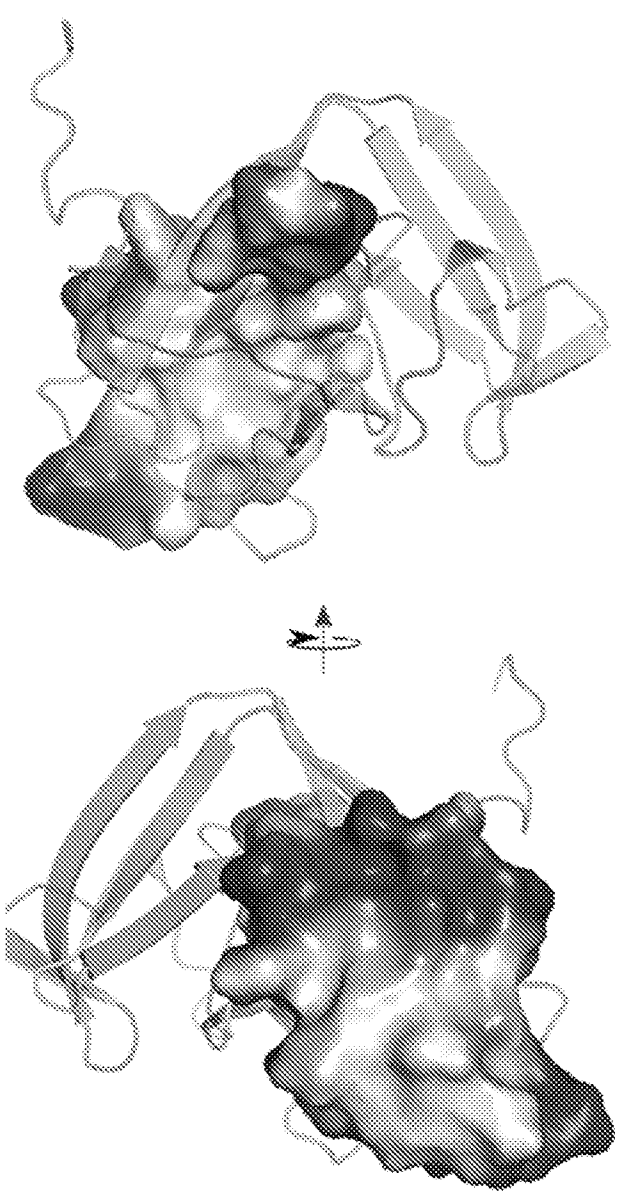


FIG. 13A

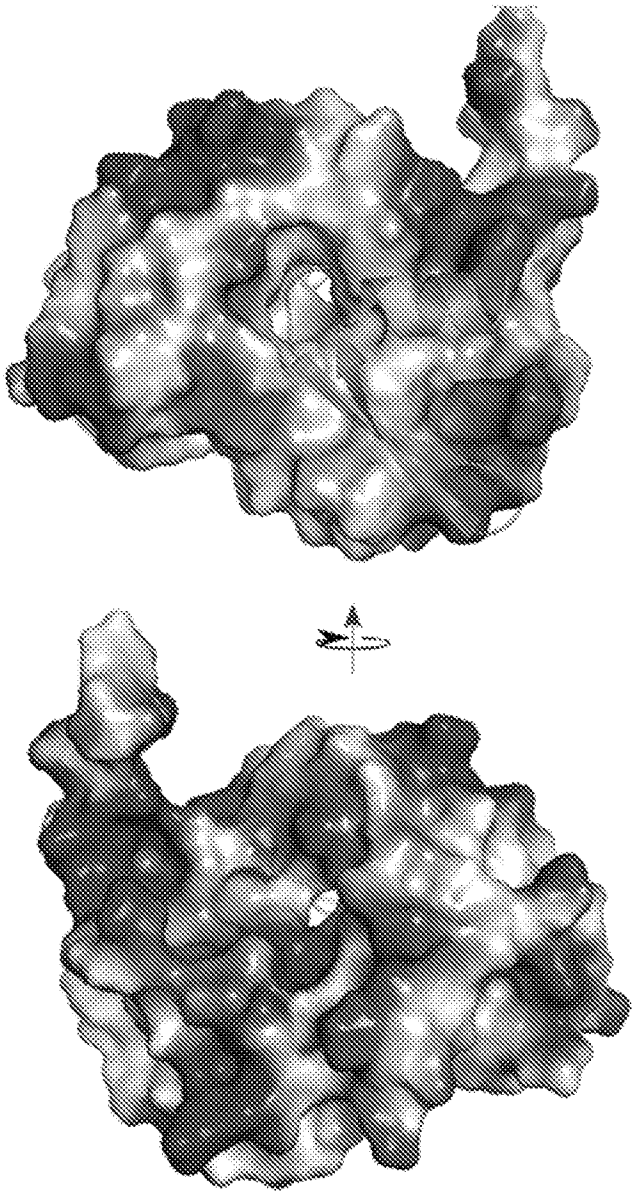


FIG. 13B

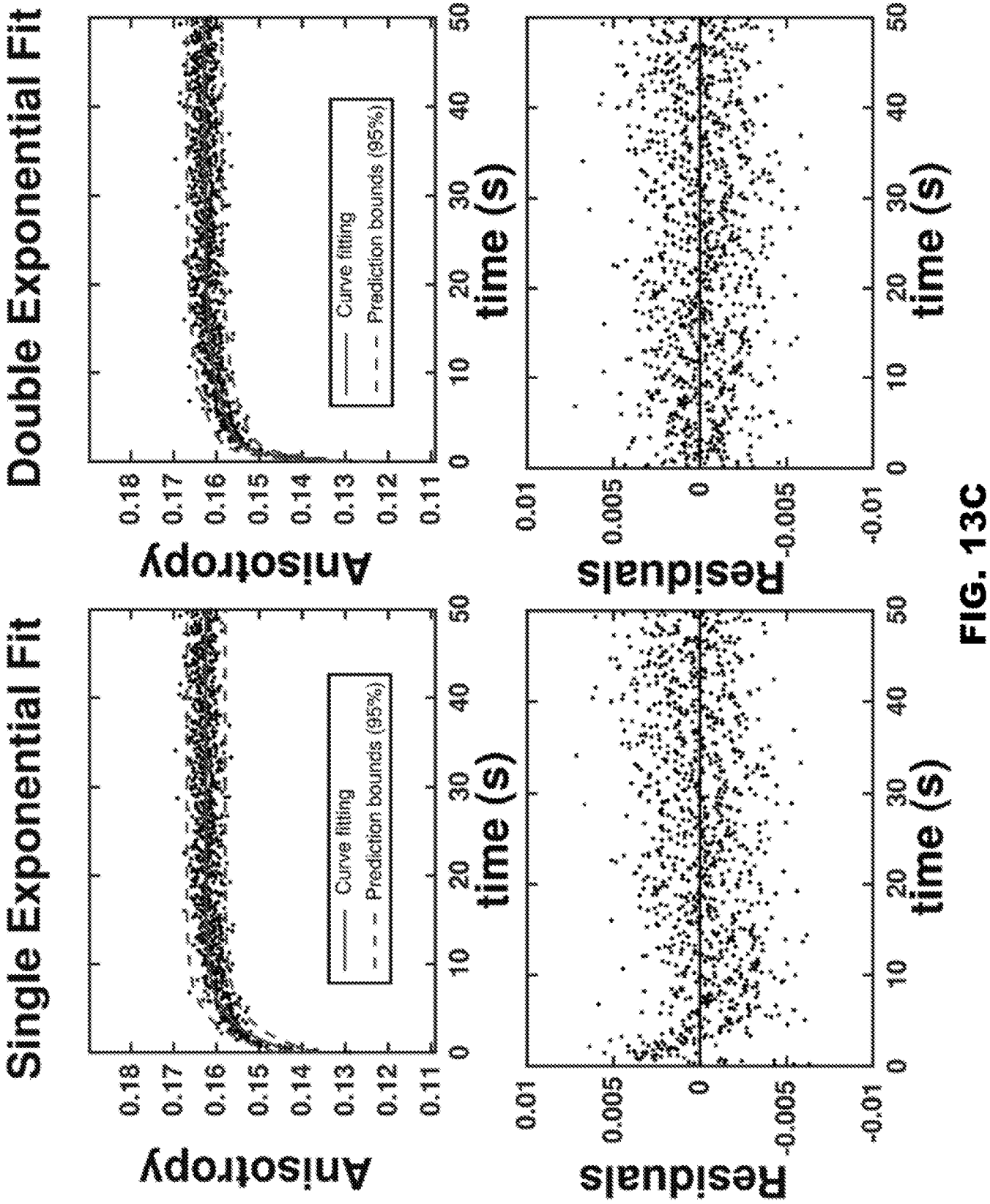

Single Exponential Fit
Double Exponential Fit
Anisotropy
Residuals
time (s)
Curve fitting
Prediction bounds (95%)
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.01
0.005
0
-0.005
-0.01
0
10
20
30
40
50

