



(51) International Patent Classification:

C12N 9/42 (2006.01) C12P 7/56 (2006.01)

C12N 9/24 (2006.01) C07K 7/08 (2006.01)

C12N 9/26 (2006.01) A61P 43/00 (2006.01)

C12P 19/14 (2006.01)

(21) International Application Number:

PCT/US2021/035968

(22) International Filing Date:

04 June 2021 (04.06.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/034,788 04 June 2020 (04.06.2020) US

(71) Applicants: **CALIFORNIA INSTITUTE OF TECHNOLOGY** [US/US]; 1200 E. California Blvd., Pasadena, California 91125 (US). **BASF SE** [DE/DE]; Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein (DE).

(72) Inventors: **LISZKA, Michael**; 3550 John Hopkins Court, San Diego, California 92121 (US). **BATZILLA, Alina**; 3550 John Hopkins Court, San Diego, California 92121 (US). **WU, Zachary**; 1200 E California Blvd, MC 6-32,

Pasadena, California 91125 (US). **ARNOLD, Frances**; 1200 E California Blvd, MC 210-41, Pasadena, California 91125 (US).

(74) Agent: **SPIEGLER, Alexander H.** et al.; Floor 42, 1301 Avenue of Americas, New York, New York 10019 (US).

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

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

(54) Title: NOVEL SIGNAL PEPTIDES GENERATED BY ATTENTION-BASED NEURAL NETWORKS

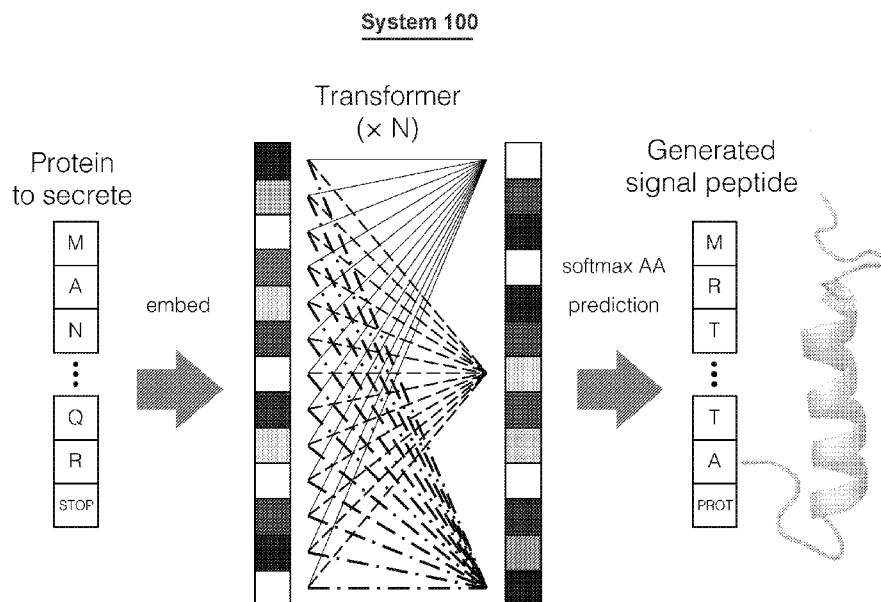


Fig. 1

(57) Abstract: The disclosure provides for artificial signal peptides generated by systems and methods utilizing deep learning.



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

Published:

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

(88) Date of publication of the international search report:

10 March 2022 (10.03.2022)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/35968

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

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☐ forming part of the international application as filed:
 - ☐ in the form of an Annex C/ST.25 text file.
 - ☐ on paper or in the form of an image file.
 - b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
 - c. ☒ furnished subsequent to the international filing date for the purposes of international search only:
 - ☒ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).
 - ☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).
2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.
3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/35968

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

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

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

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

This International Searching Authority found multiple inventions in this international application, as follows:
-***-Please See Supplemental Page-***-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Groups I+, Claims 1-14, 16-22, SEQ ID NO: 1 (signal peptide), and SEQ ID NO: 165 (mature enzyme)

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/35968

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12N 9/42; C12N 9/24; C12N 9/26; C12P 19/14; C12P 7/56; C07K 7/08; A61P 43/00 (2021.01)

CPC - C07K 7/08; C12N 1/38; C12N 9/2411; C12N 9/2477; C12N 9/2437; A61P 43/00; C07K 2319/02

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016/0108386 A1 (BP CORPORATION NORTH AMERICA, INC) 21 April 2016; paragraphs [0008], [0050], [0051], [0053], [0055], [0068], [0271], [0310], [0411]	1-7, 12, 13, 16-22
A	US 8,952,217 B2 (PUZIO, P et al.) 10 February 2015; column 86, lines 46-53; SEQ ID NO: 27,387	1-7, 12-14, 16-22
A	US 2007/0083334 A1 (MINTZ, L et al.) 12 April 2007; paragraph [0640]; SEQ ID NO: 993,084	1-7, 12-14, 16-22
A	US 2018/0020677 A1 (INDIGO AGRICULTURE, INC., et al.) 25 January 2018; SEQ ID NO: 4954	1-7, 12-14, 16-22
A	US 2019/0169586 A1 (3PLW, LTD.) 6 June 2019; paragraphs [0001], [0019], [0066], [0068], [0073], [0137]; SEQ ID NO: 23	14
A	US 2019/0031710 A1 (ENZYPEP B.V.) 31 January 2019; entire document	1-7, 12-14, 16-22

