Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2641996 C 2016/01/19

(11)(21) 2 641 996

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2007/02/13
(87) Date publication PCT/PCT Publication Date: 2007/08/23
(45) Date de délivrance/Issue Date: 2016/01/19
(85) Entrée phase nationale/National Entry: 2008/08/08
(86) N° demande PCT/PCT Application No.: US 2007/003937
(87) N° publication PCT/PCT Publication No.: 2007/095300
(30) Priorité/Priority: 2006/02/13 (US60/773,086)

(51) Cl.Int./Int.Cl. *C40B 40/08* (2006.01), *C07K 14/415* (2006.01), *C12N 15/09* (2006.01), *C12N 15/29* (2006.01), *C12Q 1/68* (2006.01), *C40B 30/04* (2006.01), *C40B 40/02* (2006.01), *C40B 40/10* (2006.01), *C40B 50/00* (2006.01), *C40B 50/06* (2006.01)

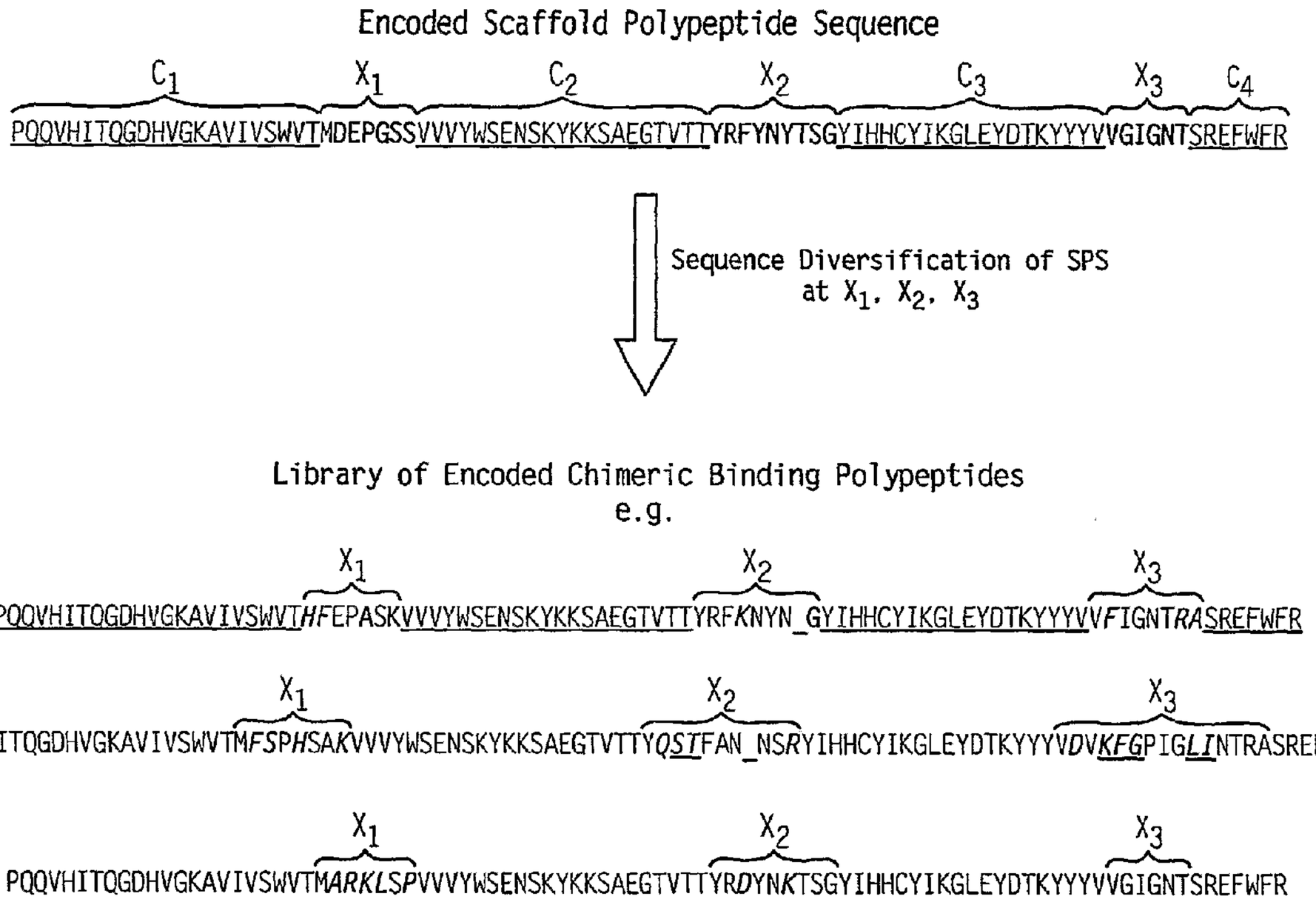
(72) Inventeur/Inventor:
JONES, JENNIFER, US

(73) Propriétaire/Owner:
MONSANTO TECHNOLOGY LLC, US

(74) Agent: GOWLING LAFLEUR HENDERSON LLP

(54) Titre : POLYPEPTIDES CHIMERES DE LIAISON VEGETAUX DE RECONNAISSANCE MOLECULAIRE UNIVERSELLE

(54) Title: PLANT CHIMERIC BINDING POLYPEPTIDES FOR UNIVERSAL MOLECULAR RECOGNITION



(57) Abrégé/Abstract:
Libraries of nucleic acids encoding chimeric binding polypeptides based on plant scaffold polypeptide sequences. Also described are methods for generating the libraries.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 August 2007 (23.08.2007)

PCT

(10) International Publication Number
WO 2007/095300 A3

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C40B 40/06* (2006.01)

(21) International Application Number:
PCT/US2007/003937

(22) International Filing Date:
13 February 2007 (13.02.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/773,086 13 February 2006 (13.02.2006) US

(71) Applicant *(for all designated States except US)*: **DIVERGENCE, INC.** [US/US]; Nidus Center, 893 N. Warson Road, St. Louis, Missouri 63141 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **JONES, Jennifer** [US/US]; 21 Homewood Drive, Kirkwood, Missouri 63122 (US).

(74) Agent: **MEIKLEJOHN, Anita, Ph.D.**; Fish & Richardson P.C., P.O. Box 1022, Minneapolis, Minnesota 55440 (US).

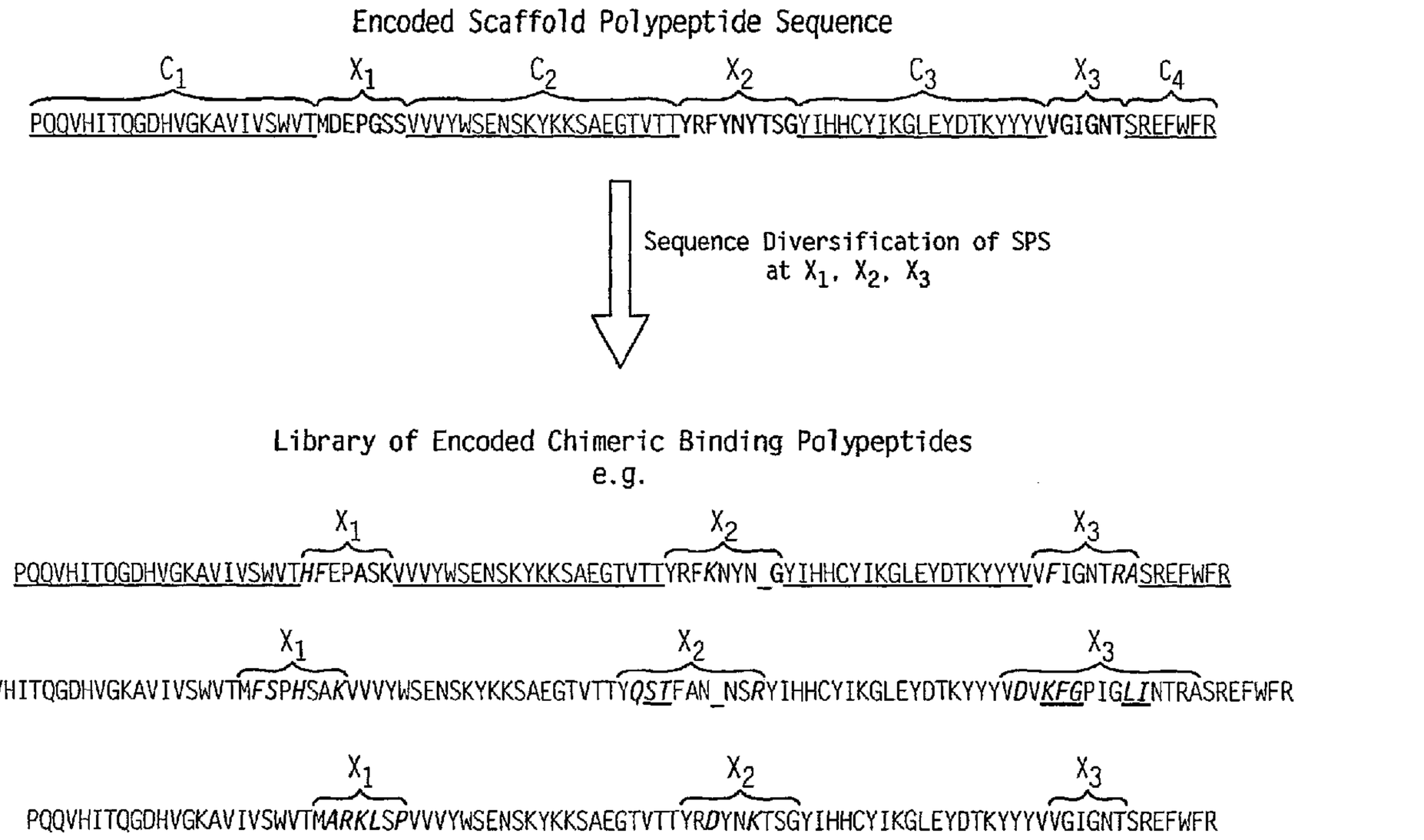
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
17 July 2008

(54) Title: PLANT CHIMERIC BINDING POLYPEPTIDES FOR UNIVERSAL MOLECULAR RECOGNITION



(57) Abstract: Libraries of nucleic acids encoding chimeric binding polypeptides based on plant scaffold polypeptide sequences. Also described are methods for generating the libraries.

WO 2007/095300 A3

PLANT CHIMERIC BINDING POLYPEPTIDES FOR UNIVERSAL MOLECULAR RECOGNITION

BACKGROUND

The binding specificity and affinity of a protein for a target are determined primarily by the protein's amino acid sequence within one or more binding regions. Accordingly, varying the amino acid sequence of the relevant regions reconfigures the protein's binding properties.

In nature, combinatorial changes in protein binding are best illustrated by the vast array of immunoglobulins produced by the immune system. Each immunoglobulin includes a set of short, virtually unique, amino acid sequences known as hypervariable regions (i.e., protein binding domains), and another set of longer, invariant sequences known as constant regions. The constant regions form β sheets that stabilize the three dimensional structure of the protein in spite of the enormous sequence diversity among hypervariable regions in the population of immunoglobulins. Each set of hypervariable regions confers binding specificity and affinity. The assembly of two heavy chain and two light chain immunoglobulins into a large protein complex (i.e., an antibody) further increases the number of combinations with diverse binding activities.

The binding diversity of antibodies has been successfully exploited in many biomedical and industrial applications. For example, libraries have been constructed that express immunoglobulins bearing artificially diversified hypervariable regions. Immunoglobulin expression libraries are very useful for identifying high affinity antibodies to a target molecule (e.g., a receptor or receptor ligand). A nucleic acid encoding the identified immunoglobulin can then be isolated and expressed in host cells or organisms.

However, despite the usefulness of immunoglobulins and antibodies in general, their expression in transgenic plants can be problematic. Immunoglobulins may not fold properly in plant cytoplasm because they require the formation of multiple disulfide bonds. Further, the large size of immunoglobulins prevents their effective uptake by some plant pests. Thus, immunoglobulins are frequently not useful as protein pesticides or pesticide targeting molecules. Finally, expressing mammalian proteins such as immunoglobulins (e.g., as so called "plantibodies") in edible plants

also raises potential issues of consumer acceptance and is thus an impediment to commercialization. This may effectively prevent use of plantibodies for many input and output traits in transgenic plants.

The above-mentioned disadvantages of immunoglobulins can be circumvented by generating diverse libraries of binding proteins from other classes of structurally tolerant proteins, preferably plant-derived proteins. These libraries can be screened to identify individual proteins that bind with desired specificity and affinity to a target of interest. Afterwards, identified binding proteins can be efficiently expressed in transgenic plants.

SUMMARY

Diverse libraries of nucleic acids encoding plant chimeric binding polypeptides, as well as methods for generating them are described herein. The chimeric binding polypeptides are conceptually analogous to immunoglobulins in that they feature highly varied binding domains in the framework of unvarying sequences that encode a structurally robust protein. However, the chimeric binding polypeptides described herein have the considerable advantage of being derived from plant protein sequences thereby avoiding many of the problems associated with immunoglobulin expression in plants. The amino acid sequences of the encoded plant chimeric binding proteins are derived from a scaffold polypeptide sequence that includes subsequences to be varied. The varied subsequences correspond to putative binding domains of the plant chimeric binding polypeptides, and are highly heterogeneous in the library of encoded plant chimeric binding proteins. In contrast the sequence of the encoded chimeric binding proteins outside of the varied subsequences is essentially the same as the parent scaffold polypeptide sequence and highly homogeneous throughout the library of encoded plant chimeric binding proteins. Such libraries can serve as a universal molecular recognition platform to select proteins with high selectivity and affinity binding for expression in transgenic plants.

Accordingly, one aspect described herein is a library of nucleic acid molecules encoding at least ten (e.g., at least 1,000, 10^5, or 10^6) different chimeric binding polypeptides. The amino acid sequence of each polypeptide includes C_1-X_1-C_2-X_2-C_3-X_3-C_4, where C_1-C_4 are backbone subsequences selected from purple acid phosphatase (i.e., SEQ ID NOs: 1-30, 31-60, 61-90, and 91-120, respectively) that can include up to

30 (e.g., 20, 10, or 5) single amino acid substitutions, deletions, insertion, or additions to the selected purple acid phosphatase sequences. The C_1-C_4 subsequences are homogeneous across many of the polypeptides encoded in the library. In contrast to the C_1-C_4 backbone subsequences, the X_1-X_3 subsequences are independent variable subsequences consisting of 2-20 amino acids, and these subsequences are heterogeneous across many of the polypeptides in the library. For example, the library of chimeric polypeptides can have the amino acid sequence of any one of SEQ ID NOs:124-126 including one to ten single amino acid substitutions, deletions, insertions, or additions to amino acid positions corresponding to 23-39, 51-49, and 79-84 of SEQ ID NOs:124-126.

Another aspect described herein is a method for generating the just-described library. The method includes providing a parental nucleic acid encoding a plant scaffold polypeptide sequence containing C_1-X_1-C_2-X_2-C_3-X_3-C_4 as defined above. The method further includes replicating the parental nucleic acid (e.g., at least one of the X_1-X_3 subsequences is selected from SEQ ID NOs: 121-123) under conditions that introduce up to 10 single amino acid substitutions, deletions, insertions, or additions to the parental X_1, X_2, or X_3 subsequences, whereby a heterogeneous population of randomly varied subsequences encoding X_1, X_2, or X_3 is generated. The population varied subsequences is then substituted into a population of parental nucleic acids at the positions corresponding to those encoding X_1, X_2, or X_3. The amino acid substitutions, deletions, insertions or additions can be introduced into the parental nucleic acid subsequences by replication *in vitro* (e.g., using a purified mutagenic polymerase or nucleotide analogs) or *in vivo* (e.g., in a mutagenic strain of *E. coli*). The just-described library can be introduced into a biological replication system (e.g., *E. coli* or bacteriophage) and amplified.

A related aspect described herein is another method for generating the above-described library of nucleic acids. The method includes selecting an amino acid sequence containing C_1-X_1-C_2-X_2-C_3-X_3-C_4 as defined above. The method further includes providing a first and second set of oligonucleotides having overlapping complementary sequences. Oligonucleotides of the first set encode the C_1-C_4 subsequences and multiple heterogeneous X_1-X_3 subsequences. Oligonucleotides of the second set are complementary to nucleotide sequences encoding the C1-C4 subsequences and multiple heterogeneous X_1-X_3 subsequences. The two sets of

oligonucleotides are combined to form a first mixture and incubated under conditions that allow hybridization of the overlapping complementary sequences. The resulting hybridized sequences are then extended to form a second mixture containing the above-described library.

Yet another aspect of the invention is a library of nucleic acids encoding chimeric binding polypeptides each of which include an amino acid sequence at least 70% (i.e., any percentage between 70% and 100%) identical to any of SEQ ID NOs: 127-129. The amino acid sequence of each of the encoded polypeptides includes amino acids that differ from those of SEQ ID NOs: 127-129 at positions 14, 15, 33, 35-36, 38, 47-48, 66, 68-69, 71, 80, 81, 99, 101-102, and 104, and the amino acid differences are heterogeneous across a plurality of the encoded polypeptides. The amino acid sequence of each of the encoded polypeptides outside of the above-listed positions is homogeneous across a plurality of the encoded chimeric polypeptides.