FIG. 13C

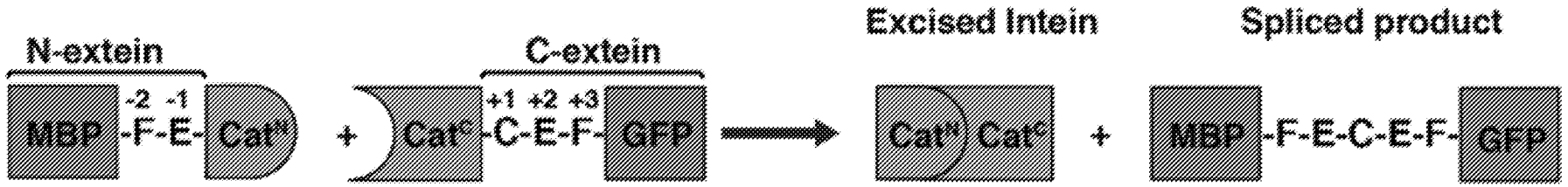


FIG. 14A

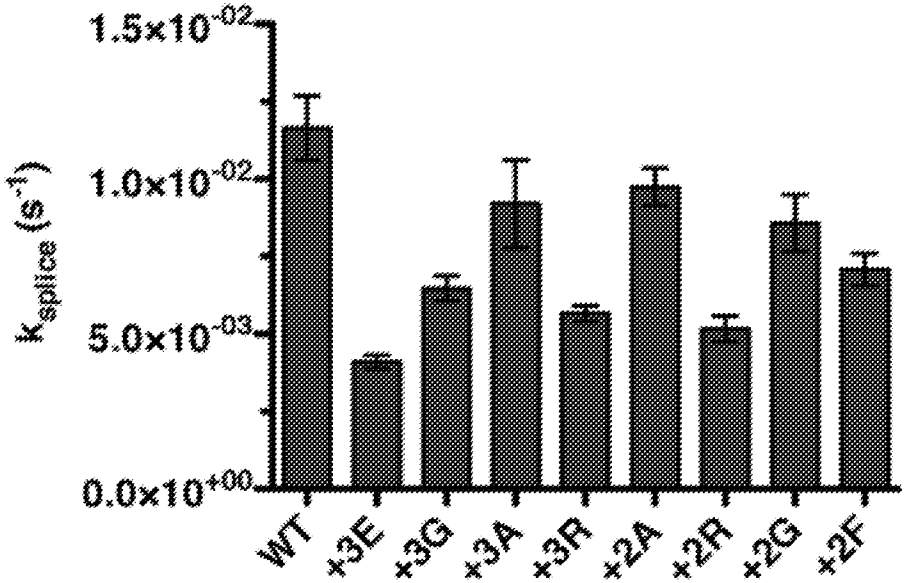


FIG. 14B

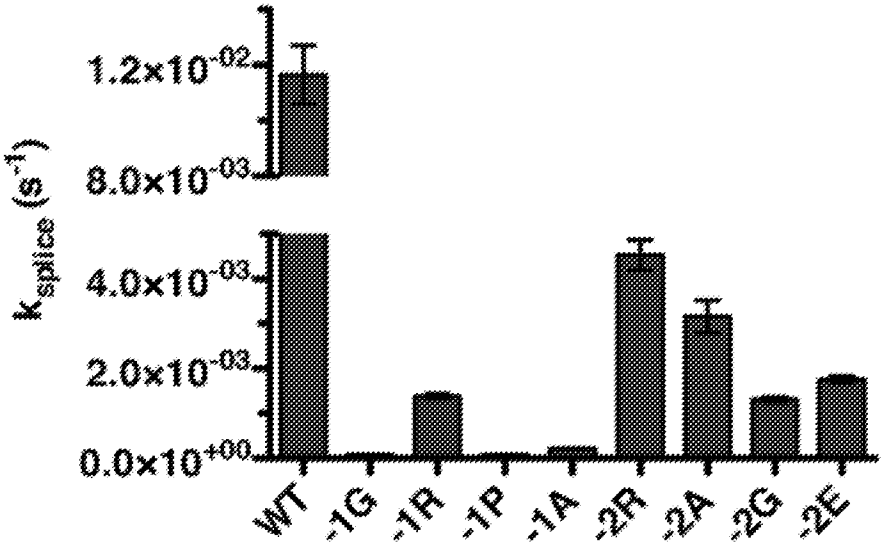


FIG. 14C

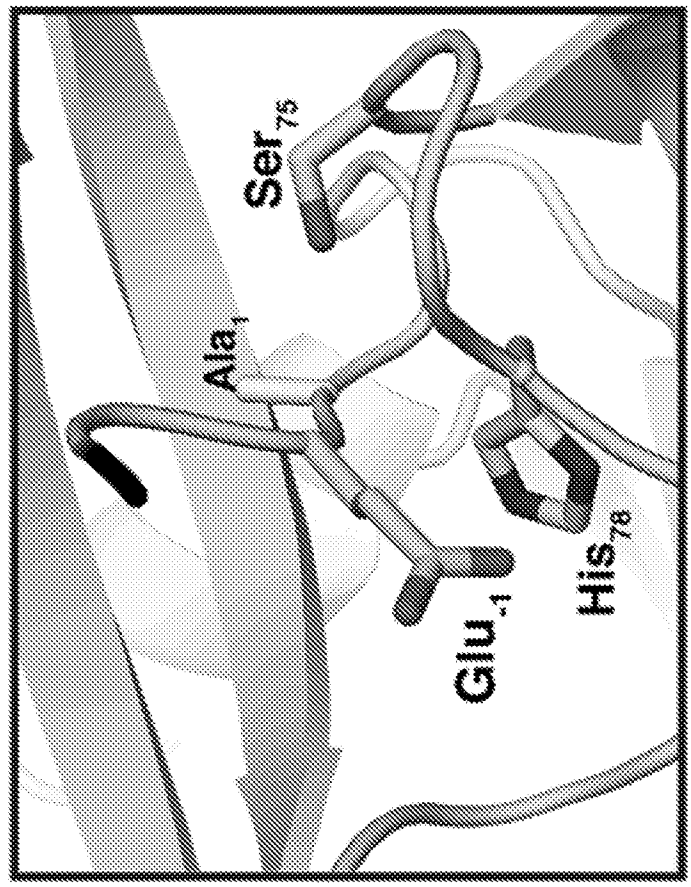


FIG. 14E

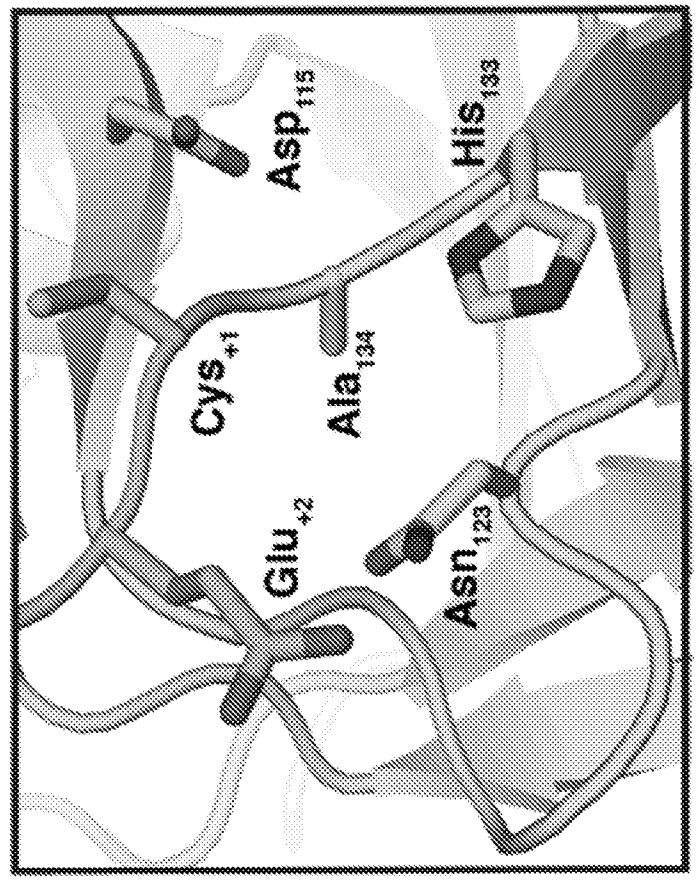


FIG. 14D

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/048508

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C07K 1/107; C07K 14/195; C07K 16/00; C07K 16/18 (2019.01)
CPC - C07K 14/195; C07K 2319/50; C07K 2319/92 (2019.08)

