



- (51) **International Patent Classification:**
A61K 47/48 (2006.01) *C07K 19/00* (2006.01)
- (21) **International Application Number:**
PCT/US2017/015455
- (22) **International Filing Date:**
27 January 2017 (27.01.2017)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/288,661 29 January 2016 (29.01.2016) US
- (71) **Applicant:** THE TRUSTEES OF PRINCETON UNIVERSITY [US/US]; 87 Prospect Avenue, Princeton, NJ 08544 (US).
- (72) **Inventors:** MUIR, Tom, W.; 35 Littlebrook Road N., Princeton, NJ 08540 (US). STEVENS, Adam, J.; 74 Spruce Street, Apt. 3, Princeton, NJ 08542 (US). SHAH, Neel, H.; 1509 Hearst Avenue, Apt. 302, Berkeley, CA 94703 (US).
- (74) **Agent:** GENIESER, Lars, H.; Venable LLP, P.O. Box 34385, Washington, DC 20043-9998 (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, 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, 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:

- as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))
- of inventorship (Rule 4.17(iv))

[Continued on next page]

(54) **Title:** SPLIT INTEINS WITH EXCEPTIONAL SPLICING ACTIVITY

(57) **Abstract:** Embodiments of the present invention relate to inteins, split inteins, compositions comprising inteins and methods for use of these.

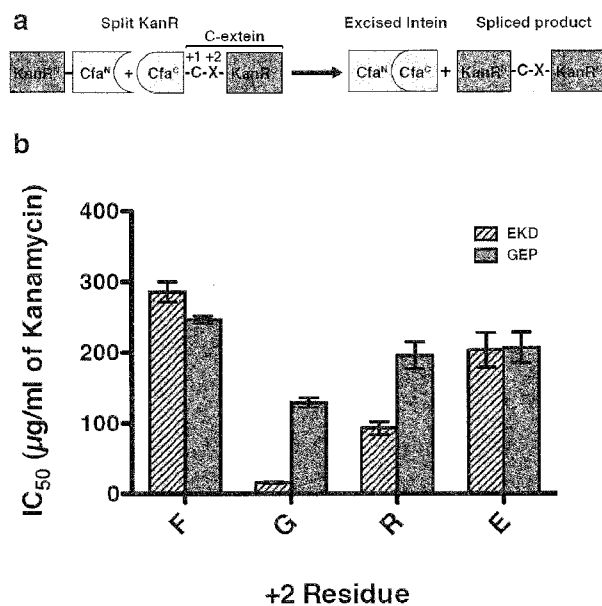


FIG. 9



Published:

- *without international search report and to be republished
upon receipt of that report (Rule 48.2(g))*

SPLIT INTEINS WITH EXCEPTIONAL SPLICING ACTIVITY

CROSS-REFERENCE OF RELATED APPLICATION

[0001] This application claims priority to U.S. Provisional Application No. 62/288,661 filed January 29, 2016, which is hereby incorporated by reference in its entirety.

[0002] This invention was made with government support under Grant No. GM086868 awarded by the National Institutes of Health. The government has certain rights in the invention.

BACKGROUND

1. Technical Field

[0003] The field of the currently claimed embodiments of the present invention relate to inteins, split inteins, compositions comprising inteins and methods for use of the like for protein engineering.

2. Discussion of Related Art

[0004] Protein splicing is a posttranslational auto-processing event in which an intervening protein domain called an intein excises itself from a host protein in a traceless manner such that the flanking polypeptide sequences (exteins) are ligated together via a normal peptide bond.¹ While protein splicing typically occurs spontaneously following translation of a contiguous polypeptide, some inteins exist naturally in a split form.¹ The two pieces of the split intein are expressed separately and remain inactive until encountering their complementary partner, upon which they cooperatively fold and undergo splicing in trans. This activity has been harnessed in a host of protein engineering methods that provide control over the structure and activity of proteins both *in vitro* and *in vivo*.¹ The first two split inteins to be characterized, from the cyanobacteria *Synechocystis* species PCC6803 (Ssp) and *Nostoc punctiforme* PCC73102 (Npu), are orthologs naturally found inserted in the alpha subunit of DNA Polymerase III (DnaE).²⁻⁴ Npu is especially notable due its remarkably fast rate of protein trans-splicing (PTS) ($t_{1/2}$ =50s at 30°C).⁵ This half-life is significantly shorter than that of Ssp ($t_{1/2}$ =80 min at 30°C),⁵ an attribute that has expanded the range of applications open to PTS.¹

[0005] Despite the ongoing discovery of new fast inteins,^{6,7} little is known about what separates them from their slower homologues. Such an understanding should help identify new inteins that are likely to splice rapidly and potentially allow for the engineering of split inteins with superior PTS properties.

SUMMARY

[0006] Embodiments of the invention include a split intein N-fragment including an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1).

[0007] Embodiments of the invention include a split intein N-fragment including an amino acid sequence, wherein said amino acid sequence comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO:2).

[0008] Embodiments of the invention include a split intein C-fragment including an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDNHFLKNGLVASN (SEQ ID NO: 3).

[0009] Embodiments of the invention include a split intein C-fragment including an amino acid sequence, wherein said amino acid sequence of said C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to MVKIISRKSLGTQNVYDIGVEKDNHFLKNGLVASN (SEQ ID NO: 4).

[0010] Embodiments of the invention include a split intein C-fragment including an amino acid sequence, wherein said amino acid sequence of said C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).

[0011] Embodiments of the invention include a composition including a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2)

and a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDNHFLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4), or
VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).

[0012] Embodiments of the invention include a nucleotide plasmid including a nucleotide sequence encoding for a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDQGMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDQGMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2).

[0013] Embodiments of the invention include a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein C-fragment including an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).

[0014] Embodiments of the invention include a method for splicing two complexes including the following: contacting a first complex comprising a first compound and a split intein N-fragment with a second complex comprising a second compound and a split intein C-fragment, with contacting performed under conditions that permit binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and reacting the intein intermediate to form a conjugate of the first compound with the second compound. The split intein N-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDQGMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDQGMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2), and

the split intein C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).

[0015] Embodiments of the invention include a method including the following: contacting a first complex comprising a first compound and a split intein N-fragment with a second complex comprising a second compound and a split intein C-fragment, with the contacting performed under conditions that permit binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and reacting the intein intermediate with a nucleophile to form a conjugate of the first compound with the nucleophile. The split intein N-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2), and

the split intein C-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLLKNGLVASN (SEQ ID NO: 389).

[0016] In some embodiments, the compound, first compound, or second compound is or includes a peptide or a polypeptide. In some embodiments, the compound, first compound, or second compound is or includes an antibody, antibody chain, or antibody heavy chain. In some embodiments, the compound, first compound, or second compound is or includes a peptide, oligonucleotide, drug, or cytotoxic molecule.

[0017] Embodiments of the invention include an intein comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLPVKIISRKSLGTQNVYDIGVEKDH
NFLLKNGLVASN (SEQ ID NO: 390).

[0018] Embodiments of the invention include a kit for splicing two complexes together including the following: a split intein N-fragment including an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2);

a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389);

reagent(s) for permitting the binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and a nucleophilic agent.

[0019] Embodiments of the invention include a method for generating a synthetic consensus intein peptide sequence including the following: generating a population of a plurality of homologous intein peptide sequences; identifying amino acids associated with fast splicing within the population of a plurality of homologous intein peptide sequences; generating a subpopulation of a second plurality of homologous intein peptide sequences, with the second plurality of homologous intein peptide sequences including amino acids associated with fast splicing; creating an alignment of at least three peptide sequences of the subpopulation; determining a most frequently occurring amino acid residue at each position of the at least three peptide sequences; and generating a synthetic consensus intein peptide sequence based on the most frequently occurring amino acid residue at each position of the at least three peptide sequences.

[0020] Embodiments of the invention include a method including the following: fusing 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. The split intein N-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2), and

the split intein C-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLLKNGLVASN (SEQ ID NO: 389).

[0021] Embodiments of the invention include a method including the following: fusing a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) including

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment) including
VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3), so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein.

[0022] Embodiments of the invention include a gene fusion including the following: a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment). The split intein N-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2), and
the split intein C-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3),

MVKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 4), or

VKIISRKSLGTQNVYDIGVGEPHNFLLKNGLVASN (SEQ ID NO: 389).

[0023] Embodiments of the invention include a gene fusion including the following: a first nucleotide sequence encoding an amino acid sequence of a first intein fragment (split intein N-fragment) including

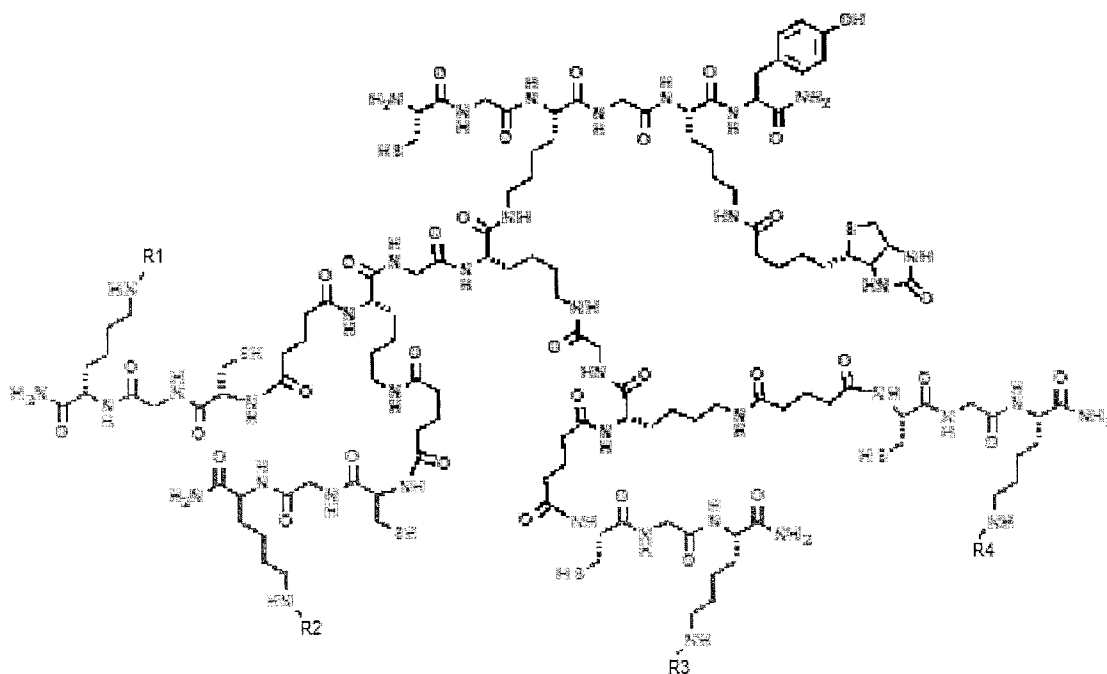
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) fused with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment (split intein C-fragment) including VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3).

[0024] Embodiments of the invention include a complex (e.g., a fusion protein) comprising a split intein N-fragment and a compound. For example, the compound can be or include a peptide, a polypeptide or an antibody chain, such as an antibody heavy chain. For example, the compound can include a peptide, oligonucleotide, drug, or cytotoxic molecule. For example, the compound can be a 1,2-amino thiol or a 1,2-amino alcohol bonded to a peptide, oligonucleotide, drug, or cytotoxic molecule. The split intein N-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

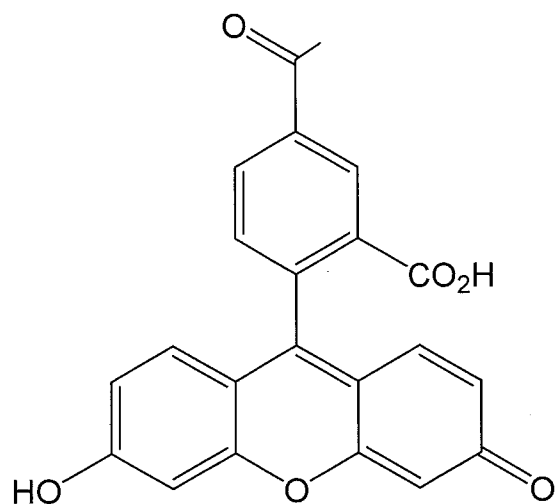
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYQTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) or an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYQTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2).

[0025] Embodiments of the invention include a complex (e.g., a fusion protein) comprising a split intein C-fragment and a compound. For example, the compound can be or include a dendrimer, peptide or polypeptide. For example, the compound can include a peptide, an oligonucleotide, a drug, or a cytotoxic molecule. For example, the compound can be a 1,2-amino thiol or a 1,2-amino alcohol bonded to a peptide, oligonucleotide, drug, or cytotoxic molecule. The split intein C-fragment includes an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDNFLLKNGLVASN (SEQ ID NO: 3), an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to MVKIISRKSLGTQNVYDIGVEKDNFLLKNGLVASN (SEQ ID NO: 4), or an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVGEPHNFLLKNGLVASN (SEQ ID NO: 389). The dendrimer can be a compound having the structure

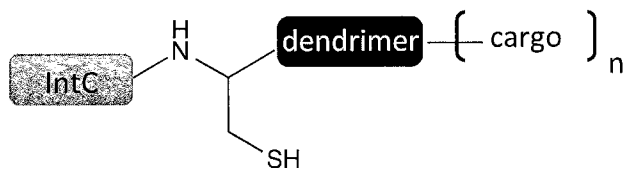


wherein R1, R2, R3, and R4 are each (independently) hydrogen (H) or a cargo molecule (the cargo molecules on R1, R2, R3, and R4 can be different from each other). R1, R2, R3, and R4 can each be a dye molecule. For example, R1, R2, R3, and R4 can each be a fluorescein derivative having the structure



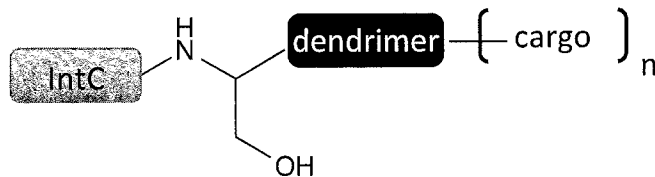
[0026]

Embodiments of the invention include a complex of the structure



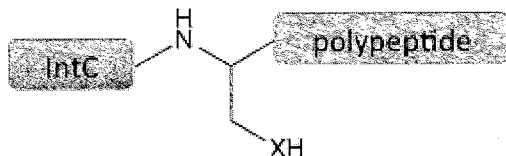
with IntC a split intein C-fragment and n from 0 to 8.

[0027] Embodiments of the invention include a complex of the structure



with IntC a split intein C-fragment and n from 0 to 8.

[0028] Embodiments of the invention include a complex of the structure



with IntC a split intein C-fragment and X sulfur (S) or oxygen (O). The split intein C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3), an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to MVKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 4), or an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVGEPHNFLLKNGLVASN (SEQ ID NO: 389).

[0029] Embodiments of the invention include a contiguous intein that can be used, for example, in traditional semi-synthesis applications such as Expressed Protein ligation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] Fig. 1 shows an alignment and a computer-generated model of the design of the Cfa split intein according to an embodiment of the invention;

[0031] Fig. 2 shows graphs showing the characterization of the Cfa intein according to an embodiment of the invention;

[0032] Fig. 3 shows expression and modification of a mouse monoclonal antibody using the Cfa intein according to an embodiment of the invention;

- [0033] Fig. 4 shows the identification of second shell ‘accelerator’ residues important for rapid protein trans-splicing according to an embodiment of the invention;
- [0034] Fig. 5 shows kinetic analysis of Batch 2 mutations and computer generated models according to an embodiment of the invention;
- [0035] Fig. 6 shows an analysis of Batch 1 mutations and computer generated models according to an embodiment of the invention;
- [0036] Fig. 7A and Fig. 7B show an alignment and refinement of the DnaE intein family according to an embodiment of the invention;
- [0037] Fig. 8 is an image of an SDS-PAGE analysis of test expression of His₆-SUMO-Npu^N and His₆-SUMO-Cfa^N according to an embodiment of the invention;
- [0038] Fig. 9 shows a schematic and graph showing increased promiscuity of Cfa^{GEP} according to an embodiment of the invention;
- [0039] Fig. 10 shows graphs and schematics showing cyclization of eGFP in *E. coli* with variable residues according to an embodiment of the invention; and
- [0040] Fig. 11 shows a table illustrating several complexes and compounds according to an embodiment of the invention.

DETAILED DESCRIPTION

[0041] Embodiments of the invention are discussed in detail below. In describing embodiments, specific terminology is employed for the sake of clarity. However, the invention is not intended to be limited to the specific terminology so selected. A person skilled in the relevant art will recognize that other equivalent parts can be employed and other methods developed without parting from the spirit and scope of the invention. All references cited herein are hereby incorporated by reference in their entirety as if each had been individually incorporated.

[0042] Embodiments of the invention include a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMITTDGQMLPIDEIFERGL (SEQ ID NO: 1).

[0043] Embodiments of the invention include a split intein N-fragment comprising an amino acid sequence, wherein said amino acid sequence comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2).

[0044] Embodiments of the invention include a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3).

[0045] Embodiments of the invention include a split intein C-fragment comprising an amino acid sequence, wherein said amino acid sequence of said C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4).

[0046] Embodiments of the invention include a composition comprising the following: a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1); and a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3).

[0047] Embodiments of the invention include a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1).

[0048] Embodiments of the invention include a nucleotide plasmid comprising a nucleotide sequence encoding for a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3).

[0049] Embodiments of the invention include a method for splicing two complexes comprising: contacting a first complex comprising a first compound and a split intein N-fragment and a second complex comprising a second compound and a split intein C-fragment, wherein contacting is performed under conditions that permit binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and

reacting the intein intermediate to form a conjugate of the first compound with the second compound, wherein said split intein N-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1), and wherein said split intein C-fragment comprises an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3). In some embodiments, reacting the intein intermediate comprises contacting the intein intermediate with a nucleophile. In some embodiments, said first compound is a polypeptide. In some embodiments, said first compound is an antibody.

[0050] Embodiments of the invention include an intein comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLPVKIISRKSLGTQNVYDIGVEKDH
NFLLKNGLVASN (SEQ ID NO: 390).

[0051] Embodiments of the invention include a kit for splicing two complexes together comprising the following: a split intein N-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to

CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED
GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1); a split intein C-fragment comprising an amino acid sequence of at least 80%, 85%, 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3); reagents for permitting the binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and a nucleophilic agent.

[0052] Embodiments of the invention include a method for generating a synthetic consensus intein peptide sequence comprising: generating a population of a plurality of homologous intein peptide sequences; identifying amino acids associated with fast splicing within said population of a plurality of homologous intein peptide sequences; generating a subpopulation of a second plurality of homologous intein peptide sequences, wherein said second plurality of homologous intein peptide sequences comprise amino acids associated with fast splicing; creating an alignment of at least three peptide sequences of said subpopulation; determining a most frequently occurring amino acid residue at each position of said at least three peptide sequences; and generating a synthetic consensus intein

peptide sequence based on said most frequently occurring amino acid residue at each position of said at least three peptide sequences.

[0053] Embodiments of the invention include a method comprising: fusing a first nucleotide sequence encoding an amino acid sequence of a first intein fragment comprising CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment comprising VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3), so that the fusion of the first nucleotide sequence and second nucleotide sequence codes for a contiguous intein.

[0054] Embodiments of the invention include a gene fusion comprising a first nucleotide sequence encoding an amino acid sequence of a first intein fragment comprising CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCLED GSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1) fused with a second nucleotide sequence encoding an amino acid sequence of a second intein fragment comprising VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3).

[0055] Embodiments of the invention include a contiguous intein that can be used, for example, in traditional semi-synthesis applications such as Expressed Protein ligation.

[0056] In some embodiments, the various intein fragments described are linked, fused, chemically bonded, complexed or coupled by conventional methods known in the art to polymers, peptides, polypeptides, oligopeptides, small molecules, nucleotides, polynucleotides, oligonucleotides, drugs, cytotoxic molecules or combinations thereof.

[0057] *Example 1*

[0058] In some embodiments, the basis of rapid protein splicing through a comparative study of the first two characterized split inteins, Npu and Ssp was investigated. The substantial difference in splicing rate between these two proteins is especially puzzling given their highly similar sequences (63% identity) and near superimposable active site structures. Previous mutagenesis studies on Npu and Ssp suggest that the difference in activity between the two is likely due to the combined effects of several residues, rather than a single site.^{6,8} However, it remains unclear just how many residues are responsible for the fast versus slow reaction rates and by extension, whether these ‘accelerator’ residues contribute equally to the individual chemical steps in the overall protein splicing process.

Consequently, we began our study by exploring these questions, in the hope that this would provide a starting point for developing an improved PTS system.

[0059] The high level of conservation within the active sites of Npu and Ssp suggests that more distal amino acid differences account for the disparity in splicing rate between the two. Thus, attention was focused on ‘second shell’ residues, those directly adjacent to the active site. To simplify this analysis, a batch mutagenesis strategy was employed in conjunction with a previously reported *in vitro* PTS assay.⁵ This assay uses split intein constructs with short native extein sequences and allows the rates of branched intermediate formation (k_1, k_2) and its resolution to final splice products (k_3) to be determined using a three state kinetic model.

[0060] The known cross-reactivity of Npu and Ssp intein fragments served as a convenient platform on which to assess which half of the split intein contributes most significantly to the difference in activity.³ Both the Ssp^N-Npu^C (chimera 1) and Npu^N-Ssp^C (chimera 2) chimeras show a decrease in the rates of branch formation and resolution compared to that of native Npu (Fig. 4C, 4D). This indicates that residues on both the N- and C-intein fragments of Npu and Ssp contribute to the difference in their splicing rate. Next, four groups of second shell positions on each of these chimeras were chosen based on their proximity to the active site, and the corresponding Ssp residues were mutated to those in Npu (Fig. 4A and 4B). From the chimera 1 mutants, Batch 2 (L56F, S70K, A83P, E85D) completely restored branch formation activity to that of native Npu (Fig. 4C), while Batch 1 (R73K, L75M, Y79G, L81M) restored the majority of branch resolution activity (Fig. 4D). The effects of mutations on the chimera 2 background were more prosaic, with no single batch able to restore splicing activity to that of native Npu (Fig. 4C and 4D). Lastly, the A136S mutation on Ssp^C has previously been shown to accelerate protein splicing and was examined separately.⁸ This A136S mutation increases the rate of branch resolution two fold, but has no impact on branch formation (Fig. 4C and 4D).

[0061] Fig. 4 shows the identification of second shell ‘accelerator’ residues important for rapid protein trans-splicing according to an embodiment of the invention. In Panels A and B, design of second shell batch mutants on chimera 1 (Ssp^N-Npu^C) and chimera 2 (Npu^N-Ssp^C) is shown. In each case, the location of the mutants (rendered as sticks) is shown using the crystal structure of Npu (pdb = 4kl5). Catalytic residues are shown in black (rendered as sticks). Panel C shows forward (k_1 , blue) and reverse (k_2 , red) rates of branched intermediate formation from starting materials for the various constructs described in this study (error = SD (n=3)). Panel D shows the rate of branch resolution (k_3) of the various constructs (error = SD (n=3)).

[0062] Next the individual contributions of residues within batch mutants 1 and 2 was investigated, since these had the most profound effect on splicing activity. For Batch 2, further mutagenesis shows that the interaction between F56, K70, and D85 is likely responsible for the increased rate of branch formation in Npu^N (Fig. 5A). Structural evidence supports this data, as K70 is a part of the highly conserved TXXH block B loop in Npu^N (residues 69-72) that catalyzes the initial N-to-S acyl shift in protein splicing.⁹ Thus, the position and dynamics of K70 (packed against F56 and D85) should directly impact the catalytic residues T69 and H72 (Fig. 5B).¹⁰⁻¹² From Batch 1, K73, M75, and M81 are responsible for the faster rate of branch resolution in Npu^N (Fig. 6A). These residues pack around the terminal asparagine of the C-intein, which must undergo succinimide formation in the final step of protein splicing (Fig. 6B). Taken together, the mutagenesis data points to the key role that second shell ‘accelerator’ residues play in tuning the activity of split inteins.

[0063] Fig. 5 shows kinetic analysis of Batch 2 mutations and computer generated models according to an embodiment of the invention. Panel A shows the equilibrium rates of branch formation (k_1 , k_2) and rates of branch resolution (k_3) for the single (A83P), double (A83P, S70K), and triple (L56F, S70K, A83P) point mutants of Ssp^N that comprise Batch 2 (L56F, S70K, A83P, E85D) (error = SD (n=3)). Panel B shows a zoom view of Batch 2 (green sticks next to labels F56, K70, P83, and D85) in the Npu active site (pdb = 4kl5). Catalytic residues are rendered as black sticks.

[0064] Fig. 6 shows an analysis of Batch 1 mutations and computer generated models according to an embodiment of the invention. Panel A shows the equilibrium rates of branch formation (k_1 , k_2) and rates of branch resolution (k_3) for the single (R73K), double (R73K, Y79G), and triple (R73K, Y79G, L81M) point mutants comprising Batch 1 (error = SD (n=3)). Panel B shows a zoom view of Batch 1 (red sticks next to labels K73, M75, G79, and M81) in the Npu structure (pdb = 4kl5). Catalytic residues are rendered as black sticks.