A related aspect described herein is a method for generating the just-described library. The method includes selecting an amino acid sequence corresponding to any of SEQ ID NOs: 127-129, in which the selected sequence differs from SEQ ID NOs:127-129 in at least one the above-mentioned positions. The method further includes providing a first and second set of oligonucleotides having overlapping complementary sequences. Oligonucleotides of the first set encode subsequences of the selected amino acid sequence, the subsequences being heterogeneous at the above-mentioned positions. Oligonucleotides of the second set are complementary to nucleotide sequences encoding subsequences of the selected amino acid sequence, the subsequences being heterogeneous at the above-mentioned positions. The two sets of oligonucleotides are combined to form a first mixture and incubated under conditions that allow hybridization of the overlapping complementary sequences. The resulting hybridized sequences are then extended to form a second mixture containing the above-described library.

Various implementations of the invention can include one or more of the following. For example, each nucleic acid in a library can include a vector sequence. Also featured is any nucleic acid isolated from one of the above-described libraries, as well as the chimeric binding polypeptide encoded by it, in pure form.

In one implementation, a population of cells (or individual cells selected from the population of cells) is provided which express chimeric binding polypeptides

encoded by one of the libraries. Another implementation features a library of purified chimeric binding polypeptides encoded by one the nucleic acid libraries. Yet another implementation provides a population of filamentous phage displaying the chimeric binding polypeptides encoded by one of the nucleic acid libraries.

In various implementations of methods for generating the above described nucleic acid libraries by oligonucleotide assembly, one or more of the following can be included. For example, the method can further include, after the second mixture that contains the nucleic acid library is generated, performing a cycle of denaturing the population of nucleic acids followed by a hybridization and an elongation step. Optionally, this cycle can be repeated (e.g., up to 100 times). The nucleic acid libraries can be amplified by a polymerase chain reaction that includes a forward and a reverse primer that hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library. In one implementation, amino acids to be encoded in variable sequence positions are selected from a subset (e.g., only 4, 6, 8, 10, 12, 14 or 16) of alanine, arginine, asparagine, aspartate, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, cysteine and valine (the 20 naturally occurring amino acids). In other cases 19 of the 20 are used (excludes cysteine). In other cases all 20 are used. In another implementation, the subset of amino acids includes at least one aliphatic, one acidic, one neutral, and one aromatic amino acid (e.g., alanine, aspartate, serine, and tyrosine).

Described herein is library of nucleic acids encoding at least ten different polypeptides, the amino acid sequence of each polypeptide comprising:

C1-X1-C2-X2-C3-X3-C4, wherein: (i) subsequence C1 is selected from SEQ. ID NOs:1-30, subsequence C2 is selected from SEQ ID NOs:31-60, subsequence C3 is selected from SEQ. ID NOs:61-90; subsequence C4 is selected from SEQ. ID NOs:91-120, and each of C1-C4 comprise up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (ii) C1-C4 are homogeneous across a plurality of the encoded polypeptides; (iii) each of X1-X3 is an independently variable subsequence consisting of 2-20 amino acids; and each of X1-X3 are heterogeneous across a plurality of the encoded polypeptides.

Also described is a library of nucleic acids encoding at least ten different polypeptides, the amino acid sequence of each polypeptide comprising:

C1-X1-C2-X2-C3-X3-C4, wherein: (i) subsequence C1 is selected from FIG. 2 or FIG. 4, subsequence C2 is selected from FIG. 2 or FIG. 4, subsequence C3 is selected from FIG. 2 or FIG. 4; subsequence C4 is selected from FIG. 2 or FIG. 4, and each of C1-C4 comprise up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (ii) C1-C4 are homogeneous across a plurality of the encoded polypeptides

(iii) each of X1-X3 is an independently variable subsequence consisting of 2-20 amino acids; and each of X1-X3 are heterogeneous across a plurality of the encoded polypeptides.

Also described is a library of nucleic acids encoding at least ten different polypeptides, the amino acid sequence of each polypeptide comprising:

C1-X1-C2-X2-C3-X3-C4, wherein (i) subsequence C1 is selected from FIG. 3 or FIG. 5, subsequence C2 is selected from FIG. 3 or FIG. 5, subsequence C3 is selected from FIG. 3 or FIG. 5; subsequence C4 is selected from FIG. 3 XX, and each of C1-C4 comprise up to 30 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (ii) C1-C4 are homogeneous across a plurality of the encoded polypeptides (iii) each of X1-X3 is an independently variable subsequence consisting of 2-20 amino acids; and each of X1-X3 are heterogeneous across a plurality of the encoded polypeptides.

In various embodiments: at least 1,000 different polypeptides are encoded; at least 100,000 different polypeptides are encoded; at least 1,000,000 different polypeptides are encoded; each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; none of C1-C4 comprise amino acid substitutions, deletions, insertions, or additions to the selected subsequence; amino acids of X1-X3 are selected from fewer than 20 amino acids genetically encoded in plants; amino acids of X1-X3 are selected from all 20 amino acids genetically encoded in plants; the fewer than 20 genetically encoded amino acids include at least one aliphatic amino acid, at least one acidic amino acid, at least one neutral amino acid, and

at least one aromatic amino acid; fewer than 20 genetically encoded amino acids comprise alanine, aspartate, serine, and tyrosine.

In some cases: the amino acid sequence of each polypeptide is selected from:

(a). a polypeptide comprising C1-X1-C2-X2-C3-X3-C4 wherein C1= SEQ. ID NO:1, C2= SEQ. ID NO: 31, C3= SEQ. ID NO: 61, and C4= SEQ. ID NO: 91;

(b). a polypeptide comprising C1-X1-C2-X2-C3-X3-C4 wherein C1= SEQ. ID NO:2, C2= SEQ. ID NO: 32, C3= SEQ. ID NO: 62, and C4= SEQ. ID NO: 92; and

(c). a polypeptide comprising C1-X1-C2-X2-C3-X3-C4 wherein C1= SEQ. ID NO:3, C2= SEQ. ID NO: 33, C3= SEQ. ID NO: 63, and C4= SEQ. ID NO: 93.

In some cases: each encoded polypeptide comprises C1-X1-C2-X2-C3-X3-C4, wherein C1= SEQ. ID NO: X1, C2= SEQ. ID NO: X2, C3= SEQ. ID NO: X3, and C4= SEQ. ID NO: X4; designated SEQ. ID NO: 130.

In some cases: each encoded polypeptide comprises C1-X1-C2-X2-C3-X3-C4, wherein C1= SEQ. ID NO: X1, C2= SEQ. ID NO: X2, C3= SEQ. ID NO: X3, and C4= SEQ. ID NO: X4; designated SEQ. ID NO: 130.

In some embodiments: wherein each of the nucleic acids comprises a vector sequence.

Also described: are an isolated nucleic acid selected from the library and a isolated cell expressing the nucleic acid as well as a purified library of purified polypeptides encoded by the library; and a population of filamentous phage displaying the polypeptides encoded by the library.

Described herein is a method of generating a library, comprising: (i) providing a parental nucleic acid encoding a parental polypeptide comprising the amino acid sequence: C1-X1-C2-X2-C3-X3-C4, wherein subsequence C1 is selected from SEQ ID NOs:1-30, subsequence C2 is selected from SEQ ID NOs:31-60, subsequence C3 is selected from SEQ ID NOs:61-90; subsequence C4 is selected from SEQ ID NOs:91 120; each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; and each of X1-X3 is an independent subsequence consisting of 2-20 amino acids; (ii) replicating the parental nucleic acid under conditions that introduce up to 10 single amino acid substitutions, deletions, insertions, or additions to the X1, X2, or X3 subsequences, whereby a population of randomly varied subsequences encoding X1', X2', or X3' is generated; and (iii) the population of randomly varied subsequences X1', X2', or X3'is

substituted, into a population of parental nucleic acids at the positions corresponding to those that encode X1, X2, or X3.

In various instances: at least one of the X1-X3 subsequences is selected from SEQ ID NOs:121-123; each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; none of C1-C4 comprise amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the replicating generates a heterogeneous population of randomly varied subsequences by introducing up to 5 amino acid substitutions in each of X1, X2, or X3; the method further comprises amplifying the library by introducing it into a biological replication system and proliferating the biological replication system; the biological replication system is a plurality of E. coli cells; the biological replication system is a plurality of bacteriophage; the replicating occurs in vitro; the replicating is performed with a purified mutagenic polymerase; the replicating is performed in the presence of a nucleotide analog; the replicating occurs in vivo; the replicating in vivo occurs in a mutagenic species of E. coli.

Also described is a method of generating the library of claim 1, comprising:(i) selecting an amino acid sequence comprising the amino acid sequence C1-X1-C2-X2 C3 X3-C4 to be encoded, wherein: (a) subsequence C1 is selected from SEQ ID NOs:1-30, subsequence C2 is selected from SEQ ID NOs:31-60, subsequence C3 is selected from SEQ ID NOs:61 90, and subsequence C4 is selected from SEQ ID NOs:91-120; (b) each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (c) each of X1, X2, and X3 consists of an amino acid sequence 2-20 amino acids in length; (ii) providing a first plurality and a second plurality of oligonucleotides, wherein: (a) oligonucleotides of the first plurality encode the C1-C4 subsequences and multiple heterogeneous X1-X3 variant subsequences X1'-X3'; (b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding the C1-C4 subsequences and to nucleotide sequences encoding multiple heterogeneous X1' X3' subsequences; and (c) the oligonucleotides of the first and second pluralities have overlapping sequences

complementary to one another; (iii) combining the population of oligonucleotides to form a first mixture; (iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and (v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

In various instances: each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises from zero and up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the method further comprises performing a cycle of steps, the cycle of steps comprising denaturing the library by increasing the temperature of the second mixture to a temperature effective for denaturing double stranded DNA, followed by steps (iv) and (v); the method comprises repeating the cycle of steps up to 100 times; the method further comprises amplifying the library by a polymerase chain reaction consisting essentially of the library, a forward primer, and a reverse primer, wherein the forward and reverse primers can hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library; the amino acid to be encoded in each position of the X1, X2, or X3 subsequences, is selected from a subset of alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine; herein the amino acid selected for each single amino acid substitution is selected from a group of amino acids consisting of at least one aliphatic, at least one one acidic, at least one one neutral, and at least one one aromatic amino acid; and the group of amino acids consists of alanine, aspartate, serine, and tyrosine.

Also described herein is a method of generating a library, comprising: (i) providing a parental nucleic acid encoding a parental polypeptide comprising the amino acid sequence: C1-X1-C2-X2-C3-X3-C4, wherein subsequence C1 is selected from FIG. 2 or FIG 4, subsequence C2 is selected from FIG. 2 or FIG 4, subsequence C3 is selected from FIG.2 or FIG 4; subsequence C4 is selected from FIG.2 or FIG 4 each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; and each of X1-X3 is an independent

subsequence consisting of 2-20 amino acids; (ii) replicating the parental nucleic acid under conditions that introduce up to 10 single amino acid substitutions, deletions, insertions, or additions to the X1, X2, or X3 subsequences, whereby a population of randomly varied subsequences encoding X1', X2', or X3' is generated; and (iii) the population of randomly varied subsequences X1', X2', or X3'is substituted, into a population of parental nucleic acids at the positions corresponding to those that encode X1, X2, or X3.

In various embodiments: at least one of the X1-X3 subsequences is selected from SEQ ID NOs:121-123; each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; none of C1-C4 comprise an amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the replicating generates a heterogeneous population of randomly varied subsequences by introducing up to 5 amino acid substitutions in each of X1, X2, or X3; the method further comprises amplifying the library by introducing it into a biological replication system and proliferating the biological replication system; the biological replication system is a plurality of E. coli cells; the biological replication system is a plurality of bacteriophage; the replicating occurs in vitro; the replicating is performed with a purified mutagenic polymerase the replicating is performed in the presence of a nucleotide analog; the replicating occurs in vivo; and the replicating in vivo occurs in a mutagenic species of E. coli.

Also described is a method of generating the library, comprising: (i) selecting an amino acid sequence comprising C1-X1-C2-X2 C3 X3-C4 to be encoded, wherein (a) subsequence C1 is selected from FIG. 2 or FIG 4, subsequence C2 is selected from FIG. 2 or FIG 4, subsequence C3 is selected from FIG. 2 or FIG 4, and subsequence C4 is selected from FIG. 2 or FIG 4; (b) each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (c) each of X1, X2, and X3 consists of an amino acid sequence 2-20 amino acids in length; (ii) providing a first plurality and a second plurality of oligonucleotides, wherein (a) oligonucleotides of the first plurality encode the C1-C4 subsequences and multiple

heterogeneous X1-X3 variant subsequences X1'-X3'; (b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding the C1-C4 subsequences and to nucleotide sequences encoding multiple heterogeneous X1' X3' subsequences; and

(c) the oligonucleotides of the first and second pluralities have overlapping sequences complementary to one another; (iii) combining the population of oligonucleotides to form a first mixture; (iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and (v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

In various cases: each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises from zero and up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the method further comprises performing a cycle of steps, the cycle of stepscomprising denaturing the library by increasing the temperature of the second mixture to a temperature effective for denaturing double stranded DNA, followed by steps (iv) and (v); the method further comprises repeating the cycle of steps up to 100 times; the method further comprises amplifying the library by a polymerase chain reaction consisting essentially of the library, a forward primer, and a reverse primer, wherein the forward and reverse primers can hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library; the amino acid to be encoded in each position of the X1, X2, or X3 subsequences, is selected from a subset of alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine; the amino acid selected for each single amino acid substitution is selected from a group of amino acids consisting of at least one aliphatic, at least one acidic, one at least one neutral, and at least one aromatic amino acid; and the group of amino acids consists of alanine, aspartate, serine, and tyrosine.

Also disclosed is a method of generating the library, comprising: (i) providing a parental nucleic acid encoding a parental polypeptide comprising the amino acid

sequence: C1-X1-C2-X2-C3-X3-C4, wherein subsequence C1 is selected from FIG. 3 or FIG 5, subsequence C2 is selected from FIG. 3 or FIG 5, subsequence C3 is selected from FIG. 3 or FIG 5; subsequence C4 is selected from FIG. 3 or FIG 5; each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; and each of X1-X3 is an independent subsequence consisting of 2-20 amino acids; (ii) replicating the parental nucleic acid under conditions that introduce up to 10 single amino acid substitutions, deletions, insertions, or additions to the X1, X2, or X3 subsequences, whereby a population of randomly varied subsequences encoding X1', X2', or X3' is generated; and (iii) the population of randomly varied subsequences X1', X2', or X3'is substituted, into a population of parental nucleic acids at the positions corresponding to those that encode X1, X2, or X3.

In various instances: at least one of the X1-X3 subsequences is selected from SEQ ID NOs:121-123; each of C1-C4 independently comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; none of C1-C4 comprise amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the replicating generates a heterogeneous population of randomly varied subsequences by introducing up to 5 amino acid substitutions in each of X1, X2, or X3; the method further comprises amplifying the library by introducing it into a biological replication system and proliferating the biological replication system; the biological replication system is a plurality of E. coli cells; the biological replication system is a plurality of bacteriophage; the replicating occurs in vitro; the replicating is performed with a purified mutagenic polymerase; the replicating is performed in the presence of a nucleotide analog; the replicating occurs in vivo; and the replicating in vivo occurs in a mutagenic species of E. coli.

Also described is a method of generating the library, comprising: (i) selecting an amino acid sequence comprising: C1-X1-C2-X2 C3 X3-C4 to be encoded, wherein (a) subsequence C1 is selected from FIG. 3 or FIG 5, subsequence C2 is selected from FIG. 3 or FIG 5, subsequence C3 is selected from FIG. 3 or FIG 5, and subsequence C4 is

selected from FIG. 3 or FIG 5; (b) each of C1-C4 comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; (c) each of X1, X2, and X3 consists of an amino acid sequence 2-20 amino acids in length; (ii) providing a first plurality and a second plurality of oligonucleotides, wherein (a) oligonucleotides of the first plurality encode the C1-C4 subsequences and multiple heterogeneous X1-X3 variant subsequences X1'-X3'; (b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding the C1-C4 subsequences and to nucleotide sequences encoding multiple heterogeneous X1' X3' subsequences; and (c) the oligonucleotides of the first and second pluralities have overlapping sequences complementary to one another; (iii) combining the population of oligonucleotides to form a first mixture; (iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and (v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

In various embodiments: each of C1-C4 comprises up to 20 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises up to 10 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; each of C1-C4 independently comprises from zero and up to 5 single amino acid substitutions, deletions, insertions, or additions to the selected subsequence; the method further comprises performing a cycle of steps, the cycle comprising denaturing the library by increasing the temperature of the second mixture to a temperature effective for denaturing double stranded DNA, followed by steps (iv) and (v); the method further comprises repeating the cycle up to 100 times; the method further comprises amplifying the library by a polymerase chain reaction consisting essentially of the library, a forward primer, and a reverse primer, wherein the forward and reverse primers can hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library; the amino acid to be encoded in each position of the X1, X2, or X3 subsequences, is selected from a subset of alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine the amino acid selected for each single amino acid substitution is selected from a group of amino acids consisting of at least

one aliphatic, one acidic, one neutral, and one aromatic amino acid; and the group of amino acids consists of alanine, aspartate, serine, and tyrosine.

Also described is a library of nucleic acids encoding at least ten different polypeptides, wherein: (i) the amino acid sequence of each of the encoded polypeptides comprises an amino acid sequence at least 70% identical to any of SEQ ID NOs:127-129; (ii) the amino acid sequence of each of the encoded polypeptides includes amino acids that differ from those of SEQ ID NOs:127-129 at positions 14, 15, 33, 35-36, 38, 47-48, 66, 68-69, 71, 80, 81, 99, 101-102, and 104, and the amino acid differences are heterogeneous across a plurality of the encoded polypeptides; and (iii) the amino acid sequence of each of the encoded polypeptides outside of the residues corresponding to positions 14, 15, 33, 35-36, 38, 47-48, 66, 68-69, 71, 80, 81, 99, 101-102, and 104 of SEQ ID NOs: 127-129 is homogeneous across a plurality of the encoded polypeptides.