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 435/7.1; 506/9; 506/16 (keyword delimited)

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6,828,112 B2 (MANFREDI et al) 07 December 2004 (07.12.2004) entire document	1, 5, 13, 17, 38, 42, 43
A	WO 2017/132580 A2 (THE TRUSTEES OF PRINCETON UNIVERSITY) 03 August 2017 (03.08.2017) entire document	1, 5, 13, 17, 38, 42, 43
A	US 2018/0057577 A1 (THE TRUSTEES OF PRINCETON UNIVERSITY) 01 March 2018 (01.03.2018) entire document	1, 5, 13, 17, 38, 42, 43
A	STEVENS et al. "An Atypical Mechanism of Split Intein Molecular Recognition and Folding," Journal of the American Chemical Society, 29 August 2018 (29.08.2018), Vol. 140, No. 37, Pgs. 11791-11799. entire document	1, 5, 13, 17, 38, 42, 43
A	US 2015/0353597 A1 (THE TEXAS A&M UNIVERSITY SYSTEM) 10 December 2015 (10.12.2015) entire document	1, 5, 13, 17, 38, 42, 43

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 January 2020	22 JAN 2020
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/048508

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

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

   a. ☒ forming part of the international application as filed:

      ☒ in the form of an Annex C/ST.25 text file.

      ☐ on paper or in the form of an image file.

   b. ☐ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

   c. ☐ furnished subsequent to the international filing date for the purposes of international search only:

      ☐ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).

      ☐ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

SEQ ID NOs: 1-7, 103-110, 125-127, and 168-170 were searched.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/048508

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 7-12, 19-37
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

See extra sheet(s).

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1, 5, 13, 17, 38, 42, and 43 to the extent that they read on a split intein N-fragment of SEQ ID NO:1 and/or a split intein C-fragment of SEQ ID NO:7.

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/048508

Continued from Box No. III Observations where unity of invention is lacking

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees need to be paid.

Group I+: claims 1-6, 13-18, and 38-45 are drawn to split inteins, and compositions and methods comprising the same.

The first invention of Group I+ is restricted to a split intein N-fragment and a split intein C-fragment, and compositions and methods comprising the same, wherein the split intein N-fragment is selected to be SEQ ID NO:1 and the split intein C-fragment is selected to be SEQ ID NO:7. It is believed that claims 1, 5, 13, 17, 38, 42, and 43 read on this first named invention and thus these claims will be searched without fee to the extent that they read on a split intein N-fragment of SEQ ID NO:1 and/or a split intein C-fragment of SEQ ID NO:7.

Applicant is invited to elect additional split intein N-fragments and/or split intein C-fragments to be searched in a specific combination by paying an additional fee for each set of election. An exemplary election would a split intein N-fragment and a split intein C-fragment, and compositions and methods comprising the same, wherein the split intein N-fragment is selected to be SEQ ID NO:3 and the split intein C-fragment is selected to be SEQ ID NO:9. Additional split inteins will be searched upon the payment of additional fees. Applicants must specify the claims that read on any additional elected inventions. Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid *additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined.*

The inventions listed in Groups I+ do not relate to a single general inventive concept under PCT Rule 13.1, because under PCT Rule 13.2 they lack the same or corresponding special technical features for the following reasons:

The Groups I+ formulas do not share a significant structural element responsible for functioning in trans-splicing reactions for expression of a gene of interest, requiring the selection of alternatives for split intein N-fragments and/or split intein C-fragments, where "...[the] split intein N-fragment comprising the amino acid sequence of SEQ ID NO:1 or a variant thereof having at least 90% sequence identity with SEQ ID NO:1" and "...[the] split intein C-fragment comprising the amino acid sequence of SEQ ID NO:7 or a variant thereof having at least 88% sequence identity with SEQ ID NO:7".