[0065] The ‘accelerator’ residues found to affect the splicing rate allow for an activity-guided approach to engineer a consensus DnaE intein. Consensus protein engineering is a tool applied to a homologous set of proteins in order to create a thermostable variant derived from the parent family.^{13,14} A multiple sequence alignment (MSA) is first generated from homologues of a particular protein, from which the most statistically frequent residue at each position is chosen as the representative in the consensus sequence. For the DnaE inteins, 105 sequences were identified through a BLAST¹⁵ search of the JGI¹⁶ and NCBI¹⁷ databases (Fig. 7A). Next, the alignment was filtered to only contain sequences bearing the second shell indicators of fast splicing: K70, M75, M81, and S136. The 73 theoretically fast inteins left in the MSA (Fig. 7B) were then used to generate a consensus fast DnaE

intein sequence (Cfa) (Fig. 1). The various sequences disclosed in Figs. 7A and 7B are presented below:

[0066] >NpuPCC73102/1-137

[0067] CLSYETEILTVEYGLLPKIGKIVEKRIECTVYSVDNNGNIYTQPVAQWHDRGEQE
VFEYCLEDGSLIRATKDHKFMFTVDGQMLPIDEIFERELDLMRVDNLPN (SEQ ID NO: 5)

[0068] IKIATRKYLGKQNVYDIGVERDHNFAKNGFIASN (SEQ ID NO: 6)

[0069] >CthPCC7203:/1-137 *Chroococcidiopsis thermalis* PCC 7203

[0070] CLSYDTEILTVEYGAIPKIGKIVEERIECTVYSVDNNGFIYTQPIAQWHNRGQGEV
FEYCLEDGSIIRATKDHKFMFTFEGKMLPIDEIFEQELDLKQVKSIQN (SEQ ID NO: 7)

[0071] VKIISRKSLGIQPVYDIGVERDHKFVLKNGLVASN (SEQ ID NO: 8)

[0072] >NspCCY9414:/1-137 *Nodularia spumigena* CCY9414 genome

[0073] CLSYDTEILTVEYGYIPKIGKIVEKAIECSVYSVDNNGNVYTQPIAQWHNRGEQE
VFEYSLEDGSTIRATKDHKFMFTTDGQMLPIDEIFAQELDLLQVHGLPK (SEQ ID NO: 9)

[0074] VKITARKFVGRENVYDIGVERYHNFAIKNGLIASN (SEQ ID NO: 10)

[0075] >AcyPCC7122:/1-137 *Anabaena cylindrica* PCC 7122

[0076] CLSYDTEVLTVEYGFIPKIGKIVEKRIECSIFSVDKNGNVYTQPIAQWHNRGRQEI
YEYCLDDGSKIRATKDHKFMFTTAGEMLPIDEIFERDLDLLKVEGLPE (SEQ ID NO: 11)

[0077] VKIISRQYLQADVYDIGVEEDHNFAIKNGFIASN (SEQ ID NO: 12)

[0078] >CspPCC7507:/1-137 *Calothrix* sp. PCC 7507, complete genome

[0079] CLSYDTEVLTVEYGLLPKIGKIVEKGIECRVFSVDNHGNVYTQPIAQWHNRGQGE
VFEYGLDDGSVIRATKDHKFMFTTDGKMLPIDEIFERGLDLLQVQGLPE (SEQ ID NO: 13)

[0080] VKVITRKYIGKENVYDIGVELDHNFAIRNGLVASN (SEQ ID NO: 14)

[0081] >NspPCC7524:/1-137 *Nostoc* sp. PCC 7524

[0082] CLSYDTEILTVEYGFIPKIGKIVEKGIECTVFSVASNGIVYTQPIAQWHNRGQGEIF
EYCLEDGSIIRATKDHKFMFTQDGQMLPIDEIFACELDLLQVQGLPE (SEQ ID NO: 15)

[0083] VKVVTRKYIGKENVYDIGVERDHNFAIRNGLVASN (SEQ ID NO: 16)

- [0084] >Naz0708:/1-137 'Nostoc azollae' 0708
[0085] CLSYKTEVLTVEYGLPIGEIVEKRIECSLFSVDENGNIYTQPIAQWHHRGVQEV
YEYCLDDGTIIRATKDHKFM TTIGEMLPIDEIFERDLNLLQVNGLPT (SEQ ID NO: 17)
[0086] VKIISRQFLGPANVYDIGVAQDHNFAIKNGLIASN (SEQ ID NO: 18)
- [0087] >NspPCC7120:/1-137 Nostoc sp. PCC 7120 DNA
[0088] CLSYDTEVLTVEYGFVPIGEIVEKGIECSVFSINNNGIVYTQPIAQWHHRGKQEV
FEYCLEDGSIKATKDHKFM TQDGKMLPIDEIFEQELDLLQVKGLPE (SEQ ID NO: 19)
[0089] IKIASRKFLGVENVYDIGVRRDHNFFIKNGLIASN (SEQ ID NO: 20)
- [0090] >AvaATCC29413/1-137 Anabaena variabilis ATCC 29413
[0091] CLSYDTEVLTVEYGFVPIGEIVDKGIECSVFSIDSNGIVYTQPIAQWHHRGKQEV
FEYCLEDGSIKATKDHKFM TQDGKMLPIDEIFEQELDLLQVKGLPE (SEQ ID NO: 21)
[0092] IKIASRKFLGVENVYDIGVGRDHNFFVKNGLIASN (SEQ ID NO: 22)
- [0093] >PspPCC7327:/1-135 Pleurocapsa sp. PCC 7327.
[0094] CLSYDTKILTVEYGAMPIGKIVEEQIDCTVYTVNQNGFVYTQPIAQWHDRGKQ
EIFEYCLEDGSIIRATKDHKFM TTDGQMLPIDKIFEKGLDLKTINCD (SEQ ID NO: 23)
[0095] VKILSRKSLGIQSVYDIGVEKDHNFLLANGLVASN (SEQ ID NO: 24)
- [0096] >CspPCC7424:/1-135 Cyanothece sp. PCC 7424
[0097] CLSYETQIMTVEYGLMPIGKIVEEQIDCTVYTVNKNGFVYTQPIAQWHYRGEQ
EVFEYCLEDGSTIRATKDHKFM TTDGQMLPIDEIFEQGLELKQIHLS (SEQ ID NO: 25)
[0098] VKIISRQSLGIQPVYDIGVEKDHNFLISDGLIASN (SEQ ID NO: 26)
- [0099] >CspPCC7822:/1-134 Cyanothece sp. PCC 7822
[00100] CLSYDTEILTVEYGPMPIGKIVEEQIECTVYTVDKNGLVYTQPIAQWHHRGQQE
VFEYCLEDGSIIRATKDHKFM TDDGQMLPIEEIFEKGLELKQIIL (SEQ ID NO: 27)
[00101] VKIISRQLAGNQTVYDLGVEKDHNFLLANGLIASN (SEQ ID NO: 28)
- [00102] >NspPCC7107:/1-137 Nostoc sp. PCC 7107

- [00103] CLSYDTQVLTVEYGLVPIGEIVEKQLECSVFTIDGHGYVYTQAIQWHNRGQQ
EVFEYGLLEDGSVIRATKDHKFMTTDGQMLPIDEIFERELDLLQVQGLRW (SEQ ID NO: 29)
- [00104] VKIITRKYIGQANVYDIGVAQDHNFVIENRLIASN (SEQ ID NO: 30)
- [00105] >TboIicb1/1-136 Tolypothrix bouteillei Iicb1
- [00106] CLSYDTEILTVEYGFLPIGKIVEKGIECNVYSVDKNGNIYTQPIAQWHDRGEQE
VFEYCLENGSVIRATKDHKFMTTSGEMLPIDEIFERGLDLIRVEDLP (SEQ ID NO: 31)
- [00107] VKILTRKSIGKQTVYDIGVERDHNFVIKNGSVASN (SEQ ID NO: 32)
- [00108] >Aov:/1-136 Aphanizomenon ovalisporum DnaE precursor (dnaE) gene
- [00109] CLSADTEILTVEYGFLPIGEIVGKAIECRVYSVDGNGNIYTQSIAQWHNRGEQEV
FEYTLLEDGSIIRATKDHKFMTTDGEMLPIDEXFARQLDLMQVQGLH (SEQ ID NO: 33)
- [00110] VKITARKFVGRENVYDIGVEHHHNFAIKNGLIASN (SEQ ID NO: 34)
- [00111] >OnvPCC7112:/1-137 Oscillatoria nigro-viridis PCC 7112
- [00112] CLSYDTKILTVEYGPMAIGKIVEEKIECTVYSVDSNGYIYTQSIAQWHRRGQQE
VFEYCLEDGSIIRATKDHKFMVGGQMLPIDEIFEQGLDLKQINSSSD (SEQ ID NO: 35)
- [00113] VKIISRKSLGTQEVYDIGVEREHNFILENSLVASN (SEQ ID NO: 36)
- [00114] >RspPCC7116:/1-135 Rivularia sp. PCC 7116, complete genome
- [00115] CLSYDTEVLTEEFGLPIGKIVEEKIDCTVYSVDVNGNVYSQPIAQWHNRGMQE
VFEYELEDGSTIRATKDHKFMVTDGEMLAIDEIFEKGLELKRVGIIY (SEQ ID NO: 37)
- [00116] VKIISRKVLKTENVYDIGLEGDHNFIKDGLIASN (SEQ ID NO: 38)
- [00117] >TerIMS101:/1-137 Trichodesmium erythraeum IMS101
- [00118] CLTYETEIMTVEYGPLPIGKIVEYRIECTVYTVDKNGYIYTQPIAQWHNRGMQE
VYEYSLEDGTVIRATPEHKFMTEDGQMLPIDEIFERNLDLKCLGTLEL (SEQ ID NO: 39)
- [00119] VKIVSRKLAKTENVYDIGVTKDHNFLANGLIASN (SEQ ID NO: 40)
- [00120] >MspPCC7113:/1-137 Microcoleus sp. PCC 7113,
- [00121] CLSYDSEILTVEYGLMPIGKIVEEGIECTVYSVDSHGYLYTQPIAQWHHRGQQE
VFEYDLEDGSVIRATKDHKFMVTDGEMLAIDEIFEKGLELKVQRSQP (SEQ ID NO: 41)

[00122] VKIVRRKSLGIQTVYDIGVERDHNFLLANGLVASN (SEQ ID NO: 42)

[00123] >ScyPCC7437:/1-137 Stanieria cyanosphaera PCC 7437

[00124] CLSYDTEILTVEYGAMPIGKIVKEQIECNVYTVNQNGFIYPQAIAQWHERGKQE
IFEYTLDNGLVIRATKDHKFMITDGQMLPIDEIFERGLELQRINDYSN (SEQ ID NO: 43)

[00125] VKIVSRKSLGKQPVYDIGVTKDHNFLLSNGVVASN (SEQ ID NO: 44)

[00126] >CspPCC6303:/1-137 Calothrix sp. PCC 6303

[00127] CLSYDTEILTWEYGFLKIGEIVEKQILCSVFSVDEQGNVYTQPIAQWHNRGLQE
LFAYQLEDGGVIRATKDHKFMITTDGQMLAIDEIFERQLDLFQVKGLPE (SEQ ID NO: 45)

[00128] VKIISRKVLKTENVYDIGLEGDHNFIKDGLIASN (SEQ ID NO: 46)

[00129] >Cst/1-134 PCC7202: Cyanobacterium stanieri PCC 7202

[00130] CLSYDTEVLTVYEGVLPYGKIVEEQIQCTVYSVDQYGFVYTQAIQWHDGRGEQ
EVFEYELENGATIKATKDHKMMTSDGQMLPIDQIFEQGLDLFMVSF (SEQ ID NO: 47)

[00131] VKIVKRRSHGIQKVYDIGVAKDHNFLHNGLVASN (SEQ ID NO: 48)

[00132] >CspATCC51142:/1-134 Cyanothece sp. ATCC 51142

[00133] CLSYDTEILTVEYGPMPYGKIVEENINCTVYTVDPNGFVYTQAIQWHYRGEQE
IFEYYLEDGATIRATKDHKFMTEGKMLPIDEIFENNLDLKQLTL (SEQ ID NO: 49)

[00134] VKIIGRQSLGVQKVYDIGVEKEHNFLHNGLIASN (SEQ ID NO: 50)

[00135] >CspPCC8801:/1-134 Cyanothece sp. PCC 8801

[00136] CLSYDTEILTVEYGAIPYGKIVEENIDCTVYTVDKNGFVYTQNIQWHLRGQQE
VFEYYLDDGSILRATKDHQFMTLEGEMLPIDEIFERGLELKKIKI (SEQ ID NO: 51)

[00137] VKIVSYRSLGKQFVYDIGVAQDHNFLLANGSIASN (SEQ ID NO: 52)

[00138] >Asp/1-136 Anabaena sp. 90 chromosome

[00139] CLSYDTEILTVEYGFLEIGEIVEKQIECKVYTIDSNGMLYTQSIAQWHNRGQQE
VYEYLLENGAIIRATKDHKFMTEAGQMLPIDEIFAQGLDLLQVGVAE (SEQ ID NO: 53)

[00140] VKIVSRTYVGQANVYDIGVESDHNFVIKNGFIASN (SEQ ID NO: 54)

- [00141] >Aha:/1-137 *Aphanothece halophytica*
- [00142] CLSYDTEIWTVEYGAMPIGKIVEEKIECSVYTVDENG FVYTQPIAQWHPRGQQE
IIEYTLLEDGRKIRATKDHKMMTESGEMPLPIEEIFQRELDLKVETFHEM (SEQ ID NO: 55)
- [00143] VKIIKRQSLGRQNVYDVCVETDHN FVLANGCVASN (SEQ ID NO: 56)
- [00144] >HspPCC7418:/1-137 *Halothece* sp. PCC 7418
- [00145] CLSYDTEIWTVEYGAMPIGKIVEEKIECSVYTVDENG FVYTQPIAQWHPRGQQE
IIEYTLLEDGRKIRATKDHKMMTESGEMPLPIEEIFQRELDLKVETFHEM (SEQ ID NO: 57)
- [00146] VKIIKRQSLGRQNVYDIGVETDHN FVLANGCVASN (SEQ ID NO:58)
- [00147] >CapPCC10605:/1-137 *Cyanobacterium aponinum* PCC 10605
- [00148] CLSYDTEILTVEYGAISIGKIVEEKINCQVYSVDKNGFIYTQNIAQWHDRGSQEL
FEYELEDGRRIKATKDHKMMTKDGQMLAINDIFEQELELYSVDDMGV (SEQ ID NO:59)
- [00149] VKIVKRRSLGVQP VYDIGVEKDHN FILANGLVASN (SEQ ID NO:60)
- [00150] >Cat:/1-133 *Candidatus Atelocyanobacterium thalassa* isolate
- [00151] CLSYDTKVLTV EYGPLPIGKV VQENIRCRVYTTNDQGLIYTQPIAQWHNRGKQ
EIFEYHLDDKTIIRATKEHQFMTVDHVMMPIDEIFEQGLELKKIK (SEQ ID NO:61)
- [00152] LKIIRKSLGMHEVFDIGLEKDHN FVLSNGLIASN (SEQ ID NO:62)
- [00153] >Oli:/1-137 *Oscillatoria limnetica* 'Solar Lake' DnaE precursor
- [00154] CLSYNTEVLTVEYGPLPIGKIVDEQIHCRVYSVDENG FVYTQAI AQWHDRGYQ
EIFAYELADGSVIRATKDHQFMTEDGQMFPIDEIWEKGLDLKKLP TVQD (SEQ ID NO:63)
- [00155] VKIVRRQSLGVQN VYDIGVEKDHN FLLASGEIASN (SEQ ID NO:64)
- [00156] >Cen:/1-137 *Cyanobacterium endosymbiont of Epithemia turgida*
- [00157] CLSYDTEVLTVEYGAIPGRMVEESLDCTVYTVDKNGFVYTQSIQQWHSRGQQ
EIFEYCFEDGSIIRATKDHKFM TAEGKMSSIH DIFEQGLELKKIIPWSG (SEQ ID NO:65)
- [00158] AKIISCKSLGKQSVYDIGVVQDHN FLLANGVVASN (SEQ ID NO:66)
- [00159] >SspPCC7502:/1-133 *Synechococcus* sp. PCC 7502

- [00160] CLGYDTPVLTVEYGFMPIGKIVEEKIQCHVYSVDQNGLVFTQAIAQWHNRGQQ
EVWEYNLDNGDIVRATKDHKFMTIDGQMLPINQIFEQGLELKVIA (SEQ ID NO:67)
- [00161] VKIVSCKPLRVQTVYDIGVEKDHNFIELDNGLVASN (SEQ ID NO:68)
- [00162] >DsaPCC8305:/1-134 *Dactylococcopsis salina* PCC 8305
- [00163] CLSYDTEVLTEEYGAIPIGKIVEERMNCHVYSVDENGFIYSQPIAQWHPRGEQE
VVEYTLLEDGKIIRATADHKMMTETGEMPLIEQIFQQQLDLKISNQ (SEQ ID NO:69)
- [00164] VKIINRQSLGKQTVYDIGVEKDHNFILGNGLVASN (SEQ ID NO:70)
- [00165] >CstPCC7417:/1-137 *Cylindrospermum stagnale* PCC 7417
- [00166] CLSYDTEILTVEYGFPIGEIVEKRIECSVYSVDNHGNVYTQPIAQWHNRGLQEV
FEYCLEDGSTIRATKDHKFMTTDKEMPLIDEIFERGLDLLRVEGLPI (SEQ ID NO:71)
- [00167] VKIIMRSYVGRENVYDIGVERDHNFVAKNGLIAAN (SEQ ID NO:72)
- [00168] >SspPCC6803:/1-137 *Synechocystis* sp. PCC 6803
- [00169] CLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGEQEV
LEYELEDGSVIRATSDHRFLT TDYQLLAIEEIFARQLDLLTLENIKQ (SEQ ID NO:73)
- [00170] VKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIAAN (SEQ ID NO:74)
- [00171] >GspPCC7407:/1-137 *Geitlerinema* sp. PCC 7407
- [00172] CLSYETPVMTVEYGPLPIGRIVEEQLDCTVYSVDEQGHVYTQPVAQWHHRGL
QEVVEYELEDGRRLRATADHRFMTETGEMPLAEIFERGLELRQVALRVP (SEQ ID NO:75)
- [00173] VKIVSRRSLGMQLVYDIGVAADHNFVLADGLIAAN (SEQ ID NO:76)
- [00174] >SspPCC6714:/1-137 *Synechocystis* sp. PCC 6714
- [00175] CLSFDAEILTVEYGPLSIGKIVGEEINCSVYSVDPQGRIYTQAIAQWHDRGVQEV
FEYELEDGSVIRATPDHRFLT TDYELLAIEEIFARQMDLLTLTNLKL (SEQ ID NO:77)
- [00176] VKVVRRLSLGMHRVFDIGLAQDHNFLLANGAIAAN (SEQ ID NO:78)
- [00177] >MaePCC7806:/1-135 *Microcystis aeruginosa* PCC 7806
- [00178] CLGGETLILTEEYGLLPIAKIVSEEVNCTVYSVDKNGFVYSQPISQWHERGLQE
VFEYTLENGQTIQATKDHKFMTNDGEMPLAIDTIFERGLDLKSSDFS (SEQ ID NO: 79)

- [00179] VKIISRQSLGRKPVYDIGVEKDHNFLLGNGLIASN (SEQ ID NO:80)
- [00180] >MaeNIES843:/1-135 *Microcystis aeruginosa* NIES-843 DNA
- [00181] CLGGETLILTEEYGLLPIAKIVSEEINCTVYTVDQNGFVYSQPISQWHERGLQEV
FEYTLENGQTIQATKDHKFMTSDGEMLAIDTIFERGLDLKSSDFS (SEQ ID NO:81)
- [00182] VKIIGRQSLGRKPVYDIGVEKDHNFLLGNGLIASN (SEQ ID NO:82)
- [00183] >AmaMBIC11017:/1-137 *Acaryochloris marina* MBIC11017,
- [00184] CLSYDTPVLTLEYGWLPIGQVVQEIECQVFSINERGHLYTQPIAQWHHRGQQ
EVFEYTLADGSTIQATAEHQFMTTDGQMYPVQQIFEEGLSLKQLPLPWQ (SEQ ID NO:83)
- [00185] VKIIQRSLGLQSVYDIGLAQDHNFMANGWVAAN (SEQ ID NO:84)
- [00186] >LspPCC7376:/1-137 *Leptolyngbya* sp. PCC 7376
- [00187] CLDGETPIVTVEYGVLPPIREIVEKELLCSVYSIDENG FVYTQPV EQWHQRGDRQ
MFEYQLDNGGVIRATPDHKFLTTEGEMVAIDEIFEKGLNLA EFAPADL (SEQ ID NO:85)
- [00188] VKILRRHSIGKAKTYDIGVSKNHNFLLANGL FASN (SEQ ID NO:86)
- [00189] >SelPCC6301:/1-137 *Synechococcus elongatus* PCC 6301
- [00190] CLAADTEVLTVEYGP IAGKLVEENIRCQVYCCNPDGYIYSQPIGQWHQRGEQE
VIEYELSDGRIIRATADHRFMTEEGEMLSLDEIFERSLELKQIPTPLL (SEQ ID NO:87)
- [00191] VKIVRRRSLGVQP VYDLGVATVHNFVLANGLVASN (SEQ ID NO:88)
- [00192] >SspPCC6312:/1-137 *Synechococcus* sp. PCC 6312
- [00193] CLSADTELYTVEYGWLPIGRLVEEQIECQVLSVNAHGHVYSQPIAQWHRRAW
QEVFEYQLETGGTIKATTDHQFLT TDGQMYRIEDIFQRGLDLWQLPPDRF (SEQ ID NO:89)
- [00194] VKIISRCSLGIQP VYDIGVAQDHN FVIRGGLVASN (SEQ ID NO:90)
- [00195] >Tel:/1-137 *Thermosynechococcus elongatus* BP-1 DNA
- [00196] CLSGETAVMTVEYGA VPIRRLVQERLSCHVYSLDGQGHL YTQPIAQWHFQGFR
PVYEQLEDGSTICATPDHRFM TTRGQMLPIEQIFQEGLELWQVAIAPR (SEQ ID NO:91)
- [00197] GKIVGRRLMGWQAVYDIGLAADHN FVLANGAIAAN (SEQ ID NO:92)

[00198] >Tsp:/1-137 Thermosynechococcus sp. NK55 genome

[00199] CLSGETAVMTVEYGAVPIRRLVQERLTCHVYSLDAQGHLYTQPIAQWHFQGF
RPVYEQLEDGSTIWATPDHRFMTTRGQMLPIEQIFQEGLELWQGPIAPS (SEQ ID NO:93)

[00200] CKIVGRQLVGWQAVYDIGVARDHNFLLANGAIAAN (SEQ ID NO:94)

[00201] >Tvu:/1-137 Thermosynechococcus vulcanus DnaE precursor

[00202] CLSGETAVMTVEYGAIPIRRLVQERLICQVYSLDPQGHLYTQPIAQWHFQGFRP
VYAYQLEDGSTICATPDHRFMTTSGQMLPIEQIFREGLELWQVAIAPP (SEQ ID NO:95)

[00203] CKIVGRRLVGWQAVYDIGLAGDHNFLLANGAIAAN (SEQ ID NO:96)

[00204] >SspPCC7002:/1-137 Synechococcus sp. PCC 7002

[00205] CLAGGTPVVTVEYGVLPITQIVEQELLCHVYSVDAQGLIYAQLIEQWHQRGDR
LLYEYELENGQMIRATPDHRFLTGTGELLPIDEIFTQNLDAAWAVPDS (SEQ ID NO:97)

[00206] VKIIRRKFIGHAPTYDIGLSQDHNFLLGQGLIAAN (SEQ ID NO:98)

[00207] >ShoPCC7110:/1-136 Scytonema hofmanni PCC 7110 contig00136

[00208] CLSYDTEVLTAEYGFLPIGKIVEKAIECTVYSVDNDGNIYTQPIAQWHDRGQQE
VFEYSLDDGSVIRATKDHKFMTTGGQMLPIDEIFERGLDLMRIDSLP (SEQ ID NO:99)

[00209] VKILTRKSIGKQTVYDIGVERDHNFVIKNGLVASN (SEQ ID NO:100)

[00210] >WinUHHT291/1-136 Westiella intricata UH HT-29-1

[00211] CLSYDTEILTVEYGFLPIGEIVEKRIECTVYTVDTNGYVYTQAIQWHNRGEQE
VFEYALDGSIIIRATKDHKFMTTSEGQMLPIDEIFVKGLDLLQVQGLP (SEQ ID NO: 101)

[00212] VKIITRKFLGIQNVYDIGVEQNHNFVIKNGLVASN (SEQ ID NO:102)

[00213] >FspPCC9605:/1-136 Fischerella sp. PCC 9605 FIS9605DRAFT

[00214] CLSYDTEILTVEYGFLPIGEIVEKGIECTVYTVDNNGNVYTQTIAQWHNRGQQE
VFEYCLEDGSVIRATKDHKFMTTDGQMLPIDEIFARGLDLLQVKNLP (SEQ ID NO:103)

[00215] VKIVTRRPLGTQNVYDIGVESDHNFVIKNGLVASN (SEQ ID NO:104)

[00216] >MrePCC10914:/1-137 Mastigocladopsis repens PCC 10914

[00217] CLSYDTEVLTV EYGFLPIGEIVEKSIECSVYTVDSNGNVYTQPIAQWHNRGQQE
VFEYCLEDGSIIRATKDHKFMTHGQMLPIDEIFERGLELMKIQGLPE (SEQ ID NO:105)

[00218] AKIITRKS LGTQNVYDIGVERDHN FVTRDGF IASN (SEQ ID NO:106)

[00219] >ShoUTEX2349:/1-137 [Scytonema hofmanni] UTEX 2349

[00220] CLSYNSEVLTV EYGFLPIGKIVEK GIECSVYSVDSYGKIYTVIAQWHNRGQQE
VFEYCLEDGTIIQATKDHKFM TVDGQMLPIDEIFERGLDLMQVQGLPD (SEQ ID NO:107)

[00221] VKIITRKS LGTQNVYDIGVSSDHN FVMKNGL IASN (SEQ ID NO:108)

[00222] >AspPCC7108:/1-137 Anabaena sp. PCC 7108 Ana7108scaffold_2_Cont3

[00223] CLSSDTEVLTV EYGLPIGEIIEKRIDCSVFSVDKNGNIYTQPIAQWHDRGIQELY
EYCLDDGSTIRATKDHKFM TTAGEMLPIDEIFERGLDLLKVHNL PQ (SEQ ID NO:109)

[00224] VKIITRNYV GKENVYDIGVERDHN FAIKNGL IASN (SEQ ID NO:110)

[00225] >FspPCC9339:/1-137 Fischerella sp. PCC 9339 PCC9339DRAFT

[00226] CLSYDTEVLTV EYGFLPIGEIVEKRIECTVYTVDHNGYVYTQPIAQWHNRGYQ
EVFEYGLE DGSVIRATKDHKFM TSEGQMLPIDEIFARELDLLQVTGLVN (SEQ ID NO:111)

[00227] VKIVTRRL LGIQNVYDIGVEQNHN FVIKNGL VASN (SEQ ID NO:112)

[00228] >Csp336:/1-137 Calothrix sp. 336/3

[00229] CLSYDTEIFTV EYGFLPIGEIVEKRLECTVLTVDNHGNIYSQPIAQWHHRGQQQI
YEYGLE DGSVIRATKDHKFM TTDGQMLPIDEIFERGLDLLQVTNLDN (SEQ ID NO:113)

[00230] VKVITRKL ADTENVYDIGVENHHN FLIKNGL VASN (SEQ ID NO:114)

[00231] >FthPCC7521:/1-136 Fischerella thermalis PCC 7521

[00232] CLSYETEILT V EYGFLPIGEIVEKRIECSVYTVDNNGYVCTQPIAQWHNRGYQE
VFEYGLE DGSVIRATKDHKFM TIDRQMLPIDEIFARGLDLLQVTGLP (SEQ ID NO:115)

[00233] VKIITRKS LGTQNVYDIGVEQNHN FVIKNGL VASN (SEQ ID NO:116)

[00234] >CyaPCC7702/1-137 cyanobacterium PCC 7702 Chl7702

[00235] CLSYDTEILT V EYGFLSIGEIVEKEIECTVYTVDSNGYIYTQPIAQWHEQGEQEIF
EYSLEDG STIRATKDHKFM TIEGEMLPIDQIFARQLDLMQITGLPQ (SEQ ID NO:117)

- [00236] VKISTKKSLGKQKVYDIGVVRDHNFIKNGFVASN (SEQ ID NO:118)
- [00237] >FspPCC9431:/1-136 *Fischerella* sp. PCC 9431
- [00238] CLSYDTEVLTVEYGFLPIGEIVEKRIECTVYTVDTNGYVYTQAIQWHNRDEQE
VFEYALDGSIIIRATKDHKFMFTSEGQMLPIDEIFAKGLDLLQVQGLP (SEQ ID NO:119)
- [00239] VKIVTRKFLGIQNVYDIGVEQNHNFVIKNGLVASN (SEQ ID NO:120)
- [00240] >FmuPCC7414:/1-137 *Fischerella muscicola* PCC 7414
- [00241] CLSYETEILTVEYGFLPIGEIVEKRIECSVYTVDNNGYVCTQTIAQWHNRGYQE
VFEYGLEDDGSVIRATKDHKFMFTIDRQMLPIDEIFARGDLLQVKGLPE (SEQ ID NO:121)
- [00242] VKIITRQSLGTQNVYDIGVEQNHNFVIKNGLVASN (SEQ ID NO:122)
- [00243] >FmuPCC73103:/1-137 *Fischerella muscicola* SAG 1427-1 = PCC 73103
- [00244] CLSYDTEVLTVEYGFLPIGEIVEKTIECNVFTVDSNGYVYTQPIAQWHNRGYQE
VFEYGLEDDGSVIRATKDHKFMFTSEGKMLPIDEIFARELDLLQVTGLIN (SEQ ID NO:123)
- [00245] VKIVTRKFLGIQNVYDIGVEQNHNFVIKNGLVASN (SEQ ID NO:124)
- [00246] >Lae:/1-137 *Lyngbya aestuarii* BL J laest3.contig.3
- [00247] CLSYDTEILTVEYGAIPIGKVVDKIECTVYSVDKNGLIYTQPIAQWHNRGKQE
VFEYSLEDGSTIRATKDHKFMFTMDNQMLPIDEILEKGLELKQVNADSV (SEQ ID NO:125)
- [00248] VKIVSRKSLDSQTVYDIGVETDHNFLLANGSVASN (SEQ ID NO:126)
- [00249] >MspPCC7126:/1-135 *Microchaete* sp. PCC 7126
- [00250] CLSYKTQVLTVEYGLLAIGEIVEKNIECSVFSVDIHGNVYTQPIAQWHHRGQQE
VFEYGLEDDGSIIIRATKDHKFMFTTQGEMLPIDEIFARGDLLQVKGV (SEQ ID NO:127)
- [00251] VKIITRKYIGKENVYDIGVEQDHNFAIKNGLIAAN (SEQ ID NO:128)
- [00252] >Lsp:/1-137 *Leptolyngbya* sp. JSC-1
- [00253] CLSYDTEILTVEYGALPIGKIVENQMICSVYSIDNNGYIYIQPIAQWHNRGQQEV
FEYILEDGSIIRSTKDHKFMFTKGGEMLPIDEIFERGLELAQVTRLEQ (SEQ ID NO:129)
- [00254] VKIISRRSVGVQSVYDIGVKQDHNFFLRNGLIASN (SEQ ID NO:130)