In various embodiments: the amino acid sequence of the polypeptides has at least 75% identity to any of SEQ ID NOs 127-129; the amino acid sequence of the polypeptides has at least 80% identity to any of SEQ ID NOs 127-129; and the amino acid sequence of the polypeptides has at least 85% identity to any of SEQ ID NOs 127-129 each of the nucleic acids comprises a vector sequence. Also disclosed: an isolated nucleic acid encoding a polypeptide, selected from the library; a purified polypeptide encoded by the nucleic acid; a population of cells expressing the polypeptides encoded by the library; .a cell selected from the population of cells; a purified library of polypeptides encoded by the library; a population of filamentous phage displaying the library of polypeptides encoded by the library.

Also disclosed is a method of generating the library, comprising: (i) selecting an amino acid sequence corresponding to any one of SEQ ID NOs: 127 129 to be encoded, wherein the selected sequence differs from those of SEQ ID NOs:127-129 in at least one of variable positions 14, 15, 33, 35-36, 38, 47-48, 66, 68-69, 71, 80, 81, 99, 101-102, and 104; (ii) chemically providing a first and a second plurality of oligonucleotides, wherein (a) oligonucleotides of the first plurality encode amino acid subsequences of the selected amino acid sequence; the subsequences being heterogeneous at the encoded variable positions; (b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding subsequences of the selected amino acid sequence, the subsequences being heterogeneous at the encoded variable positions; and (c) the first and second pluralities comprise

oligonucleotides have overlapping sequences complementary to one another; (iii) combining the population of oligonucleotides to form a first mixture; (iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and (v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

In various instances: the method further comprises performing a cycle of denaturing the library by increasing the temperature of the second mixture to a temperature effective for denaturing double stranded DNA, followed by steps (iv) and (v); the method further comprises repeating the cycle up to 100 times; the method further comprises amplifying the library by a polymerase chain reaction consisting essentially of the library, a forward primer, and a reverse primer, wherein the forward and reverse primers can hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library; the amino acids to be encoded or the variable positions, are selected from a subset of alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine the amino acids selected for the variable positions are selected from a group consisting of an aliphatic, an acidic, a neutral, and an aromatic amino acid; the group of amino acids consists of alanine, aspartate, serine, and tyrosine.

Another aspect described herein is a library of nucleic acids encoding at least ten different polypeptides, the amino acid sequence of each polypeptide comprising:

C_1-X_1-C_2-X_2-C_3-X_3-C_4, wherein

(i) subsequence C_1 is selected from SEQ ID NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90; subsequence C_4 is selected from SEQ ID NOs:91-120, and each of C_1-C_4 independently comprises up to 10 single conservative substitutions to the selected subsequence; (ii) C_1-C_4 are homogeneous across the encoded polypeptides; (iii) each of X_1-X_3 is an independently variable subsequence consisting of 2-20 amino acids; and (iv) each of X_1-X_3 are heterogeneous across the encoded polypeptides.

Also described is a method of generating the above-described library comprising: (i) providing a parental nucleic acid encoding a parental polypeptide comprising the amino acid sequence: C_1-X_1-C_2-X_2-C_3-X_3-C_4, wherein subsequence C_1 is selected from SEQ ID

NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90; subsequence C_4 is selected from SEQ ID NOs:91-120; each of C_1-C_4 independently comprises up to 10 single conservative substitutions to the selected subsequence; and each of X_1-X_3 is an independent subsequence consisting of 2-20 amino acids; (ii) replicating the parental nucleic acid under conditions that introduce up to 10 single conservative substitutions to the X_1, X_2, or X_3 subsequences, whereby a population of randomly varied subsequences encoding X_1', X_2', or X_3' is generated; and (iii) the population of randomly varied subsequences encoding X_1', X_2', or X_3'is substituted into a population of parental nucleic acids at the positions corresponding to those that encode X_1, X_2, or X_3.

Also described is a method of generating the library of any one of claims 1 to 9, comprising:

(i) selecting an amino acid sequence comprising the amino acid sequence C_1-X_1-C_2-X_2-C_3-X_3-C_4 to be encoded, wherein (a) subsequence C_1 is selected from SEQ ID NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90, and subsequence C_4 is selected from SEQ ID NOs:91-120; (b) each of C_1-C_4 independently comprises up to 10 conservative substitutions to the selected subsequence; (c) each of X_1, X_2, and X_3 consists of an amino acid sequence 2-20 amino acids in length;

(ii) providing a first plurality and a second plurality of oligonucleotides, wherein (a) oligonucleotides of the first plurality encode the C_1-C_4 subsequences and multiple heterogeneous X_1-X_3 variant subsequences X_1'-X_3'; (b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding the C_1-C_4 subsequences and to nucleotide sequences encoding multiple heterogeneous X_1'-X_3' subsequences; and (c) the oligonucleotides of the first and second pluralities have overlapping sequences complementary to one another;

(iii) combining the first and second pluralities of oligonucleotides to form a first mixture;

(iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and

(v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

The details of one or more embodiments of the invention are set forth in the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

FIG 1 is a schematic representation depicting the generation of a library of nucleic acids encoding chimeric binding polypeptides by diversifying subsequences within an encoded polypeptide scaffold sequence. Encoded scaffold polypeptide sequence is SEQ ID NO:124. Library of encoded chimeric binding polypetides are SEQ ID NOs: 838, 839, and 840, respectively (i.e., from top to bottom).

FIG 2 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. These proteins are homologous to oryzacystatin. The Cl, C2, C3 and C4 are boxed and labeled. The sequences shown are SEQ ID No:132 (i.e., homologous sequences Q2V8I6_CUCMA_1441/1-18 – Q2V8I4_CUCMO_734/1-28); SEQ ID NO:133 (i.e. Q2V8H9_LAGLE_431/1-28); SEQ ID NO:134 (i.e., Q6DKU9_CUCMA_1441/1-28 and Q6DLC8_CUCMA_1441/1-28); SEQ ID NO: 131 (i.e., 080389_CUCSA_795/1-89); SEQ ID NOs: 136-150 (i.e., QIRVW3_MEDTR_2578/1-54 – Q8GZV2_CHEMJ_340/1-38); SEQ ID NO:130 (i.e., Reference/1-102); and SEQ ID NOs: 151-330 (i.e., CYT1_ORYSA_1097/1-88 to end).

FIG 3 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. These proteins are homologous to C2. The Cl, C2, C3 and C4 are boxed and labeled. Sheets 1-3 show SEQ ID NOs: 331-367 (i.e. Q9M366_ARATH_43120/1-78 - Q9FJG3_ARATH_325405/1-81); SEQ ID NO: 130 (i.e., Reference/1-156); and SEQ ID NOs: 368-384 (i.e., ERG1_ORYSA_795/1-89 – Q4JHI8_CUCMA_692/1-87. Sheets 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24 and 25-27 show SEQ ID NOs: 385-821.

FIG 4 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. The sequences shown are SEQ ID NO: 130 (i.e., oryza full) and SEQ ID NOs 822-832. These proteins are homologous to oryzacystatin. The Cl, C2, C3 and C4 are boxed and labeled.

FIG 5 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. The sequences shown are, from top to

FIG. 3 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. These proteins are homologous to C2. The C1, C2, C3 and C4 are boxed and labeled.

FIG. 4 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. These proteins are homologous to oryzacystatin. The C1, C2, C3 and C4 are boxed and labeled.

FIG. 5 is an alignment of the sequences of a number of proteins that have regions which can be used as a scaffold. These proteins are homologous to C2. The C1, C2, C3 and C4 are boxed and labeled.

DETAILED DESCRIPTION

Diverse libraries of nucleic acids (e.g., cDNA libraries) encoding plant chimeric binding polypeptides, as well as methods for generating them are described below. The amino acid sequences of the library of encoded plant chimeric binding proteins are derived from a scaffold polypeptide sequence that includes subsequences to be varied. The varied subsequences correspond to putative binding domains of the plant chimeric binding proteins, and are highly heterogeneous in the library of plant chimeric binding proteins. In contrast, the sequence of the encoded chimeric binding proteins outside of the varied subsequences is essentially the same as the parent scaffold polypeptide sequence and highly homogeneous throughout the library of encoded plant chimeric binding proteins. Thus, libraries of plant chimeric binding proteins can serve as a universal molecular recognition library platform for selection of specialized binding proteins for expression in transgenic plants. Libraries of plant chimeric binding proteins can be expressed by transfected cells (i.e., as expression libraries) and tested for interaction with a molecular target of interest. For example, expression libraries can be screened to identify polypeptides that bind with high specificity and affinity to polypeptides expressed by plant pests, including nematodes. Ultimately, individual chimeric binding proteins with desired target binding properties can be expressed in a transgenic plant.

bottom, SEQ ID NO: 131 and SEQ ID NOs 833-837. These proteins are homologous to C2. The Cl, C2, C3 and C4 are boxed and labeled.

DETAILED DESCRIPTION

Diverse libraries of nucleic acids (e.g., cDNA libraries) encoding plant chimeric binding polypeptides, as well as methods for generating them are described below. The amino acid sequences of the library of encoded plant chimeric binding proteins are derived from a scaffold polypeptide sequence that includes subsequences to be varied. The varied subsequences correspond to putative binding domains of the plant chimeric binding proteins, and are highly heterogeneous in the library of plant chimeric binding proteins. In contrast, the sequence of the encoded chimeric binding proteins outside of the varied subsequences is essentially the same as the parent scaffold polypeptide sequence and highly homogeneous throughout the library of encoded plant chimeric binding proteins. Thus, libraries of plant chimeric binding proteins can serve as a universal molecular recognition library platform for selection of specialized binding proteins for expression in transgenic plants. Libraries of plant chimeric binding proteins can be expressed by transfected cells (i.e., as expression libraries) and tested for interaction with a molecular target of interest. For example, expression libraries can be screened to identify polypeptides that bind with high specificity and affinity to polypeptides expressed by plant pests, including nematodes. Ultimately, individual chimeric binding proteins with desired target binding properties can be expressed in a transgenic plant.

I. Plant Scaffold Polypeptide Sequences

A plant scaffold polypeptide sequence is an amino acid sequence based on a plant protein that is structurally tolerant of extreme sequence variation within one or more regions. The regions to be varied within the scaffold polypeptide sequence are conceptually analogous to the hypervariable regions of immunoglobulins, and form putative binding domains in a chimeric binding polypeptide. Thus, a large library of nucleic acid sequences encoding diverse plant chimeric binding polypeptides is produced by diversifying specific sequences within a scaffold polypeptide sequence, as is described in detail below.

Plant scaffold polypeptide sequences are selected to have a number of properties, e.g., they: (i) are derived from sequences that are of plant origin; (ii) encode proteins that tolerate the introduction of sequence diversity structurally; (iii) only contain disulfide bonds that do not interfere with folding of the polypeptide when expressed in a plant; (iv) express at high levels in diverse plant tissues; and (v) can be targeted to different subcellular locations (e.g., cytoplasm, mitochondria, plastid) or secreted from the cell. Based on these properties, plant scaffold polypeptide sequences permit the generation of large libraries of chimeric binding polypeptides with highly diverse binding activities. Libraries of chimeric binding polypeptides can be screened for binding to a target molecule. Chimeric binding proteins having the desired binding activity can subsequently be expressed in plants to confer input traits (e.g., pest or pathogen resistance, drought tolerance) or output traits (e.g. modified lipid composition, heavy metal binding for phytoremediation, medicinal uses). Such binding proteins can also be used in various affinity-based applications, e.g., diagnostic detection of an antigen using a sandwich ELISA; histochemical detection of antigens; generation of protein biochips; and affinity purification of antigens.

It is helpful to select the scaffold polypeptide sequence based on the sequence of a plant protein or protein domain of known three dimensional structure (see, e.g., Nygren *et al.* (2004) "Binding Proteins from Alternative Scaffolds," *J. of Immun. Methods* 290:3-28). However, even without experimentally determined structural data for a potential scaffold polypeptide sequence, valuable inferences can be gleaned from computational structural analysis of a candidate amino acid sequence. Useful programs for structure prediction from an amino acid sequence include, e.g., the "SCRATCH Protein Predictor" suite of programs available to the public on the world wide web at

ics.uci.edu/~baldig/scratch/index. It is important that introduction of sequence variation not destabilize the known or predicted secondary structure of the scaffold polypeptide sequence. Accordingly, the known or predicted secondary structure of the scaffold polypeptide sequence informs the selection of amino acid subsequences that can be varied within a scaffold polypeptide sequence to form putative binding domains. The structural adequacy of a particular scaffold polypeptide sequence can be readily tested, e.g., by phage display expression analysis methods that are commonly known in the art. For example, a scaffold polypeptide sequence containing 0, 1, 2, 3, or more disulfide bonds can be tested for its ability to fold into a stable protein. Since proteins that do not fold properly will not be incorporated into a phage coat, they will not be displayed. Thus, without undue effort, many candidate scaffold polypeptide sequences can be rapidly screened for their ability to fold into stable proteins once expressed.

The plant scaffold polypeptide sequences can be based on the accessory domain from purple acid phosphatases (PAPs). The crystal structure of the PAP accessory domain of kidney bean, *Phaseolus vulgaris*, has been determined (Strater *et al.* (1995), *Science* 268(5216):1489-1492). Three exposed loops within the protein are reminiscent of the hypervariable domains found in immunoglobulins. The loops are brought together by the rigid anti-parallel β-sheet framework of the protein. The subsequences that form each loop form the putative binding domains of a chimeric binding protein derived from a PAP. These subsequences are diversified by substituting, deleting, inserting, or adding up to 10 (e.g., up to 3, 4, 6, 8) amino acids. The loops that form the putative binding domains are particularly well suited to binding target molecules containing pockets or clefts.

PAP-based scaffold polypeptide sequences take the general form:

$$\mathbf{C_1\text{-}X_1\text{-}C_2\text{-}X_2\text{-}C_3\text{-}X_3\text{-}C_4}$$

where C_1, C_2, C_3, and C_4 correspond to "backbone" subsequences which can include some introduced variation, but are not highly diversified. On the other hand, X_1, X_2, and X_3 correspond to highly varied subsequences that form the putative binding domains of each PAP-based chimeric binding protein. Table 1 shows a list of suitable C_1-C_4 backbone subequences derived from the amino acid sequences of 30 PAPs.

C_1, C_2, C_3, and C_4 correspond to SEQ ID NOs: 1-30, 31-60, 61-90, and 91-120, respectively, in Table 1.

X_1, X_2, and X_3 can be based on naturally occurring variants of corresponding PAP sequences, e.g., those shown in Table 2 as SEQ ID NOs: 121-123. Table 2 shows the range variation at each amino acid position in subsequences corresponding, respectively, to X_1, X_2, and X_3, within 30 naturally occurring PAP sequences.

Alternatively, the parent variable subsequences, X_1,-X_3, can be arbitrary sequences 2-20 amino acids in length.

In some implementations, C_1, C_2, C_3, and C_4 of a scaffold polypeptide sequence can be selected from multiple PAP-based scaffold polypeptide sequence sequences listed in Table 1, in any combination, e.g., $C_{1(SEQ\ ID\ NO:5)}$, $C_{2(SEQ\ ID\ NO:12)}$, $C_{3(SEQ\ ID\ NO:7)}$, and $C_{4(SEQ\ ID\ NO:19)}$; $C_{1(SEQ\ ID\ NO:5)}$, $C_{2(SEQ\ ID\ NO:12)}$, $C_{3(SEQ\ ID\ NO:5)}$, and $C_{4(SEQ\ ID\ NO:12)}$; $C_{4(SEQ\ ID\ NO:22)}$; $C_{1(SEQ\ ID\ NO:17)}$, $C_{2(SEQ\ ID\ NO:17)}$, $C_{3(SEQ\ ID\ NO:19)}$, and $C_{4(SEQ\ ID\ NO:1)}$, and so forth.