Additionally, even if Groups I+ were considered to share the technical features of a split intein N-fragment;a split intein C-fragment; a composition comprising: (a) a first polynucleotide encoding a first fusion protein comprising, from the N terminus to the C-terminus: - a first polypeptide of interest and - a split intein N-fragment and (b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: - an AceL-TerL split intein C-fragment or a variant thereof or a split intein C-fragment and - a second polypeptide of interest; a method for expressing a gene of interest in a cell comprising: (i) contacting the cell with (a) a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus: a first polypeptide of interest and - a split intein N-fragment, and (b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment, and a second polypeptide of interest; a method for expressing a gene of interest comprising: (i) contacting a first cell with a first polynucleotide encoding a first fusion protein comprising, from the N-terminus to the C-terminus: - a first polypeptide of interest and - a split intein N-fragment, wherein the first fusion protein comprises a signal peptide, and (ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: - an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment, and a - second polypeptide of interest wherein the second fusion protein comprises a signal peptide; these shared technical features do not represent a contribution over the prior art.

Specifically, WO 2017/132580 A2 to The Trustees of Princeton University discloses a split intein N-fragment (split intein, Abstract; a split intein N-fragment, Para. [0006]); a composition comprising: (a) a first polynucleotide encoding a first fusion protein comprising, from the N terminus to the C-terminus: - a first polypeptide of interest and - a split intein N-fragment (a composition including a split intein N-fragment, Para. [0011]; a nucleotide plasmid including a nucleotide sequence encoding for a split intein N-fragment, Para. [0012]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide... the compound can be... oligonucleotide, Para. [0024]) and (b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: a split intein C-fragment and a second polypeptide of interest (a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein C-fragment, Para. [0013]; using a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) including with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment), so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein, Para. [0020]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide, Para. [0024]); a method for expressing a gene of interest in a cell (the DDec205-HC-Int were co-transfected into HEK293T cells, Para. [00742]) comprising: (i) contacting the cell with (a) a first polynucleotide (the Intein-dendrimer fusion was expressed and purified exactly as described, Para. [00723]; the DDec205-HC-Int were co-transfected into HEK293T cells, Para. [00742]) encoding a first fusion protein comprising, from the N-terminus to the C-terminus: a first polypeptide of interest and - a split intein N-fragment (a composition including a split intein N-fragment, Para. [0011]; a nucleotide plasmid including a nucleotide sequence encoding for a split intein N-fragment, Para. [0012]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide... the compound can be... oligonucleotide, Para. [0024]), and (b) a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: an AceL-TerL split intein C-fragment or a functionally equivalent variant thereof, or a split intein C-fragment, and a second polypeptide of interest (a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein C-fragment, Para. [0013]; using a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) including with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment), so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein, Para. [0020]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide, Para. [0024]); a method for expressing a gene of interest (the Intein-dendrimer fusion was expressed and purified exactly as described, Para. [00723]; the DDec205-HC-Int were co-transfected into HEK293T cells,

Para. [00742]; the expression and purification protocols... methods, Para. [00709]) comprising: (i) contacting a first cell with a first polynucleotide (the Intein-dendrimer fusion was expressed and purified exactly as described, Para. [00723]; the DDec205-HC-Int were co-transfected into HEK293T cells, Para. [00742]) encoding a first fusion protein comprising, from the N-terminus to the C-terminus: - a first polypeptide of interest and - a split intein N-fragment, wherein the first fusion protein comprises a peptide (a composition including a split intein N-fragment, Para. [0011]; a nucleotide plasmid including a nucleotide sequence encoding for a split intein N-fragment, Para. [0012]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide... the compound can be... oligonucleotide, Para. [0024]), and (ii) contacting a second cell with a second polynucleotide encoding a second fusion protein comprising, from the N-terminus to the C-terminus: a split intein C-fragment, and a - second polypeptide of interest wherein the second fusion protein comprises a peptide (a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein C-fragment, Para. [0013]; using a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) including with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment), so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein, Para. [0020]; fusion protein comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide, Para. [0024]; the DDec205-HC-Int were co-transfected into HEK293T cells, Para. [00742]).

US 6,828,112 B2 to Manfredi et al. discloses a signal peptide (a pair of fusion proteins can have a signal peptide, Col. 4, Lns. 33-34).

The inventions listed in Groups I+ therefore lack unity under Rule 13 because they do not share a same or corresponding special technical features.