- [00255] >CwaWH8501:/1-137 *Crocospaera watsonii* WH8501
- [00256] CLSYDTEILTVEYGAMYIGKIVEENINCTVYTVDKNGFVYTQTIAQWHNRGEQ
EIFEYDLEDGSKIKATKDHKFMTIDGEMPLIDEIFEKNLDLKQVVSHPD (SEQ ID NO:131)
- [00257] VKIIGCRSLGTQKVYDIGVEKDHNFLLANGSIASN (SEQ ID NO:132)
- [00258] >CchPCC7420:/1-135 *Coleofasciculus chthonoplastes* PCC 7420 (Mcht)
- [00259] CLSYDTQILTVEYGAVAIGEIVEKQIECTVYSVDENGYVYTQPIAQWHNRGEQE
VFEYLLLEDGATIRATKDHKFMTDEDQMLPIDQIFEQGLELKQVEVL (SEQ ID NO:133)
- [00260] VKIIGRKPLGTQPVYDIGVERDHNFLLFNGSVASN (SEQ ID NO:134)
- [00261] >CspPCC6712/1-133
- [00262] CLSYDTEVLTVEYGAIPIGKIVEEKIACNVYSVDKNGFVYTQPIAQYHDRGIQE
VFEYRLENGSVIRATKDHKMMTADGQMLPIDEIFKQNLDLKQLN (SEQ ID NO:135)
- [00263] VKIISRQSLGKQSVFDIGVAKDHNFLLANGLVASN (SEQ ID NO:136)
- [00264] >AfiNIES81:/1-132 *Aphanizomenon flos-aquae* NIES-81
- [00265] CLSYDTEILTVEYGFLQIGEIVEKQIECKVYTVDSNGILYTQSIAQWHNRGQQEV
YEYLLENGAIIRATKDHKFMTTEEGQMLPIDEIFSQGLDLLQV (SEQ ID NO:137)
- [00266] VKIISRTYVGQANVYDIGVENDHNFVIKNGFIAAN (SEQ ID NO:138)
- [00267] >Rbr:/1-137 *Raphidiopsis brookii* D9 D9_5,
- [00268] CLSYETEVLTLLEYGFLPIGEIVDKQMVCTVFSVNDSGNVYTQPIGQWHDRGVQ
ELYEYCLDDGSTIRATKDHKFMTTQGEMVPIDEIFHQGWELVQVSGTMN (SEQ ID
NO:139)
- [00269] VKIVSRRYLKGADVYDIGVAKDHNFIKNGLVASN (SEQ ID NO:140)
- [00270] >CspCCy0110:/1-134 *Cyanothece* sp. CCY0110 1101676644604
- [00271] CLSYDTEILTVEYGPMPIGKIVEENINCSVYTVNKNNGFVYTQSIAQWHHRGEQE
VFEYYLEDDGETIRATKDHKFMTTEGKMLPIDEIFENNLDLKKLTV (SEQ ID NO:141)
- [00272] VKIIEERSLKGQNVYDIGVEKDHNFLLSNNLIASN (SEQ ID NO:142)
- [00273] >XspPCC7305:/1-135 *Xenococcus* sp. PCC 7305

[00274] CLSADTEVLTV EYGAISIGKIVEERIECTVYSVDANGFVYTQEIAQWHNRGEQE
VFEYMLDDG SVIRATKDHKLMTIDGQMV AIDEIFSQGLELKQVLGL (SEQ ID NO:143)

[00275] VKIVSRKSLGTQTVYDLGVARDHNFLLANGTVASN (SEQ ID NO:144)

[00276] >PspPCC7319:/1-135 *Pleurocapsa* sp. PCC 7319

[00277] CLSYDTEIYTV EYGALPIGKIVESRIKCTVLTVDKNGLVYSQPIVQWHDRGIQEV
FEYTLDN GATIRATKDHKFMTVEGQMLPIDEIFELGLELKEIQQF (SEQ ID NO:145)

[00278] VKIISRQSLGKQSVYDIGVAKDHNFLLANGMVASN (SEQ ID NO:146)

[00279] >CraCS505:/1-137 *Cylindrospermopsis raciborskii* CS-505

[00280] CLSYETEVLTV EYGFVPIGEIVNKQMVCTVFSLNDSGNVYTQPIGQWHDRGVQ
DLYEYCLDD GSTIRATKDHKFMTTQGEMVPIDEIFHQGWELVQVSGISK (SEQ ID NO:147)

[00281] VKIVSRRYL GKADVYDIGVAKDHNFIKNGLVASN (SEQ ID NO:148)

[00282] >SmaPCC6313/1-129 *Spirulina major* PCC 6313

[00283] CLTYDTLVLTV EYGPVPIGKLVEAQINCQVYSVDANGFIYTQAIAQWHDRGQR
QVYEYTL EDGSTIRATPDHKFMTATGEMLPIDQIFEQGLDL (SEQ ID NO:149)

[00284] VKIIHRRALPPQSVYDIGVERDHNFLPSGWVASN (SEQ ID NO:150)

[00285] >SsuPCC9445:/1-131 *Spirulina subsalsa* PCC 9445

[00286] CLSYDTKIITVEYGAIAIGTIVEQGLHCHVYSVDPNGFIYTQPIAQWHQRGEQEV
FAYTLENGSIIQATKDHKFMTQQGKMLPIDTIFEQGLDLLQ (SEQ ID NO:151)

[00287] VKIIKRTSLGVRPVYDIGVIQDHNFLLENGLVASN (SEQ ID NO:152)

[00288] >MaePCC9807:/1-135 *Microcystis aeruginosa* 9807

[00289] CLGGETLILTEEYGLLP IAKIVSEEINCTVYSVDKNNGFIYSQPISQWHERGLQEVF
EYTLENGQTIQATKDHKFMTSDGEMLAIDTIFERGLDLKSSDFS (SEQ ID NO:153)

[00290] VKIISRQFLGRKPVYDIGVEKDHNFLNGLIASN (SEQ ID NO:154)

[00291] >MspGI1:/1-130 *Myxosarcina* sp. GI1 contig_13

[00292] CLSYDTEVLTLKYGALPIGEIVEKRINCHVYTRAESGFFYIQSIEQWHDRGEQEV
FEYTLENGATIKATKDHKFMTSGGQMLPIDEIFERGLDLL (SEQ ID NO:155)

[00293] VKIVSRKSLGKQPVYDLGVAKDHNFLLANGTVASN (SEQ ID NO:156)

[00294] >LspPCC6406:/1-136 Leptolyngbya sp. PCC 6406

[00295] CLSADTQLLTVEYGPLEIGRIVEEQIACHVYSVDANGFVYTQPIAQWHSRGEQE
IFEYQLEDGRTLRTADHKFMTTGTGEMGRINDIFEQGLDLKQIDLPO (SEQ ID NO:157)

[00296] VKVVSQRSLGVQPVYDIGVATDHNFLADGLVASN (SEQ ID NO:158)

[00297] >AspCCMEE5410:/1-132 Acaryochloris sp. CCMEE 5410

[00298] CLSYDTPVLTLEYGWLPIGQVVQEIECQVFSINERGHLYTQPIAQWHHRGQQ
EVFEYTLTDGSTIQATAEHQFMTTDGQMYPIQQIFEEGLSLKQL (SEQ ID NO:159)

[00299] VKITQRRSLGLQSVYDIGLAQDHNFVIANGWVAAN (SEQ ID NO:160)

[00300] >GhePCC6308:/1-133 Geminocystis herdmanii PCC 6308

[00301] CLSYDTEVLTVEFGAIPMGKIVEERLNCQVYSVDKNGFIYTQNIQWHDGRGVQ
EVFEYELEDGRIIKATKDHKMMIENCCEMVEIDRIFEEGLELFEVN (SEQ ID NO:161)

[00302] VKILKRRSISSQQVYDIGVEKDHNFLLANGLVAAN (SEQ ID NO:162)

[00303] >NnoPCC7104:/1-133 Nodosilinea nodulosa PCC 7104

[00304] CLSADTELLTLEYGPLTIGEIVAKRIPCHVFSVDESGYVYTQPVAAQWHQRGHQE
VFEYQLDDGTTIRATADHQFMTTELGEMMAIDEIFQRGLELQVE (SEQ ID NO: 163)

[00305] VKIISRQSLGVQPVYDIGVARDHNFLADGQVASN (SEQ ID NO:164)

[00306] >RlaKORDI51-2:/1-137 Rubidibacter lacunae KORDI 51-2

[00307] CLSYDTEVLTVEYGPLAIGTIVSERLACTVYTVDRSGFLYAQAISQWHERGRQD
VFEYALDNGMTIRATKDHKLMTADGQMVAIDDIFTQGLTLKAIDTAAN (SEQ ID NO:165)

[00308] MKIVSRKSLGVQHVYDIGVARDHNFLANGAIASN (SEQ ID NO:166)

[00309] >CfrPCC9212/1-136 Chlorogloeopsis fritschii PCC 9212

[00310] CLSYDTAILTVEYGFLPIGEIVEKGIECTVYTVDSNGYIYTQPIAQWHNRGEQEL
FEYSLEDGSIIRATKDHKFMIDGQMLPIDEIFARKLELMQVKGLP (SEQ ID NO:167)

[00311] VKIIAKKSLGTQNVYDIGVERDHNFVIKNGLVAAN (SEQ ID NO:168)

- [00312] >RinHH01:/1-137 *Richelia intracellularis* HH01 WGS project
- [00313] CLSYDTQILTVEHGPMSIGEIVEKCLECHVYTVNKNNGNICIQTTTQWHFRGEQEI
FEYELEDGSFIQATKDHKFMTTTGEMPLPIHEIFTNGLEILQLSKSLL (SEQ ID NO:169)
- [00314] VKILARKSLGTQKVYDIGVNDDHNFALSNSFIASN (SEQ ID NO:170)
- [00315] >SspPCC7117/1-137
- [00316] CLAGDTPVVTVEYGVLPPIQTIVEQELLQVYSVDAQGLIYTQPIEQWHNRGDR
LLYEYELENGQMIRATPDHKFLTTTGELLPIDEIFTQNLDLAAWAVPDS (SEQ ID NO:171)
- [00317] VKIIRRKFIGHAPTYDIGLSQDHNFLLGQGLIAAN (SEQ ID NO:172)
- [00318] >SspPCC8807/1-137
- [00319] CLAGDTPVVTVEYGVLPPIQTIVEQELLCHVYSVDAQGLIYTQPIEQWHQRGDR
LYEYELENGQMIRATPDHKFLTTTGKLLPIDEIFTQNLDLAAWAVPDS (SEQ ID NO:173)
- [00320] VKIIRRKFIGHAPTYDIGLSQDHNFLLGQGFIAN (SEQ ID NO:174)
- [00321] >SspNKBG042902:/1-137 *Synechococcus* sp. NKBG 042902
- [00322] CLAGDTPVVTVEYGVLPPIQTIVEQELLCHVYSVDAQGLIYTQPIEQWHQRGDR
LLYEYELENGQMIRATPDHKFLTTTGELLPIDEIFTQNLDLAAWAVPDS (SEQ ID NO:175)
- [00323] VKILRRKFIGRAPTYDIGLSQDHNFLLGQGLVAAN (SEQ ID NO:176)
- [00324] >SspNKBG15041:/1-129 *Synechococcus* sp. NKBG15041
- [00325] CLAGDTPVVTVEYGVLPPIRTIVDQELLCHVYSLDPQGFIYAQPVEQWHRRGDR
LLYEYELETGAVIRATPDHKFLTATGEMPLPIDEIFVRNLDL (SEQ ID NO:177)
- [00326] VKIIRRNLI GEAA TYDIGLGKDHNFLLGQGLIASN (SEQ ID NO:178)
- [00327] >SspPCC73109/1-130
- [00328] CLAGGTPVVTVEYGVLPPIQTIVEQELLCHVYSVDAQGLIYTQPIEQWHQRGDR
LLYEYELENGQMIRATPDHKFLTTTGELLPIDEIFTQNLDLL (SEQ ID NO:179)
- [00329] VKIIRRKFIGHAPTYDIGLSQDHNFLLGQGLIAAN (SEQ ID NO:180)
- [00330] >SspPCC7003/1-130

[00331] CLAGDTPVVTVEYGVLPITQIVEQELLCHVYSVDAQGLIYTQPIEQWHKRGDR
LLYEYELENGQIIRATPDHKFLTTTGEMRPIDEIFAKNLSLL (SEQ ID NO: 181)

[00332] VKIIRRK FVGHAPTYDIGLSQDHNFLLGQGLIAAN (SEQ ID NO:182)

[00333] >CspPCC8802/1-134 : *Cyanothece* sp. PCC 8802

[00334] CLSYDTEILTVEYGAIPIGKVVEENIDCTVYTVDKNGFVYTQNIQWHLRGQQE
VFEYYLDDGSILRATKDHQFMTLEGEMLPIDEIFERGLELKKIKI (SEQ ID NO:183)

[00335] VKIVSYRSLGKQFVYDIGVAQDHNFLLANGSIASN (SEQ ID NO:184)

[00336] >SelPCC7942:/1-137 *Synechococcus elongatus* PCC 7942

[00337] CLAADTEVLTVEYGPPIAGKLVEENIRCQVYCCNPDGYIYSQPIGQWHQRGEQE
VIEYELSDGRIIRATADHRFMTEEGEMLSLDEIFERSLELKQIPTPLL (SEQ ID NO:185)

[00338] VKIVRRRSLGVQP VYDLGVATVHNFVLANGLVASN (SEQ ID NO:186)

[00339] >CfrPCC6912:/1-137 *Chlorogloopsis fritschii* PCC 6912

[00340] CLSYDTAILTVEYGFPIGEIVEKGIECTVYTVDSNGYIYTQPIAQWHNRGEQEL
FEYSLEDGSIIRATKDHKFM TIDGQMLPIDEIFARKLELMQVKGLPE (SEQ ID NO:187)

[00341] VKIIAKKSLGTQNVYDIGVERDHN FVIKNGLVASN (SEQ ID NO:188)

[00342] >CspATC51472:/1-132 *Cyanothece* sp. ATCC 51472

[00343] CLSYDTEILTVEYGPMPIGKIVEENINCTVYTVDPNGFVYTQAIQWHYRGEQE
IFEYYLEDGATIRATKDHKFM TMEGKMLPIDEIFENNLDLKQL (SEQ ID NO:189)

[00344] VKIIGRQSLGVQKVYDIGVEKEHNFL LHNGLIASN (SEQ ID NO:190)

[00345] >Lma:/1-132 *Lyngbya majuscula*

[00346] CLSYDTEIITVEYGPPIAGEIVEKGIPCTVYSVDSNGYVYTQPIAQWHNRGEQEV
FEYTLDDGSVIRATKDHKFM TIDGQMLPIDEIFEGGLELKQL (SEQ ID NO:191)

[00347] VKIISRKSLGTQP VYDIGVKDDHNFILANGMVASN (SEQ ID NO:192)

[00348] >CspESFC/1-137

[00349] CLSYDTEVLTVEYGAVPIGKLVEEKLNC SVYTVDPNGYIYTQAIQWHDRGIQ
EVFEYQLEDNTIIRATKDHKFM TEDHQMLPIDEIFERGLELKKCPQPQQ (SEQ ID NO:193)

- [00350] VKIIRRRSLGFQPVYDIGLEQDHNFLNQGAIASN (SEQ ID NO:194)
- [00351] >SspPCC7002:/1-129 *Synechococcus* sp. PCC 7002
- [00352] CLAGGTPVVTVEYGVLPPIQTIVEQELLCHVYSVDAQGLIYAQLIEQWHQRGDR
LLYEYELENGQMIRATPDHRFLTTTGELLPIDEIFTQNLDL (SEQ ID NO:195)
- [00353] VKIIRRKFIGHAPTYDIGLSQDHNFLLGQGLIAAN (SEQ ID NO:196)
- [00354] >AmaMBIC11017:/1-132 *Acaryochloris marina* MBIC11017
- [00355] CLSYDTPVLTLEYGWLPIGQVVQEIECQVFSINERGHLYTQPIAQWHHRGQQ
EVFEYTLADGSTIQATAEHQFMTTDGQMYPVQQIFEEGLSLKQL (SEQ ID NO:197)
- [00356] VKIIQRRSLGLQSVYDIGLAQDHNFMANGWVAAN (SEQ ID NO:198)
- [00357] >Mae905:/1-129 *Microcystis aeruginosa* DIANCHI905
- [00358] CLGGETLILTEEYGLLPIAKIVSEEVNCTVYSVDKNGFVYSQPISQWHERGLQE
VFEYTLENGQTIQATKDHKFMTNDGEMLAIDTIFERGLDL (SEQ ID NO:199)
- [00359] VKIISRQSLGRKPVYDIGVEKDHNFLGNGLIASN (SEQ ID NO:200)
- [00360] >AciAWQC310F:/1-125 AWQC: *Anabaena circinalis* AWQC310F
- [00361] CLSYDTEILTVEYGFLEIGEIVEKQIECKVYTVDSNGILYTQPIAQWHHRGQQEV
YEYLLENGAIIRATKDHKFMTEAGEMLPIDDIFTQ (SEQ ID NO:201)
- [00362] VKIISRTYVGQANVYDIGVENDHNFVIKNGFVAAN (SEQ ID NO:202)
- [00363] >AciAWQC131C:/1-125 *Anabaena circinalis* AWQC131C
- [00364] CLSYDTEILTVEYGFLEIGEIVEKQIECRVYTVDSNGILYTQPIAQWHYRGQQEV
YEYLLENGAIIRATKDHNFMTAGEMLPIDDIFTQ (SEQ ID NO: 203)
- [00365] IKIISRKYVGQANVYDIGVENDHNFVIKNGFVAAN (SEQ ID NO: 204)
- [00366] >CspUCYN:/1-124 *Cyanobacterium* sp. UCYN-A2
- [00367] CLSYDTKVLTVYEGPLPIGKVVQENIRCRVYTTNDQGLIYTQPIAQWHNRGKQ
EIFEYHLDDKTIIRATKEHQFMTVDHVMMPIDEIFEQ (SEQ ID NO:205)
- [00368] KIIIRKSLGMHEVFDIGLEKDHNFLVLSNGLIASN (SEQ ID NO:206)

- [00369] >Pst:/1-129 Planktothrix st147 : st147_cleanDRAFT_c6
- [00370] CLSYDTEVL TVEYGLIPISKIVEEKIECTVYTVNNQGYVYTQPIAQWHNRGEQE
VFEYYLEDG SVIRATKDHKFMTVEGQMLPIDEIFEKELDL (SEQ ID NO:207)
- [00371] VKIISRKSLGTQPVYDIGVQEDHNFVLNNGLVASN (SEQ ID NO:208)
- [00372] >PlaCYA98/1-129 : Planktothrix NIVA-CYA 98
- [00373] CLSYDTEIL TVEYGLMPIGKIVKEKIECTVYTVNNQGYVYTQPIAQWHHRGEQ
EVFEYCLE DGSVIRATKDHKFMTVQGQMLPIDEIFEKELDL (SEQ ID NO:209)
- [00374] VKIISRKSLGTQPVYDIGVQEDHNFLLNNGLVASN (SEQ ID NO:210)
- [00375] >FdiUTEX481:/1-137 Fremyella diplosiphon UTEX 481
- [00376] CLSYDTEVL TVEYGLPIGEIVEKRLECSVYSVDINGNVYTQPIAQWHHRGQQE
VFEYAL EDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLLQVPHLPE (SEQ ID NO:211)
- [00377] VKIVTRRAIGAANVYDIGVEQDHNFAIKNGLIAAN (SEQ ID NO:212)
- [00378] >Pst585:/1-129 Planktothrix sp. 585: Length=1586997
- [00379] CLSYDTEIL TVEYGLIPISKIVEEKIECTVYTVNNQGYVYTQPIAQWHNRGEQEV
FEYYLEDG SVIRATKDHKFMTVDGQMLPIDEIFEKELDL (SEQ ID NO:213)
- [00380] VKIISRKSLGTQPVYDIGVQEDHNFVLNNGLVASN (SEQ ID NO:214)
- [00381] >NpuPCC73102/1-137
- [00382] CLSYETEIL TVEYGLLPIGKIVEKRIECTVYSVDNNGNIYTQPVAQWHDRGEQE
VFEYCLE DGSILIRATKDHKFMTVDGQMLPIDEIFERELDLMRVDNLPN (SEQ ID NO:215)
- [00383] IKIATRKYLGKQNVYDIGVERDHNFALKNGFIASN (SEQ ID NO:216)
- [00384] >CthPCC7203:/1-137 Chroococcidiopsis thermalis PCC 7203
- [00385] CLSYDTEIL TVEYGAIPIGKIVEERIECTVYSVDNNGFIYTQPIAQWHNRGQQEV
FEYCLE DGSIIIRATKDHKFMTFEGKMLPIDEIFEQELDLKQVKSIQN (SEQ ID NO:217)
- [00386] VKIISRKSLGIQPVYDIGVERDHFV LKNGLV ASN (SEQ ID NO:218)
- [00387] >NspCCY9414:/1-137 Nodularia spumigena CCY9414 genome

[00388] CLSYDTEILTVEYGYPIGEIVEKAIECSVYSVDNNGNVYTQPIAQWHNRGEQE
VFEYSLEDGSTIRATKDHKFMTTDQGMLPIDEIFAQELDLLQVHGLPK (SEQ ID NO:219)

[00389] VKITARKFVGRENVYDIGVERYHNFAIKNGLIASN (SEQ ID NO:220)

[00390] >AcyPCC7122:/1-137 *Anabaena cylindrica* PCC 7122

[00391] CLSYDTEVLTVEYGFPIGEIVEKRIECSIFSVDKNGNVYTQPIAQWHNRGRQEI
YEYCLDDGSKIRATKDHKFMTTAGEMPLPIDEIFERDLDLLKVEGLPE (SEQ ID NO:221)

[00392] VKIISRQYLGQADVYDIGVEEDHNFAIKNGFIASN (SEQ ID NO:222)

[00393] >CspPCC7507:/1-137 *Calothrix* sp. PCC 7507, complete genome

[00394] CLSYDTEVLTVEYGLPIGEIVEKGIECRVFSVDNHGNVYTQPIAQWHNRGQQE
VFEYGLDDGSVIRATKDHKFMTTDGMPLPIDEIFERGLDLLQVQGLPE (SEQ ID NO:223)

[00395] VKVITRKYIGKENVYDIGVELDHNFAIRNGLVASN (SEQ ID NO:224)

[00396] >NspPCC7524:/1-137 *Nostoc* sp. PCC 7524

[00397] CLSYDTEILTVEYGFLPIGEIVEKGIECTVFSVASNGIVYTQPIAQWHNRGQQEIF
EYCLEDGSIIRATKDHKFMTTQDQGMLPIDEIFACELDLLQVQGLPE (SEQ ID NO:225)

[00398] VKVVTRKYIGKENVYDIGVERDHNFVIRNGLVASN (SEQ ID NO:226)

[00399] >Naz0708:/1-137 '*Nostoc azollae*' 0708

[00400] CLSYKTEVLTVEYGLPIGEIVEKRIECSLFSVDENGNIYTQPIAQWHHRGVQEV
YEYCLDDGTIIRATKDHKFMTTIGEMPLPIDEIFERDLNLLQVNGLPT (SEQ ID NO:227)

[00401] VKIISRQFLGPANVYDIGVAQDHNFAIKNGLIASN (SEQ ID NO:228)

[00402] >NspPCC7120:/1-137 *Nostoc* sp. PCC 7120 DNA

[00403] CLSYDTEVLTVEYGFVPIGEIVEKGIECSVFSINNNGIVYTQPIAQWHHRGKQEV
FEYCLEDGSIKATKDHKFMTTQDGMPLPIDEIFEQELDLLQVKGLPE (SEQ ID NO:229)

[00404] IKIASRKFLGVENVYDIGVRRDHNFFIKNGLIASN (SEQ ID NO:230)

[00405] >AvaATCC29413/1-137 *Anabaena variabilis* ATCC 29413

[00406] CLSYDTEVLTVEYGFVPIGEIVDKGIECSVFSIDSNGIVYTQPIAQWHHRGKQEV
FEYCLEDGSIKATKDHKFMTTQDGMPLPIDEIFEQELDLLQVKGLPE (SEQ ID NO:231)

- [00407] IKIASRKFLGVENVYDIGVGRDHNFFVKNGLIASN (SEQ ID NO:232)
- [00408] >PspPCC7327:/1-135 *Pleurocapsa* sp. PCC 7327.
- [00409] CLSYDTKILTVEYGAMPIGKIVEEQIDCTVYTVNQNGFVYTQPIAQWHDRGKQ
EIFEYCLEDGSIIRATKDHKFMTTDGQMLPIDKIFEKGLDLKTINCD (SEQ ID NO: 233)
- [00410] VKILSRKSLGIQSVYDIGVEKDHNFLLANGLVASN (SEQ ID NO:234)
- [00411] >CspPCC7424:/1-135 *Cyanothece* sp. PCC 7424
- [00412] CLSYETQIMTVEYGLMPIGKIVEEQIDCTVYTVNKNGFVYTQPIAQWHYRGEQ
EVFEYCLEDGSTIRATKDHKFMTTDGQMLPIDEIFEQGLELKQIHLS (SEQ ID NO:235)
- [00413] VKIISRQSLGIQPVYDIGVEKDHNFLISDGLIASN (SEQ ID NO:236)
- [00414] >CspPCC7822:/1-134 *Cyanothece* sp. PCC 7822
- [00415] CLSYDTEILTVEYGPMPIGKIVEEQIECTVYTVDKNGLVYTQPIAQWHHRGQQE
VFEYCLEDGSIIRATKDHKFMTTDDGQMLPIEEIFEKGLELKQIIL (SEQ ID NO:237)
- [00416] VKIISRQLAGNQTVYDLGVEKDHNFLLANGLIASN (SEQ ID NO:238)
- [00417] >NspPCC7107:/1-137 *Nostoc* sp. PCC 7107
- [00418] CLSYDTQVLTVEYGLVPIGEIVEKQLECSVFTIDGHGYVYTQAIQWHNRGQQ
EVFEYGLEDGSVIRATKDHKFMTTDGQMLPIDEIFERELDLLQVQGLRW (SEQ ID NO:239)
- [00419] VKIITRKYIGQANVYDIGVAQDHNFVIENRLIASN (SEQ ID NO:240)
- [00420] >TboIicb1/1-136 *Tolypothrix bouteillei* Iicb1
- [00421] CLSYDTEILTVEYGFLPIGKIVEKGIECNVYSVDKNGNIYTQPIAQWHDRGEQE
VFEYCLENGSVIRATKDHKFMTTSGEMLPIDEIFERGLDLIRVEDLP (SEQ ID NO:241)
- [00422] VKILTRKSIGKQTVYDIGVERDHNFVIKNGSVASN (SEQ ID NO:242)
- [00423] >Aov:/1-136 *Aphanizomenon ovalisporum* DnaE precursor (dnaE) gene
- [00424] CLSADTEILTVEYGFPLPIGEIVGKAIECRVYSVDGNGNIYTQSIAQWHNRGEQEV
FEYTLLEDGSIIRATKDHKFMTTDGEMLPIDEXFARQLDLMQVQGLH (SEQ ID NO:243)
- [00425] VKITARKFVGRENVYDIGVEHHHNFAIKNGLIASN (SEQ ID NO:244)

- [00426] >OnvPCC7112:/1-137 *Oscillatoria nigro-viridis* PCC 7112
[00427] CLSYDTKILTVEYGPMAIGKIVEEKIECTVYSVDSNGYIYTQSIAQWHRRGQQE
VFEYCLEDGSIIRATKDHKFMTVGGQMLPIDEIFEQGLDLKQINSSSD (SEQ ID NO:245)
[00428] VKIISRKSLGTQEVYDIGVEREHNFILENSLVASN (SEQ ID NO:246)
- [00429] >RspPCC7116:/1-135 *Rivularia* sp. PCC 7116, complete genome
[00430] CLSYDTEVLTEEFGLPIGKIVEEKIDCTVYSVDVNGNVYSQPIAQWHNRGMQE
VFEYELEDGSTIRATKDHKFMTVDGEMLAIDEIFEKGLELKRVGIIY (SEQ ID NO:247)
[00431] VKIISRKVLKTENVYDIGLEGDHNFIKDGLIASN (SEQ ID NO:248)
- [00432] >MspPCC7113:/1-137 *Microcoleus* sp. PCC 7113,
[00433] CLSYDSEILTVEYGLMPIGKIVEEGIECTVYSVDSHGPLYTQPIAQWHHRGQQE
VFEYDLEDGSVIRATKDHKFMTSEGQMLAIDEIFERGLELKQVKRSQP (SEQ ID NO:249)
[00434] VKIVRRKSLGIQTVYDIGVERDHNFLLANGLVASN (SEQ ID NO:250)
- [00435] >ScyPCC7437:/1-137 *Stanieria cyanosphaera* PCC 7437
[00436] CLSYDTEILTVEYGAMPIGKIVKEQIECNVYTVNQNGFIYPQAIAQWHERGKQE
IFEYTLDNGLVIRATKDHKFMTIDGQMLPIDEIFERGLELQRINDYSN (SEQ ID NO:251)
[00437] VKIVSRKSLGKQPVDIGVTKDHNFLLSNGVVASN (SEQ ID NO:252)
- [00438] >CspPCC6303:/1-137 *Calothrix* sp. PCC 6303
[00439] CLSYDTEILTWEYGFLKIGEIVEKQILCSVFSVDEQGNVYTQPIAQWHNRGLQE
LFAYQLEDGGVIRATKDHKFMTTDGQMLAIDEIFERQLDLFQVKGLPE (SEQ ID NO:253)
[00440] VKIISRKVLKTENVYDIGLEGDHNFIKDGLIASN (SEQ ID NO:254)
- [00441] >Cst:/1-134 PCC7202: *Cyanobacterium stanieri* PCC 7202
[00442] CLSYDTEVLTVYGVLPKIGKIVEEQIQCTVYSVDQYGFVYTQAIAQWHDRGEQ
EVFEYELENGATIKATKDHKMMTSDGQMLPIDQIFEQGLDLFMVSF (SEQ ID NO:255)
[00443] VKIVKRRSHGIQKVYDIGVAKDHNFLHNLVASN (SEQ ID NO:256)
- [00444] >CspATCC51142:/1-134 *Cyanothece* sp. ATCC 51142