Table 1: SPSs Based on the Accessory Domain of PAPs

Seq ID	C_1	Seq ID	C_2
1	PQQVHITQGDHVGKAVIVSWVT	31	VVVYWSENSKYKKSAEGTVTT
2	PQQVHITQGDLVGKAVIVSWVT	32	EVHYWSENSDKKKIAEGKLVT
3	PQQVHITQGDLVGRAMIISWVT	33	AVRYWSEKNGRKRIAKGKMST
4	PQQVHITQGDLVGKAVIVSWVT	34	EVHYWSENSDKKKIAEGKLVT
5	PQQVHITQGDHVGKAVIVSWVT	35	AVRYWSKNSKQKRLAKGKIVT
6	PQQVHITQGDHVGKAMIVSWVT	36	KVVYWSENSQHKKVAKGNIRT
7	PQQVHITQGDHVGKAMIVSWVT	37	KVVYWSENSQHKKVARGNIRT
8	PQQVHITQGDHEGKTVIVSWVT	38	TVLYWSEKSKQKNTAKGKVTT
9	PQQVHITQGDLVGQAMIISWVT	39	QVIYWSDSSLQNFTAEGEVFT
10	PQQVHITQGDLVGQAMIISWVT	40	QVIYWSDSSLQNFTAEGEVFT
11	PQQVHITQGDHVGKAMIVSWVT	41	TVLYWSNNSKQKNKATGAVTT
12	PQQVHITQGDLEGEAMIISWVR	42	KVLYWIDGSNQKHSANGKITK
13	PQQVHITQGDHVGKAVIVSWVT	43	TVVYWSEKSKLKNKANGKVTT
14	PQQVHITQGDHVGQAMIISWVT	44	EVIYWSNSSLQNFTAEGEVFT
15	PQQVYITQGDHEGKGVIASWTT	45	SVLYWAENSNVKSSAEGFVVS
16	PQQVHITQGDYEGKGVIISWVT	46	TVVYWAENSSVKRRADGVVVT
17	PQQVHITQGDLVGRAMIISWVT	47	AVRYWSEKNGRKRIAKGKMST
18	PQQVHLTQGDHVGKGVIVSWVT	48	KVLYWEFNSKIKQIAKGTVST
19	PQQVHITQGDVEGKAVIVSWVT	49	KVIYWKENSTKKHKAHGKTNT
20	PQQVHVTQGNHEGNGVIISWVT	50	TVRYWCENKKSRKQAEATVNT
21	PQQVHVTQGNHEGNGVIISWVT	51	TVQYWCENEKSRKQAEATVNT
22	PQQVHITQGDYDGKAVIVSWVT	52	KVQFGTSENKFQTSAEGTVSN
23	PQQVHITQGDHEGRSIIVSWIT	53	TVFYGTSENKLDQHAEGTVTM
24	PQQVHITLGDQTGTAMTVSWVT	54	TVRYGSSPEKLDRAAEGSHTR
25	PQQVHITQGDYDGKAVIVSWVT	55	EVVYGTSPNSYDHSAQGKTTN
26	PQQVHITQGDYDGKAVIISWVT	56	HIQYGTSENKFQTSEEGTVTN
27	PQQVHITQGDYDGEAVIISWVT	57	EVRYGLSEGKYDVTVEGTLNN
28	PQQVHITQGDYDGKAVIISWVT	58	QVHYGAVQGKYEFVAQGTYHN
29	PQQVHITQGDYDGKAVIISWVT	59	QVHYGAVQGKYEFVAQGTYHN
30	PQQVHITQGDYNGKAVIVSWVT	60	EVLYGKNEHQYDQRVEGTVTN

Table 1 continued

Seq ID	C_3	Seq ID	C_4
61	YIHHCYIKGLEYDTKYYYV	91	SREFWFR
62	FIHHTTIRNLEYKTKYYYE	92	TRQFWFV
63	FIHHTTIRKLKYNTKYYYE	93	TRRFSFI
64	FIHHTTIRNLEYKTKYYYE	94	TRQFWFV
65	FIHHTTIRNLEYNTKYYYE	95	TRQFWFV
66	YIHHCTIRNLEYNTKYYYE	96	TRSFWFT
67	YIHHCTIRNLEYNTKYYYE	97	TRSFWFT
68	YIHHSTIRHLEFNTKYYYK	98	ARTFWFV
69	FIHHTTITNLEFDTTYYYE	99	TRQFWFI
70	FIHHTTITNLEFDTTYYYE	100	TRQFWFI
71	YIHHCIIKHLKFNTKYYYE	101	PRTFWFV
72	FIHHCTIRRLKHNTKYHYE	102	VRSFWFM
73	YIHHCNIKNLKFDTKYYYK	103	ARTFWFT
74	FIHHTNITNLEFNTTYFYV	104	TRQFWFI
75	YIHHCTIKDLEFDTKYYYE	105	TRKFWFV
76	YIHHCTIKDLEYDTKYYYE	106	KRQFWFV
77	YIHHCTIKNLEYNTKYFYE	107	TRQFWFT
78	YIHHCTIQNLKYNTKYYYM	108	RRTFWFV
79	FIHHCPIRNLEYDTKYYYV	109	ERKFWFF
80	YIHHCLIDDLEFDTKYYYE	110	SRRFWFF
81	YIHHCLIDDLEFDTKYYYE	111	SRRFWFF
82	YVHHCLIEGLEYKTKYYYR	112	SREFWFE
83	YIHHCVLTDLKYDRKYFYK	113	ARLFWFK
84	FIHHCTLTGLTHATKYYYA	114	VRTFSFT
85	YIHHCLLDKLEYDTKYYYK	115	AREFWFH
86	YIHHCLIEGLEYETKYYYR	116	SREFWFK
87	YIHQCLVTGLQYDTKYYYE	117	ARKFWFE
88	FIHHCLVSDLEHDTKYYYK	118	SREFWFV
89	FIHHCLVSDLEHDTKYYYK	119	SREFWFV
90	YIHHCLVDGLEYNTKYYYK	120	AREFWFE

Table 2: Naturally Occurring Residue Variation in PAP Subsequences Corresponding to X_1, X_2, and X_3 (SEQ ID NOs:121-123)

X_1 (SEQ ID NO: 121)							X_2 (SEQ ID NO: 122)									X_3 (SEQ ID NO: 123)					
Position							Position									Position					
a	b	c	d	e	f	g	a	b	c	d	e	f	g	h	i	a	b	c	d	e	f
M	D	E	P	G	S	S	Y	K	Y	Y	N	Y	T	S	G	V	G	L	R	N	T
V	E	A	K		P	N		R	F	F	T		S		P	I	E	I	G	H	
E	N	K	L		K	K		T		H	K		N			L		V	E	D	
P	V	D				T		F			D		K			M		E	D	Q	
Q	S				H						E		E					T		K	
T	I				T													S		S	
A	A																	E		E	
F																		F			
																		K			

After diversification of the above-listed subsequences of the scaffold polypeptide sequence, the diversified X_1', X_2', and X_3' subsequences are highly heterogeneous within the library of encoded plant chimeric binding polypeptides, and can each contain up to 10 (e.g., 8, 6, 4, 3) single amino acid substitutions, deletions, insertions, or additions with respect to SEQ ID NOs: 121-123 listed in Tables 1, respectively (see, e.g., Fig. 1). For example, the length of the amino acid sequences corresponding to regions X_1, X_2, or X_3 can be unaltered, shortened, or lengthened relative to SEQ ID NOs: 121-123.

The regions outside of the putative binding domains are referred to as "backbone" regions (i.e., C_1, C_2, C_3, and C_4). Unlike the amino acid sequences for X_1, X_2, and X_3, the amino acid sequences of the backbone regions are generally not substantially diversified within the library of encoded chimeric binding proteins, although some sequence variation in these regions within the library is permissible. The backbone regions of a plant scaffold polypeptide sequence can be at least 70% (i.e.,

80, 85, 90, 95, 98, or 100%) identical to any of SEQ ID NOs: 1-120. Alternatively, the backbone regions can contain up to 30 (i.e., 28, 26, 24, 22, 20, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1) single amino acid substitutions, deletions, insertions or additions. For example, C_1, C_2, C_3, and C_4 can each include 0, 1, 2, 3, 4, or 5 or more single amino acid changes. If amino acid substitutions are to be introduced into the backbone regions, it is preferable to make conservative substitutions. A conservative substitution is one that preserves the substitutes an amino acid with one that has similar chemical properties (e.g., substitution of a polar amino acid such as serine with another polar amino acid such as threonine).

In one embodiment, the plant scaffold polypeptide sequence is one of SEQ ID NOs: 124-126 shown below. Sequences corresponding to X_1, X_2, and X_3 are in bold and underlined.

SEQ ID NO: 124

PQQVHITQGDHVGKAVIVSWVT**MDEPGSS**VVVYWSENSKYKKSAEGTVTT**YRFY**
NYTSGYIHHCYIKGLEYDTKYYYV**VGIGNT**SREFWFR

SEQ ID NO: 125

PQQVHITQGDLVGKAVIVSWVT**VDEPGSS**EVHYWSENSDKKKIAEGKLVT**YRF**
FNYSSGFIHHTTIRNLEYKTKYYYE**VGLGNT**TRQFWFV

SEQ ID NO: 126

PQQVHITQGDLVGRAMIISWVT**MDEPGSS**AVRYWSEKNGRKRIAKGKMST**YR**
FFNYSSGFIHHTTIRKLKYNTKYYYE**VGLRNT**TRRFSFI

In other embodiments, a plant scaffold polypeptide sequence is based on the amino acid sequence of plant proteins that have ankyrin-like repeats. Ankryin-like repeats are small turn-helix-helix (THH) repeats consisting of approximately 33 amino acids. The number of THH repeats within a scaffold polypeptide sequence can vary from 2 to 20. The putative binding sites within the THH repeats are typically non-contiguous, but clustered on the same side of the protein of which they are a part.

A plant THH repeat-containing scaffold polypeptide sequence can have an amino acid sequence that is based on any of SEQ ID NOs: 127-129 listed below. High levels of amino acid sequence variation are introduced at the bolded/underlined residues. The plant THH repeat-containing scaffold polypeptide sequences can contain substitutions of up to 3 amino acids or a deletion in the place of the amino acids

corresponding to residues 12-13, 33, 35-36, 38, 46-47, 66, 68-69, 71, 79-80, 99, 101-102, 104, and 112-113 (residues in bold and underlined) of SEQ ID NOs:127-129.

SEQ ID NO: 127

GDDLGKKLHLAA**SR**GHLEIVRVLVEAGADVNA**L**D**KF**G**R**TALHIAA**SR**GHLEV
VKLLLEAGADVNA**L**D**KF**G**R**TALHLAA**SR**GHLEVVKLLLEAGADVNA**L**D**KF**G
DTALHVSI**DN**GNEDIAEILQ

SEQ ID NO: 128

GDDLGKKLHLAA**SR**GHLEIVRVLVEAGADVNA**L**D**KF**G**R**TPLHIAA**SK**GNEQV
VKLLLEAGADPNA**L**D**KF**G**R**TPLHIAA**SK**GNEQVVKLLLEAGADPNA**Q**D**KF**G**D**
TALHVSI**DN**GNEDIAEILQ

SEQ ID NO: 129

GSDLGKKLLEAA**RA**GQDDEVRILMANGADVNA**L**D**KF**G**R**TPLHIAA**SK**GNEQ
VVKLLLEAGADPNA**L**D**KF**G**R**TPLHIAA**SK**GNEQVVKLLLEAGADPNA**Q**D**KF**G
KTAFDISI**DN**GNED**L**AEILQ

The sequence of the scaffold polypeptide sequences can be at least 70% (i.e., 80, 85, 90, 95, 98, or 100%) identical to the sequence outside of the foregoing amino acid positions (in bold) of SEQ ID NOS: 127-129. Alternatively, the sequence of the scaffold polypeptide sequences outside of the foregoing amino acid positions (in bold) of SEQ ID NOS:127-129 can contain up to 30 (i.e., 28, 26, 24, 22, 20, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1) single amino acid substitutions, deletions, insertions or additions. In some cases it can be desirable to include additional repeating units. SEQ ID NOs: 127-129 have an amino-terminal cap, two internal repeats and a carboxy-terminal cap. It might be desirable to have 1-6 internal repeats. The amino-terminal cap sequence is aa 1-33. The first internal repeat is 34-66 and the second internal repeat is 67- 99. The carboxy-terminal cap sequence is aa 100-123. The first or the second internal repeats or both can be independently repeated 1, 2, 3, 4, 5 or 6 times.

The putative binding sites are formed by amino acid side chains protruding from the rigid secondary structure formed by the scaffold polypeptide sequence. These proteins may typically form a larger, flatter binding surface and are particularly useful for binding to targets that do not have deep clefts or pockets.

Another suitable scaffold can be based on oryzacystatin (*J Biol Chem* 262:16793 (1987); *Biochemistry* 39:14753 (2000)), a member of the cystatin/Papain

Family (Pfam Identifier PF00031) that is identified as a cysteine proteinase inhibitor of rice. The sequence of oryzacystatin is depicted below. A scaffold having the amino acid sequence C1-X1-C2-X2-C3-X3-C4 where each of Xl, X2, X3 and X4 is a variable region and Cl, C2, C3 and C4 are the backbone regions can be created based on oryzacystatin.

MSSVGGPVLGGVEPVGNENDLHLVDLARFAVTEHNKKANSLLEFEKLV SVKQQVVAGTLYYFTLEVKEGDAKKLYEAKVWEKPWMDFKELOEFKPVDAS ANA (SEQ ID NO:130)

Cl-MSS (aa 1-3 of SEQ ID NO:130)

Xl-VGGP (aa 4-7 of SEQ ID NO:130)

C2- VLGGVEPVGNENDLHLVDLARFAVTEHNKKANSLLEFEKLVSV (aa-8-50 of SEQ ID NO:130)

X2-KQQ VVAGT (aa 51-58 of SEQ ID NO:130)

C3-LYYFTLEVKEGD AKKLYEAKVWE (aa 59-81 of SEQ ID NO:130)

X3-KPWM (aa 82-85 of SEQ ID NO:130)

C4-DFKELQEFKPVDASANA (aa 86-102 of SEQ ID NO:130)

FIG. 2 depicts the sequences of a large number of plant proteins aligned with oryzacystatin. Examples of suitable C1-C4 regions are indicated. FIG. 4 depicts the sequences of a small number of plant proteins aligned with oryzacystatin. Examples of suitable C1-C4 regions are indicated. In general, Xl can be a sequence of 2-20 random amino acids (e.g., 3 amino acids). X2 can be a sequence of 2-20 random amino acids (e.g., 4 amino acids). X3 can be a sequence of 2-20 random amino acids (e.g., 4 amino acids).

Yet another suitable can be based on the C2 protein of rice (*Biochemistry* 42:11625 (2003)), a member of the C2 domain family (Pfam Identifier PF00168) that is thought to be be involved in plant defense signaling systems. The sequence of rice C2 is depicted below. A scaffold having the amino acid sequence Cl-Xl - C2-X2-C3-X3-C4 where each of Xl , X2, X3 and X4 is a variable region and Cl, C2, C3 and C4 are the backbone regions can be created based on rice C2.

MAGSGVLEVHLVDAKGLTGNDFLGKIDPYVVVOYRSOERKSSVARDQ GKNPSWNEVFKFOINSTAATGOHKLFLRLMDHDTFSRDDFLGEATINVTDLISL

GMEHGTWEMSESKHRWLADKTYHGEIRVSLTFTASAKAODHAEOVGGW

AH SFRQ (SEQ ID NO:131)

Cl -MAGSGVLEVHLVDAKG (aa 1-16 of SEQ ID NO:131)

Xl -LTGNDFLGKID (aa 17-27 of SEQ ID NO:131)

C2-PYWVQYRSQERK (aa 28-40 of SEQ ID NO:131)

X2-SSVARDQGKNP (aa 41-51 of SEQ ID NO:131)

C3-SWNEVFKFQINSTAATGQHKLFLRL (aa 52-76 of SEQ ID NO:131)

X3- MDHDTFSRDDFL (aa 77-88 of SEQ ID NO:131)

C4-

GEATINVTDLISLGMEHGTWEMSESKHRVVLADKTYHGEIRVSLTFTASAKAQ

DHAEQVGGWAHSFRQ (aa 89-156 of SEQ ID NO:131)

FIG. 3 depicts the sequences of a large number of plant proteins aligned with rice C2. Examples of suitable C1-C4 regions are indicated. FIG. 5 depicts the sequences of a small number of plant proteins aligned with oryzacystatin. Examples of suitable C1-C4 regions are indicated. In general, Xl can be a sequence of 2-20 random amino acids (e.g., 11 amino acids). X2 can be a sequence of 2-20 random amino acids (e.g., 11 amino acids). X3 can be a sequence of 2-20 random amino acids (e.g., 12 amino acids).

The following sections disclose methods for generating libraries of nucleic acids encoding chimeric binding proteins based on plant scaffold polypeptide sequences.

II. Generation of Nucleic Acid Libraries based on a Plant scaffold polypeptide sequence

A large library of nucleic acid sequence variants encoding the plant scaffold polypeptide sequence is created based on one or more plant scaffold polypeptide sequences. The library of nucleic acids encodes at least 5 (e.g., 1,000, 10^5, 10^6, 10^7, 10^9, 10^{12}, 10^{15} or more) different chimeric binding protein sequences. It is recognized that not every member of a library generated by the methods described herein will encode a unique amino acid sequence. Nevertheless, it is desirable that at least 10%

(e.g., 25%, 30%, 40%, 50%, 60%, 70%, 75%, or 90%) of the encoded chimeric binding proteins represented in the library be unique.

Prior to diversifying a plant scaffold polypeptide sequence, it may be useful to estimate computationally the expected sequence diversity to be generated with a given set of sequence variation parameters. A method for estimating sequence diversity is described, e.g., in Volles *et al.* (2005), 33(11): 3667-3677. For example, the number of different sequences expected in a library of nucleic acids generated by PCR can be estimated based on the mutation frequency of the mutagenic polymerase used for the amplification. Useful algorithms for estimating sequence diversity in randomized protein-encoding libraries can also be found on the world wide web, e.g., at guinevere.otago.ac.nz/mlrgd/STATS/index.

Libraries of nucleic acids encoding plant chimeric binding proteins can be generated by a number of known methodologies. Sequence diversity is introduced into a plant scaffold polypeptide sequence by substitution, deletion, insertion, or addition of amino acids at the highly variable positions of a scaffold polypeptide sequence as described above. Since the set of 20 amino acids that are genetically encoded in plants have somewhat redundant chemical and structural properties, a subset of amino acids (e.g., a subset of 4 types of amino acids) that encompasses this structural diversity can be adopted for substitutions. For example, amino acids to be used for substitution or insertion can be selected to include an acidic amino acid, a neutral amino acid, an aliphatic amino acid, and an aromatic amino acid (see Table 3). For example, the amino acids used for substitution could be limited to aspartate, serine, alanine, and tyrosine. Limiting the redundancy of amino acid substitutions will increase the overall structural and binding diversity of the library of chimeric binding proteins.