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

23 October 2021 (23.10.2021)

Date of mailing of the international search report

DEC 20 2021

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/35968

****Continued From Box No. III: Observations where unity of invention is lacking****

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

Groups I+, Claims 1-14, 16-22, SEQ ID NO: 1 (signal peptide), and SEQ ID NO: 165 (mature enzyme) are directed toward signal peptides, nucleic acids encoding the signal peptides, and associated proteins and methods.

The peptides, nucleic acids, proteins and methods will be searched to the extent they encompass SEQ ID NO: 1 (first exemplary signal peptide) and SEQ ID NO: 165 (first exemplary mature enzyme). Applicant is invited to elect additional signal peptide(s) and/or mature enzyme(s), with specified SEQ ID NO: for each, to be searched. Additional signal peptide and/or mature enzyme sequence(s) will be searched upon the payment of additional fees. It is believed that claims 1-7 (each in-part), 12-14 (each in-part), and 16-22 (each in-part) encompass this first named invention and thus these claims will be searched without fee to the extent that they encompass SEQ ID NO: 1 (signal peptide) and SEQ ID NO: 165 (mature enzyme). Applicants must specify the searchable claims that encompass any additionally elected signal peptide and/or mature enzyme sequence(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be SEQ ID NO: 2 (signal peptide).

No technical features are shared between the signal peptide and/or mature enzyme sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Groups I+ share the technical features including: a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence; wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a *Bacillus* cell; a protein sequence comprising a signal peptide conjugated to a mature enzyme; a nucleic acid sequence encoding the signal peptide amino acid sequence; and a method of expressing a recombinant protein in a host cell, the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence, and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; and expressing the recombinant protein in the host cell.

Group II, Claim 15 is directed toward a signal peptide generated by a deep machine learning method.

The inventions listed as Groups I+ and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: the special technical features of Groups I+ include a nucleic acid sequence, not present in Group II; the special technical features of Group II include machine learning, not present in any of Groups I+.

Groups I+ and II share the technical features including: a signal peptide (SP) comprising an amino acid sequence selected from any one of SEQ ID Nos: 1-164

However, these shared technical features are previously disclosed by US 2009/0120555 A1 to Breves et al. (hereinafter 'Breves').

Breves discloses a signal peptide comprising SEQ ID NO: 18 (a signal peptide comprising the first 31 N-terminal amino acids of SEQ ID NO: 2 (SEQ ID NO: 18); paragraph [0036], wherein the first 31 amino acids of SEQ ID NO: 2 are 100% identical to Applicant's SEQ ID NO: 18).

No technical features are shared between the signal peptide and/or mature enzyme sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Groups I+ share the technical features including: a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence; wherein the variant is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a *Bacillus* cell; a protein sequence comprising a signal peptide conjugated to a mature enzyme; a nucleic acid sequence encoding the signal peptide amino acid sequence; and a method of expressing a recombinant protein in a host cell, the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence, and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; and expressing the recombinant protein in the host cell.

However, these shared technical features are previously disclosed by US 2016/0108386 A1 to BP Corporation North America Inc. (hereinafter 'BP').

BP discloses a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence (a peptide comprising a signal sequence; paragraph [0050]); wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a *Bacillus* cell (wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a *Bacillus* cell; paragraphs [0051], [0092], [00291]); a protein sequence comprising a signal peptide conjugated to a mature enzyme (a protein sequence comprising a heterologous signal peptide attached to an enzyme (a signal peptide conjugated to a mature enzyme); paragraph [0051]); a nucleic acid sequence encoding the signal peptide amino acid sequence (a nucleic acid sequence encoding the signal peptide amino acid sequence; paragraphs [0050], [0051]); and a method of expressing a recombinant protein in a host cell (a method of expressing a recombinant protein in a host cell; paragraphs [0051], [0092]), the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence (the method comprising: a sequence encoding a polypeptide comprising a heterologous

****Continued Within the Next Supplemental Box****

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/35968

Continued from previous Supplemental Box

signal sequence (cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence; paragraph [0051]), and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP (and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; paragraph [0051]); and expressing the recombinant protein in the host cell (expressing the recombinant protein in the host cell; paragraphs [0051], [0092]).