[00445] CLSYDTEILTVEYGPMPIGKIVEENINCTVYTVDPNGFVYTQAIQWHYRGEQE
IFEYYLEDGATIRATKDHKFMTEGKMLPIDEIFENNLDLKQLTL (SEQ ID NO:257)

[00446] VKIIGRQSLGVQKVYDIGVEKEHNFLHNGLIASN (SEQ ID NO:258)

[00447] >CspPCC8801:/1-134 *Cyanothece* sp. PCC 8801

[00448] CLSYDTEILTVEYGAIPIGKVVEENIDCTVYTVDKNGFVYTQNIQWHLRGQQE
VFEYYLDDGSILRATKDHQFMTLEGEMLPHEIFERGLELKKIKI (SEQ ID NO:259)

[00449] VKIVSYRSLGKQFVYDIGVAQDHNFLLANGSIASN (SEQ ID NO:260)

[00450] >Asp:/1-136 *Anabaena* sp. 90 chromosome

[00451] CLSYDTEILTVEYGFLEIGEIVEKQIECKVYTIDSNGLYTSIAQWHNRGQQE
VYEYLLENGAIIRATKDHKFMTEAGQMLPIDEIFAQGLDLLQVGVAE (SEQ ID NO:261)

[00452] VKIVSRTYVGQANVYDIGVESDHNFKNGFIASN (SEQ ID NO:262)

[00453] >Aha:/1-137 *Aphanothece halophytica*

[00454] CLSYDTEIWTVEYGAMPIGKIVEEKIECSVYTVDENG FVYTQPIAQWHPRGQQE
IIEYTLLEDGRKIRATKDHKMMTESGEMLPIDEEIFQRELDLKVETFHEM (SEQ ID NO:263)

[00455] VKIIKRQSLGRQNVYDVCVETDHNFLANGCVASN (SEQ ID NO:264)

[00456] >HspPCC7418:/1-137 *Halothece* sp. PCC 7418

[00457] CLSYDTEIWTVEYGAMPIGKIVEEKIECSVYTVDENG FVYTQPIAQWHPRGQQE
IIEYTLLEDGRKIRATKDHKMMTESGEMLPIDEEIFQRELDLKVETFHEM (SEQ ID NO:265)

[00458] VKIIKRQSLGRQNVYDIGVETDHNFLANGCVASN (SEQ ID NO:266)

[00459] >CapPCC10605:/1-137 *Cyanobacterium aponinum* PCC 10605

[00460] CLSYDTEILTVEYGAISIGKIVEEKINCQVYSVDKNGFIYTQNIQWHDRGSQEL
FEYELEDGRIIKATKDHKMMTKDGQMLAINDIFEQELELYSVDDMGV (SEQ ID NO:267)

[00461] VKIVKRRLSLGVQP VYDIGVEKDHNFILANGLVASN (SEQ ID NO:268)

[00462] >Cat:/1-133 *Candidatus Atelocyanobacterium thalassa* isolate

[00463] CLSYDTKVLTV EYGPLPIGKV VQENIRCRVYTTNDQGLIYTQPIAQWHNRGKQ
EIFEYHLDDKTIIRATKEHQFMTVDHVMMPIDEIFEQGLELKKIK (SEQ ID NO:269)

- [00464] LKIIRKSLGMHEVFDIGLEKDHNFLSNGLIASN (SEQ ID NO:270)
- [00465] >Oli:/1-137 *Oscillatoria limnetica* 'Solar Lake' DnaE precursor
- [00466] CLSYNTEVLTVEYGPLPIGKIVDEQIHCRVYSVDENG FVYTQAIAQWHDRGYQ
EIFAYELADGSVIRATKDHQFMTEDGQMFPIDEIWEKGLDLKKLP TVQD (SEQ ID NO:271)
- [00467] VKIVRRQSLGVQNVYDIGVEKDHNFLASGEIASN (SEQ ID NO:272)
- [00468] >Cen:/1-137 *Cyanobacterium endosymbiont of Epithemia turgida*
- [00469] CLSYDTEVLTVEYGAIPIGRMVEESLDCTVYTVDKNGFVYTQSIQQWHSRGQQ
EIFEYCFEDGSIIRATKDHKFMTAEGKMSSIHDFEQGLELKKIIPWSG (SEQ ID NO:273)
- [00470] AKIISCKSLGKQSVYDIGVVQDHNFLLANGVVASN (SEQ ID NO:274)
- [00471] >SspPCC7502:/1-133 *Synechococcus* sp. PCC 7502
- [00472] CLGYDTPVLTVEYGFMPIGKIVEEKIQCHVYSVDQNGLVFTQAIAQWHNRGQQ
EVWEYNLDNGDIVRATKDHKFMTIDGQMLPINQIFEQGLELK VIA (SEQ ID NO:275)
- [00473] VKIVSCKPLRVQTVYDIGVEKDHNFI LDNGLVASN (SEQ ID NO:276)
- [00474] >CspUCYN:/1-124 *Cyanobacterium* sp. UCYN-A2
- [00475] CLSYDTKVLTV EYGPLPIGKV VQENIRCRVYTTNDQGLIYTQPIAQWHNRGKQ
EIFEYHLDDKTIIRATKEHQFMTVDHVMMPIDEIFEQ (SEQ ID NO:277)
- [00476] KIIIRKSLGMHEVFDIGLEKDHNFLSNGLIASN (SEQ ID NO:278)
- [00477] >Pst:/1-129 *Planktothrix st147* : st147_cleanDRAFT_c6
- [00478] CLSYDTEVLTVEYGLIPISKIVEEKIECTVYTVNNQGYVYTQPIAQWHNRGEQE
VFEYYLEDGSVIRATKDHKFMTVEGQMLPIDEIFEKELDL (SEQ ID NO:279)
- [00479] VKIISRKSLGTQPVYDIGVQEDHNFLNGLVASN (SEQ ID NO:280)
- [00480] >PlaCYA98/1-129 : *Planktothrix NIVA-CYA 98*
- [00481] CLSYDTEILTVEYGLMPIGKIVKEKIECTVYTVNNQGYVYTQPIAQWHHRGEQ
EVFEYCLEDGSVIRATKDHKFMTVQGQMLPIDEIFEKELDL (SEQ ID NO:281)
- [00482] VKIISRKSLGTQPVYDIGVQEDHNFLNGLVASN (SEQ ID NO:282)

- [00483] >Pst585:/1-129 Planktothrix sp. 585: Length=1586997
- [00484] CLSYDTEILTVEYGLIPISKIVEEKIECTVYTVNNQGYVYTQPIAQWHNRGEQEV
FEYYLEDGSGVIRATKDHKFMTVDGQMLPIDEIFEKELDL (SEQ ID NO:283)
- [00485] VKIISRKSLGTQPVDIGVQEDHNFVLNGLVASN (SEQ ID NO:284)
- [00486] >CspPCC8802/1-134 : Cyanothece sp. PCC 8802
- [00487] CLSYDTEILTVEYGAIPIGKVVEENIDCTVYTVDKNGFVYTQNIQWHLRGQQE
VFEYYLDDGSILRATKDHQFMTLEGEMLPHEIFERGLELKKIKI (SEQ ID NO:285)
- [00488] VKIVSYRSLGKQFVYDIGVAQDHNFLLANGSIASN (SEQ ID NO:286)
- [00489] >CfrPCC6912:/1-137 Chlorogloopsis fritschii PCC 6912
- [00490] CLSYDTAILTVEYGFLPIGEIVEKGIECTVYTVDSNGYIYTQPIAQWHNRGEQEL
FEYSLEDGSIIRATKDHKFMTIDGQMLPIDEIFARKLELMQVKGLPE (SEQ ID NO:287)
- [00491] VKIIAKKSLGTQNVYDIGVERDHNFVIKNGLVASN (SEQ ID NO:288)
- [00492] >CspATC51472:/1-132 Cyanothece sp. ATCC 51472
- [00493] CLSYDTEILTVEYGPMPIGKIVEENINCTVYTVDPNGFVYTQAIAQWHYRGEQE
IFEYYLEDGATIRATKDHKFMTMEGKMLPIDEIFENNLDLKQL (SEQ ID NO:289)
- [00494] VKIIGRQSLGVQKVYDIGVEKEHNFLHNGLIASN (SEQ ID NO:290)
- [00495] >Lma:/1-132 Lyngbya majuscula
- [00496] CLSYDTEILTVEYGPIAIGEIVEKGIPCTVYSVDSNGYVYTQPIAQWHNRGEQEV
FEYTLDDGSGVIRATKDHKFMTIDGQMLPIDEIFEGLLELKQL (SEQ ID NO:291)
- [00497] VKIISRKSLGTQPVDIGVKDDHNFILANGMVASN (SEQ ID NO:292)
- [00498] >CspESFC/1-137
- [00499] CLSYDTEVLTVEYGAVPIGKLVEEKLNCVYTVDPNGYIYTQAIAQWHDRGIQ
EVFEYQLEDNTIIRATKDHKFMTEDHQMLPIDEIFERGLELKKCPQPQQ (SEQ ID NO:293)
- [00500] VKIIRRRSLGFQPVYDIGLEQDHNFLNQGAIASN (SEQ ID NO:294)
- [00501] >Mae905:/1-129 Microcystis aeruginosa DIANCHI905

[00502] CLGGETLILTEEYGLLPIAKIVSEEVNCTVYSVDKNGFVYSQPISQWHERGLQE
VFEYTLENGQTIQATKDHKFMNTDGEMLAIDTIFERGLDL (SEQ ID NO:295)

[00503] VKIISRQSLGRKPVYDIGVEKDHNFLNGLIASN (SEQ ID NO:296)

[00504] >RlaKORDI51-2:/1-137 *Rubidibacter lacunae* KORDI 51-2

[00505] CLSYDTEVLTV EYGPLAIGTIVSERLACTVYTVDRSGFLYAQAISQWHERGRQD
VFEYALDNGMTIRATKDHKLMTADGQMVAIDIFTQGLTLKAIDTAAF (SEQ ID NO:297)

[00506] MKIVSRKSLGVQH VYDIGVARDHNFLLANGAIASN (SEQ ID NO:298)

[00507] >CfrPCC9212/1-136 *Chlorogloeopsis fritschii* PCC 9212

[00508] CLSYDTAILTV EYGFLPIGEIVEKGIECTVYTVDSNGYIYTQPIAQWHNRGEQEL
FEYSLEDGSIIRATKDHKFMNTIDGQMLPIDEIFARKLELMQVKGLP (SEQ ID NO:299)

[00509] VKIIAKKSLGTQNVYDIGVERDHN FVIKNGLVASN (SEQ ID NO:300)

[00510] >RinHH01:/1-137 *Richelia intracellularis* HH01 WGS project

[00511] CLSYDTQILTV EHGPM SIGEIVEKCLECHVYTVNKNGNICIQTITQWHFRGEQEI
FEYELEDGSGFIQATKDHKFMNTTTGEMPLPIHEIFTNGLEILQLSKSLL (SEQ ID NO:301)

[00512] VKILARKSLGTQKVYDIGVNDDHNFALSNSFIASN (SEQ ID NO:302)

[00513] >GhePCC6308:/1-133 *Geminocystis herdmanii* PCC 6308

[00514] CLSYDTEVLTV EFGAIPMGKIVEERLNCQVYSVDKNGFIYTQPIAQWHDRGVQ
EVFEYELEDGRIIKATKDHKMMIENC EMVEIDRIFEEGLELFEVN (SEQ ID NO:303)

[00515] VKILKRRSISSQQVYDIGVEKDHNFLLANGLVASN (SEQ ID NO:304)

[00516] >SsuPCC9445:/1-131 *Spirulina subsalsa* PCC 9445

[00517] CLSYDTKIITVEYGAIAIGTIVEQGLHCHVYSVDPNGFIYTQPIAQWHQRGEQEV
FAYTLENGSIIQATKDHKFMNTQQGKMLPIDTIFEQGLDLLQV (SEQ ID NO:305)

[00518] KIIKRTSLGVRPVYDIGVIQDHNFLLENGLVASN (SEQ ID NO:306)

[00519] >MaePCC9807:/1-135 *Microcystis aeruginosa* 9807

[00520] CLGGETLILTEEYGLLPIAKIVSEEINCTVYSVDKNGFIYSQPISQWHERGLQEVEF
EYTLENGQTIQATKDHKFMNTSDGEMLAIDTIFERGLDLKSSDFS (SEQ ID NO:307)

- [00521] VKIISRQFLGRKPVYDIGVEKDHNFLNGLIASN (SEQ ID NO:308)
- [00522] >MspGI1:/1-130 Myxosarcina sp. GI1 contig_13
- [00523] CLSYDTEVLTLKYGALPIGEIVEKRINCHVYTRAESGFFYIQSIEQWHDRGEQEV
FEYTLENGATIKATKDHKFMTSGGQMLPIDEIFERGLDLL (SEQ ID NO:309)
- [00524] VKIVSRKSLGKQPVYDLGVAKDHNFLLANGTVASN (SEQ ID NO:310)
- [00525] >ShoPCC7110:/1-136 Scytonema hofmanni PCC 7110 contig00136
- [00526] CLSYDTEVLTAEYGFLPIGKIVEKAIECTVYSVDNDGNIYTQPIAQWHDRGQQE
VFEYSLDDGSVIRATKDHKFMTTGGQMLPIDEIFERGLDLMRIDSLP (SEQ ID NO:311)
- [00527] VKILTRKSIGKQTVYDIGVERDHNFVIKNGLVASN (SEQ ID NO:312)
- [00528] >WinUHHT291/1-136 Westiella intricata UH HT-29-1
- [00529] CLSYDTEILTVEYGFLPIGEIVEKRIECTVYTVDTNGYVYTQAIQWHNRGEQE
VFEYALDGSIIIRATKDHKFMTSEGQMLPIDEIFVKGLDLLQVQGLP (SEQ ID NO:313)
- [00530] VKIITRKFLGIQNVYDIGVEQNHNFVIKNGLVASN (SEQ ID NO:314)
- [00531] >FspPCC9605:/1-136 Fischerella sp. PCC 9605 FIS9605DRAFT
- [00532] CLSYDTEILTVEYGFLPIGEIVEKGIECTVYTVDNNGNVYTQTIAQWHNRGQQE
VFEYCLEDDGSVIRATKDHKFMTTDGQMLPIDEIFARGLDLLQVKNLP (SEQ ID NO:315)
- [00533] VKIVTRRPLGTQNVYDIGVESDHNFVIKNGLVASN (SEQ ID NO:316)
- [00534] >MrePCC10914:/1-137 Mastigocladopsis repens PCC 10914
- [00535] CLSYDTEVLTV EYGFLPIGEIVEKSIECSVYTVDSNGNVYTQPIAQWHNRGQQE
VFEYCLEDDGSIIIRATKDHKFMTIHGQMLPIDEIFERGLELMKIQGLPE (SEQ ID NO:317)
- [00536] AKIITRKS LGTQNVYDIGVERDHNFVTRDGF IASN (SEQ ID NO:318)
- [00537] >ShoUTEX2349:/1-137 [Scytonema hofmanni] UTEX 2349
- [00538] CLSYNSEVLTV EYGFLPIGKIVEKGIECSVYSVDSYGKIYTQVIAQWHNRGQQE
VFEYCLEDDGTIIQATKDHKFMTVDGQMLPIDEIFERGLDLMQVQGLPD (SEQ ID NO:319)
- [00539] VKIITRKS LGTQNVYDIGVSSDHNFMKNGLIASN (SEQ ID NO:320)

- [00540] >AspPCC7108:/1-137 *Anabaena* sp. PCC 7108 Ana7108scaffold_2_Cont3
[00541] CLSSDTEVLTV EYGLPIGEIIEKRIDCSVFSVDKNGNIYTQPIAQWHDRGIQELY
EYCLDDGSTIRATKDHKFM TTAGEMLPIDEIFERGLDLLKVHNL PQ (SEQ ID NO:321)
[00542] VKIITRNYVGKENVYDIGVERDHNFAIKNGLIASN (SEQ ID NO:322)
- [00543] >FspPCC9339:/1-137 *Fischerella* sp. PCC 9339 PCC9339DRAFT
[00544] CLSYDTEVLTV EYGFLPIGEIVEKRIECTVYTVDHNGYVYTQPIAQWHNRGYQ
EVFEYGLDGSVIRATKDHKFM TSEGQMLPIDEIFARELDLLQVTGLVN (SEQ ID NO:323)
[00545] VKIVTRRL LGIQNVYDIGVEQNHN FVIKNGLVASN (SEQ ID NO:324)
- [00546] >Csp336:/1-137 *Calothrix* sp. 336/3
[00547] CLSYDTEIFTV EYGFLPIGEIVEKRLECTVLTVDNHGNIYSQPIAQWHHRGQQQI
Y EYGLDGSVIRATKDHKFM TTDGQMLPIDEIFERGLDLLQVTNLDN (SEQ ID NO:325)
[00548] VKVITRKLADTENVYDIGVENHHN FLIKNGLVASN (SEQ ID NO:326)
- [00549] >FthPCC7521:/1-136 *Fischerella thermalis* PCC 7521
[00550] CLSYETEILTVEY GFLPIGEIVEKRIECSVYTVDNNGYVCTQPIAQWHNRGYQE
VFEYGLDGSVIRATKDHKFM TIDRQMLPIDEIFARGLDLLQVTGLP (SEQ ID NO:327)
[00551] VKIITRKS LGTQNVYDIGVEQNHN FVIKNGLVASN (SEQ ID NO:328)
- [00552] >CyaPCC7702/1-137 cyanobacterium PCC 7702 Chl7702
[00553] CLSYDTEILTVEY GFLSIGEIVEKEIECTVYTVDSNGYIYTQPIAQWHEQGEQEIF
EYSLEDGSTIRATKDHKFM TIEGEMLPIDQIFARQLDLMQITGLPQ (SEQ ID NO:329)
[00554] VKISTKKS LGKQKVYDIGVVRDHN FIIKNGFVASN (SEQ ID NO:330)
- [00555] >FspPCC9431:/1-136 *Fischerella* sp. PCC 9431
[00556] CLSYDTEVLTV EYGFLPIGEIVEKRIECTVYTVDTNGYVYTQAIAQWHNRDEQE
VFEYALDGSII RATKDHKFM TSEGQMLPIDEIFAKGLDLLQVQGLP (SEQ ID NO:331)
[00557] VKIVTRKFLGIQNVYDIGVEQNHN FVIKNGLVASN (SEQ ID NO:332)
- [00558] >FmuPCC7414:/1-137 *Fischerella muscicola* PCC 7414

[00559] CLSYETEILTVEYGFPLPIGEIVEKRIECSVYTVDNNGYVCTQTIAQWHNRGYQE
VFEYGLDGSVIRATKDHKFM TIDRQMLPIDEIFARGLDLLQVKGLPE (SEQ ID NO:333)

[00560] VKIITRQSLGTQNVYDIGVEQNHN FVIKNGLVASN (SEQ ID NO:334)

[00561] >FmuPCC73103:/1-137 *Fischerella muscicola* SAG 1427-1 = PCC 73103

[00562] CLSYDTEVLTVEYGFPLPIGEIVEKTIECNVFTVDSNGYVYTQPIAQWHNRGYQE
VFEYGLDGSVIRATKDHKFM TSEGKMLPIDEIFARELDLLQVTGLIN (SEQ ID NO:335)

[00563] VKIVTRKFLGIQNVYDIGVEQNHN FVIKNGLVASN (SEQ ID NO:336)

[00564] >Lae:/1-137 *Lyngbya aestuarii* BL J laest3.contig.3

[00565] CLSYDTEILTVEYGAIPIGKVVDEKIECTVYSVDKNGLIYTQPIAQWHNRGKQE
VFEYSLEDGSTIRATKDHKFM TMDNQMLPIDEILEKGLELKQVNADSV (SEQ ID NO:337)

[00566] VKIVSRKSLDSQTVYDIGVETDHN FLLANGSVASN (SEQ ID NO:338)

[00567] >Lsp:/1-137 *Leptolyngbya* sp. JSC-1

[00568] CLSYDTEILTVEYGALPIGKIVENQMICSVYSIDNNGYIYIQPIAQWHNRGQQEV
FEYILEDGSIIRSTKDHKFM TKGGEMLPIDEIFERGLELAQVTRLEQ (SEQ ID NO:339)

[00569] VKIISRRSVGVQSVYDIGVKQDHN FFLRNGLIASN (SEQ ID NO:340)

[00570] >CwaWH8501:/1-137 *Crocospaera watsonii* WH8501

[00571] CLSYDTEILTVEYGAMYIGKIVEENINCTVYTVDKNGFVYTQTIAQWHNRGEQ
EIFEYDLEDGSKIKATKDHKFM TIDGEMLPIDEIFEKNLDLKQVVSHPD (SEQ ID NO:341)

[00572] VKIIGCRSLGTQKVYDIGVEKDHN FLLANGSIASN (SEQ ID NO:342)

[00573] >CchPCC7420:/1-135 *Coleofasciculus chthonoplastes* PCC 7420

[00574] CLSYDTQILTVEYGAVAIGEIVEKQIECTVYSVDENGYVYTQPIAQWHNRGEQE
VFEYLLEDGATIRATKDHKFM TDEDQMLPIDQIFEQGLELKQVEVL (SEQ ID NO:343)

[00575] VKIIGRKPLGTQPVYDIGVERDHN FLLFNGLSVASN (SEQ ID NO:344)

[00576] >CspPCC6712/1-133

[00577] CLSYDTEVLTVEYGAIPIGKIVEEKIACNVYSVDKNGFVYTQPIAQYHDRGIQE
VFEYRLENGSVIRATKDHKMM TADGQMLPIDEIFKQNLDLKQLN (SEQ ID NO:345)

- [00578] VKIISRQSLGKQSVFDIGVAKDHNFLLANGLVASN (SEQ ID NO:346)
- [00579] >Rbr:/1-137 Raphidiopsis brookii D9 D9_5,
- [00580] CLSYETEVLTLLEYGFLPIGEIVDKQMVCTVFSVNDSGNVYTQPIGQWHDRGVQ
ELYEYCLDDGSTIRATKDHKFM TTQGEMVPIDEIFHQGWELVQVSGTMN (SEQ ID
NO:347)
- [00581] VKIVSRRYL GKADVYDIGVAKDHNFIKNGLVASN (SEQ ID NO:348)
- [00582] >CspCCy0110:/1-134 Cyanothece sp. CCY0110 1101676644604
- [00583] CLSYDTEILTVEYGPMPIGKIVEENINCSVYTVNKNGFVYTQSIAQWHHRGEQE
VFEYYLEDGETIRATKDHKFM TTEGKMLPIDEIFENNLDLKKLTV (SEQ ID NO:349)
- [00584] VKIIEERSLGKQNVYDIGVEKDHNFLSNNLIASN (SEQ ID NO:350)
- [00585] >XspPCC7305:/1-135 Xenococcus sp. PCC 7305
- [00586] CLSADTEVL TVEYGAISIGKIVEERIECTVYSVDANGFVYTQEIAQWHNRGEQE
VFEYMLDDGSVIRATKDHKLM TIDGQMVAIDEIFSQGLELKQVLGL (SEQ ID NO:351)
- [00587] VKIVSRKSLGTQTVYDLGVARDHNFLLANGTVASN (SEQ ID NO:352)
- [00588] >PspPCC7319:/1-135 Pleurocapsa sp. PCC 7319
- [00589] CLSYDTEIYTVEY GALPIGKIVESRIKCTVLTVDKNGLVYSQPIVQWHDRGIQEV
FEYTLDN GATIRATKDHKFM TVEGQMLPIDEIFELGLELKEIQQF (SEQ ID NO:353)
- [00590] VKIISRQSLGKQSVYDIGVAKDHNFLLANGMVASN (SEQ ID NO:354)
- [00591] >CraCS505:/1-137 Cyindrospermopsis raciborskii CS-505
- [00592] CLSYETEVLTLLEYGFVPIGEIVNKQMVCTVFSLNDSGNVYTQPIGQWHDRGVQ
DLYEYCLDDGSTIRATKDHKFM TTQGEMVPIDEIFHQGWELVQVSGISK (SEQ ID NO:355)
- [00593] VKIVSRRYL GKADVYDIGVAKDHNFIKNGLVASN (SEQ ID NO:356)
- [00594] >MaePCC7806:/1-135 Microcystis aeruginosa PCC 7806
- [00595] CLGGETLILTEEYGLLPIAKIVSEEVNCTVYSVDKNGFVYSQPISQWHERGLQE
VFEYTLENGQTIQATKDHKFM TNDGEMLAIDTIFERGLDLKSSDFS (SEQ ID NO:357)
- [00596] VKIISRQSLGRKPVYDIGVEKDHNFLLGNGLIASN (SEQ ID NO:358)

[00597] >MaeNIES843:/1-135 *Microcystis aeruginosa* NIES-843 DNA
 [00598] CLGGETLILTEEYGLLPIAKIVSEEINCTVYTVDQNGFVYSQPISQWHERGLQEV
 FEYTLENGQTIQATKDHKFMTSDGEMLAIDTIFERGLDLKSSDFS (SEQ ID NO:359)
 [00599] VKIIGRQSLGRKPVYDIGVEKDHNFLGNGLIASN (SEQ ID NO:360)

[00600] Fig. 1 shows an alignment and a computer-generated model of the design of the Cfa split intein according to an embodiment of the invention. Panel A shows a sequence alignment of Npu DnaE and Cfa DnaE. The sequences share 82% identity with the differences (underlined, cyan) evenly distributed through the primary sequence. Catalytic residues and second shell 'accelerator' residues are shown in caret, orange and asterisk, green, respectively. Panel B shows the same residues highlighted in panel a mapped on to the Npu structure (pdb = 4kl5).

[00601] The Cfa intein has high sequence similarity to Npu (82%), and the non-identical residues are spread throughout the 3D structure of the protein.

[00602] Cfa intein fragments fused to model exteins were generated and their PTS activity was measured using the aforementioned *in vitro* assay (Fig. 2). This revealed that the Cfa intein splices 2.5 fold faster at 30°C than Npu ($t_{1/2}$ 20s vs. 50s), a notable enhancement in activity since the latter is the fastest characterized DnaE split intein (Fig. 2A). This accelerated rate manifests itself both in branch formation (3-fold increase) and branch resolution (2-fold increase). In line with parent DnaE inteins, Cfa retains the preference for a bulky hydrophobic residue at the +2 position of the C-extein. Strikingly, Cfa shows an increased splicing rate as a function of temperature and is consistently faster than Npu (Fig. 2A). The Cfa intein even maintains activity at 80°C, albeit with reduced yield of splice products, while Npu is inactive at this temperature. These results demonstrate that consensus engineering is effective in producing an intein that is highly active across a broad range of temperatures.

[00603] Applications of PTS typically require fission of a target protein and fusion of the resulting fragments to the appropriate split intein segments.¹ As a consequence, the solubility of these fusion proteins can sometimes be poor. Because protein denaturants such as guanidine hydrochloride (GuHCl) and urea are frequently used to keep these less soluble fragments in solution, the ability of Cfa to splice in the presence of these chaotropic agents was tested. Cfa intein was found to splice in the presence of up to 4M GuHCl (with little decrease in activity seen up to 3M), while no activity was

observed for Npu in $\geq 3\text{M}$ GuHCl (Fig. 2B). Remarkably, the splicing of Cfa is largely unaffected up to 8M urea, while splicing of Npu falls off dramatically above 4M urea (Fig. 2C).

[00604] Fig. 2 shows graphs showing the characterization of the Cfa intein according to an embodiment of the invention. In Panel A, splicing rates for Cfa and Npu as a function of temperature are shown. Npu is inactive at 80°C (error = SD (n=3)). In Panels B and C, splicing rates for Cfa and Npu as a function of added chaotrope are shown. Npu is inactive in 3M GuHCl or 8M Urea. Note, Cfa has residual activity in 4M GuHCl ($k=7 \times 10^{-5}$) (error = SD (n=3)).

[00605] The unprecedented and unexpected tolerance of Cfa to high concentrations of GuHCl and urea suggests the intein might retain activity directly following chaotropic extraction of insoluble proteins from bacterial inclusion bodies, thereby expediting PTS-based studies. Accordingly, the model fusion protein, His₆-Sumo-Cfa^N, was overexpressed in *E. coli* cells and extracted the protein from inclusion bodies with 6M urea. The protein was purified from this extract by nickel affinity chromatography and then directly, and efficiently, modified by PTS under denaturing conditions, i.e. without the need for any intervening refolding steps. In general, it is expected that the robust activity of Cfa in the presence of chaotropic agents will prove useful when working with protein fragments that demonstrate poor solubility under native conditions.