Table 3 Chemical Properties of Amino Acids Genetically Encoded in Plants

Acidic	Neutral	Aliphatic	Aromatic	Basic
Aspartate, Glutamate,	Asparagine, Cysteine Glutamine, Methionine, Proline, Serine, Threonine,	Alanine, Glycine, Isoleucine, Leucine, Valine	Histidine, Phenylalanine, Tryptophan, Tyrosine	Arginine, Lysine

The library of nucleic acids can be generated *in vitro* by assembly of sets of oligonucleotides with overlapping complementary sequences. First, a scaffold polypeptide sequence sequence is selected that is to be encoded by sets of assembled oligonucleotides. The sequences to be encoded in the variable regions of a given scaffold polypeptide sequence will include a multitude of heterogeneous sequences containing substitutions, insertions, deletions in additions in accordance with the library of chimeric binding polypeptides to be generated as described above. The scaffold polypeptide sequences to be encoded can include the C_1-C_4 subequences corresponding to any of SEQ ID NOs:1-30, 31-60, 61-90, and 91-120, respectively.

One set of oligonucleotides encodes regions of the plant scaffold polypeptide sequence where diversity is to be introduced (e.g., at X_1, X_2, and X_3). In contrast, regions of the scaffold polypeptide sequence in which little or no variation is to be introduced (e.g., in backbone domains of PAP scaffold polypeptide sequences) are encoded by a set of oligonucleotides encoding amino acid sequences with no less than 70% (i.e., 75%, 80%, 85%, 90%, 95%, or 100%) identity to any one of the above-mentioned scaffold polypeptide sequences. The details of this method are described, e.g., in U.S. patent No. 6,521,453.

Sequence-varied oligonucleotides used to generate libraries of nucleic acids are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers (1981), Tetrahedron Letts., 22(20):1859-1862, e.g., using an automated synthesizer, as described in Needham-VanDevanter *et al.* (1984) Nucleic Acids Res., 12:6159-6168. A wide variety of equipment is commercially available for automated oligonucleotide synthesis. Multi-nucleotide synthesis approaches (e.g., tri-nucleotide synthesis), as discussed, supra, are also useful.

Nucleic acids can be custom ordered from a variety of commercial sources, such as Sigma-Genosys (at sigma-genosys.com/oligo.asp); The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (at genco.com), ExpressGen Inc. (at expressgen.com), Operon Technologies Inc. (Alameda, Calif.) and many others.

The oligonucleotides can have a codon use optimized for expression in a particular cell type (e.g., in a plant cell, a mammalian cell, a yeast cell, or a bacterial cell). Codon usage frequency tables are publicly available, e.g., on the world wide web at kazusa.or.jp/codon. Codon biasing can be used to optimize expression in a cell or

on the surface of a cell in which binding of a plant chimeric binding protein is to be assessed, and can also be used to optimize expression of the chimeric binding protein in a transgenic organism of commercial interest (e.g., a transgenic plant). In general, codons with a usage frequency of less than 10% are not used. Before synthesis oligonucleotide sequences are checked for potentially problematic sequences, e.g, restriction sites useful for subcloning, potential plant splice acceptor or donor sites (see, e.g., cbs.dtu.dk/services/FeatureExtract/), potential mRNA destabilization sequences (e.g., "ATTTA"), and stretches of more than four occurrences of the same nucleotide. Potentially problematic sequences are changed accordingly.

Populations of oligonucleotides are synthesized that encode amino acid variations in the putative binding regions of the selected scaffold polypeptide sequence (e.g., in regions X_1, X_2, and X_3 of a PAP scaffold polypeptide sequence).

Preferably, all of the oligonucleotides of a selected length (e.g., about 10, 12, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 or more nucleotides) that correspond to regions where sequence diversity is to be introduced in the scaffold polypeptide sequence encode all possible amino acid variations from a diverse set of amino acids as described above. This includes N oligonucleotides per N sequence variations, where N is the number of different sequences at a locus. The N oligonucleotides are identical in sequence, except for the nucleotide(s) encoding the variant amino acid(s). In generating the sequence-varied oligonucleotides, it can be advantageous to utilize parallel or pooled synthesis strategies in which a single synthesis reaction or set of reagents is used to make common portions of each oligonucleotide. This can be performed e.g., by well-known solid-phase nucleic acid synthesis techniques, or, e.g., utilizing array-based oligonucleotide synthetic methods (see e.g., Fodor *et al.* (1991) Science, 251: 767-777; Fodor (1997) "Genes, Chips and the Human Genome" FASEB Journal. 11:121-121; Fodor (1997) "Massively Parallel Genomics" Science. 277:393-395; and Chee *et al.* (1996) "Accessing Genetic Information with High-Density DNA Arrays" Science 274:610-614).

In typical synthesis strategies the oligonucleotides have at least about 10 bases of sequence identity to either side of a region of variance to ensure reasonably efficient recombination. However, flanking regions with identical bases can have fewer identical bases (e.g., 4, 5, 6, 7, 8, or 9) and can, of course, have larger regions of identity (e.g., 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 50, or more).

The oligonucleotides to be assembled together are incubated to allow hybridization between oligonucleotides containing overlapping complementary sequences. Each set of hybridizing overlapping oligonucleotides thereby forms a contiguous nucleic acid interrupted by small gaps. These small gaps can be filled to form full length sequences using any of a variety of polymerase-mediated reassembly methods, e.g.,·as described herein and as known to one of skill. The greatest sequence diversity is introduced in oligonucleotides encoding the plant scaffold polypeptide sequence putative binding regions and residues. However, oligonucleotides encoding specific sequence variations can be "spiked" in the recombination mixture at any selected concentration, thus causing preferential incorporation of desirable modifications into the encoded plant chimeric binding proteins in regions outside of the putative binding domains.

For example, during oligonucleotide elongation, hybridized oligonucleotides are incubated in the presence of a nucleic acid polymerase, e.g., Taq, Klenow, or the like, and dNTP's (i.e., dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq or other high-temperature polymerase can be used with a hybridization temperature of between about room temperature (i.e., about 25 °C) and, e.g., about 65 °C. If the areas of identity are small, Klenow, Taq or polymerases can be used with a hybridization temperature of below room temperature. The polymerase can be added to the assembly reaction prior to, simultaneously with, or after hybridization of the oligonucleotides. Afterwards, the resulting elongated double-stranded nucleic acid sequences are denatured, hybridized, and elongated again. This cycle can be repeated for any desired number of times. The cycle is repeated e.g., from about 2 to about 100 times.

Optionally, after multiple cycles of combinatorial nucleic acid assembly, the resulting products can be amplified, e.g., by standard polymerase chain reaction (PCR). A portion of the volume of the above-described assembly reaction is incubated with unique forward and reverse primers that hybridize universally to the ends of the nucleic acids, as well as dNTPs and a suitable polymerase (e.g., *pfu* polymerase). The PCR reaction is then carried out for about 10 to 40 cycles.

To determine the extent of oligonucleotide incorporation any approach which distinguishes similar nucleic acids can be used. For example, the nucleic acids can be cloned and sequenced, or amplified (*in vitro* or by cloning, e.g., into a standard cloning

or expression vector) and cleaved with a restriction enzyme which specifically recognizes a particular oligonucleotide sequence variant.

It is useful to include rare restriction sites (e.g., Not I) in the 5' ends of the 5' and 3' most primers used either in the assembly or PCR reactions. Inclusion of restriction sites in these primers facilitates subcloning of the nucleic acids into a vector by restriction digestion and subsequent ligation. Alternatively, the assembly reaction or PCR products can also be subcloned, without being restriction digested, using standard methods, e.g.,"TA" cloning.

Other methods for introducing diversity into a plant scaffold polypeptide sequence can also be used. For example, a scaffold polypeptide sequence can be encoded in a nucleic acid template, e.g., a plasmid contruct. Alternatively, a PCR product, mRNA or genomic DNA from an appropriate plant species such as soybean may also serve as a template encoding a plant scaffold polypeptide sequence. One or more scaffold polypeptide sequence subsequences to be diversified (e.g., the X_2 region of a PAP scaffold polypeptide sequence) can be diversified during or after amplification from the scaffold polypeptide sequence nucleic acid template by any of a number of error-prone PCR methods. Error-prone PCR methods can be divided into (a) methods that reduce the fidelity of the polymerase by unbalancing nucleotides concentrations and/or adding of chemical compounds such as manganese chloride (see, e.g., Lin-Goerke *et al.* (1997) *Biotechniques*, 23, 409–412), (b) methods that employ nucleotide analogs (see, e.g., U.S. Patent No. 6,153,745), (c) methods that utilize 'mutagenic' polymerases (see, e.g., Cline, J. and Hogrefe,H.H. (2000) *Strategies* (Stratagene Newsletter), 13, 157–161 and (d) combined methods (see, e.g., Xu, H., Petersen, E.I., Petersen, S.B. and el-Gewely, M.R. (1999) *Biotechniques*, **27**, 1102–1108. Other PCR-based mutagenesis methods include those, e.g., described by Osuna J, Yanez J, Soberon X, and Gaytan P. (2004), *Nucleic Acids Res.* 2004, 32(17):e136 and Wong TS, Tee KL, Hauer B, and Schwaneberg, Nucleic Acids Res. 2004 Feb 10;32(3):e26), and others known in the art.

After generating a population of sequence variants, these can be substituted into the appropriate region of a chosen plant scaffold polypeptide sequence nucleic acid (e.g., a plasmid containing a scaffold polypeptide sequence) by subcloning which thereby effectively acts as a vector for the library of diversified sequences.

Yet another approach to mutagenizing specific plant scaffold polypeptide sequence regions is the use of a mutagenic *E. coli* strain (see, e.g., Wu *et al.* (1999), *Plant Mol. Biol.*, 39(2):381-386). A nucleic acid vector containing a target sequence to be mutated is introduced into the mutator strain, which is then propagated. Error-prone DNA replication in the mutator *E. coli* strain introduces mutations into the introduced target sequence. The population of altered target sequences is then recovered and subcloned into the appropriate position of a nucleic acid encoding the selected plant scaffold polypeptide sequence to generate a diverse library of nucleic acids encoding plant chimeric binding proteins.

III. Expression and Screening of Plant chimeric binding proteins

The library of nucleic acids based on a plant scaffold polypeptide sequence and encoding plant chimeric binding polypeptides are subcloned into an expression vector and introduced into a biological replication system to generate an expression library. The expression library can be propagated and screened to identify plant chimeric binding proteins that bind a target molecule (TM) of interest (e.g., a nematode, insect, fungal, viral or plant protein).

The biological replication system on which screening of plant chimeric binding proteins will be practiced should be capable of growth in a suitable environment, after selection for binding to a target. Alternatively, the nucleic acid encoding the selected plant chimeric binding protein can be isolated by in vitro amplification. During at least part of the growth of the biological replication system, the increase in number is preferably approximately exponential with respect to time. The frequency of library members that exhibits the desired binding properties may be quite low, for example, one in 10^6 or less.

Biological replication systems can be bacterial DNA viruses, vegetative bacterial cells, bacterial spores. Eukaryotic cells (e.g., yeast cells) can also be used as a biological replication system.

In a particularly useful embodiment, a chimeric binding protein-phage coat protein fusion is encoded in a phagemid construct. The phagemid constructs are transformed into host bacteria, which are subsequently infected with a helper phage that expresses wild type coat proteins. The resulting phage progeny have protein coats that include both fusion protein and wild-type coat proteins. This approach has the

advantage that phage viability is greater compared to viability of phage that have exclusively chimeric binding protein-coat fusion proteins. Phagemid-based display library construction and screening kits are commercially available, e.g., the EZnet™ Phage Display cDNA Library Construction Kit and Screening Kit (Maxim Biotech, Inc., San Francisco, CA).

Nonetheless, a strain of any living cell or virus is potentially useful if the strain can be: 1) genetically altered with reasonable facility to encode a plant chimeric binding protein, 2) maintained and amplified in culture, 3) manipulated to display the potential binding protein domain where it can interact with the target material, and 4) selected while retaining the genetic information encoding the expressed plant chimeric binding protein in recoverable form. Preferably, the biological replication system remains viable after affinity-based selection.

When the biological replication system is a bacterial cell or a phage which is assembled in the periplasm, the expression vector for display of the plant chimeric binding protein encodes the chimeric binding protein itself fused to two additional components. The first component is a secretion signal which directs the initial expression product to the inner membrane of the cell (a host cell when the package is a phage). This secretion signal is cleaved off by a signal peptidase to yield a processed, mature, plant chimeric binding protein. The second component is an outer surface transport signal which directs the biological replication system to assemble the processed protein into its outer surface. This outer surface transport signal can be derived from a surface protein native to the biological replication system (e.g., the M13 phage coat protein gIII).

For example, the expression vector comprises a DNA encoding a plant chimeric binding protein operably linked to a signal sequence (e.g., the signal sequences of the bacterial phoA or bla genes or the signal sequence of M13 phage gene III) and to DNA encoding a coat protein (e.g., the M13 gene III or gene VIII proteins) of a filamentous phage (e.g., M13). The expression product is transported to the inner membrane (lipid bilayer) of the host cell, whereupon the signal peptide is cleaved off to leave a processed hybrid protein. The C-terminus of the coat protein-like component of this hybrid protein is trapped in the lipid bilayer, so that the hybrid protein does not escape into the periplasmic space. As the single-stranded DNA of the nascent phage particle passes into the periplasmic space, it collects both wild-type coat protein and the hybrid

protein from the lipid bilayer. The hybrid protein is thus packaged into the surface sheath of the filamentous phage, leaving the plant chimeric binding protein exposed on its outer surface. Thus, the filamentous phage, not the host bacterial cell, is the biological replication system in this embodiment. If a secretion signal is necessary for the display of the plant chimeric binding protein, a "secretion-permissive" bacterial strain can be used for growth of the filamentous phage biological replication system.

It is unnecessary to use an inner membrane secretion signal when the biological replication system is a bacterial spore, or a phage whose coat is assembled intracellularly. In these cases, the display means is merely the outer surface transport signal, typically a derivative of a spore or phage coat protein.

Filamentous phage in general are attractive as biological replication systems for display of plant chimeric binding proteins, and M13 in particular, is especially attractive because: 1) the 3D structure of the virion is known; 2) the processing of the coat protein is well understood; 3) the genome is expandable; 4) the genome is small; 5) the sequence of the genome is known; 6) the virion is physically resistant to shear, heat, cold, urea, guanidinium Cl, low pH, and high salt; 7) the phage is a sequencing vector so that sequencing is especially easy; 8) antibiotic-resistance genes have been cloned into the genome; 9) It is easily cultured and stored, with no unusual or expensive media requirements for the infected cells, 10) it has a high burst size, each infected cell yielding 100 to 1000 M13 progeny after infection; and 11) it is easily harvested and concentrated by standard methods.

For example, when the biological replication system is M13 the gene III or the gene VIII proteins can be used as an outer surface targeting signal. Alternatively, the proteins from genes VI, VII, and IX may also be used.

The encoded plant chimeric binding protein can be fused to the surface targeting signal (e.g., the M13 gene III coat protein) at its carboxy or amino terminal. The fusion boundary between the plant chimeric binding protein and the targeting signal can also include a short linker sequence (e.g., up to 20 amino acids long) to avoid undesirable interactions between the chimeric binding protein and the fused targeting signal. In some embodiments it is advantageous to include within the linker sequence a specific proteolytic cleavage site. In addition, the amino terminal or carboxy terminal of the fused protein can include a short epitope tag (e.g., a hemaglutinin tag). Inclusion of a proteolytic cleavage site or a short epitope tag is particularly useful for purification of a

library of chimeric binding proteins from a population of cells expressing the library. Epitope-tagged chimeric binding proteins can be conveniently purified by proteolytic cleavage of linker sequence followed by affinity chromatography utilizing an antibody or other binding agent that recognizes the epitope tag.

Many methods exist for screening phage display libraries (see, e.g., Willats (2002), *Plant Mol. Biol.*, 50:837-854). As commonly practiced, the target molecule of interest is adsorbed to a support and then exposed to solutions of phage displaying plant chimeric binding proteins. The target molecule can be immobilized by passive adsorption on a support medium, e.g, tubes, plates, columns, or magnetic beads. Generally, the adsorptive support medium is pre-blocked, e.g., with bovine serum albumin, milk, or gelatin, to reduce non-specific binding of the phage during screening. Alternatively, the target molecule can be biotinylated, so interaction between chimeric binding protein-bearing phage and the target molecule can be carried out in solution. Phage that bind to the target can then be selected using avidin or streptavidin bound to a solid substrate (e.g., beads or a column).

After phage are allowed to interact with the target molecule, non-interacting phage are removed by washing. The remaining, specifically binding phage are then eluted by one of any number of treatments including, e.g., lowering or increasing pH, application of reducing agents, or use of detergents. In one embodiment, a specific proteolytic cleavage site is introduced between the plant chimeric binding protein sequence and the phage coat protein sequence. Thus, phage elution can be accomplished simply by addition of the appropriate protease.

Eluted phage are then amplified by infection of host cells and can subsequently be re-screened by the method just outlined to reduce the number of false positive binders. During each round of phage screening, care should be taken to include growth of the phage on a solid medium rather than exclusively in a liquid medium as this minimizes loss of phage clones that grow sub-optimally.