Since none of the special technical features of the Groups I+ and II inventions is found in more than one of the inventions, and since all of the shared technical features are previously disclosed by the Breves and BP references, unity of invention is lacking.

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

Groups I+, Claims 1-14, 16-22, SEQ ID NO: 1 (signal peptide), and SEQ ID NO: 165 (mature enzyme) are directed toward signal peptides, nucleic acids encoding the signal peptides, and associated proteins and methods.

The peptides, nucleic acids, proteins and methods will be searched to the extent they encompass SEQ ID NO: 1 (first exemplary signal peptide) and SEQ ID NO: 165 (first exemplary mature enzyme). Applicant is invited to elect additional signal peptide(s) and/or mature enzyme(s), with specified SEQ ID NO: for each, to be searched. Additional signal peptide and/or mature enzyme sequence(s) will be searched upon the payment of additional fees. It is believed that claims 1-7 (each in-part), 12-14 (each in-part), and 16-22 (each in-part) encompass this first named invention and thus these claims will be searched without fee to the extent that they encompass SEQ ID NO: 1 (signal peptide) and SEQ ID NO: 165 (mature enzyme). Applicants must specify the searchable claims that encompass any additionally elected signal peptide and/or mature enzyme sequence(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be SEQ ID NO: 2 (signal peptide).

No technical features are shared between the signal peptide and/or mature enzyme sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Groups I+ share the technical features including: a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence; wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a Bacillus cell; a protein sequence comprising a signal peptide conjugated to a mature enzyme; a nucleic acid sequence encoding the signal peptide amino acid sequence; and a method of expressing a recombinant protein in a host cell, the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence, and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; and expressing the recombinant protein in the host cell.

Group II, Claim 15 is directed toward a signal peptide generated by a deep machine learning method.

The inventions listed as Groups I+ and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: the special technical features of Groups I+ include a nucleic acid sequence, not present in Group II; the special technical features of Group II include machine learning, not present in any of Groups I+.

Groups I+ and II share the technical features including: a signal peptide (SP) comprising an amino acid sequence selected from any one of SEQ ID Nos: 1-164

However, these shared technical features are previously disclosed by US 2009/0120555 A1 to Breves et al. (hereinafter 'Breves').

Breves discloses a signal peptide comprising SEQ ID NO: 18 (a signal peptide comprising the first 31 N-terminal amino acids of SEQ ID NO: 2 (SEQ ID NO: 18); paragraph [0036], wherein the first 31 amino acids of SEQ ID NO: 2 are 100% identical to Applicant's SEQ ID NO: 18).

No technical features are shared between the signal peptide and/or mature enzyme sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Groups I+ share the technical features including: a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence; wherein the variant is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a Bacillus cell; a protein sequence comprising a signal peptide conjugated to a mature enzyme; a nucleic acid sequence encoding the signal peptide amino acid sequence; and a method of expressing a recombinant protein in a host cell, the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence, and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; and expressing the recombinant protein in the host cell.

However, these shared technical features are previously disclosed by US 2016/0108386 A1 to BP Corporation North America Inc. (hereinafter 'BP').

BP discloses a peptide sequence comprising an amino acid sequence, wherein the peptide is a signal peptide or a peptide sequence comprising an amino acid sequence that is a variant a signal peptide sequence, wherein the variant comprises: a) a truncated subsequence, and/or a homologous sequence (a peptide comprising a signal sequence; paragraph [0050]); wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a Bacillus cell (wherein the peptide is capable of mediating secretion of an enzyme when covalently linked to the enzyme and expressed in a Bacillus cell; paragraphs

Continued Within the Next Supplemental Box

-Continued from previous Supplemental Box-

[0051], [0092], [00291]); a protein sequence comprising a signal peptide conjugated to a mature enzyme (a protein sequence comprising a heterologous signal peptide attached to an enzyme (a signal peptide conjugated to a mature enzyme); paragraph [0051]); a nucleic acid sequence encoding the signal peptide amino acid sequence (a nucleic acid sequence encoding the signal peptide amino acid sequence; paragraphs [0050], [0051]); and a method of expressing a recombinant protein in a host cell (a method of expressing a recombinant protein in a host cell; paragraphs [0051], [0092]), the method comprising: cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence (the method comprising: a sequence encoding a polypeptide comprising a heterologous signal sequence (cloning in frame a first nucleotide sequence encoding an SP having an amino acid sequence; paragraph [0051]), and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP (and a second nucleotide sequence encoding a mature enzyme protein, wherein the mature enzyme protein lacks a natural SP; paragraph [0051]); and expressing the recombinant protein in the host cell (expressing the recombinant protein in the host cell; paragraphs [0051], [0092]).

Since none of the special technical features of the Groups I+ and II inventions is found in more than one of the inventions, and since all of the shared technical features are previously disclosed by the Breves and BP references, unity of invention is lacking.