[00606] Fusing a protein of interest to a split intein can result in a marked reduction in cellular expression levels compared to the protein alone.⁶ This situation is more frequently encountered for fusions to N-inteins than to C-inteins, which is likely due to the larger size of the former and their partially folded state.¹⁸ It was therefore investigated whether the improved thermal and chaotropic stability of Cfa would translate to increased expression levels of Cfa^N fusions. Indeed, model studies in *E. coli* revealed a significant (30-fold) increase in soluble protein expression for a Cfa^N fusion compared to the corresponding Npu^N fusion (Fig. 8). Given this result, it was investigated whether Cfa^N fusions would also exhibit increased protein expression levels in mammalian cells. In particular, intein fusions to the heavy chain (HC) of monoclonal antibodies (mAbs) have emerged as a powerful tool for site-specific conjugation of synthetic cargoes.¹⁹⁻²¹ The expression levels in HEK293 cells of a mAb (α Dec205) as a function of the N-intein fused to its HC was explored. Consistent with the bacterial expression results, production of the HC-Cfa^N fusion was significantly higher than for the other inteins examined; for example, the secreted levels of the mAb-Cfa construct were ~10-fold higher than for the corresponding Npu fusion (Fig. 3A and 3B). Importantly, mAb-Cfa retained PTS activity and could be site-specifically modified with a synthetic peptide by splicing directly in the growth medium following the four-day expression at 37°C.

[00607] Fig. 8 is an SDS-PAGE analysis of test expression of His₆-SUMO-Npu^N and His₆-SUMO-Cfa^N. Coomassie brilliant blue stained gel from a 4 mL column volume (CV) Ni-NTA purification of the soluble fraction of 1L of *E. coli* culture. Lanes correspond to (P) the inclusion body pellet, (FT) flow through of batch bound Ni-NTA solution, (W1) a 5CV wash with 5 mM imidazole, (W2) a 5CV wash of 25 mM imidazole, (E1-E4) and four 1.5CV elutions of 250 mM imidazole.

[00608] Finally, to further explore the utility of the Cfa intein in the context of antibody conjugation, whether the PTS system could be used to attach multiple copies of a synthetic cargo to the heavy chain of the mAb was investigated. Accordingly, semisynthesis was used to prepare a construct in which the C-terminal half of Cfa (Cfa^C) was fused to a C-extein containing a dendrimeric scaffold allowing multimeric attachment of cargo, in this case fluorescein (Fig. 3C). This dendritic cargo was successfully linked to the αDec205 antibody via Cfa-mediated PTS, again performed directly in situ within the cellular growth medium (Fig. 3D and 3E). This represents the first time that PTS has been used to attach a branched extein construct to a target protein, highlighting the potential of the system for manipulating the payload quantity of antibody drug conjugates.²²

[00609] Fig. 3 shows expression and modification of a mouse monoclonal antibody using the Cfa intein according to an embodiment of the invention. Panel A shows test expression in HEK293T cells of various Int^N homologues (Npu, Mcht, Ava and Cfa) fused to the C-terminus of the heavy chain of a mouse αDec205 monoclonal antibody. Top: Western blot analysis (αMouse IgG) of antibody levels present in the medium following the 96 hour expression. Bottom: α-actin western blot of cell lysate as a loading control. Panel B shows quantification of normalized expression yield by densitometry of αDEC205 HC-Int^N signal in panel A (error = SD (n=4)). Panel C shows the structure of the Cfa^C-dendrimer construct used in PTS reactions with the αDEC205 HC-Int^N fusion. For simplicity, the Cfa^C peptide sequence is depicted symbolically in green (as a rectangle with a triangular cut-out on the left). Panel D is a schematic of the *in situ* PTS approach used to modify the HC of a mAb with a multivalent cargo. Panel E is an SDS-PAGE analysis of PTS reaction. Lane 1: Wild type mouse αDEC205 mAB. Lane 2: Mouse αDEC205-Cfa^N mAB fusion. Lane 3: addition of the Cfa^C-dendrimer to the media containing the αDEC205-Cfa^N mAB. The splicing reaction was analyzed by fluorescence (bottom) and western blot (top, αMouse IgG).

[00610] The discovery of fast split inteins has revolutionized the applications of protein trans splicing. The remarkable robustness of the Cfa intein described in this study should extend the utility of many of these technologies by allowing PTS to be performed in a broader range of reaction conditions. Moreover, the ability of Cfa to increase the expression yields of N-intein fusions should

encourage further use of split inteins for protein semisynthesis. The activity-guided approach we use to engineer this intein may be applied to other intein families or act as a general strategy for the refinement of multiple sequence alignments used for consensus engineering.

[00611] *Materials and Methods*

[00612] *Materials*

[00613] Oligonucleotides and synthetic genes were purchased from Integrated DNA Technologies (Coralville, IA). The QuickChange XL II site directed mutagenesis kit and Pfu Ultra II Hotsart fusion polymerase were purchased from Agilent (La Jolla, CA). All restriction enzymes and 2x Gibson Assembly Master Mix were purchased from New England Biolabs (Ipswich, MA). "In-house" high-competency cells used for cloning and protein expression were generated from One Shot BL21 (DE3) chemically competent *E. coli* and sub-cloning efficiency DH5 α competent cells purchased from Invitrogen (Carlsbad, CA). Dulbecco's Modified Eagle Medium (DMEM), Lipofectamine 2000, and low IgG fetal bovine serum were purchased from Invitrogen as well. DNA purification kits were purchased from Qiagen (Valencia, CA). All plasmids were sequenced by GENEWIZ (South Plainfield, NJ). N,N-diisopropylethylamine (DIPEA), 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), tetrakis(triphenylphosphine)palladium(0) (Pd(PPh₃)₄), and 5(6)-carboxyfluorescein were purchased from Sigma-Aldrich (Milwaukee, WI) and used without further purification. Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and isopropyl- β -D-thiogalactopyranoside (IPTG) were purchased from Gold Biotechnology (St. Louis, MO). The protease inhibitor used was the Roche Complete Protease Inhibitor (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). (7-Azabenzotriazol-1-yl)oxy)tripyrrolidinophosphonium hexafluorophosphate (PyAOP) and O-(Benzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate (HBTU) were purchased from Genscript (Piscataway, NJ). Rink Amide-ChemMatrix resin was purchased from Biotage (Charlotte, NC). Trifluoroacetic acid (TFA) was purchased from Halocarbon (North Augusta, SC). Immun-blot PVDF membrane (0.2 μ m) and Criterion XT Bis-Tris gels (12% polyacrylamide) were purchased from Bio-Rad (Hercules, CA). MES-SDS running buffer was purchased from Boston Bioproducts (Ashland,

MA). Anti-Mouse IgG secondary antibody (Licor mouse 800) and Mouse α Actin primary antibody were purchased from Li-COR biotechnology (Lincoln, NE).

[00614] Equipment

[00615] Analytical RP-HPLC was performed on Hewlett-Packard 1100 and 1200 series instruments equipped with a C₁₈ Vydac column (5 μ m, 4.6 x 150 mm) at a flow rate of 1 mL/min. Preparative RP-HPLC was performed on a Waters prep LC system comprised of a Waters 2545 Binary Gradient Module and a Waters 2489 UV detector. Purifications were carried out on a C₁₈ Vydac 218TP1022 column (10 μ m; 22 x 250 mm) at a flow rate of 18 mL/min. All runs used 0.1 % TFA (trifluoroacetic acid) in water (solvent A) and 90 % acetonitrile in water with 0.1 % TFA (solvent B). Unless otherwise stated, peptides and proteins were analyzed using the following gradient: 0% B for 2 minutes (isocratic) followed by 0-73% B over 30 minutes. Electrospray ionization mass spectrometric analysis (ESI-MS) was performed on a Bruker Daltonics MicroTOF-Q II mass spectrometer. Size-exclusion chromatography was carried out on an AKTA FPLC system (GE Healthcare) using a Superdex S75 16/60 (CV=125 mL) column. Coomassie-stained gels and western blots were imaged using a LI-COR Odyssey Infrared Imager. Fluorescent gels were imaged using a GE ImageQuant LAS 4000 Imager. The splicing-dependent *E. coli* growth assay was performed on a VersaMax tunable microplate reader from Molecular Devices. Cell lysis was carried out using a S-450D Branson Digital Sonifier.

[00616] Cloning of DNA plasmids

[00617] All N-intein constructs for *E. coli* expression were cloned into previously used pET and pTXB1 vectors.¹ Plasmids encoding for WT pet30-His₆-SUMO-AEY-Ssp^N, pet30-His₆-SUMO-AEY-Npu^N, pTXB1-Ssp^C-MxeGyrA-His₆, and pTXB1-Npu^C-MxeGyrA-His₆ plasmids were cloned as previously described¹ and encode for the following protein sequences. Protein products after either SUMO cleavage (N-inteins) or thiolysis (C-inteins) are shown in bold for all plasmids.

[00618] **Plasmid 1:**

[00619] WT Ssp^N: pet30- His₆-SUMO-AEY- Ssp^N

[00620] MGSSHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHD
RGEQEVLEYELEDGSVIRATSDHRFLT TDYQLLAIEEIFARQLDLLTLENIKQTEEALD
NHRLPFPLLDAGTIK (SEQ ID NO:361)

[00621] **Plasmid 2:**

[00622] WT Npu^N: pet30- His₆-SUMO-AEY-Npu^N

[00623] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLFLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYALS^{**Y**}ETEILTVEYGLLP**IGK**IVEKRIECTVYSVDNNGNIYTQPVAQWHD
RGEQEVFEYCLEDGSLIRATKDHKFMTVDGQMLPIDEIFERELDLMRVDNLPN (SEQ
ID NO:362)

[00624] **Plasmid 3:**

[00625] WT Ssp^C: pTXB1-Ssp^C-MxeGyrA- His₆

[00626] **MVKVIGRRSLGVQRIFDIGLPQDHNFL**ANGAIAANCITGDALVALPEGE
SVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTAN
HPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
LVRFL^{**E**}AHHRDPDAQAIAD^{**E**}LT^{**D**}GRFY^{**Y**}AKVASVTDAGVQPVYSLRVDTADHAFITNGF
VSHAHHHHHH (SEQ ID NO:363)

[00627] **Plasmid 4:**

[00628] WT Npu^C: pTXB1-Npu^C-MxeGyrA- His₆

[00629] **MIKIATRKYLGKQNVYDIGVERDHN**FALKNGFIASNCITGDALVALPEG
ESVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTA
NHPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
GLVRFL^{**E**}AHHRDPDAQAIAD^{**E**}LT^{**D**}GRFY^{**Y**}AKVASVTDAGVQPVYSLRVDTADHAFITNG
FVSHAHHHHHH (SEQ ID NO:364)

[00630] All Ssp^N batch mutants were cloned using the QuikChange site directed mutagenesis kit using plasmid 1 as a template and encode for the protein sequences shown below. The N-intein sequence is shown in bold with the residues corresponding to the batch mutation underlined.

[00631] **Plasmid 5:**

[00632] Batch 1: Pet30- His₆-SUMO-AEY-Ssp^N (R73K, L75M, Y79G, L81M)

[00633] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLFLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCL^{**S**}FGTEILTVEYGPL**IGK**IVSEEINCSVYSVDPEGRVYTQAIAQWHD**RGE**
QEVLEYELEDG**S**VIRATSDH**KF**MTT**DG**QMLAIEEIFARQLDLLTLEN**IKQ**TEEALDNHRLPF
PLLDAGTIK (SEQ ID NO:365)

[00634] **Plasmid 6:**

[00635] Ssp^N R73K: Pet30- His₆-SUMO-AEY- Ssp^N (R73K)

[00636] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAIEYCLSFGEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVLEYELEDGSGVIRATSDHKFLTTDYQLLAIEEIFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:366)

[00637] **Plasmid 7:**

[00638] Ssp^N R73K Y79G: Pet30- His₆- SUMO -AEY- Ssp^N (R73K, Y79G)

[00639] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAIEYCLSFGEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVLEYELEDGSGVIRATSDHKFLTTDGQLLAIEEIFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:367)

[00640] **Plasmid 8:**

[00641] Ssp^N R73K Y79G L81M: Pet30- His₆-SUMO-AEY- Ssp^N (R73K, Y79G, L81M)

[00642] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAIEYCLSFGEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVLEYELEDGSGVIRATSDHKFLTTDGQMLAIEEIFARQLDLLTLENIKQTEEALDNHRLPFP
PLLDAGTIK (SEQ ID NO:368)

[00643] **Plasmid 9:**

[00644] Batch 2: Pet30- His₆-SUMO-AEY- Ssp^N (L56F, S70K, A83P, E85D)

[00645] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAIEYCLSFGEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVFEYELEDGSGVIRATKDHRFLTTDYQLLPIDEIFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:369)

[00646] **Plasmid 10:**

[00647] Ssp^N A83P: Pet30- His₆-SUMO-AEY- Ssp^N (A83P)

[00648] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRFYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAIEYCLSFGEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE

QEVLEYELEDGSVIRATSDHRFLTTDYQLLP^{IEE}IFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:370)

[00649] **Plasmid 11:**

[00650] Ssp^N S70K A83P: Pet30- His₆-SUMO-AEY- Ssp^N (S70K, A83P)

[00651] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVLEYELEDGSVIRATK^DHHRFLTTDYQLLP^{IEE}IFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:371)

[00652] **Plasmid 12:**

[00653] Ssp^N L56, S70K, A83P: Pet30- His₆-SUMO-AEY- Ssp^N (L56F, S70K, A83P)

[00654] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDPEGRVYTQAIAQWHDRGE
QEVFEYELEDGSVIRATK^DHHRFLTTDYQLLP^{IEE}IFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:372)

[00655] **Plasmid 13:**

[00656] Batch 3: Pet30- His₆-SUMO-AEY- Ssp^N (S23E, E24K, E25R, N27E)

[00657] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSFGTEILTVEYGPLPIGKIVEK^RIECSVYSVDPEGRVYTQAIAQWHDRGE
QEVLEYELEDGSVIRATSDHRFLTTDYQLLA^{IEE}IFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:373)

[00658] **Plasmid 14:**

[00659] Batch 4: Pet30- His₆-SUMO-AEY- Ssp^N (P35N, E36N, R38N, V39I)

[00660] MGSSHHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSFGTEILTVEYGPLPIGKIVSEEINCSVYSVDN^{NG}N^IYTQAIAQWHDRGE
QEVLEYELEDGSVIRATSDHRFLTTDYQLLA^{IEE}IFARQLDLLTLENIKQTEEALDNHRLPFP
LLDAGTIK (SEQ ID NO:374)

[00661] The four batch mutants (Batches 5-8) and A136S point mutant on the Ssp^C intein were cloned by inverse PCR using Pfu Ultra II HS Polymerase (Agilent) using plasmid 3 as a template and code the protein sequences shown below:

[00662] **Plasmid 15:**

[00663] Batch 5: pTXB1-Ssp^C-MxeGyrA-His₆ (V103I, V105I, I106A, G107T)

[00664] **MIKIATRRSLGVQRIFDIGLPQDHNFLLANGAIAANCITGDALVALPEGE**
SVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTAN
HPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
LVRFLFAHHRDPDAQAIADDELTDGRFYYAKVASVTDAGVQPVYSLRVDADHAFITNGF
VSHAHHHHHH (SEQ ID NO:375)

[00665] **Plasmid 16:**

[00666] Batch 6: pTXB1- Ssp^C -MxeGyrA-His₆ (R115N, I116V, F117Y)

[00667] **MVKVIGRRSLGVQNVYDIGLPQDHNFLLANGAIAANCITGDALVALPEG**
ESVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTA
NHPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
GLVRFLFAHHRDPDAQAIADDELTDGRFYYAKVASVTDAGVQPVYSLRVDADHAFITNG
FVSHAHHHHHH (SEQ ID NO:376)

[00668] **Plasmid 17:**

[00669] Batch 7 pTXB1- Ssp^C -MxeGyrA-His₆ (L121V, P122E, Q123R)

[00670] **MVKVIGRRSLGVQRIFDIGVERDHNFLLANGAIAANCITGDALVALPEGE**
SVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTAN
HPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
LVRFLFAHHRDPDAQAIADDELTDGRFYYAKVASVTDAGVQPVYSLRVDADHAFITNGF
VSHAHHHHHH (SEQ ID NO:377)

[00671] **Plasmid 18:**

[00672] Batch 8: pTXB1- Ssp^C -MxeGyrA-His₆ (L128A, A130K, A133F)

[00673] **MVKVIGRRSLGVQRIFDIGLPQDHNFAKNGFIAANCITGDALVALPEGE**
SVRIADIVPGARPNSDNAIDLKVLDRHGPNVLADRLFHSGEHPVYTVRTVEGLRVTGTAN
HPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
LVRFLFAHHRDPDAQAIADDELTDGRFYYAKVASVTDAGVQPVYSLRVDADHAFITNGF
VSHAHHHHHH (SEQ ID NO:378)

[00674] **Plasmid 19:**

[00675] Ssp^C A136S: pTXB1- Ssp^C-MxeGyrA-His₆ (A136S)

[00676] **MVKVIGRRSLGVQRIFDIGLPQDHNFLLANGAIA_SNCITGDALVALPEGE**
SVRIADIVPGARPNSDNAIDLKVLDRHGNPVLADRLFHSGEHPVYTVRTVEGLRVTGTAN
HPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGVPG
LVRFLAEHHRDPDAQAIADELTDGRFYYAKVASVTDAGVQPVYSLRVDTADHAFITNGF
VSHAHHHHHH (SEQ ID NO:379)

[00677] The gene for the fused Consensus DnaE sequence was codon-optimized for *E. coli* expression through IDT DNA and purchased as a gBlock. The DNA gBlock sequence is shown below:

[00678] **TGCCTGTCTTACGACACAGAGATTCTGACCGTTGAATATGGATTCCTTCC**
TATCGGTAAGATCGTGGAGGAACGGATTGAATGCACAGTCTATACGGTAGATAAAAA
TGGCTTTGTGTATACACAACCTATTGCTCAGTGGCATAACCGGGGAGAACAGGAAGTT
TTCGAATACTGCTTAGAAGACGGTTCGATTATCCGTGCAACGAAAGATCACAAATTTA
TGACGACCGACGGTCAGATGTTACCGATTGATGAGATTTTCGAACGGGGGTTAGACCT
GAAACAAGTTGATGGTTTGCCGATGGTCAAGATCATTAGTCGTAAGAGTCTGGGCACT
CAAAACGTCTACGATATTGGAGTAGAAAAAGATCATAATTTTTTGCTGAAGAATGGGC
TGGTGGCCTCTAAC (SEQ ID NO:380)

[00679] The expression plasmid for Cfa^N was cloned using Gibson assembly into **plasmid 1**, yielding a vector coding for the following protein shown below:

[00680] **Plasmid 20:**

[00681] Cfa^N: pET30- His₆-SUMO-AEY-Cfa^N

[00682] **MGSSHHHHHGSGLVPRGSASMSDSEVNQEAKPEVKPEVKPETHINLKVS**
DGSSEIFFKIKKTTPLRRLMEAFKRQKGEMDSLRLYDGIRIQADQTPEDLDMEDNDIIEA
HREQIGGAEYCLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHN
RGEQEVFEYCLEDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID
NO:381)

[00683] The expression plasmid for the Consensus C-intein was cloned using Gibson Assembly into **plasmid 3**, yielding a vector coding for the following gene:

[00684] **Plasmid 21:**

[00685] Cfa^C: pTXB1-Cfa^C-MxeGyrA-H6

[00686] **MVKIISRKSLGTQNVYDIGVEKDHNFLLNGLVASNCITGDALVALPEG**
ESVRIADIVPGARPNSDNAIDLKVLDRHGNPVLADRLFHSGEHPVYTVRTVEGLRVTGTA

NHPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGFARGKPEFAPTTYTVGV
GLVRFLEAHHRDPDAQIADELTDGRFYAKVASVTDAGVQPVYSLRVDTADHAFITNG
FVSHAHHHHHH (SEQ ID NO:382)

[00687] Cfa constructs used for *E. coli* growth screen.

[00688] Cfa plasmids used to screen the dependency of splicing at the +2 position of the C-extein were generated using restriction cloning into a previously generated plasmid² containing a dual expression system of the split aminoglycoside phosphotransferase (Kan^R) gene. The Cfa dual expression construct is shown below:

[00689] **Plasmids 22-25**

[00690] [KanR promoter]-[RBS]-[KanR^N]-[Cfa^N]-[iRBS]-[Cfa^C]-[CXN-KanR^C]

[00691] Following the promoter sequence, there are two separate *E. coli* ribosomal binding sites in this vector (RBS and iRBS). Each RBS is followed by one half of the split KanR-Intein construct, whose protein sequences are shown below (the Cfa intein is highlighted in bold).

[00692] KanR^N-Cfa^N:

[00693] MEQKLISEEDLSHIQRETSCSRPRLNSNMDADLYGYKWARDNVGQSGATI
YRLYGKPDAPFLKHKGSVANDVTDEMVRNLNWLTEFMPLPTIKHFIRTPDDAWLLTTA
IPGKTAFQVLEEYPDSGENIVDALAVFLRRLHSIPVCNCPFNSDRVFRLAQAQSRMNGLV
DASDFDDERNGWPVEQVWKEMHKLLPF**CLSYDTEILTVEYGFLPIGKIVEERIECTVYT**
VDKNGFVYTQPIAQWHNRGEQEVFEYCLEDGSIIRATKDHKFMTTDQGMLPIDEIFE
RGLDLKQVDGLP (SEQ ID NO:384)

[00694] Cfa^C-KanR^C

[00695] **MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASNCXNSVVTHGDFSL**
DNLIFDEGKLIGCIDVGRVGIADRYQDLAILWNCLGEFSPSLQKRLFQKYGIDNPDMNKLQ
FHLMLDEFF (SEQ ID NO:385)

[00696] The +2 position of the C-extein is underlined, and is either phenylalanine, glycine, arginine, or glutamate.

[00697] αDEC205-HC-Cfa^N

[00698] pCMV Plasmids containing the αDEC205 antibody light chain (LC), heavy chain (HC), and HC-intein fusions (HC-Npu^N, HC-Mcht^N, HC-Ava^N) were obtained as previously described.³ A codon-optimized Cfa DnaE sequence for mammalian cell expression was generated using JCAT⁴ and purchased as a gBlock through IDT DNA. The sequence is shown below:

[00699] TGCCTGAGCTACGACACCGAGATCCTGACCGTGGAGTACGGCTTCCTGC
CCATCGGCAAGATCGTGGAGGAGCGCATCGAGTGCACCGTGTACACCGTGGACAAGA
ACGGCTTCGTGTACACCCAGCCCATCGCCCAGTGGCACAACCGCGGGCGAGCAGGAGG
TGTTGAGTACTGCCTGGAGGACGGCAGCATCATCCGCGCCACCAAGGACCACAAGT
TCATGACCACCGACGGCCAGATGCTGCCCATCGACGAGATCTTCGAGCGCGGCCTGG
ACCTGAAGCAGGTGGACGGCCTGCCCCGTGAAGATCATCAGCCGCAAGAGCCTGGGCA
CCCAGAACGTGTACGACATCGGCGTGGAGAAGGACCACAACCTTCCTGCTGAAGAACG
GCCTGGTGGCCAGCAAC (SEQ ID NO:386)

[00700] The mammalian codon-optimized Cfa^N sequence was then cloned into the pCMV HC-Npu^N plasmid using restriction cloning to give a sequence coding for the following protein:

[00701] **Plasmid 26:**

[00702] HC-Cfa^N: pCMV-HC-Cfa^N

[00703] MGWSCHLFLVATATGVHSEVKLLESGGGLVQPGSLRLSCAASGFTFND
YMNWIRQPPGQAPEWLGVIRNKGNGYTTEVNTSVKGRFTISRDNQNILYLQMNSLRAED
TAIYYCARGGPYYYSGDDAPYWGQGMVTVSSATTKGPSVYPLAPGSAAQTNSMVTLG
CLVKGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSETVTCNVAHP
ASSTKVDKKIVPRDCGCKPCICTVPEVSSVFIFPPKPKDVLITLTPKVTCTVVAISKDDPEV
QFSWFVDDDEVHTAQTPREEQFNSTFRSVSELPIMHQDWLNGKEFKCRVNSAAFPAPIE
KTISKTKGRPKAPQVYTIPPPKEQMAKDKVSLTCMITDFFPEDITVEWQWNGQPAENYKN
TQPIMDTDGSYFVYSKLVQKSNWEAGNTFTCSVLHEGLHNHHTKSLSHSPGKASGGC
LSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVEYFC
LEDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLPGHHHHHHG (SEQ ID
NO:387)

[00704] Cfa^C intein for ligation of dendrimer:

[00705] A plasmid containing the Cfa C-intein with a C-extein linker was cloned by inverse PCR into **plasmid 21** and codes for the protein sequence shown below:

[00706] **Plasmid 27:**

[00707] Cfa^C-link: pTXB1-H6-Cfa^C-CFNSGG-MxeGyrA-H6

[00708] **MGHHHHHHSGVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASNCF**
NSGGCITGDALVALPEGESVRIADIVPGARPNSDNAIDLKVLDRHGNPVLADRLFHSGEHP
VYTVRTVEGLRVTGTANHPLLCLVDVAGVPTLLWKLIDEIKPGDYAVIQRSAFSVDCAGF

ARGKPEFAPTTYTVGVPLVRFLEAHHRDPDAQIADELTDGRFYAKVASVTDAGVQP
VYSLRVDTADHAFITNGFVSHAHHHHHH (SEQ ID NO:388)

[00709] The expression and purification protocols of all His₆-SUMO-AEY-Int^N (plasmids 1, 2, 5-14, 20) and Int^C-GyrA-His₆ (plasmids 3,4, 15-19, 21-27) constructs were adapted from previously described methods.¹

[00710] Expression of all His₆-SUMO-AEY-Int^N constructs

[00711] *E. coli* BL21(DE3) cells were transformed with an N-intein plasmid and grown at 37°C in 1L of LB containing 50 µg/mL of kanamycin. Once the culture had reached an OD₆₀₀=0.6, 0.5 mM IPTG was added to induce expression (0.5 mM final concentration, 3hr at 37°C). The cells were pelleted via centrifugation (10,500 rcf, 30 min) and stored at -80°C.

[00712] Purification of all His₆-SUMO-AEY-Int^N Constructs

[00713] Purification of N-intein constructs for batch mutagenesis

[00714] The cell pellets (from expression of plasmids 1,2, 5-14) were resuspended in 30 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, pH 8.0) containing Roche Complete protease inhibitor cocktail. The resuspended cells were then lysed by sonication on ice (35% amplitude, 8 x 20 second pulses on / 30 seconds off). The insoluble inclusion body containing the N-intein was recovered by centrifugation (35,000 rcf, 30 min). The supernatant was discarded and the pellet was resuspended in 30 mL of Triton wash buffer (lysis buffer with 0.1% triton X-100) and incubated at room temperature for 30 minutes. The Triton wash was then centrifuged at 35,000 rcf for 30 minutes. The supernatant was discarded, the inclusion body pellet was resuspended in 30 mL of lysis buffer containing 6M Urea, and the suspension was incubated overnight at 4°C to extract and resolubilize the protein. This mixture was then centrifuged at 35,000 rcf for 30 minutes.

[00715] The supernatant was then mixed with 4 mL of Ni-NTA resin (for affinity purification using the His₆ tag) and incubated at 4°C for 30 minutes to batch bind the protein. This mixture was loaded on a fritted column, the flow through was collected, and the column was washed with 5 column volumes (CV) of lysis buffer with 6M Urea and 5 CV of lysis buffer with 25 mM imidazole and 6M urea. The protein was then eluted in four 1.5 CV fractions of lysis buffer with 250 mM imidazole and 6M Urea. The first two elution fractions were generally found by SDS-PAGE (12% Bis-Tris gel, run for 50 minutes at 170V) to contain the expressed protein and were combined for refolding.

[00716] The N-inteins were refolded by stepwise dialysis into lysis buffer with 0.5 mM DTT at 4°C. This refolded protein was then treated with 10 mM TCEP and Ulp1 protease (overnight, RT)

to cleave the His₆-SUMO expression tag. The solution was then mixed with 4 mL Ni-NTA resin and incubated for 30 minutes at 4°C. The slurry was applied to a fritted column and the flow through was collected together with a 3 CV wash with lysis buffer. The protein was then treated with 10 mM TCEP, concentrated to 10 mL, and further purified by size exclusion chromatography using an S75 16/60 gel filtration column employing degassed splicing buffer (100 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, pH 7.2) as the mobile phase. Fractions were analyzed by SDS-PAGE, analytical RP-HPLC, and ESI-MS. Pure protein was stored by flash-freezing in liquid N₂ following the addition of glycerol (20 % v/v). Note: during the refolding step, significant protein precipitation was observed for Batch 3, suggesting it is prone to aggregation.

[00717] Purification of Cfa^N:

[00718] The cell pellet (from expression of plasmid 20) was first resuspended in 30 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, pH 8.0) containing the Roche Complete protease inhibitor cocktail. The cells were then lysed by sonication (35% amplitude, 8 x 20 second pulses on / 30 seconds off), and the lysate was pelleted by centrifugation (35,000 rcf, 30 min). The supernatant was incubated with 4 mL of Ni-NTA resin for 30 minutes at 4°C to enrich for the soluble Cfa^N protein. The slurry was then loaded onto a fritted column, and the column was washed with 20 mL of wash buffer 1 (lysis buffer) followed by 20 mL of wash buffer 2 (lysis buffer with 25 mM imidazole). Finally, the protein was eluted from the column with 4 x 1.5 CV of elution buffer (lysis buffer +250 mM imidazole).