Plant chimeric binding proteins can also be expressed and screened for binding solely *in vitro* using ribosomal display. An exclusively *in vitro* approach circumvents the requirement to introduce the library of nucleic acids encoding plant chimeric binding proteins into a biological replication system. Methods for screening polypeptides *in vitro* by ribosomal protein display are described in detail, e.g., in U.S. Patent No. 6,589,741. The nucleic acids described in the section above are modified by

adding a phage promoter sequence (e.g., a T7 promoter) enabling *in vitro* transcription, a ribosome binding sequence upstream to the start of translation of the encoded plant chimeric binding protein, and a transcription termination sequence (e.g., from phage T3). The modified library of nucleic acids is then transcribed *in vitro* to generate a corresponding mRNA population encoding plant chimeric binding proteins. Plant chimeric binding proteins are then expressed *in vitro* by translating the population of mRNA molecules devoid of stop codons in the correct reading frame in an in vitro translation system, under conditions that allow the formation of polysomes. The polysomes so formed are then brought into contact with a target molecule under conditions that allow the interaction of plant chimeric binding proteins with the target molecule. Polysomes displaying chimeric binding proteins that interact with the target molecule are then separated from non-interacting polysomes displaying no such (poly)peptides; and the mRNA associated with the interacting polysome is then amplified (e.g., by PCR) and sequenced.

Interaction of a plant chimeric binding protein with a target protein can also be detected in a genetic screen. In the screen, the target protein functions as a "bait protein" and each plant chimeric binding protein functions as a potential "prey" protein in a binding assay that utilizes a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos *et al.* (1993) Cell 72:223-232; Madura *et al.* (1993) J. Biol. Chem. 268:12046-12054; Bartel *et al.* (1993) Biotechniques 14:920-924; Iwabuchi *et al.* (1993) Oncogene 8:1693-1696; Hubsman *et al.* (2001) Nuc. Acids Res. Feb 15;29(4):E18; and Brent WO94/10300).

A two-hybrid assay can be carried out using a target polypeptide as the bait protein. In sum, the target polypeptide is fused to the LexA DNA binding domain and used as bait. The prey is plant chimeric binding protein library cloned into the active site loop of TrxA as a fusion protein with an N-terminal nuclear localization signal, a LexA activation domain, and an epitope tag (Colas *et al.* 1996 Nature 380:548; and Gyuris *et al.* Cell 1993 75:791). Yeast cells are transformed with bait and prey genes. When the target fusion protein binds to a plant chimeric binding protein fusion protein, the LexA activation domain is brought into proximity with the LexA DNA binding domain and expression of reporter genes or selectable marker genes having an appropriately positioned LexA binding site increases. Suitable reporter genes include fluorescent proteins (e.g., EGFP), enzymes (e.g., luciferase, β-galactosidase, alkaline

phosphatase, etc.) Suitable selectable marker genes include, for example, the yeast LEU2 gene.

After identification of one or more target-binding chimeric binding proteins, the isolated nucleic acids encoding the chimeric binding proteins can be mutagenized by the methods described herein, to generate small expression libraries expressing variant chimeric binding proteins. The chimeric binding protein-variant expression libraries can be screened to identify chimeric binding protein variants with improved target binding properties (e.g., increased affinity or specificity).

The following specific examples are to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever. Without further elaboration, it is believed that one skilled in the art can, based on the description herein, utilize the present invention to its fullest extent.

EXAMPLES

Example 1 Design and Expression of Plant Scaffold Polypeptide Sequences

Several protein domain families were analyzed for their potential use as scaffolds. A search of PFAM domains (pfam.wustl.edu; see Bateman *et al.* (2004)), restricting the output to *Viridiplantae*, was conducted to limit domains only to those present in green plants. Four protein domain families were selected to develop plant universal molecular recognition libraries; the accessory domain of purple acid phosphatase (PAP), plant cystatins, plant C2 domains and the turn-helix-helix (THH) motif found in ankyrin repeat proteins.

Three purple acid phosphatase scaffolds were designed having the sequence of SEQ ID NOs:34-36. The amino acid sequence of the accessory domain from kidney bean PAP was used as a query sequence to BLAST the NCBI database. When the output was restricted to proteins found in Viridiplantae, 62 unique sequences were identified. From an alignment of these sequences, a consensus plant PAP sequence was generated (SEQ ID NO:34) by selecting the most frequent amino acid at each position in the alignment. The kidney bean (*Phaseolus vulgaris*) PAP was selected as a parental scaffold (SEQ ID NO:35), because of its known structure. A PAP from soybean,

Glycine max, was also chosen (SEQ ID NO:36), as this species represents a common crop species in which transgenic products are generated.

A set of scaffold polypeptide sequences which contain plant ankyrin-like repeats was also designed. Ankyrin-like repeats are small turn-helix-helix (THH) motifs consisting of approximately 33 amino acids. They are common elements of proteins from all organisms and are often found in tandem arrays of 2 to 20 repeats within a protein.

Three THH scaffolds were generated. These proteins are similar in structure to GA binding protein (GABP-β). This protein consists of THH like amino and carboxy terminal caps with 3 THH internal repeats. In this protein, it is thought that the caps help stabilize the protein by shielding hydrophobic residues found in the internal repeats.

Three hundred and twelve Viridiplantae ankyrin repeats proteins found in PFAM were aligned to aid in designing plant-specific THH scaffolds. A plant consensus THH sequence was generated by selecting the most frequently occurring amino acid at each position. This sequence was termed the plant consensus internal repeat sequence. This sequence was used to search the NCBI databases by BLAST alignment to find the closest natural THH sequence found in plants. A sequence from wheat (*Triticum aestivum*) was found. The designed repeat based on *T. aestivum* contains a substitution of valine for the single cysteine occurring in the *T. aestivum* sequence. Two sets of N and C terminal caps were generated. One set consists of sequences derived from GABP-β and the second set was derived from the plant THH consensus sequence and optimized to resemble the structure of GABP-β. In particular, the N terminal cap has an extended alpha-helical structure, while the C terminal cap has a truncated helix compared to the typical THH repeat.

Three THH scaffolds were designed, one consists of plant consensus N and C caps and two plant consensus internal THH repeats (SEQ ID NO:37). Another consists of plant consensus N and C caps and two wheat internal repeats (SEQ ID NO:38) and the third consists of ankyrin like N and C caps with two wheat internal repeats (SEQ ID NO:39).

The genes encoding the plant scaffold polypeptide sequences were designed for expression testing in plants, bacteria, and on the surface of phage. Codons were selected for plant expression using a publicly available *Glycine max* codon usage table

(at kazusa.or.jp/codon, codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nakamura, Y, Gojobori, T and Ikemura, T (2000) *Nucl. Acids Res.* 28:292.). Codon selection was done manually with the aim for the final codon frequency to roughly reflect the natural frequency for *Glycine max*. Rarely used codons (<10% frequency) were not used. Final sequences were checked for potential problematic sequences, including removal of restriction sites needed for cloning, potential plant splice acceptor or donor sites (see website at cbs.dtu.dk/services/NetPgene/), potential mRNA destabilization sequences (ATTTA) and stretches of more than 4 occurrences of the same nucleotide. Any potential problematic sequences were altered in the genes by modifying codon usage. Since the THH sequences have 4 similar repeat sequences within each protein, steps were taken to reduce nucleotide similarity within repeats; the average repeat identity was reduced 10-15% by these means.

Seven constructs were produced using synthetic gene assembly, (three based on THH scaffold polypeptide sequences, two based on PAP scaffold polypeptide sequences, one plant cystatin and one plant C2 domain protein). The three THH scaffold polypeptide sequences were placed into a phagemid vector as fusion sequences with the gene III coat protein (gIII) at its carboxy terminus (Phage 3.2, Maxim Biotech, Inc., South San Francisco, CA). A 6-His tag was included at the 5' end of the gene as well as a c-Myc tag between the scaffold gene and the encoded amino terminus of gIII. The phagemid constructs were then packaged into phage particles and the phage were tested for expression and surface display of the THH scaffold. A phage ELISA using either anti-His and anti-Myc indicated that the THH scaffold proteins were expressed on the surface of phage in phage ELISAs, suggesting that all 3 THH scaffold polypeptide sequence constructs are folding and expressing well on the phage surface. The selected scaffold polypeptide sequences were then used to generate expression vectors to evaluate their expression in transgenic plants by immunoblotting.

Tobacco leaves were injected with *agrobacterium*, LB4404 transformed with THH containing plant expression vectors. Two days later, sections of leaves injected with agrobacterium were harvested, frozen on dry ice, then ground into a fine powder with a pestle. PBS containing 0.2% Tween-20 was added to the fine powder at a 1:1 weight to volume ratio and additional grinding was done. Insoluble material was removed by centrifugation and 10 ul of the remaining supernatant was loaded onto a 4-

12% acrylamide SDS page gel (NuPage, Intvitrogen). Proteins were transferred to PVDF membranes. Proteins were detected using a rat anti-HA antibody (Roche) and an anti-rat HRP conjugated secondary antibody (Chemicon). HRP was detected using Amerham Lumigen reagents.

All three THH scaffold were found to be expressed, with the relative level of expression of the three scaffolds being TA-THH > CC-THH > TC-THH.

OTHER EMBODIMENTS

All of the features disclosed in this specification may be combined in any combination. Each feature disclosed in this specification may be replaced by an alternative feature serving the same, equivalent, or similar purpose. Thus, unless expressly stated otherwise, each feature disclosed is only an example of a generic series of equivalent or similar features.

The scope of the claims should not be limited by the preferred embodiments set forth in the examples, but should be given the broadest interpretation consistent with the description as a whole.

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. A library of nucleic acids encoding at least ten different polypeptides, the amino acid sequence of each polypeptide comprising:

C_1-X_1-C_2-X_2-C_3-X_3-C_4, wherein

(i) subsequence C_1 is selected from SEQ ID NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90; subsequence C_4 is selected from SEQ ID NOs:91-120, and each of C_1-C_4 independently comprises up to 10 single conservative substitutions to the selected subsequence;

(ii) C_1-C_4 are homogeneous across the encoded polypeptides;

(iii) each of X_1-X_3 is an independently variable subsequence consisting of 2-20 amino acids; and

(iv) each of X_1-X_3 are heterogeneous across the encoded polypeptides.

2. The library of claim 1, wherein at least 1,000 different polypeptides are encoded.

3. The library of claim 1, wherein at least 100,000 different polypeptides are encoded.

4. The library of claim 1, wherein at least 1,000,000 different polypeptides are encoded.

5. The library of any one of claims 1 to 4, wherein the amino acids of X_1-X_3 are selected from fewer than 20 amino acids.

6. The library of claim 5, wherein the fewer than 20 amino acids include at least one aliphatic amino acid, at least one acidic amino acid, at least one neutral amino acid, and at least one aromatic amino acid.

7. The library of claim 6, wherein the fewer than 20 amino acids comprise alanine, aspartate, serine, and tyrosine.

8. The library of any one of claims 1 to 7, wherein the amino acid sequence of each polypeptide is selected from:

(a). a polypeptide comprising C_1-X_1-C_2-X_2-C_3-X_3-C_4 wherein C_1= SEQ ID NO:1, C_2= SEQ ID NO: 31, C_3= SEQ ID NO: 61, and C_4= SEQ ID NO: 91;

(b). a polypeptide comprising C_1-X_1-C_2-X_2-C_3-X_3-C_4 wherein C1= SEQ ID NO:2, C2= SEQ ID NO: 32, C3= SEQ ID NO: 62, and C4= SEQ ID NO: 92; and

(c). a polypeptide comprising C_1-X_1-C_2-X_2-C_3-X_3-C_4 wherein C1= SEQ ID NO:3, C2= SEQ ID NO: 33, C3= SEQ ID NO: 63, and C4= SEQ ID NO: 93.

9. The library of any one of claims 1 to 8, wherein each of the nucleic acids comprises a vector sequence.

10. A population of cells expressing the polypeptides encoded by the library of any one of claims 1 to 9.

11. A library of purified polypeptides encoded by the library of any one of claims 1 to 9.

12. A population of filamentous phage displaying the polypeptides encoded by the library of any one of claims 1 to 9.

13. A method of generating the library of any one of claims 1 to 9, comprising:

(i) providing a parental nucleic acid encoding a parental polypeptide comprising the amino acid sequence: C_1-X_1-C_2-X_2-C_3-X_3-C_4, wherein subsequence C_1 is selected from SEQ ID NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90; subsequence C_4 is selected from SEQ ID NOs:91-120; each of C_1-C_4 independently comprises up to 10 single conservative substitutions to the selected subsequence; and each of X_1-X_3 is an independent subsequence consisting of 2-20 amino acids;

(ii) replicating the parental nucleic acid under conditions that introduce up to 10 single conservative substitutions to the X_1, X_2, or X_3 subsequences, whereby a population of randomly varied subsequences encoding X_1', X_2', or X_3' is generated; and

(iii) the population of randomly varied subsequences encoding X_1', X_2', or X_3'is substituted into a population of parental nucleic acids at the positions corresponding to those that encode X_1, X_2, or X_3.

14. The method of claim 13, wherein at least one of the X_1-X_3 subsequences is selected from SEQ ID NOs:121-123.

15. The method of claim 13 or 14, wherein the replicating generates a heterogeneous population of randomly varied subsequences by introducing up to 5 amino acid substitutions in each of X_1, X_2, or X_3.

16. The method of any one of claims 13 to 15, further comprising amplifying the library by introducing it into a biological replication system and proliferating the biological replication system.

17. The method of claim 16, wherein the biological replication system is a plurality of *E. coli* cells.

18. The method of claim 16, wherein the biological replication system is a plurality of bacteriophage.

19. The method of any one of claims 13 to 18, wherein the replicating occurs *in vitro*.

20. The method of claim 19, wherein the replicating is performed with a purified mutagenic polymerase.

21. The method of claim 19 or 20, wherein the replicating is performed in the presence of a nucleotide analog.

22. The method of any one of claims 13 to 18, wherein the replicating occurs *in vivo*.

23. The method of claim 22, wherein the replicating *in vivo* occurs in a mutagenic species of *E. coli*.

24. A method of generating the library of any one of claims 1 to 9, comprising:

(i) selecting an amino acid sequence comprising the amino acid sequence C_1-X_1-C_2-X_2-C_3-X_3-C_4 to be encoded, wherein

(a) subsequence C_1 is selected from SEQ ID NOs:1-30, subsequence C_2 is selected from SEQ ID NOs:31-60, subsequence C_3 is selected from SEQ ID NOs:61-90, and subsequence C_4 is selected from SEQ ID NOs:91-120;

(b) each of C_1-C_4 independently comprises up to 10 conservative substitutions to the selected subsequence;

(c) each of X_1, X_2, and X_3 consists of an amino acid sequence 2-20 amino acids in length;

(ii) providing a first plurality and a second plurality of oligonucleotides, wherein

(a) oligonucleotides of the first plurality encode the C_1-C_4 subsequences and multiple heterogeneous X_1-X_3 variant subsequences X_1'-X_3';

(b) oligonucleotides of the second plurality are complementary to nucleotide sequences encoding the C_1-C_4 subsequences and to nucleotide sequences encoding multiple heterogeneous X_1'-X_3' subsequences; and

(c) the oligonucleotides of the first and second pluralities have overlapping sequences complementary to one another;

(iii) combining the first and second pluralities of oligonucleotides to form a first mixture;

(iv) incubating the mixture under conditions effective for hybridizing the overlapping complementary sequences to form a plurality of hybridized complementary sequences; and

(v) elongating the plurality of hybridized complementary sequences to form a second mixture containing the library.

25. The method of claim 24, further comprising performing a cycle of steps, the cycle of steps comprising denaturing the library by increasing the temperature of the second mixture to a temperature effective for denaturing double stranded DNA, followed by steps (iv) and (v).

26. The method of claim 25, comprising repeating the cycle of steps up to 100 times.

27. The method of claim 24, further comprising amplifying the library by a polymerase chain reaction consisting essentially of the library, a forward primer, and a reverse primer, wherein the forward and reverse primers can hybridize to the 5' and 3' end sequences, respectively, of all nucleic acids in the library.

28. The method of any one of claims 24 to 27, wherein the amino acid to be encoded in each position of the X_1, X_2, or X_3 subsequences, is selected from a subset of alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine.

29. The method of claim 28, wherein the amino acid selected for each single amino acid substitution is selected from a group of amino acids consisting of at least one aliphatic, at least one acidic, at least one neutral, and at least one aromatic amino acid.