[00719] The desired protein, which was present in elution fractions 1 and 2 as determined by SDS-PAGE (12% bis-tris gel run in MES-SDS running buffer at 170V for 50 minutes), was then dialyzed into lysis buffer for 4 hours at 4°C. Following dialysis, the protein was treated with 10 mM TCEP and Ulp1 protease overnight at room temperature to cleave the His₆-SUMO expression tag. The solution was then incubated with 4 mL Ni-NTA resin for 30 minutes at 4°C. The slurry was applied to a fritted column and the flow through was collected together with a 3 CV wash with lysis buffer. The protein was then treated with 10 mM TCEP, concentrated to 10 mL, and purified over an S75 16/60 gel filtration column employing degassed splicing buffer (100 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, pH 7.2) as the mobile phase. Fractions were analyzed by SDS-PAGE (12% bis-tris gel run in MES-SDS running buffer at 170V for 60 minutes), analytical RP-HPLC, and ESI-MS. Pure Protein was stored in glycerol (20% v/v) and flash-frozen in liquid N₂.

[00720] Semisynthesis of Int^C-CFN Constructs

[00721] *E. Coli* BL21 (DE3) cells were transformed with the appropriate pTXB1-Int^C-GyrA-H₆ plasmid (plasmids 3, 4, 15-19, 21) and grown in 2L of LB media containing ampicillin (100 µg/mL) at 37°C. Once the culture had reached an OD₆₀₀ = 0.6, expression was induced by the addition of IPTG (0.5 mM, 3 hours, 37°C). Cell pellets were harvested by centrifugation (10,500 rcf, 30 min), resuspended in lysis buffer, and lysed by sonication on ice (35% amplitude, 10 x 20 second pulses on / 30 seconds off). The protein in the soluble fraction was isolated by centrifugation (35,000 rcf, 30 min) and then enriched by Ni-NTA purification (4 mL beads, carried out as described for N-intein constructs). Following elution in lysis buffer with 250 mM imidazole, the imidazole was removed by dialysis into fresh lysis buffer. The ligation was then carried out overnight at room temperature with the addition of 10 mM TCEP, the Roche Complete protease inhibitor cocktail, 100 mM MESNa, 5 mM EDTA, and 5 mM CFN-NH₂ (pH 7.0). The ligated Int^C-CFN peptide was acidified with 0.5% TFA and purified via RP-HPLC on a C₁₈ preparative column: Gradient = 10% B for 10 minutes (isocratic) followed by 20-60% B over 60 minutes. The purity of each protein was determined by analytical RP-HPLC and its identity was confirmed by ESI-MS.

[00722] Isolation of Cfa^C-link-MESNa

[00723] The Cfa^C-link-MESNa peptide used for the semisynthesis of the Intein-dendrimer fusion was expressed and purified exactly as described above for the Int^C-CFN constructs (expression from plasmid 27). However, no tripeptide was added during the final ligation step, instead resulting in thiolysis of the intein and formation of an α-thioester. This Cfa^C-MESNa α -thioester was then purified by preparative RP-HPLC. Fractions were analyzed by ESI-MS, combined, and lyophilized.

[00724] Analysis of protein *trans*-splicing by RP-HPLC and ESI-MS for Batch Mutants.

[00725] Splicing reactions were carried out as adapted from a previously described protocol.¹ Briefly, N- and C-inteins (15 µM Int^N, 10µM 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 minutes. All splicing reactions were carried out at 30°C unless otherwise indicated. Splicing reactions comparing the tolerance of Npu and Cfa to chaotropic agents were carried out with the indicated concentration of either Urea or guanidine hydrochloride. Splicing was initiated by mixing equal volumes of N- and C- inteins with aliquots removed at the indicated times and quenched by the addition of 8M guanidine hydrochloride, 4% TFA (3:1 v/v). For all splicing reactions containing either Npu^C-CFN or Cfa^C-CFN, reaction progress was monitored by RP-HPLC. For all splicing reactions containing Ssp^C-CFN, reaction progress was monitored by ESI-MS (samples desalted with ZipTip prior to injection) due to poor chromatographic resolution of each state as seen previously.¹

Splicing for Batch 3 and for Cfa at 80°C (15 minute preincubation) were both observed to be inefficient, reaching ~50% completion. This is likely due to aggregation (and inactivation) of the N-intein. Note, shorter preincubations of Cfa at 80°C led to more efficient splicing.

[00726] Kinetic analysis of trans-splicing reactions of Batch Mutants:

[00727] Kinetic analysis was carried out as previously described.¹ Briefly, five species (1-5) are separated by RP-HPLC, and peak areas are determined. For ESI-MS, peak areas are calculated for species 1-4. Each individual peak was normalized against the total area of all peaks combined and reaction progress curves were plotted (n=3). The data were then fit in ProFit to the analytical solution to the coupled differential rate equation for the three state kinetic splicing model. Because the starting material cannot be separated from the linear thioester using this assay, the three state kinetic model collapses the binding step and the first two steps of the splicing reaction into one equilibrium. Each splicing reaction was carried out in triplicate with each replicate analyzed separately. The mean and standard deviation for all values (n=3) are reported.

[00728] Kinetic analysis of overall trans-splicing reactions for Npu and Cfa

[00729] All splicing reactions comparing Npu and Cfa were separated by RP-HPLC with peak areas once again calculated using the manufacturer's software. For these reactions, peak areas for the starting material and branched intermediate (species 1 and 2) and product (species 3, 4, 5) were calculated. The data was then fit to the first order rate equation using the GraphPad Prism software.

[00730] $[P](t) = [P]_{max} \cdot (1 - e^{-kt})$

[00731] Where [P] is the normalized intensity of product, [P]_{max} is this value at t=∞ (the reaction plateau), and k is the rate constant (s⁻¹). The mean and standard deviation (n = 3) are reported.

[00732] Generation and refinement of the DnaE Intein Multiple Sequence alignment.

[00733] Homologues of Npu DnaE were identified through a BLAST⁵ search of the NCBI⁶ (nucleotide collection) and JGI⁷ databases using the Npu DnaE protein sequences. This led to the identification of 105 proteins with >60% sequence identity. For N-inteins with long C-terminal tails, the proteins were truncated to 102 residues, the length of Npu. For N-inteins from the JGI database, the point of truncation was determined by the results of the BLAST program (the last residue identified in the Blast search was selected as the truncation point). Next, a multiple sequence alignment (MSA) was generated of the fused sequence (i.e. the N-intein connected to the C-intein) of all 105 inteins in Jalview (Fig. 7A).⁸ To refine the MSA for inteins predicted to splice quickly, all sequences not containing K70, M75, M81, and S136 (the 'accelerator' residues) were removed from the alignment, leaving behind 73 inteins predicted to have fast splicing kinetics (Fig. 7B). The consensus sequence

of this refined alignment of fast inteins (Cfa) was calculated in Jalview by determining the amino acid that appeared most frequently at each position. A consensus residue was not identified at positions 98 and 102 due to lack of homology in the alignment, and thus the consensus sequence was truncated to 101 amino acids and position 98 was fixed to the residue found in Npu DnaE. This consensus sequence was then aligned with Npu DnaE in Jalview to calculate its percent identity. Non-identical residues were mapped onto the crystal structure of Npu DnaE (pdb = 4Kl5) (Fig. 1).

[00734] Figs. 7A and 7B show an alignment and refinement of the DnaE intein family. Fig. 7A shows the multiple sequence alignment (MSA) of the 105 members of the DnaE intein family found from a BLAST search of the JGI and NCBI sequences databases. The locations of the 'accelerator' residues used to filter the alignment are indicated with black arrows. Fig. 7B shows MSA of the 73 DnaE inteins predicted to demonstrate fast splicing kinetics due to the presence of all four accelerator residues.

[00735] *E. coli* Kan^R screen for Cfa extein dependency.

[00736] The protein splicing coupled kanamycin resistance (Kan^R) assay was carried out as previously described.^{2,9} Briefly, a plasmid coding for a fragmented aminoglycoside phosphotransferase fused to a split intein (Cfa) with either F, G, R, or E present at the +2 position of the C-extein (plasmids 22-25) was transformed into DH5α competent cells and grown in starter cultures overnight (LB Broth, 100 µg/mL ampicillin, 18hrs). These cultures were then diluted twenty-fold into a 96 well plate, and *E. coli* growth was measured at various concentrations of kanamycin (2.5, 10, 25, 50, 100, 250, 1000 µg/mL kanamycin with 100 µg/mL ampicillin). The cell optical density at 650 nm (OD₆₅₀) at the 24-hour end point was fit to a dose response curve with variable slope.

$$[00737] \quad OD_{obs} = OD_{min} + \frac{(OD_{Max} - OD_{min})}{1 + 10^{[(\log IC_{50} - \log [Kan]) \cdot HillSlope]}}$$

[00738] Where OD_{min} was fixed to background absorbance at 650 nm. Each assay was carried out in triplicate, fit separately, and IC₅₀ values are reported as the mean and standard deviation of IC₅₀ for these three separate measurements.

[00739] Protein Trans Splicing of Extracted Inclusion body

[00740] *E. coli* inclusion bodies containing His₆-Sumo-Cfa^N expression (plasmid 20) were resuspended and extracted overnight at 4°C in lysis buffer containing 6M urea. Following centrifugation (35,000 rcf, 30 min), the supernatant was removed and the protein enriched by Ni-NTA under denaturing conditions (as described above). However, instead of refolding the protein, trans-

splicing was directly initiated by the addition of Cfa^C-CFN (10 μ M Cfa^C, 2mM TCEP, 2mM EDTA, 2hrs, RT). Reaction progress was monitored by SDS-PAGE.

[00741] α Dec205-HC-Int^N test expression and splicing

[00742] Test Expression of HC-Npu^N, HC-Mcht^N, HC-Ava^N, HC-Cfa^N

[00743] Expression of all mAb constructs was carried out as previously described.³ Briefly, plasmids encoding the α Dec205-LC and the α Dec205-HC-Int^N were co-transfected into HEK293T cells and incubated for 96 hr (5% CO₂). The cells were spun down (5 minutes, 1,000 rcf), 15 μ L of media for each intein fusion was mixed with 5 μ L of 4x loading dye, and run on a 12% Bis-Tris gel in MES-SDS running buffer (170V for 50 minutes). The protein was then analyzed by western blot (transferred to a PVDF membrane, blotting against α Mouse IgG). Expression yield was measured as the amount of HC-Int^N in the media as determined by densitometry. To account for varying cell growth and survival, the yield was normalized using an α -actin blot of the HEK293T cell lysate (5s sonication, 35% amplitude, in 1x loading dye) and then represented relative to the expression of HC-Cfa^N. Four replicates of this test expression were carried out, and the mean was calculated with error represented as the standard deviation.

[00744] Protein trans-splicing in growth media

[00745] Following the 96 hr expression at 37°C of the mAB-Ava^N and mAB-Cfa^N constructs described above, the media was spun down (1,000 rcf, 5 minutes). The supernatant was then mixed with the Cfa^C-CFN peptide (semisynthesis of expressed plasmid 21) and incubated for 2 hours at room temperature (1 μ M Cfa^C-CFN, 2mM TCEP, 2mM EDTA). The splicing reactions were analyzed by SDS-PAGE (12% Bis-Tris run in MES-SDS running buffer at 170V for 50 minutes) followed by western blot (α Mouse IgG).

[00746] Peptide and Dendrimer Synthesis

[00747] **Cys-Gly-Lys(Fluorescein).** This peptide was synthesized by manual addition of reagents on the Rink Amide resin according to a previously published procedure.²



62

[00750] The Trityl protecting group was selectively removed using 1% TFA, 5% TIS in DCM using a total of 30 mL (10x 3 mL) of deprotection cocktail. Thorough washing of the resin with DCM both during and after these cycles ensured the removal of any liberated Trityl species. The resin was also neutralized with 5% DIPEA in DMF before the next coupling was undertaken. The Alloc group was deprotected using 0.1 eq of tetrakis(triphenylphosphine) palladium(0), 20 eq of phenylsilane in DCM for 3x 45 min each. Thorough washing of the resin with DCM during and after these cycles was used, as well as a 5% DIPEA in DMF wash before the next coupling. The glutaric anhydride monomer was used as a preactivated dicarboxylic acid to allow the formation of the thioesters (i.e. to have a free resin-bound carboxylic acid to functionalize). 20 eq of glutaric anhydride and 10 eq of DIPEA (relative to the number of amines to be acylated) was added to the resin and allowed to react for one hour. The resin was then washed and the coupling was repeated to ensure complete reaction of the resin bound primary amines. To form the resin bound thioesters, 30 eq of methyl thioglycolate, 5 eq of PyAOP and 10 eq of DIPEA (relative to the number of carboxylates) in DMF was added to the resin and allowed to react for one hour. The resin was washed with excess DMF and the coupling procedure was repeated an additional two times.

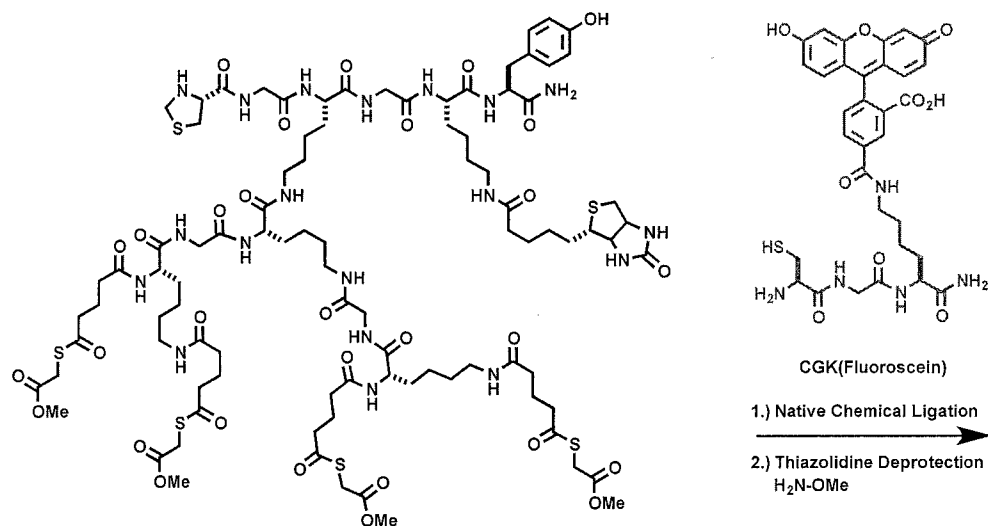
[00751] Cleavage was performed with 95% TFA, 2.5% TIS and 2.5% H₂O for two hours at room temperature. The peptide was then precipitated with diethyl ether, dissolved in water with 0.1% TFA and analyzed via RP-HPLC. The crude material was purified via semi-preparative scale RP-HPLC, and the desired fractions were analyzed, pooled and lyophilized. RP-HPLC characterization: gradient 0-73% B, t_r = 18.4 min. Expected Mass: 2198.86 Da. Found: 2198.82 Da.

[00752] Compound 3 (dendrimer fluorescein).

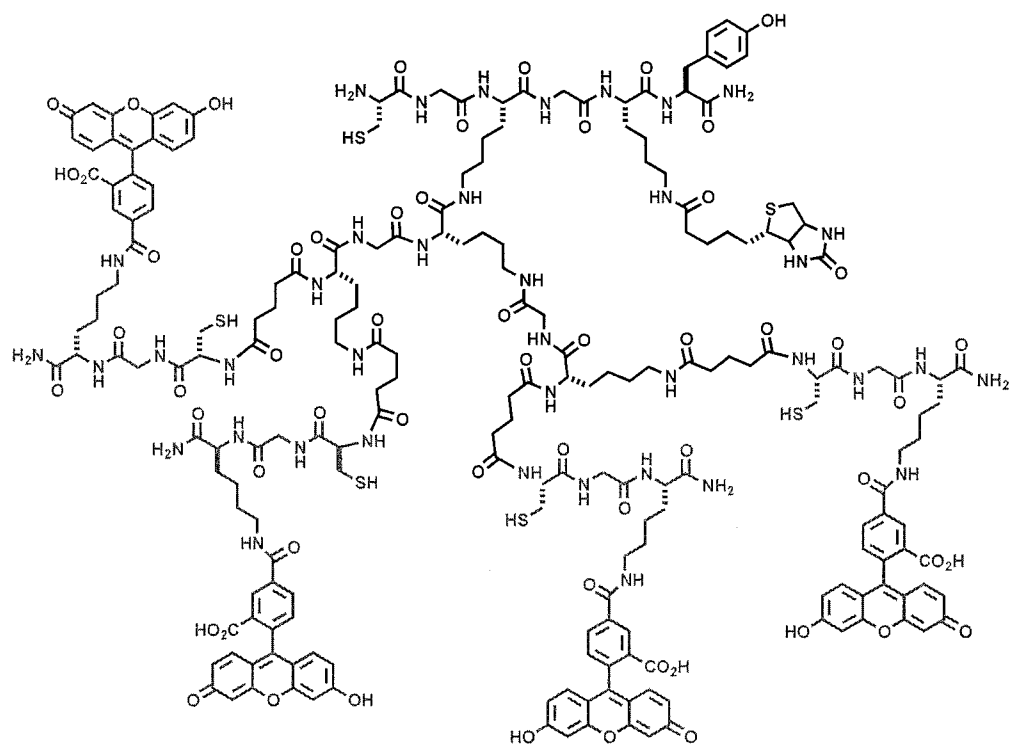
[00753] Compound 3 was synthesized by native chemical ligation (scheme 2). Compound 2 was dissolved in ligation buffer and mixed with five eq. of Cys-Gly-Lys(Fluorescein) (1 mM 2, 5 mM peptide, 4M Guanidine, 100 mM phosphate, 150 mM NaCl, 100mM MPAA, 20 mM TCEP, pH 7.0) and allowed to react overnight at room temperature. Deprotection of the thiazolidine was then accomplished by the addition of 0.1M methoxyamine (final concentration) and decreasing the pH of the ligation buffer to 4.0 (overnight, RT).

[00754] When attempting to purify compound 3 by RP-HPLC, we noticed that it displayed poor solubility when acidified and diluted in water. However, Cys-Gly-Lys(Fluorescein), MPAA, and methoxyamine all remained in solution. From this observation, we purified 3 by selective precipitation following 10-fold dilution in water with 0.1% TFA. The precipitated powder was isolated

by centrifugation (17,000 rcf, 5min), and then redissolved (100 mM phosphate, 150 mM NaCl, pH 7.2) to wash away any remaining contaminants. Once again, the solution was precipitated by acidification and isolated by centrifugation (17,000 rcf, 5 min). This isolated powder was then lyophilized. Expected mass: 4417.8 Da. Found: 4417.5 Da.



Compound 2



Compound 3

[00755] **Supplemental Scheme 2.** Native chemical ligation was used to elaborate tetrathioester-containing compound **2** with a fluorescent tripeptide. Subsequent deprotection using methoxylamine was used to expose the N-terminal cysteine for further ligations yielded tetra-functionalized dendrimer, compound **3**.

[00756] **Compound 1:** (Cfa^C-Dendrimer)

[00757] Compound **1** was synthesized by expressed protein ligation. Compound **3** was dissolved in ligation buffer and mixed with 1.5 eq of the Cfa^C-MESNa thioester (100 μ M **3**, 150 μ M Cfa^C-MESNa, 4M Guanidine, 100 mM phosphate, 150 mM NaCl, 20 mM TCEP, 100 mM MPAA). The reaction was allowed to proceed overnight at room temperature. The ligated product was then purified by semi-preparative RP-HPLC. Desired fractions were pooled and lyophilized. Expected mass: 9860.8 Da. Found: 9860.3 Da.

[00758] Protein Trans-Splicing of dendrimer with α Dec205 mAb.

[00759] The α Dec205 mAb with Cfa^N fused to its C-terminus was expressed as described above. Following the 96 hr expression, the media was concentrated 10-fold in an Amicon 30K concentrator (0.5 mL). Compound **1** was dissolved in splicing buffer (100 mM phosphate, 150 mM NaCl, 1 mM EDTA, pH 7.2) and then mixed with the concentrated media (2 μ M compound **1**, 2 mM TCEP, 1 mM EDTA) and the reaction allowed to proceed for 2 hrs at room temperature. The splicing mixture was then analyzed by SDS-PAGE (12% Bis-Tris run in MES-SDS running buffer at 170V for 50 minutes) and imaged on a fluorescence imager. This was followed by transfer to a PVDF membrane and western blot analysis (α Mouse IgG).

[00760] The invention allows for the formation of various complexes between a split intein fragment and a compound. Several such complexes and compounds are illustrated in the table of Fig. 11. IntC is a split intein fragment, for example, a split intein C-fragment. For example, the dendrimer can have the form of Compound 2, Compound 3, or portions of these. For example, the cargo can be a dye (e.g., fluorescein), another marker molecule, a drug (e.g., a cytotoxic molecule, such as used in the treatment of cancer), or a nucleotide. For example, the polypeptide can be a wholly or partially synthetic or a naturally occurring polypeptide or portion thereof. A dendrimer can be a molecule having a branched chemical structure onto which one or more "cargo" molecules can be "loaded". A "cargo" molecule can be a synthetic or naturally occurring molecule. The cargo molecule can be structured to have no free 1,2-amino thiols or 1,2-amino alcohols. When the intein is bonded

through an amino thiol or amino alcohol to a polypeptide, as shown in row 3 of the table of Fig. 11, the complex formed can be considered to be a recombinant fusion protein.

[00761] *Example 2*

[00762] A major caveat to splicing-based methods is that all characterized inteins exhibit a sequence preference at extein residues adjacent to the splice site. In addition to a mandatory catalytic Cys, Ser, or Thr residue at position +1 (i.e., the first residue within the C-extein), there is a bias for residues resembling the proximal N- and C-extein sequence found in the native insertion site. Deviation from this preferred sequence context leads to a marked reduction in splicing activity, limiting the applicability of PTS-based methods.^{23, 24} Accordingly, there is a need for split inteins whose activities are minimally affected by local sequence environment. For DnaE inteins, extein sequence preferences are largely confined to the catalytic cysteine at the +1 position and large hydrophobic residues that are preferred at the +2 position.²⁵

[00763] In this example, a “EKD” to “GEP” loop mutation into residues 122-124 of Cfa (Cfa_{GEP}) was engineered and resulted in increased promiscuity at the +2 position of the C-extein in a kanamycin resistance assay (Fig. 9). The EKD → GEP mutation increases the activity of Cfa under a wide range of extein contexts. In addition, it can be reasonably expected that these same (or similar) mutations will increase promiscuity among other members of the DnaE intein family (including Npu and those listed in Figs. 7A and 7B).

[00764] The following sequences represent the engineered inteins:

[00765] The Cfa C-intein with the “GEP” mutation that imparts more “promiscuous” activity according to an embodiment of the invention is:

[00766] VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).

[00767] An example of a fusion intein of the Cfa N-intein and Cfa C-intein with the “GEP” mutation of SEQ ID: 389) is:

[00768] CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGE
QEVFEYCLEDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLPVKIIISRKSLGTQN
VYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 390).

[00769] Fig. 9 shows a schematic and a table showing the increased promiscuity of Cfa_{GEP}. Panel A shows a schematic depicting the PTS-dependent *E. coli* selection system with the Cfa split intein. The kanamycin resistance protein, KanR, is split and fused to N- and C-intein fragments (Cfa^N and Cfa^C). The +2 C-extein residue (red X) is varied in the system. In Panel B, IC₅₀ values for

kanamycin resistance of the Cfa_{EKD} (WT) and Cfa_{GEP} (GEP) inteins with indicated +2 C-extein residue are shown (error = standard error (n = 3)).

[00770] Furthermore, this same tolerance for varying extein sequences was also observed in the cyclization of eGFP in *E. coli* (Fig. 10). The Cfa_{GEP} intein demonstrated improved yields of cyclized product in all unfavorable +2 C-extein contexts tested (Fig. 10 panel A, Fig. 10 panel B). In addition, Cfa_{GEP} maintains this improved cyclization activity even when the -1 and +3 extein positions are varied (Fig. 10 panel C, Fig. 10 panel D). This engineered “GEP” loop sequence, which has not been identified in a wild type naturally split DnaE intein, should thus expand the breadth of proteins and peptides accessible to PTS-based technologies.

[00771] Fig. 10 shows schematics and graphs showing eGFP Cyclization with the Cfa_{GEP} split intein. Panel A is a schematic depicting cyclization of eGFP in *E. coli* with variable residues at the +2 C-extein position (red X). In panel B, the fraction of cyclized eGFP formed after overnight expression in *E. coli* for Cfa_{EKD} (WT) and Cfa_{GEP} (GEP) with the indicated +2 C-extein residue is shown (mean $\pm$ standard deviation, n = 3). Panel C is a schematic depicting the cyclization of eGFP in *E. coli* with variable residues at the +3 C-extein position (blue X) and -1 N-extein position (red X). Panel D shows a fraction of cyclized eGFP formed after overnight expression in *E. coli* for Cfa_{EKD} (WT) and Cfa_{GEP} (GEP) with the indicated +3 C-extein and -1 N-extein residues (mean $\pm$ standard deviation, n = 3).

[00772] The following claims are thus to be understood to include what is specifically illustrated and described above, what is conceptually equivalent, what can be obviously substituted and also what essentially incorporates the essential idea of the invention. Those skilled in the art will appreciate that various adaptations and modifications of the just-described preferred embodiment can be configured without departing from the scope of the invention. The illustrated embodiment has been set forth only for the purposes of example and that should not be taken as limiting the invention. Therefore, it is to be understood that, within the scope of the appended claims, the invention may be practiced other than as specifically described herein.

References

- (1) Shah, N. H.; Muir, T. W. *Chem. Sci.* **2014**, 5, 15.
- (2) Wu, H.; Hu, Z.; Liu, X. Q. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, 95, 9226.
- (3) Iwai, H.; Zuger, S.; Jin, J.; Tam, P. H. *FEBS Lett.* **2006**, 580, 1853.
- (4) Zettler, J.; Schutz, V.; Mootz, H. D. *FEBS Lett.* **2009**, 583, 909.

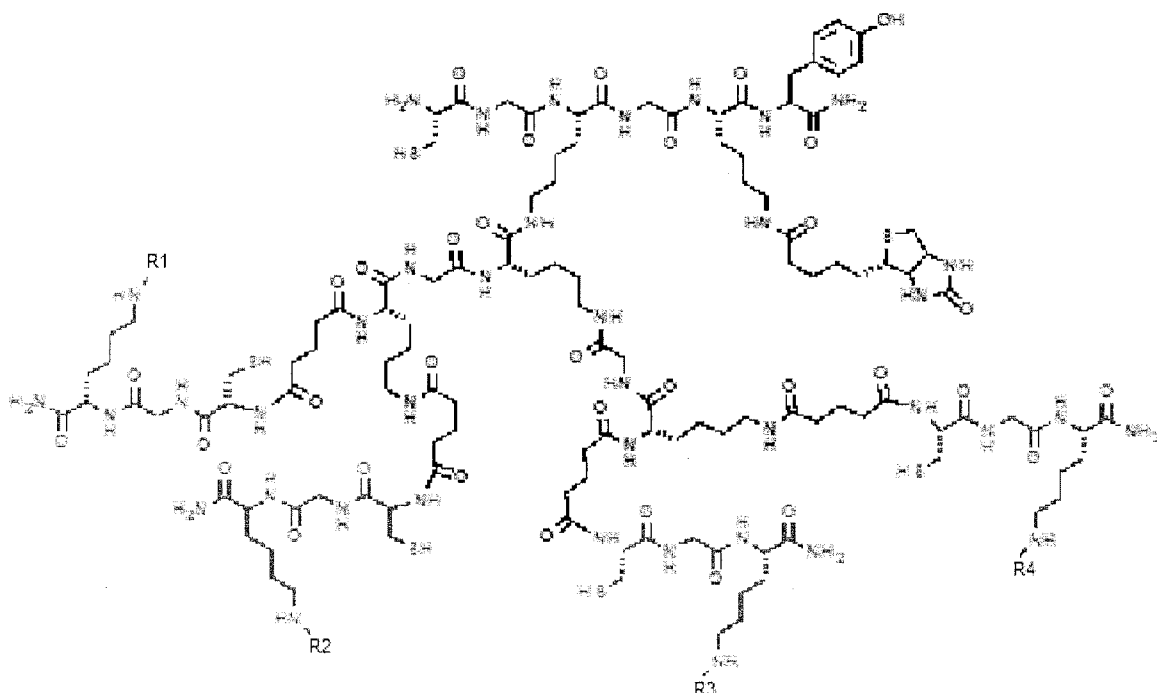
- (5) Shah, N. H.; Eryilmaz, E.; Cowburn, D.; Muir, T. W. *J. Am. Chem. Soc.* **2013**, *135*, 5839.
- (6) Shah, N. H.; Dann, G. P.; Vila-Perello, M.; Liu, Z.; Muir, T. W. *J. Am. Chem. Soc.* **2012**, *134*, 11338.
- (7) Carvajal-Vallejos, P.; Pallisse, R.; Mootz, H. D.; Schmidt, S. R. *J. Biol. Chem.* **2012**, *287*, 28686.
- (8) Wu, Q.; Gao, Z.; Wei, Y.; Ma, G.; Zheng, Y.; Dong, Y.; Liu, Y. *Biochem. J.* **2014**, *461*, 247.
- (9) Aranko, A. S.; Oeemig, J. S.; Kajander, T.; Iwai, H. *Nat. Chem. Biol.* **2013**, *9*, 616.
- (10) Pietrokovski, S. *Protein Sci.* **1994**, *3*, 2340.
- (11) Dearden, A. K.; Callahan, B.; Roey, P. V.; Li, Z.; Kumar, U.; Belfort, M.; Nayak, S. K. *Protein Sci.* **2013**, *22*, 557.
- (12) Du, Z.; Shemella, P. T.; Liu, Y.; McCallum, S. A.; Pereira, B.; Nayak, S. K.; Belfort, G.; Belfort, M.; Wang, C. *J. Am. Chem. Soc.* **2009**, *131*, 11581.
- (13) Lehmann, M.; Kostrewa, D.; Wyss, M.; Brugger, R.; D'Arcy, A.; Pasamontes, L.; van Loon, A. P. *Protein Eng.* **2000**, *13*, 49.
- (14) Steipe, B. *Methods Enzymol.* **2004**, *388*, 176.
- (15) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403.
- (16) Grigoriev, I. V.; Nordberg, H.; Shabalov, I.; Aerts, A.; Cantor, M.; Goodstein, D.; Kuo, A.; Minovitsky, S.; Nikitin, R.; Ohm, R. A.; Otilar, R.; Poliakov, A.; Ratnere, I.; Riley, R.; Smirnova, T.; Rokhsar, D.; Dubchak, I. *Nucleic Acids Res.* **2012**, *40*, D26.
- (17) Tatusova, T.; Ciufu, S.; Fedorov, B.; O'Neill, K.; Tolstoy, I. *Nucleic Acids Res.* **2014**, *42*, D553.
- (18) Shah, N. H.; Eryilmaz, E.; Cowburn, D.; Muir, T. W. *J. Am. Chem. Soc.* **2013**, *135*, 18673.
- (19) Mohlmann, S.; Bringmann, P.; Greven, S.; Harrenga, A. *BMC Biotechnol.* **2011**, *11*, 76.
- (20) Barbuto, S.; Idoyaga, J.; Vila-Perello, M.; Longhi, M. P.; Breton, G.; Steinman, R. M.; Muir, T. W. *Nat. Chem. Biol.* **2013**, *9*, 250.
- (21) Vila-Perello, M.; Liu, Z.; Shah, N. H.; Willis, J. A.; Idoyaga, J.; Muir, T. W. *J. Am. Chem. Soc.* **2013**, *135*, 286.