30. The method of claim 29, wherein the group of amino acids consists of alanine, aspartate, serine, and tyrosine.

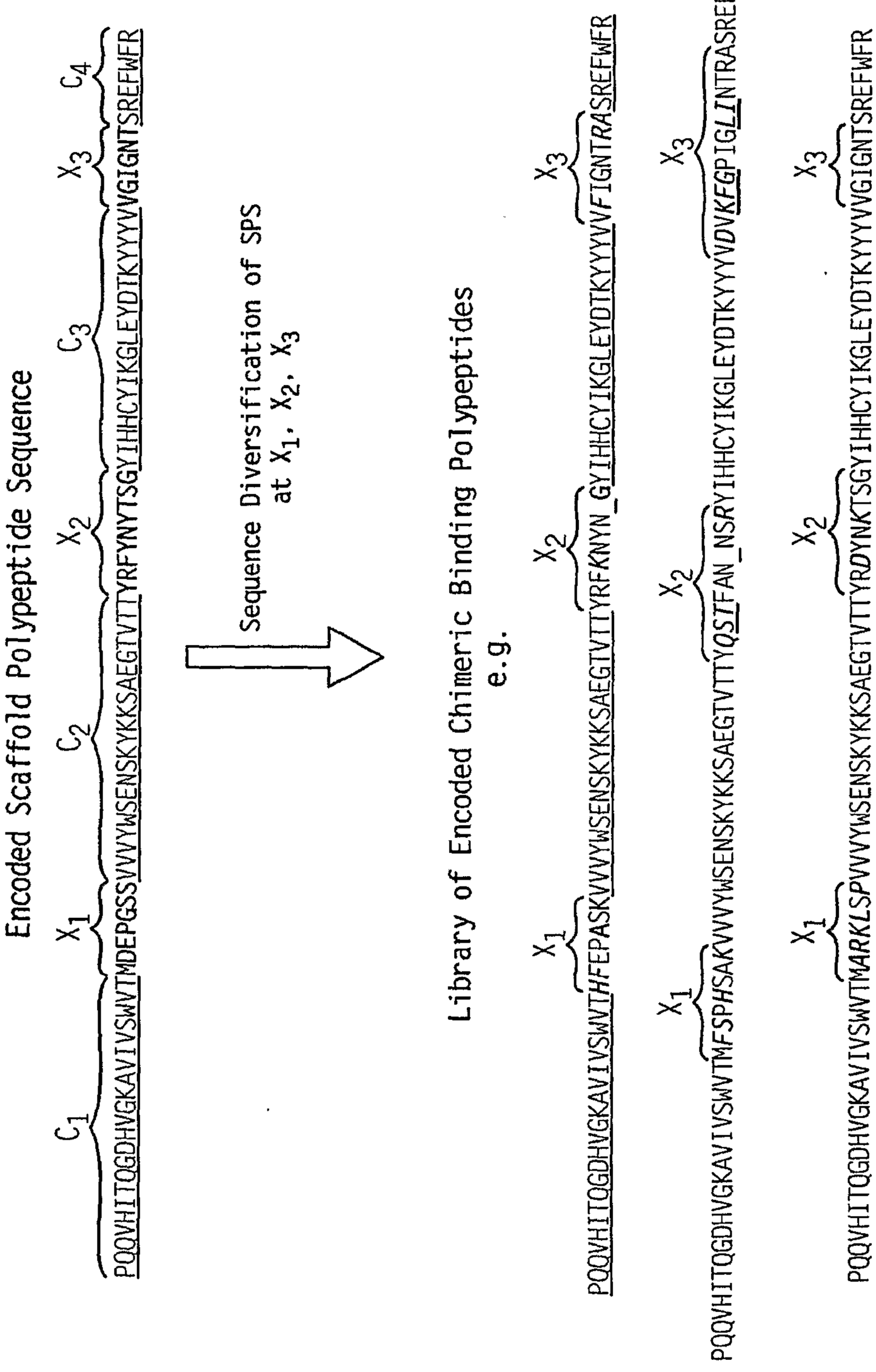

Encoded Scaffold Polypeptide Sequence
C1 X1
PQQVHITQGDHVGKAVIVSWVTMDEPGSVVVYWSENSKYKKSAEGTVTTYRFYNYTSGYIHHCYIKGLEYDTKYYVTDKFGPIGLINTRASREFWFR
Sequence Diversification of SPS at X1, X2, X3
Library of Encoded Chimeric Binding Polypeptides e.g.
PQQVHITQGDHVGKAVIVSWVTHFEPASKVVVYWSENSKYKKSAEGTVTTYRFKNYN_GYIHHCYIKGLEYDTKYYVTDKFGPIGLINTRASREFWFR
PQQVHITQGDHVGKAVIVSWVTMFSPHSAKVVVYWSENSKYKKSAEGTVTTYQSTFAN_NSRYIHHCYIKGLEYDTKYYVDVKFGPIGLINTRASREFWFR
PQQVHITQGDHVGKAVIVSWVTMARKLSPVVVYWSENSKYKKSAEGTVTTYRDYNKTSGYIHHCYIKGLEYDTKYYVTDKFGPIGLINTRASREFWFR

FIG. 1

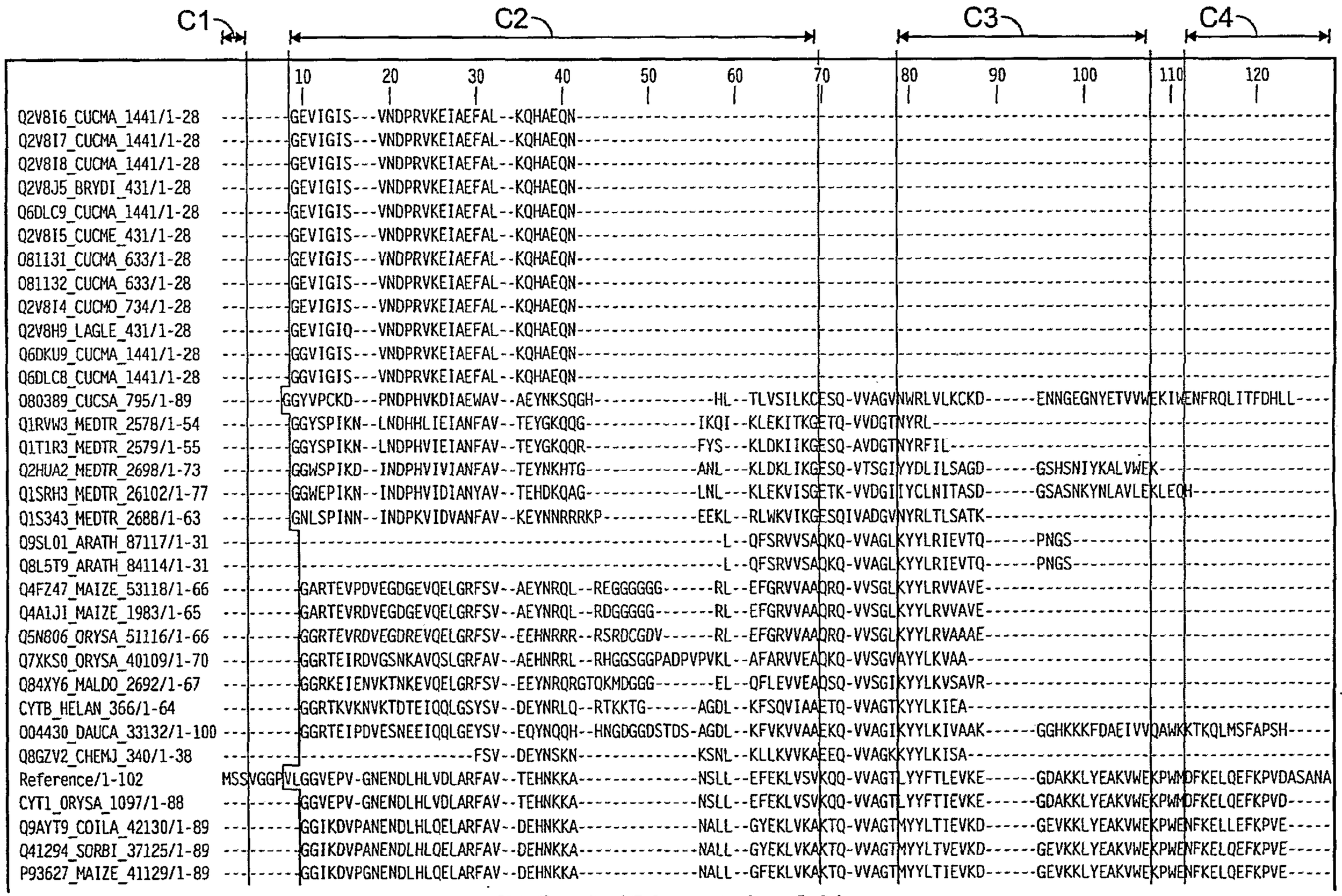


FIG. 2 (Sheet 1 of 6)

SUBSTITUTE SHEET (RULE 26)

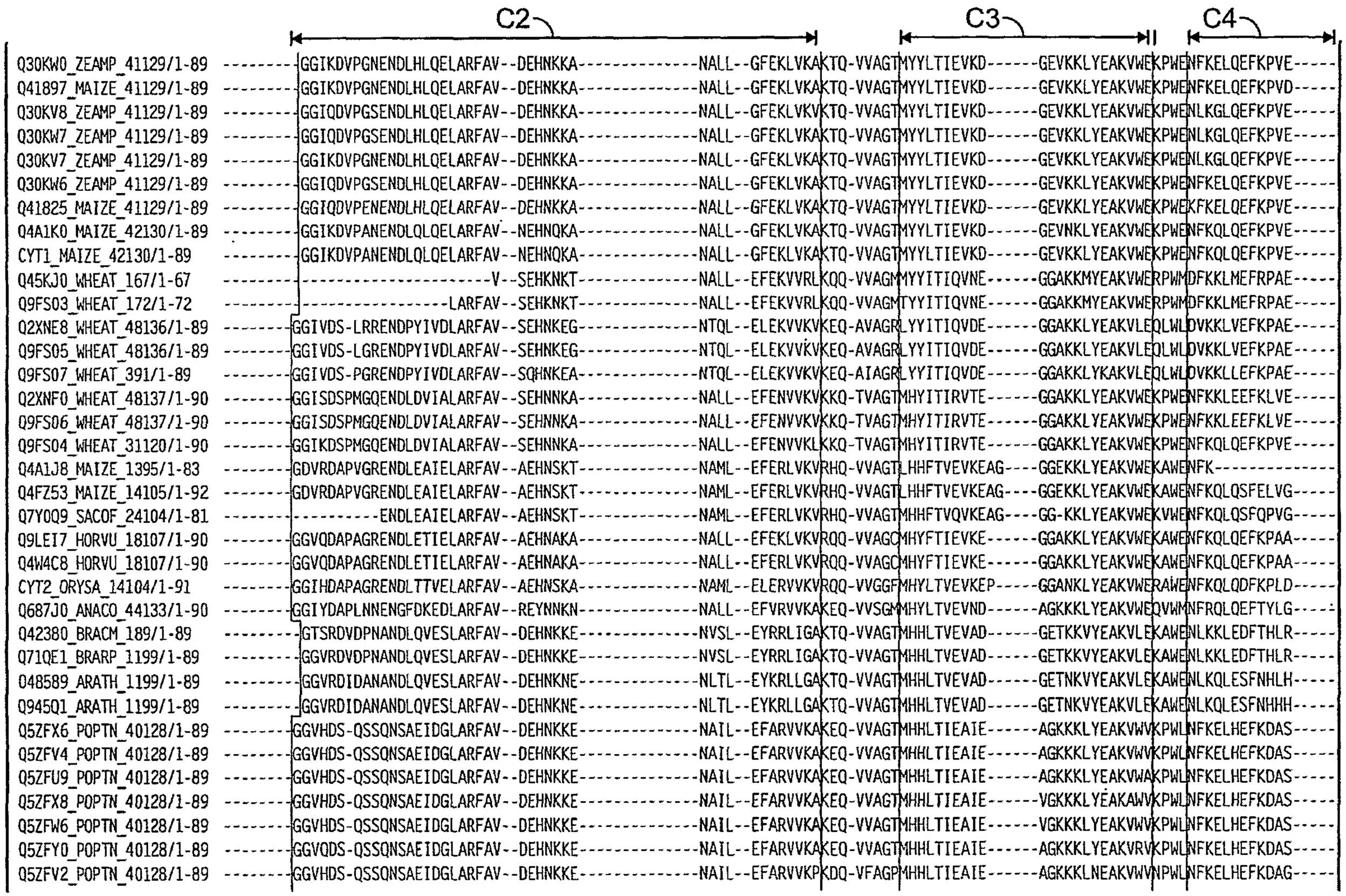


FIG. 2 (Sheet 2 of 6)

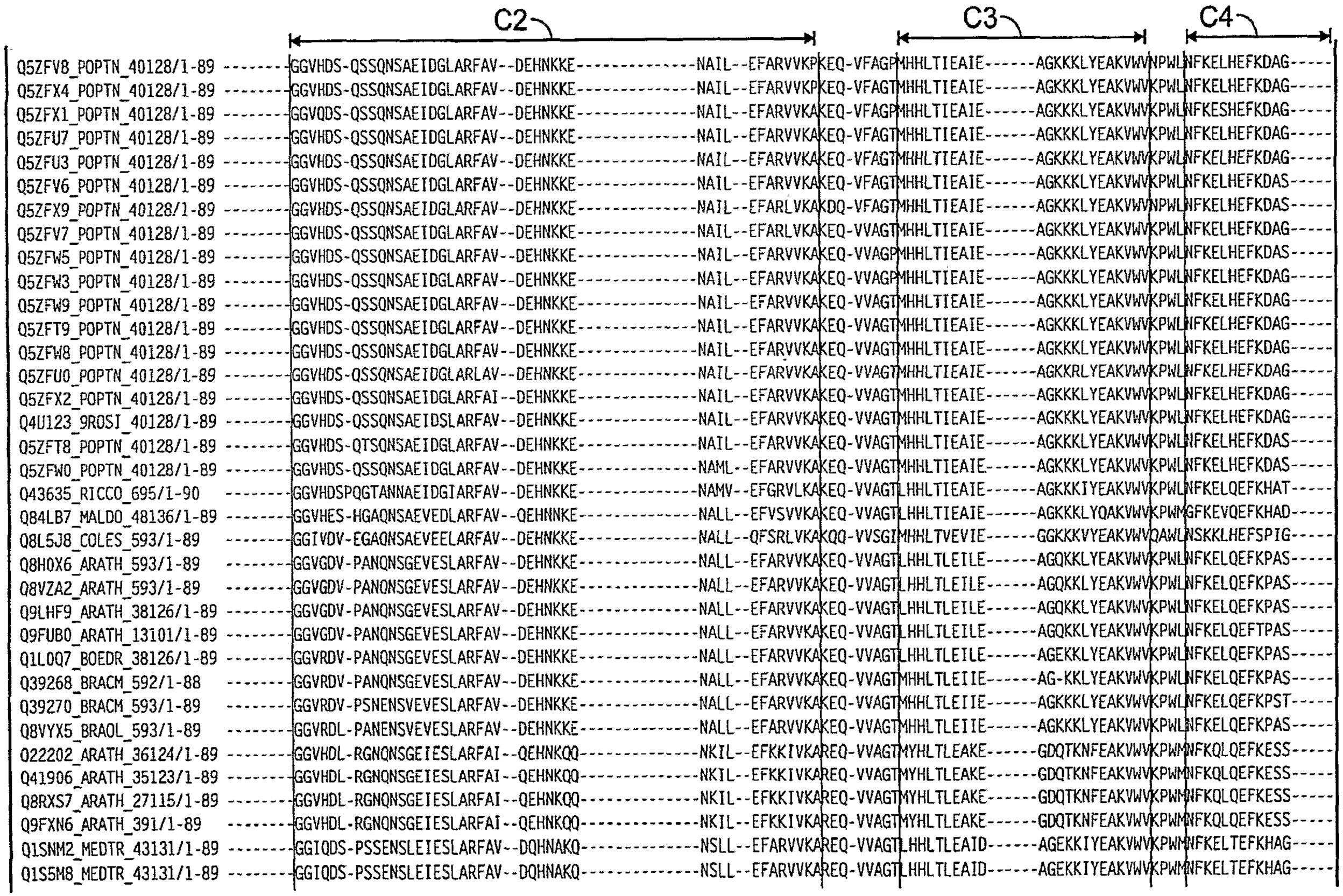


FIG. 2 (Sheet 3 of 6)

SUBSTITUTE SHEET (RULE 26)