- (22) Shah, N. D.; Parekh, H. S.; Steptoe, R. J. *Pharm. Res.* **2014**, *31*, 3150.
- (23) Iwai, H.; Zuger, S.; Jin, J.; Tam, P. H. *FEBS Lett.* **2006**, *580*, 1853.
- (24) Amitai, G.; Callahan, B. P.; Stanger, M. J.; Belfort, G.; Belfort, M. *Proc Natl Acad Sci U S A* **2009**, *106*, 11005.
- (25) Cheriyan, M.; Pedomallu, C. S.; Tori, K.; Perler, F. *J Biol Chem* **2013**, *288*, 6202.

What is claimed is:

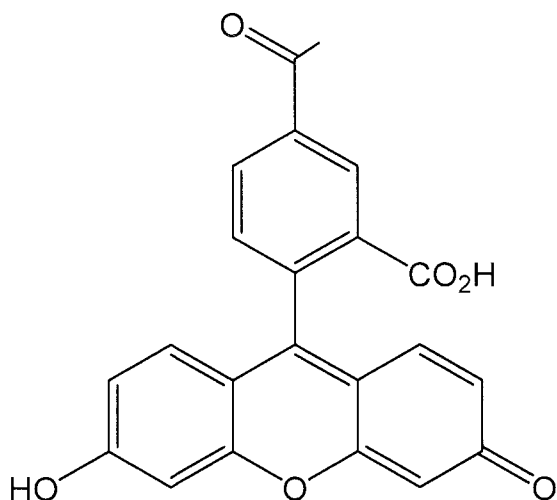
1. A split intein N-fragment comprising an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDGQMLPIDEIFERGL (SEQ ID NO: 1).
2. The split intein N-fragment of claim 1, wherein the amino acid sequence comprises an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDGQMLPIDEIFERGLDLKQVDGLP (SEQ ID NO: 2).
3. A complex comprising the split intein N-fragment of claim 1 or 2 and a compound.
4. The complex of claim 3, wherein the compound is a peptide or a polypeptide.
5. The complex of claim 3, wherein the compound is an antibody chain.
6. The complex of claim 3, wherein the compound is an antibody heavy chain.
7. The complex of claim 3, wherein the compound comprises a peptide, an oligonucleotide, a drug, or a cytotoxic molecule.
8. A split intein C-fragment comprising an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 3).
9. The split intein C-fragment of claim 8, wherein the amino acid sequence of said C-fragment comprises an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to MVKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN (SEQ ID NO: 4).

10. A split intein C-fragment comprising an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to VKIISRKSLGTQNVYDIGVGEPHNFLKNGLVASN (SEQ ID NO: 389).
11. A complex comprising the split intein C-fragment of any one of claims 8, 9, and 10 and a compound.
12. The complex of claim 11, wherein the compound is a dendrimer.
13. The complex of claim 11, wherein the compound is a dendrimer having the structure

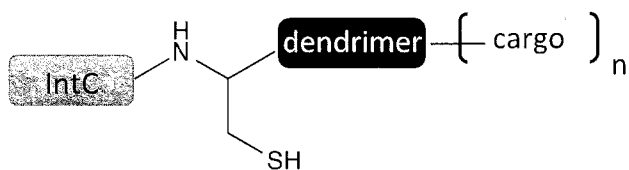


wherein R1, R2, R3, and R4 are independently selected from the group consisting of hydrogen (H) and cargo molecules.

14. The complex of claim 13, wherein R1, R2, R3, and R4 are each a dye molecule.
15. The complex of claim 13, wherein R1, R2, R3, and R4 are each a fluorescein derivative having the structure

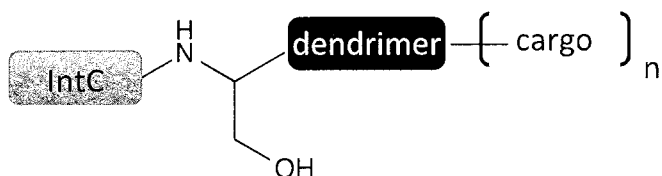


16. The complex of claim 11, wherein the compound is a peptide or a polypeptide.
17. The complex of claim 11, wherein the compound comprises a peptide, an oligonucleotide, a drug, or a cytotoxic molecule.
18. The complex of claim 11, wherein the compound is a 1,2-amino thiol bonded to a peptide, an oligonucleotide, a drug, or a cytotoxic molecule.
19. The complex of claim 11, wherein the compound is a 1,2-amino alcohol bonded to a peptide, an oligonucleotide, a drug, or a cytotoxic molecule.
20. A complex of the structure



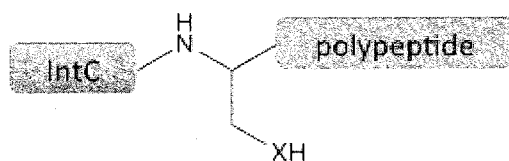
wherein IntC is the split intein C-fragment of any one of claims 8, 9, and 10 and wherein n is from 0 to 8.

21. A complex of the structure



wherein IntC is the split intein C-fragment of any one of claims 8, 9, and 10 and
wherein n is from 0 to 8.

22. A complex of the structure



wherein IntC is the split intein C-fragment of any one of claims 8, 9, and 10 and
wherein X is sulfur (S) or oxygen (O).

23. A composition comprising:

the split intein N-fragment of any one of claims 1 and 2; and
the split intein C-fragment of any one of claims 8, 9, and 10.

24. A nucleotide plasmid comprising a nucleotide sequence encoding for the split intein N-fragment of any one of claims 1 and 2.

25. A nucleotide plasmid comprising a nucleotide sequence encoding for the split intein C-fragment of any one of claims 8, 9, and 10.

26. A method for splicing two complexes comprising:

contacting a first complex comprising a first compound and the split intein N-fragment of any one of claims 1 and 2 and a second complex comprising a second compound and the split intein C-fragment of any one of claims 8, 9, and 10, wherein the contacting is performed under conditions that permit binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and

reacting the intein intermediate to form a conjugate of the first compound with the second compound.

27. A method comprising:
contacting a first complex comprising a first compound and the split intein N-fragment of any one of claims 1 and 2 and a second complex comprising a second compound and the split intein C-fragment of any one of claims 8, 9, and 10, wherein the contacting is performed under conditions that permit binding of the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and
reacting the intein intermediate with a nucleophile to form a conjugate of the first compound with the nucleophile.
28. The method of claim 27, wherein the first compound is a polypeptide or an antibody.
29. The method of claim 27, wherein the second compound is a dendrimer or a polypeptide.
30. An intein comprising an amino acid sequence of at least 90%, 95%, 98%, 99%, or 100% sequence identity to
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYQTQPIAQWHNRGEQEVFEYCL
EDGSIIRATKDHKFMTTDQGMLPIDEIFERGLDLKQVDGLPVKIISRKSLGTQNVYDIGV
EKDHNFLKNGLVASN (SEQ ID NO: 390).
31. A kit for splicing two complexes together comprising:
the split intein N-fragment of any one of claims 1 and 2;
the split intein C-fragment of any one of claims 8, 9, and 10;
a reagent for binding the split intein N-fragment to the split intein C-fragment to form an intein intermediate; and
a nucleophilic agent.
32. A method for generating a synthetic consensus intein peptide sequence comprising:
generating a population of a plurality of homologous intein peptide sequences;

identifying amino acids associated with fast splicing within said population of a plurality of homologous intein peptide sequences;
 generating a subpopulation of a second plurality of homologous intein peptide sequences, wherein said second plurality of homologous intein peptide sequences comprise amino acids associated with fast splicing;
 creating an alignment of at least three peptide sequences of said subpopulation;
 determining a most frequently occurring amino acid residue at each position of said at least three peptide sequences; and
 generating a synthetic consensus intein peptide sequence based on said most frequently occurring amino acid residue at each position of said at least three peptide sequences.

33. A method comprising:
 fusing a first nucleotide sequence encoding an amino acid sequence of the split intein N-fragment of any one of claims 1 and 2
 with a second nucleotide sequence encoding an amino acid sequence of the split intein C-fragment of any one of claims 8, 9, and 10,
 so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein.

34. The method of claim 33, comprising:
 fusing a first nucleotide sequence encoding an amino acid sequence of a split intein N-fragment comprising
 CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRG
 EQEVFEYCLEDGSIIRATKDHKFMTTDGQMLPIDEIFERGL
 (SEQ ID NO: 1)
 with a second nucleotide sequence encoding an amino acid sequence of a split intein C-fragment comprising VKIISRKSLGTQNVYDIGVEKDHNFLKNGLVASN
 (SEQ ID NO: 3),
 so that the fusion of the first nucleotide sequence and the second nucleotide sequence codes for a contiguous intein.

35. A gene fusion comprising:
a first nucleotide sequence encoding an amino acid sequence of the split intein
N-fragment of any one of claims 1 and 2
fused with a second nucleotide sequence encoding an amino acid sequence of the split
intein C-fragment of any one of claims 8, 9, and 10.
36. The gene fusion of claim 35, comprising:
a first nucleotide sequence encoding an amino acid sequence of a split intein N-fragment
comprising
CLSYDTEILTVEYGFLPIGKIVEERIECTVYTVDKNGFVYTQPIAQWHNRG
EQEVFEYCLEDGSIIRATKDHKFMTTDGQMLPIDEIFERGL
(SEQ ID NO: 1)
fused with a second nucleotide sequence encoding an amino acid sequence of a split
intein C-fragment comprising
VKIISRKSLGTQNVYDIGVEKDHNFLLKNGLVASN (SEQ ID NO: 3).

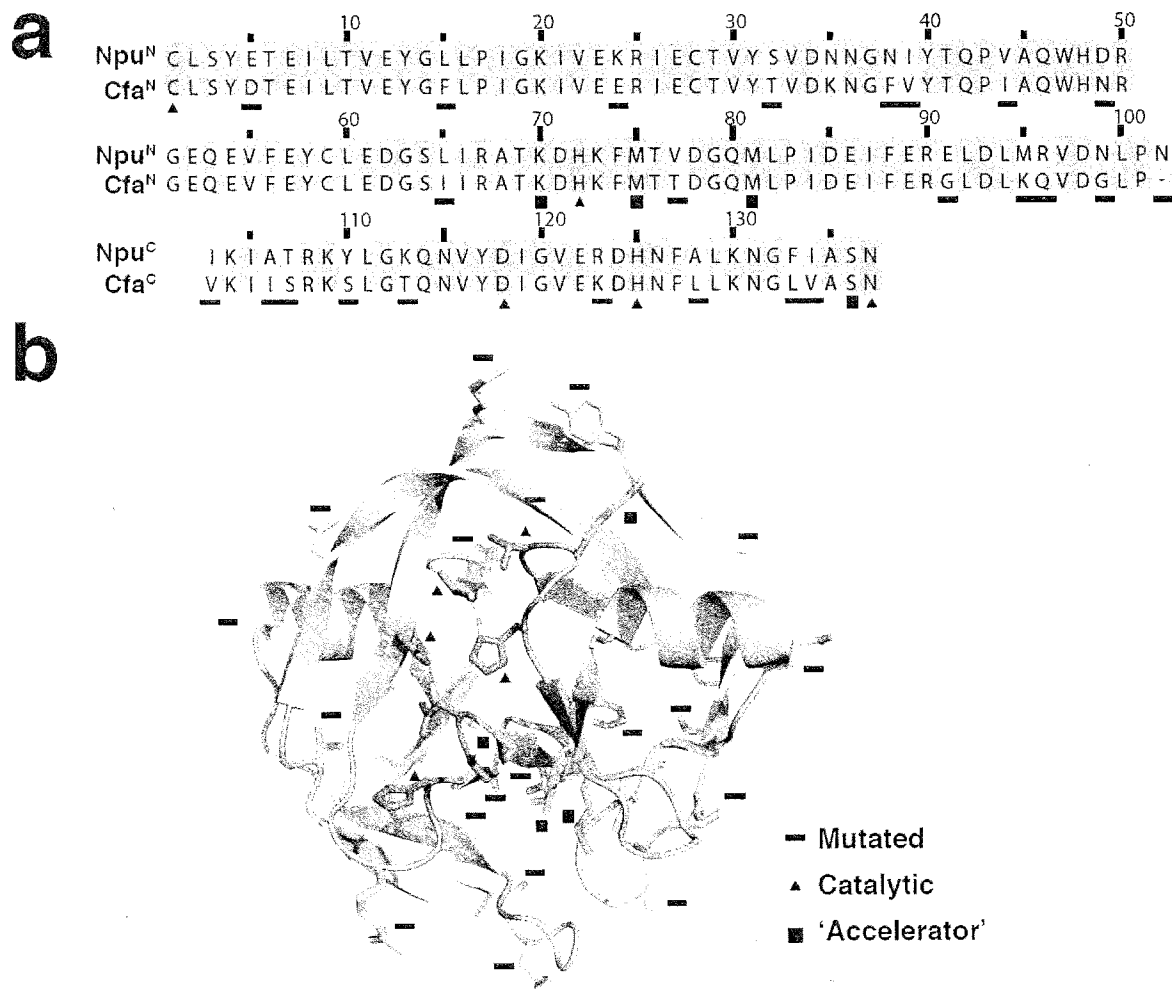


FIG. 1

2/14

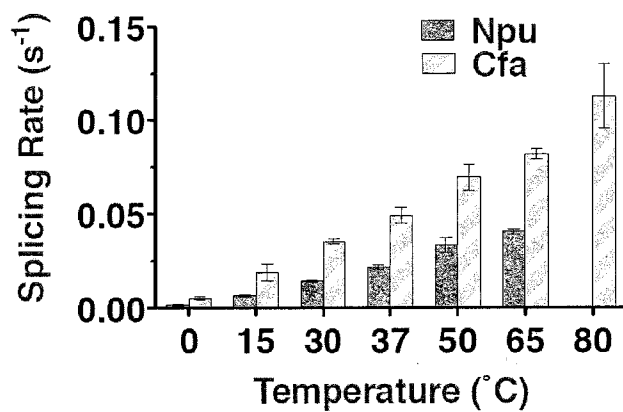
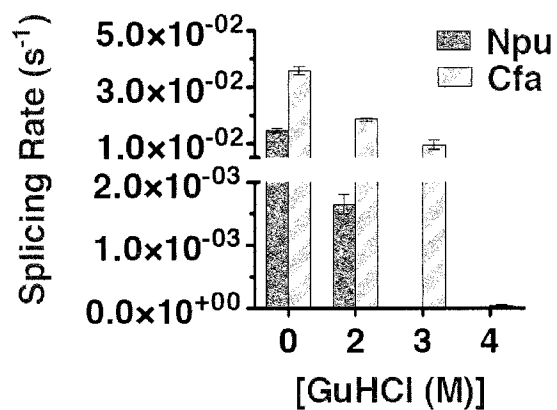
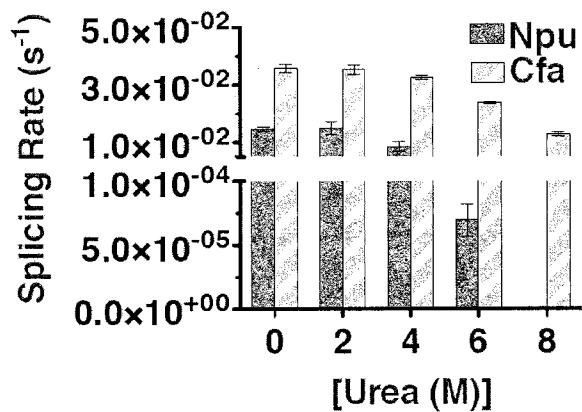
a**b****c**

FIG. 2

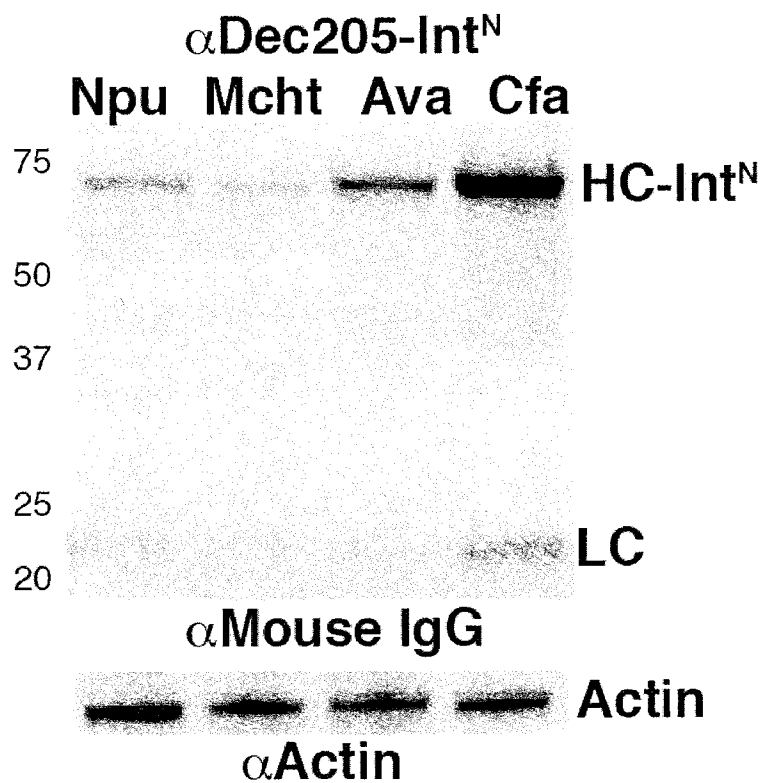
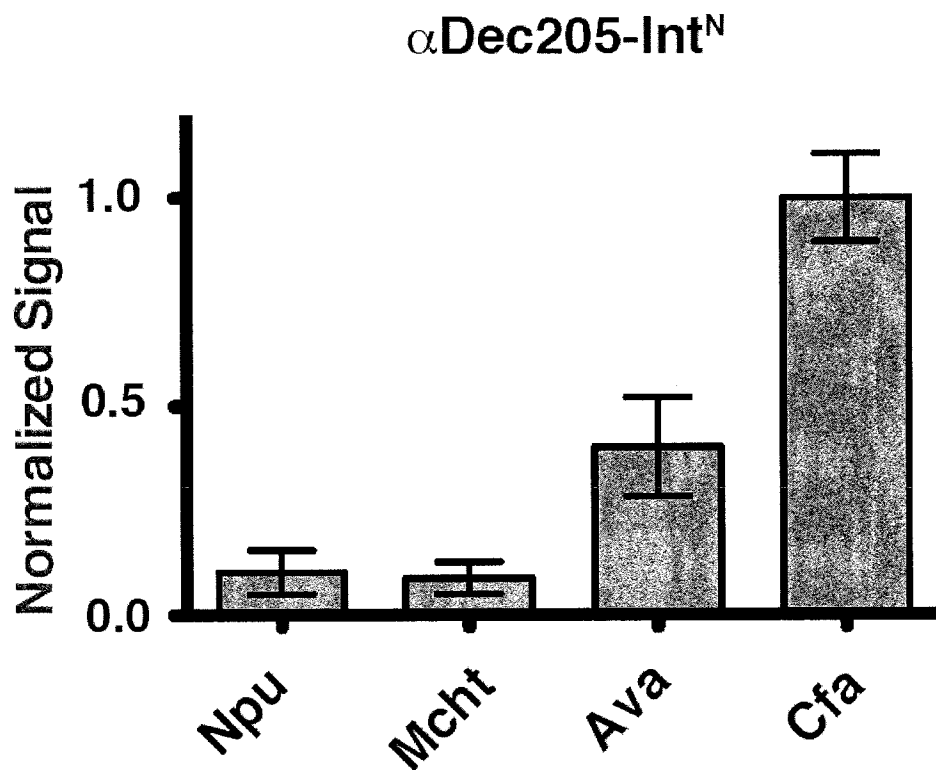
a**b**

FIG. 3

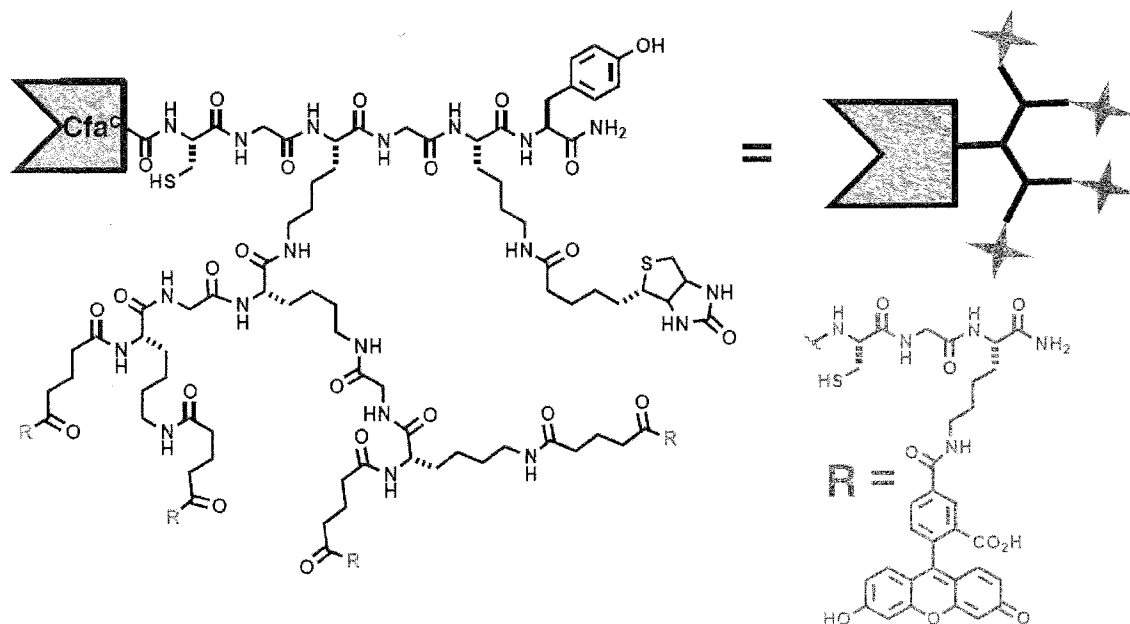
C

FIG. 3

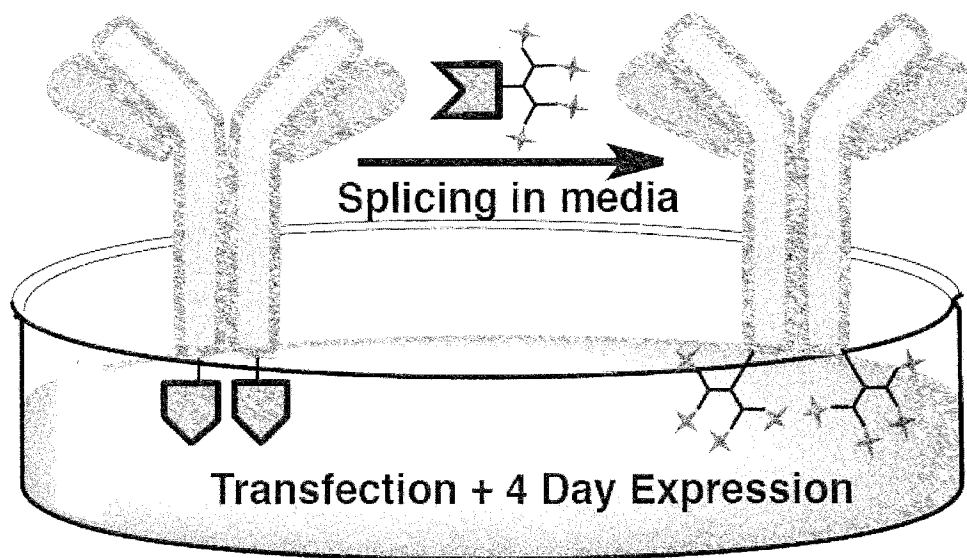
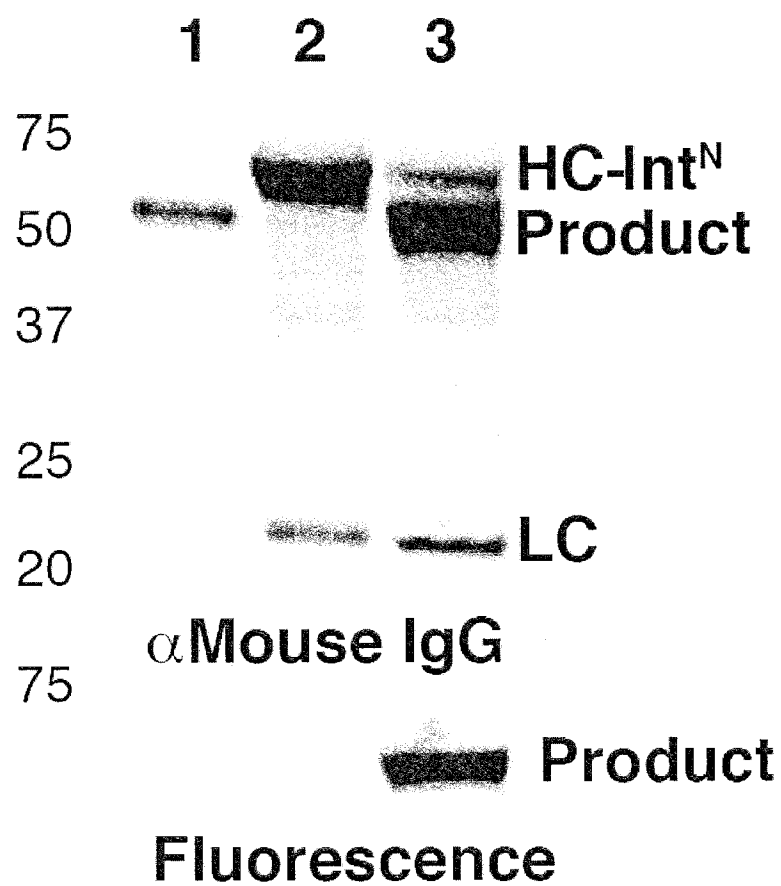
d**e**

FIG. 3

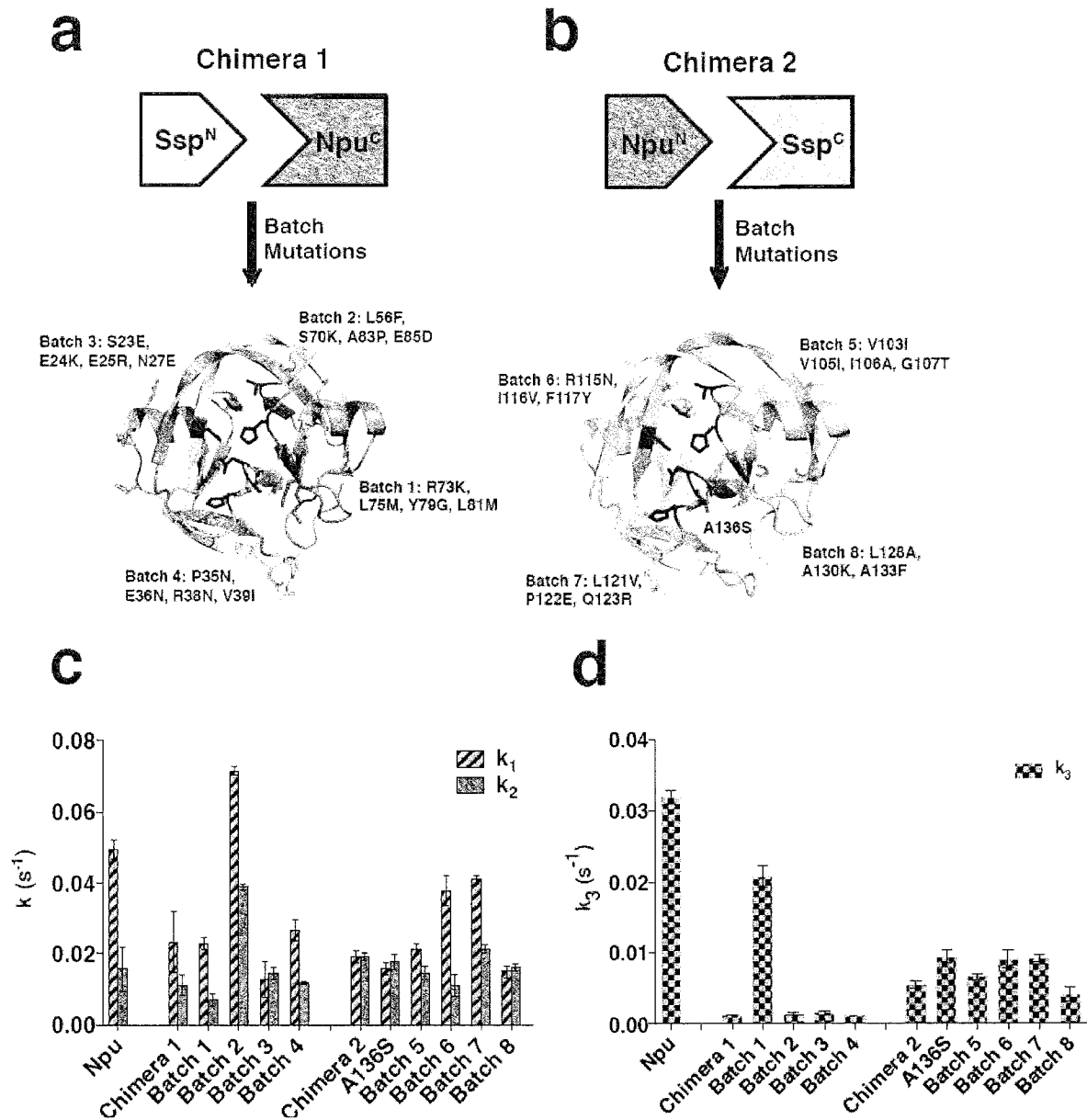


FIG. 4

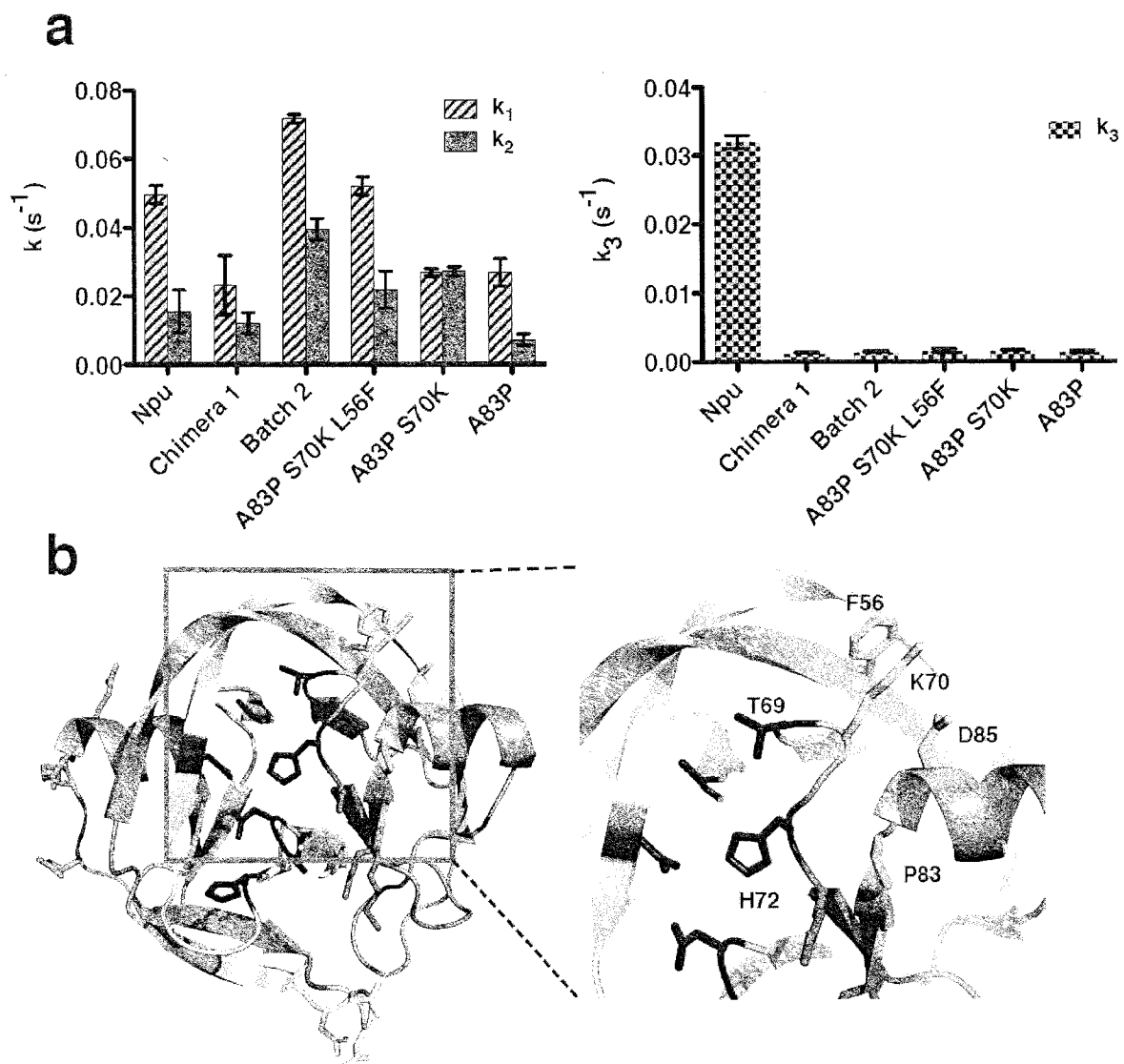


FIG. 5

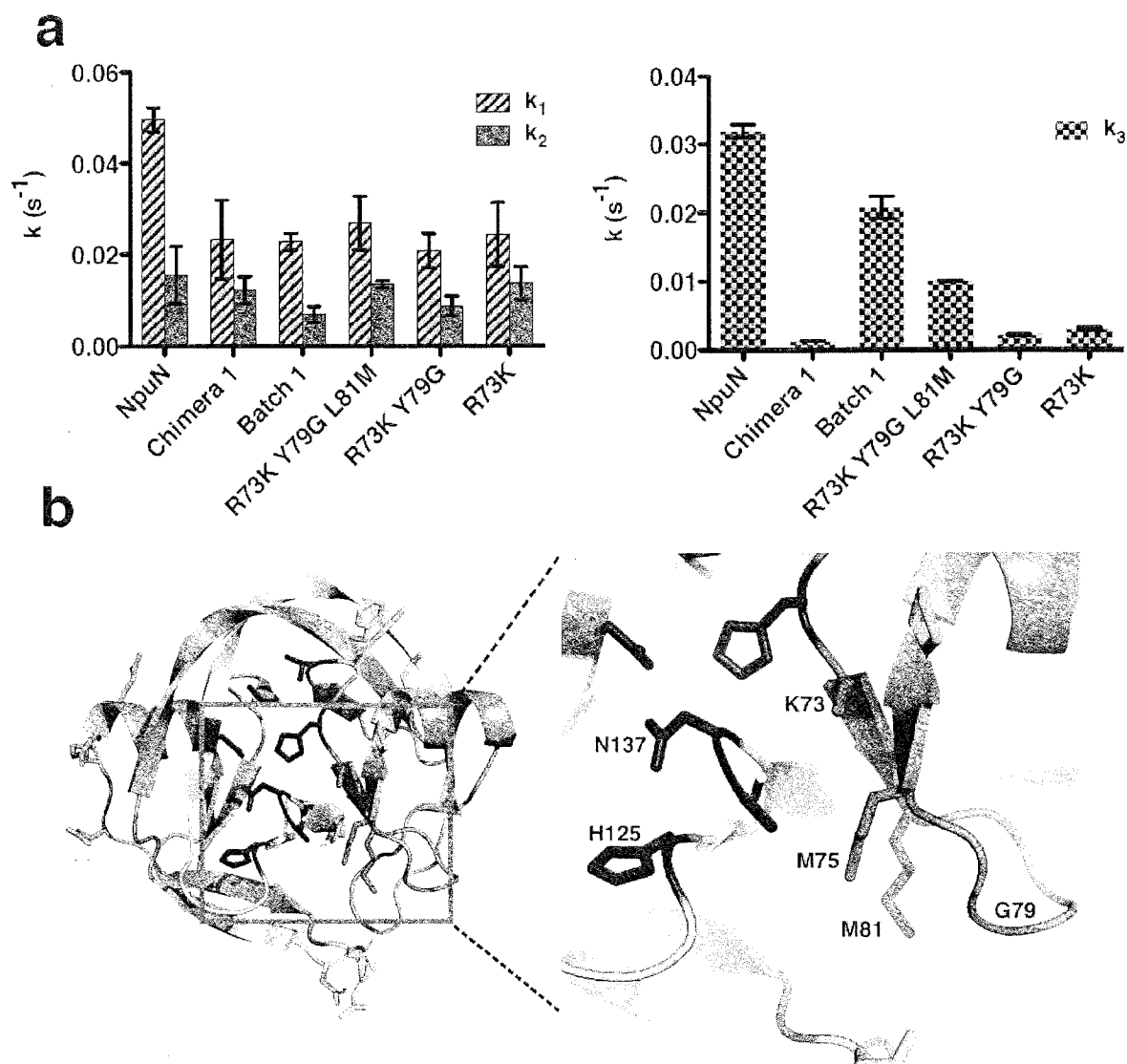


FIG. 6

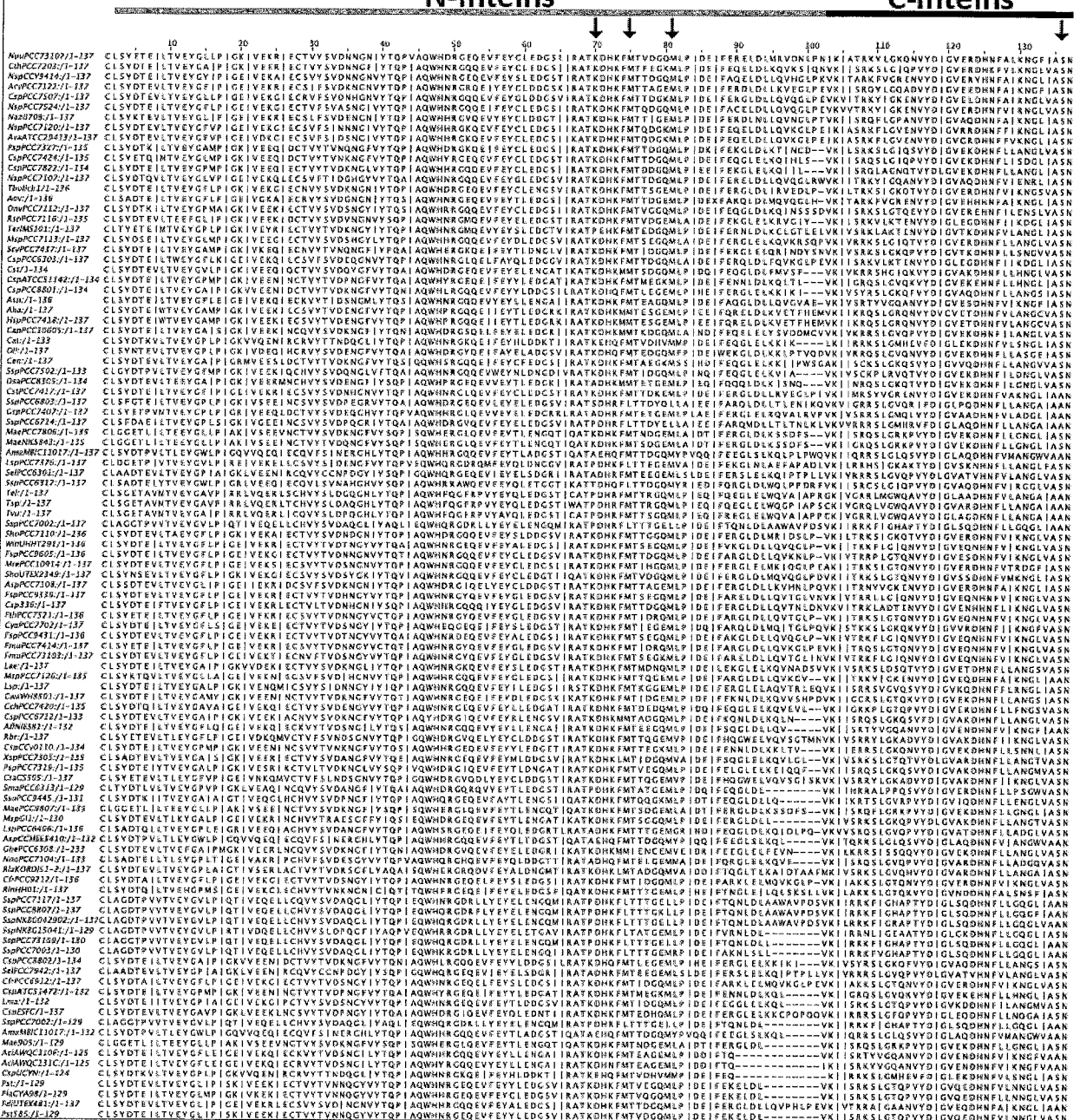


FIG. 7A

C-Inteins

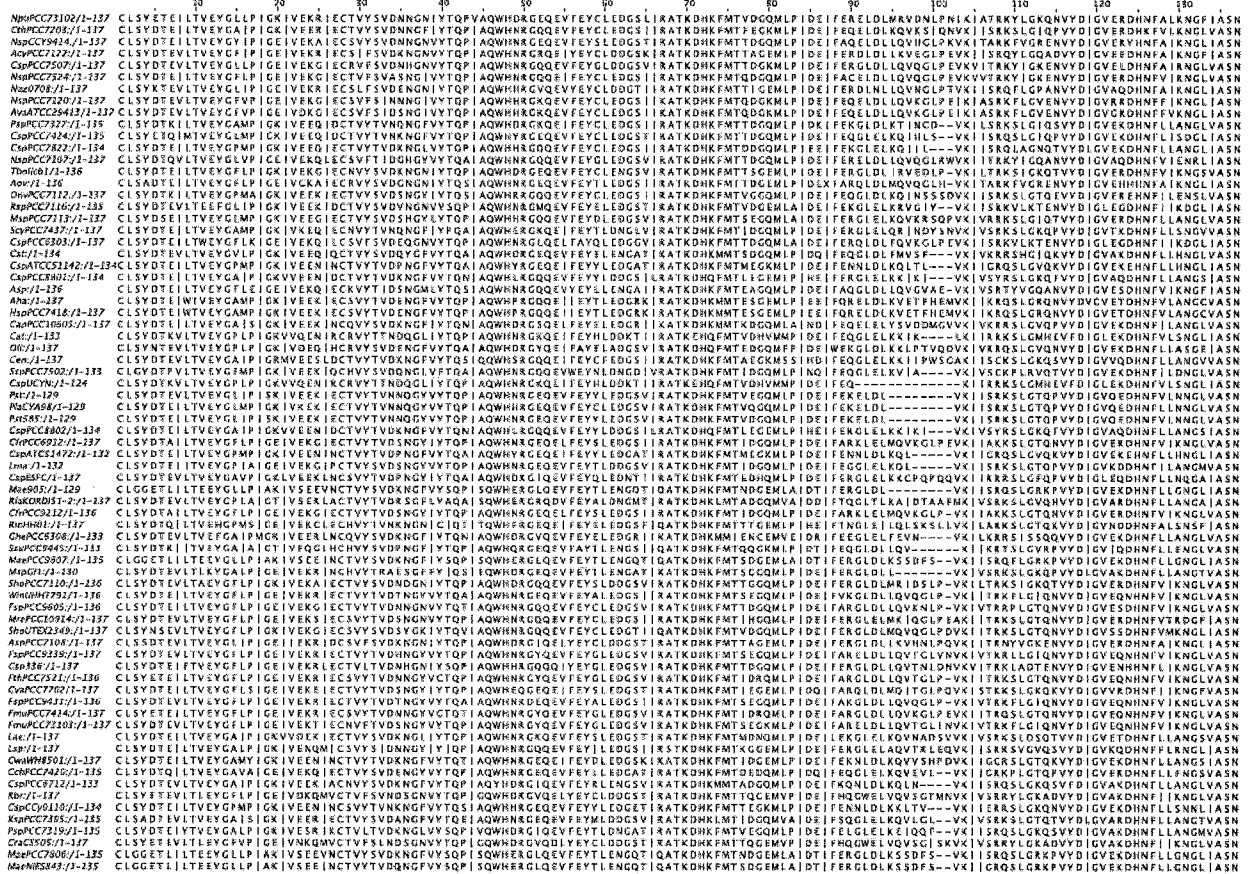


FIG. 7B



FIG. 8

12/14

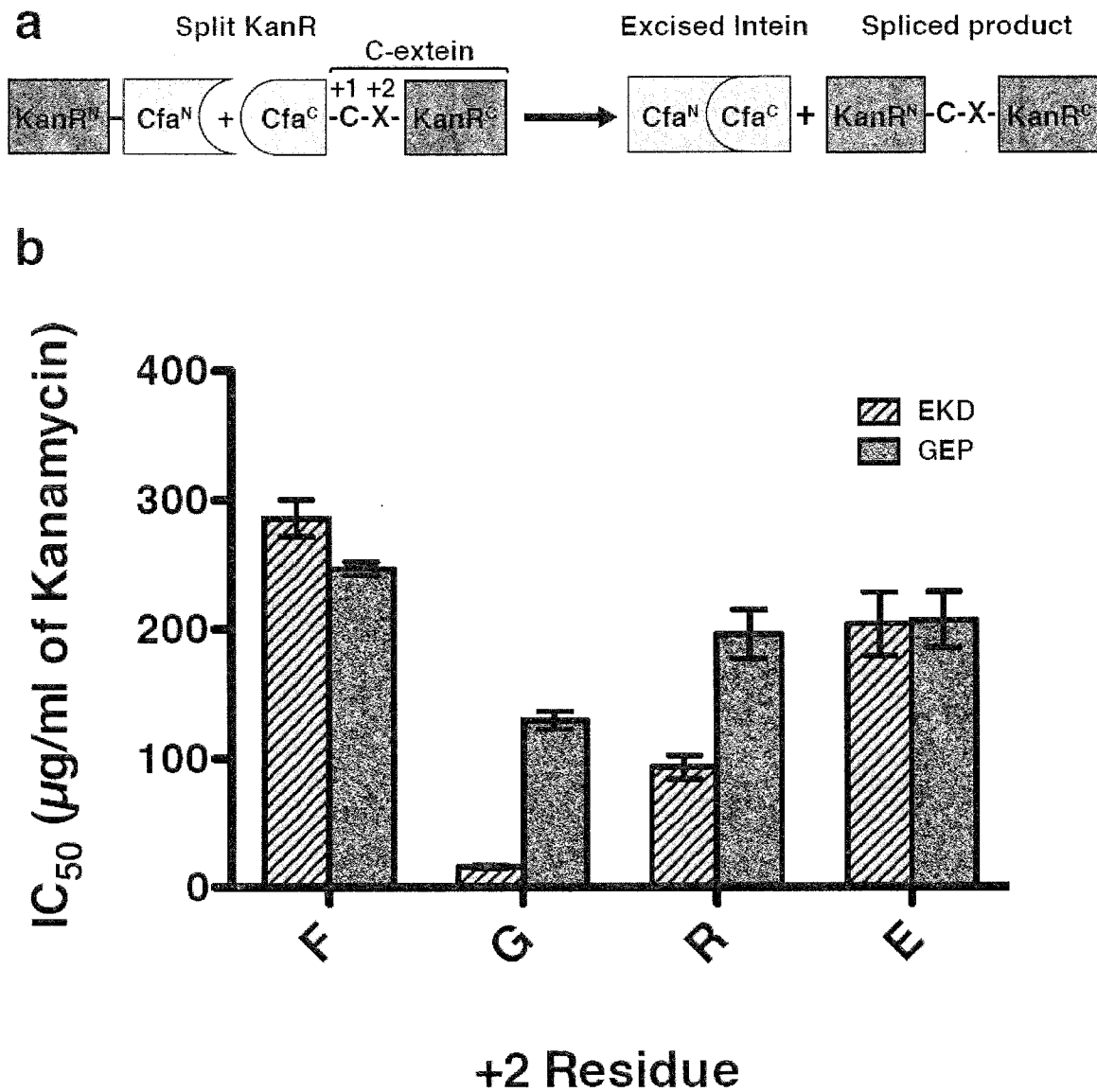


FIG. 9

13/14

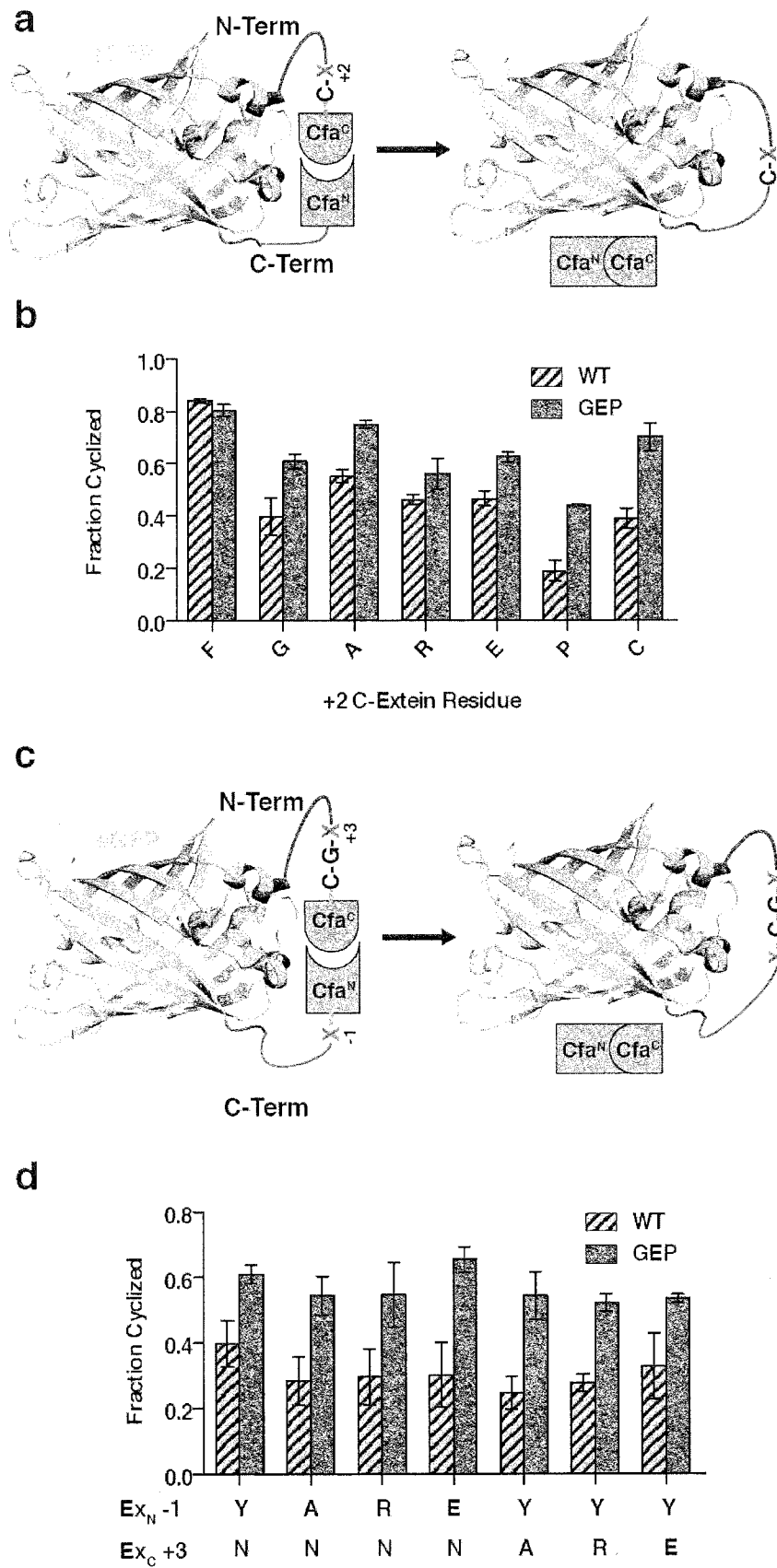


FIG. 10

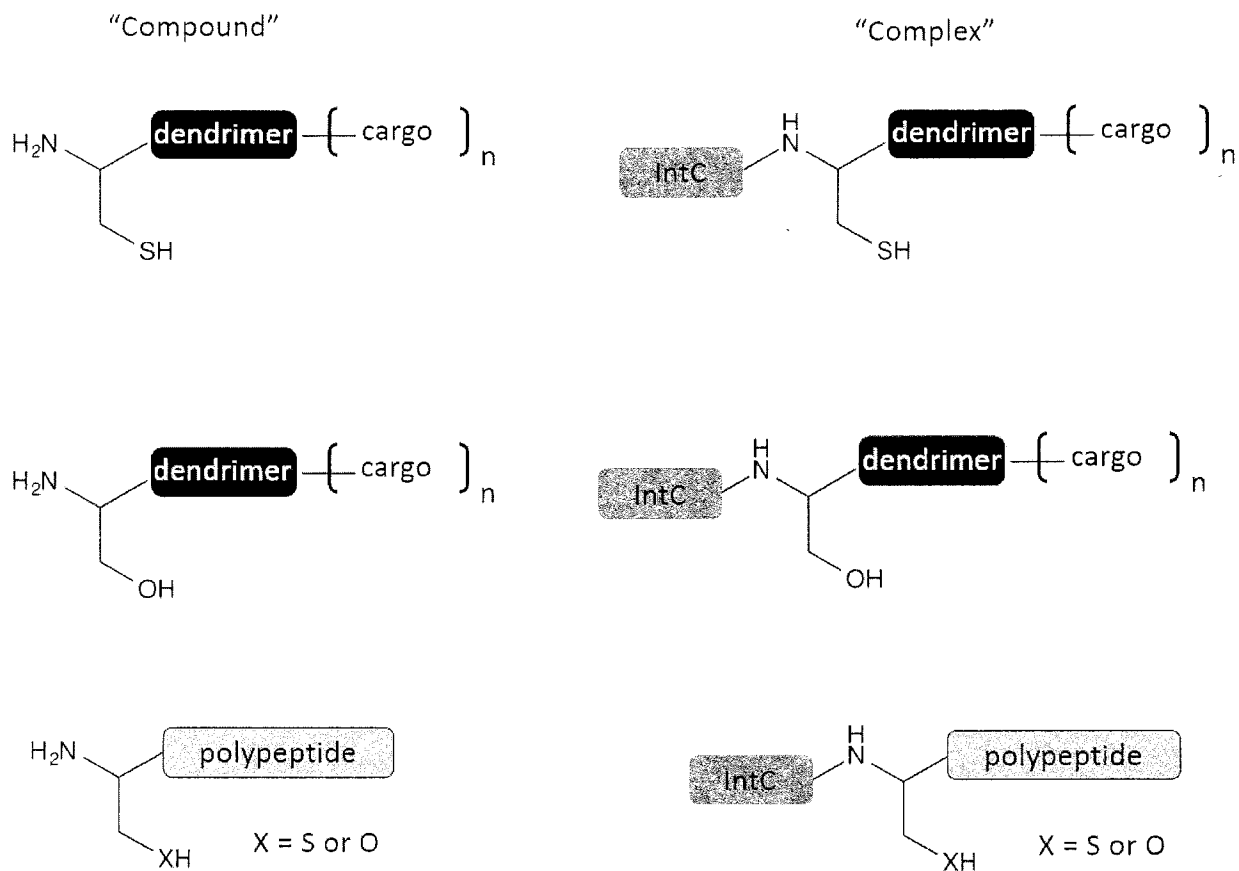


FIG. 11