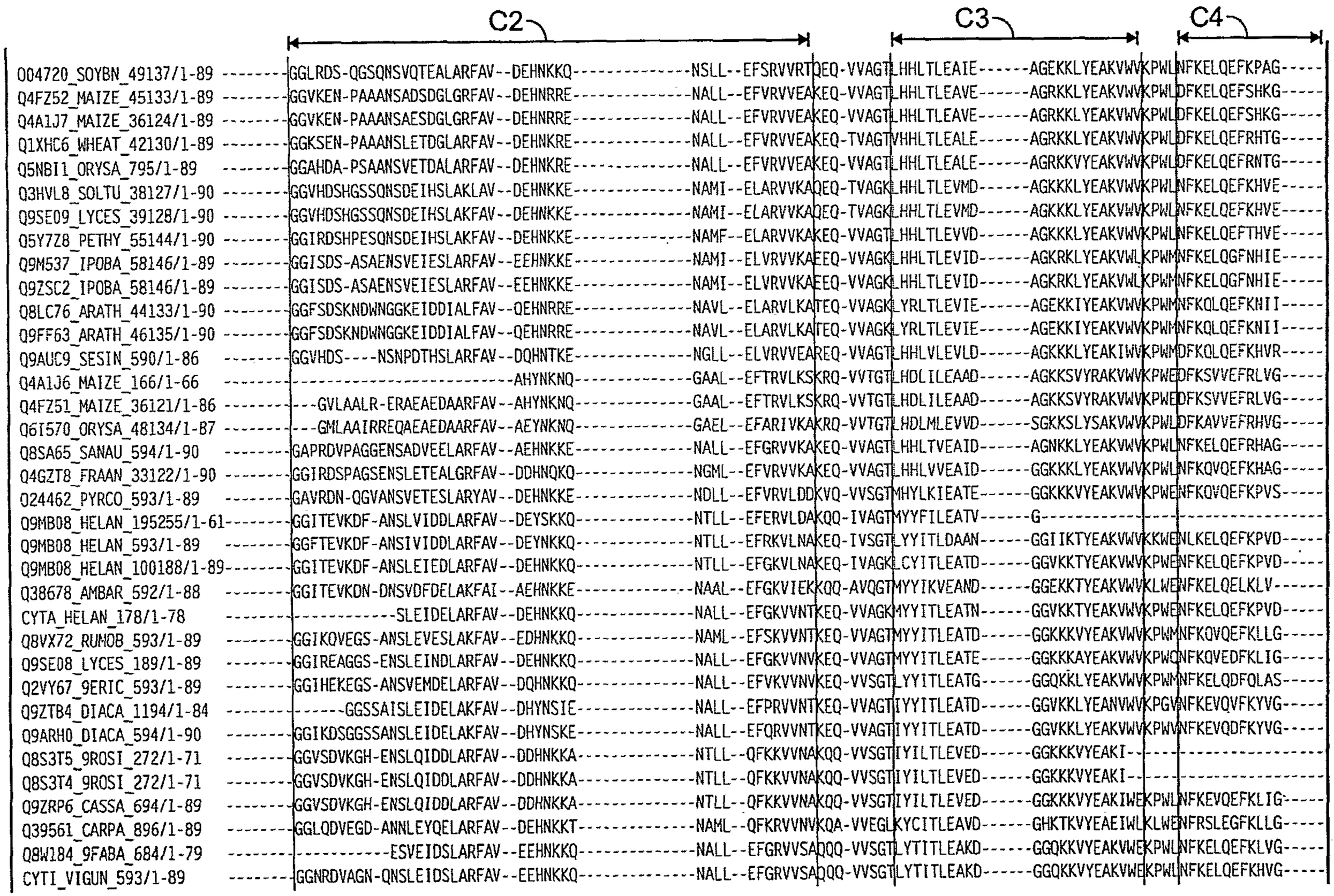


FIG. 2 (Sheet 4 of 6)

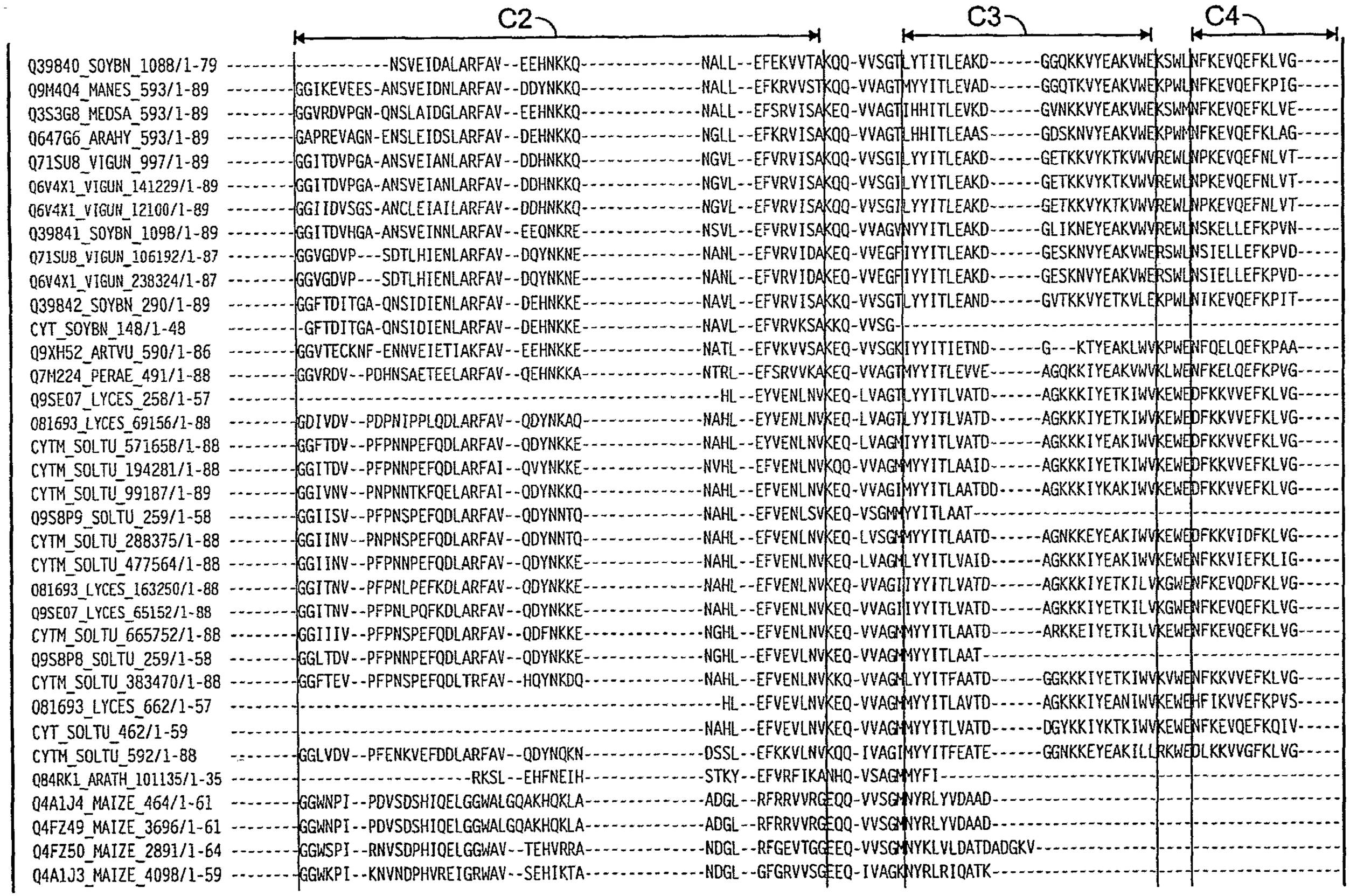


FIG. 2 (Sheet 5 of 6)

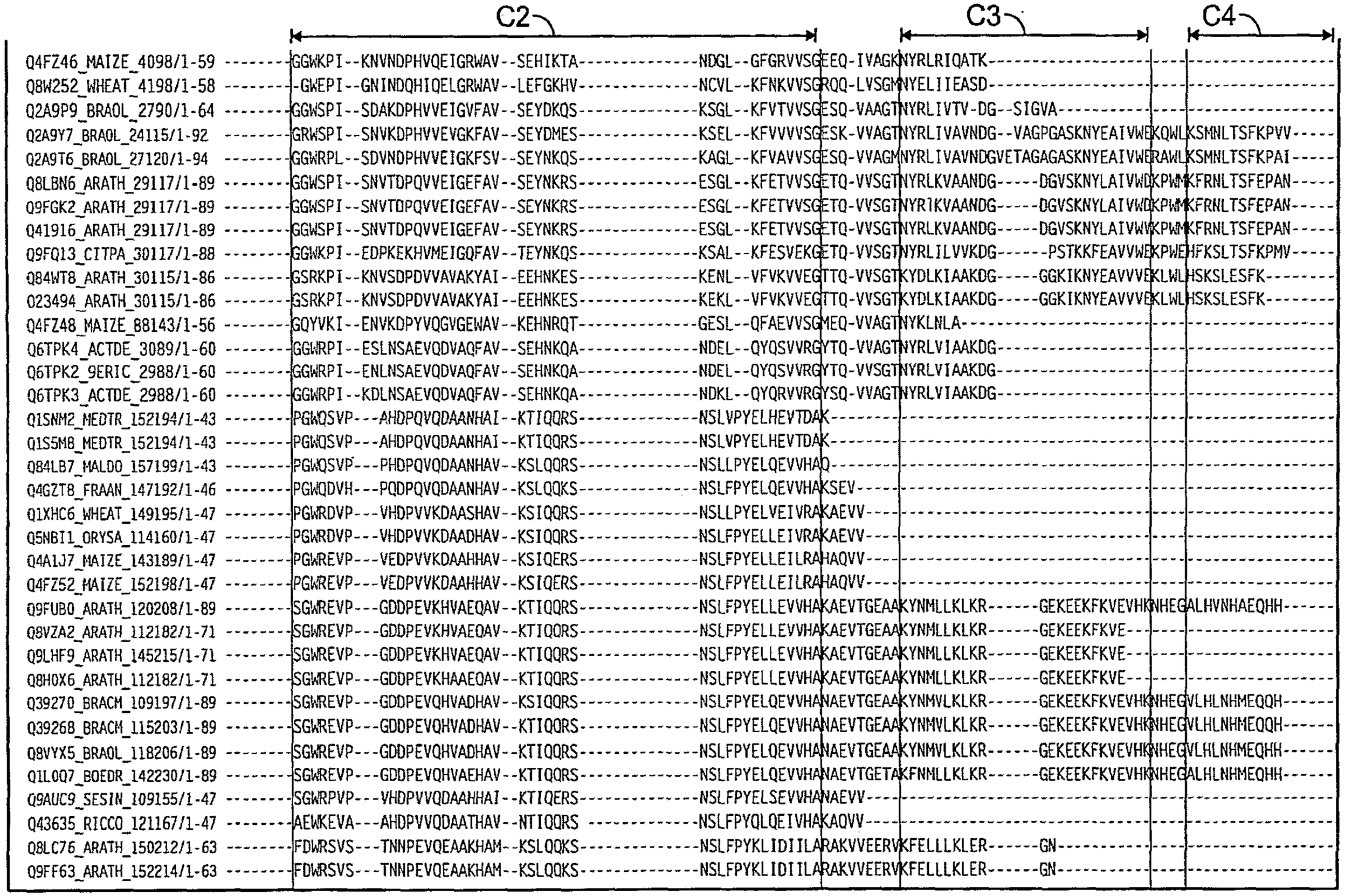


FIG. 2 (Sheet 6 of 6)

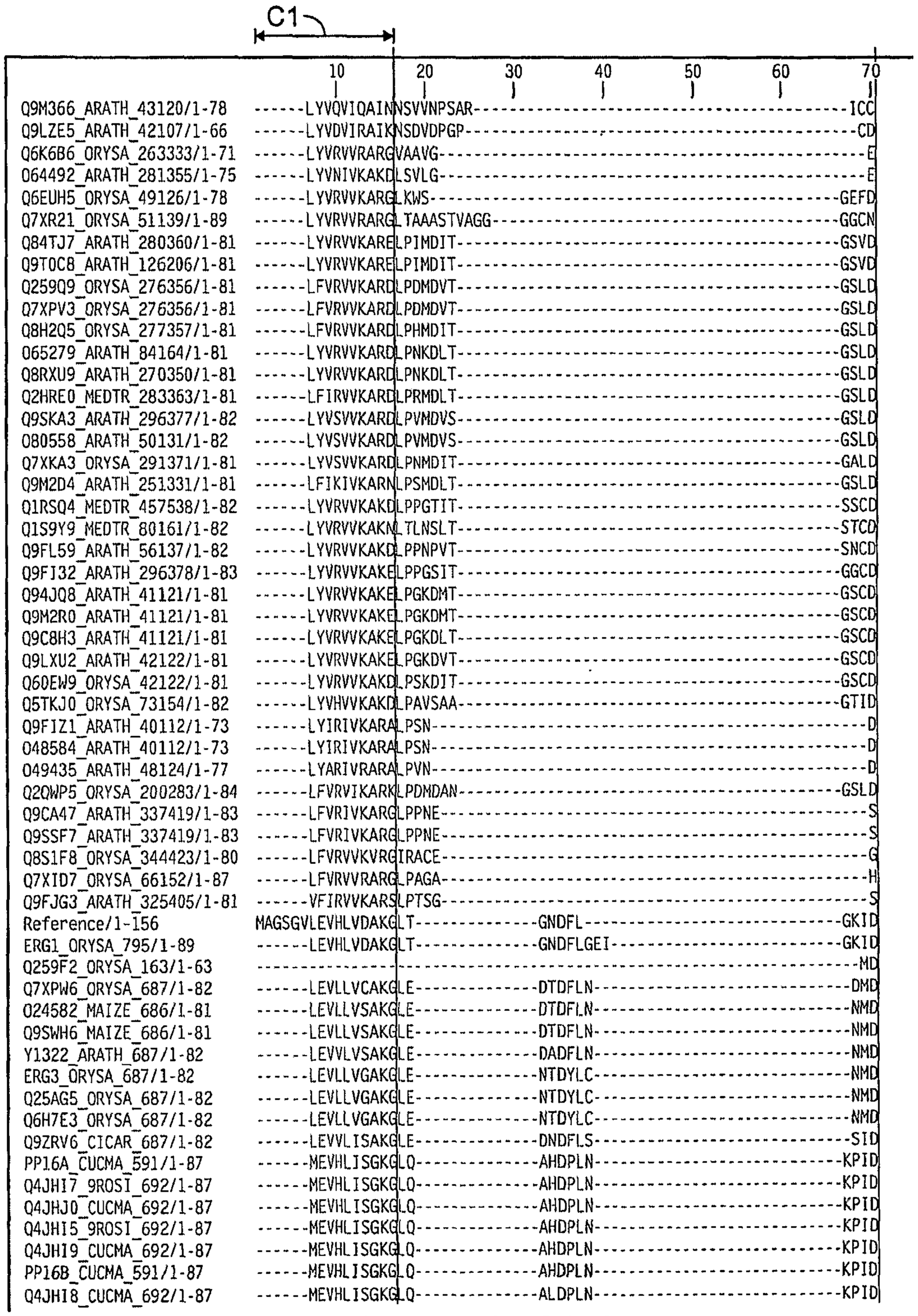


FIG. 3 (Sheet 1 of 27)

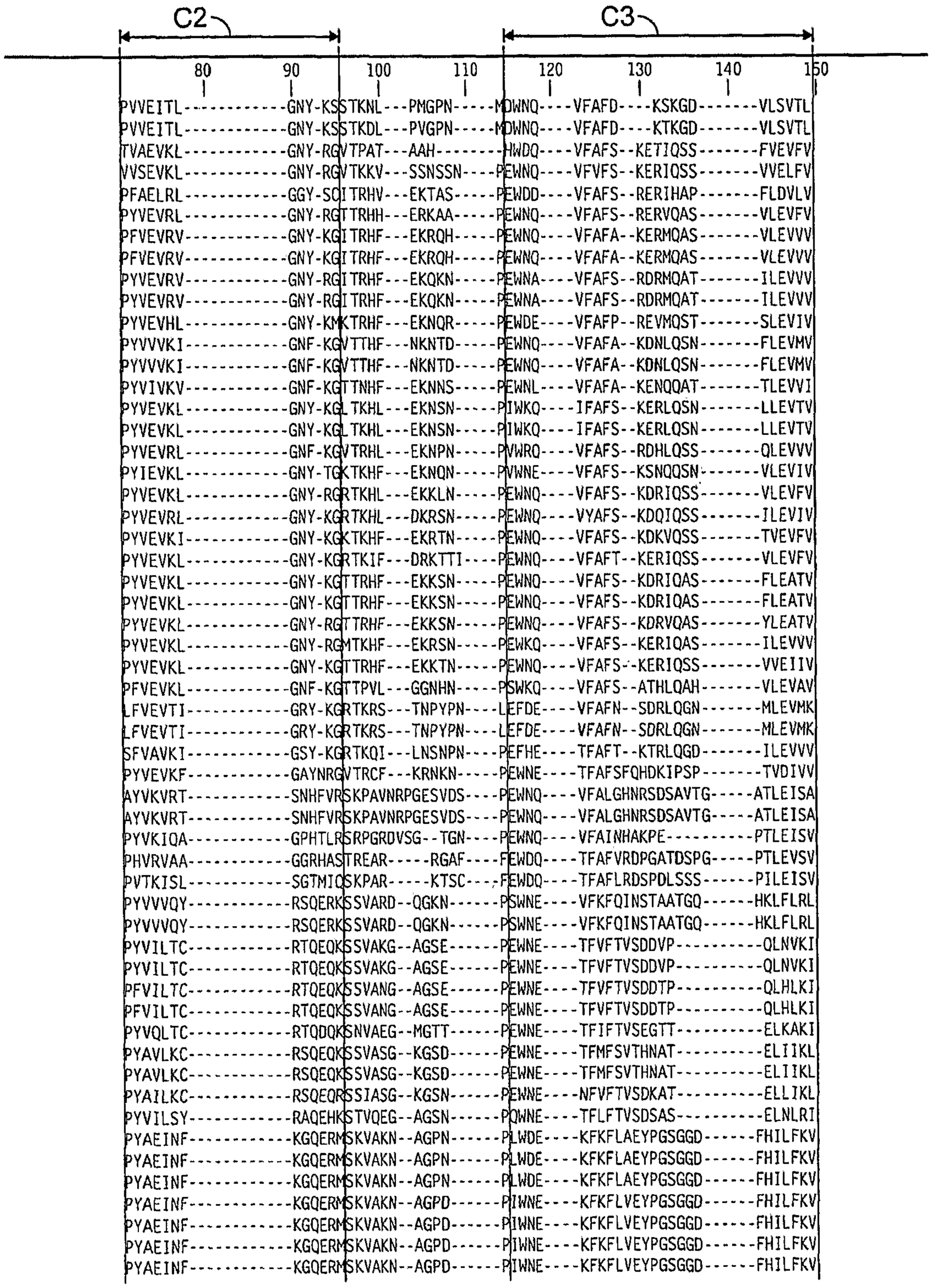


FIG. 3 (Sheet 2 of 27)

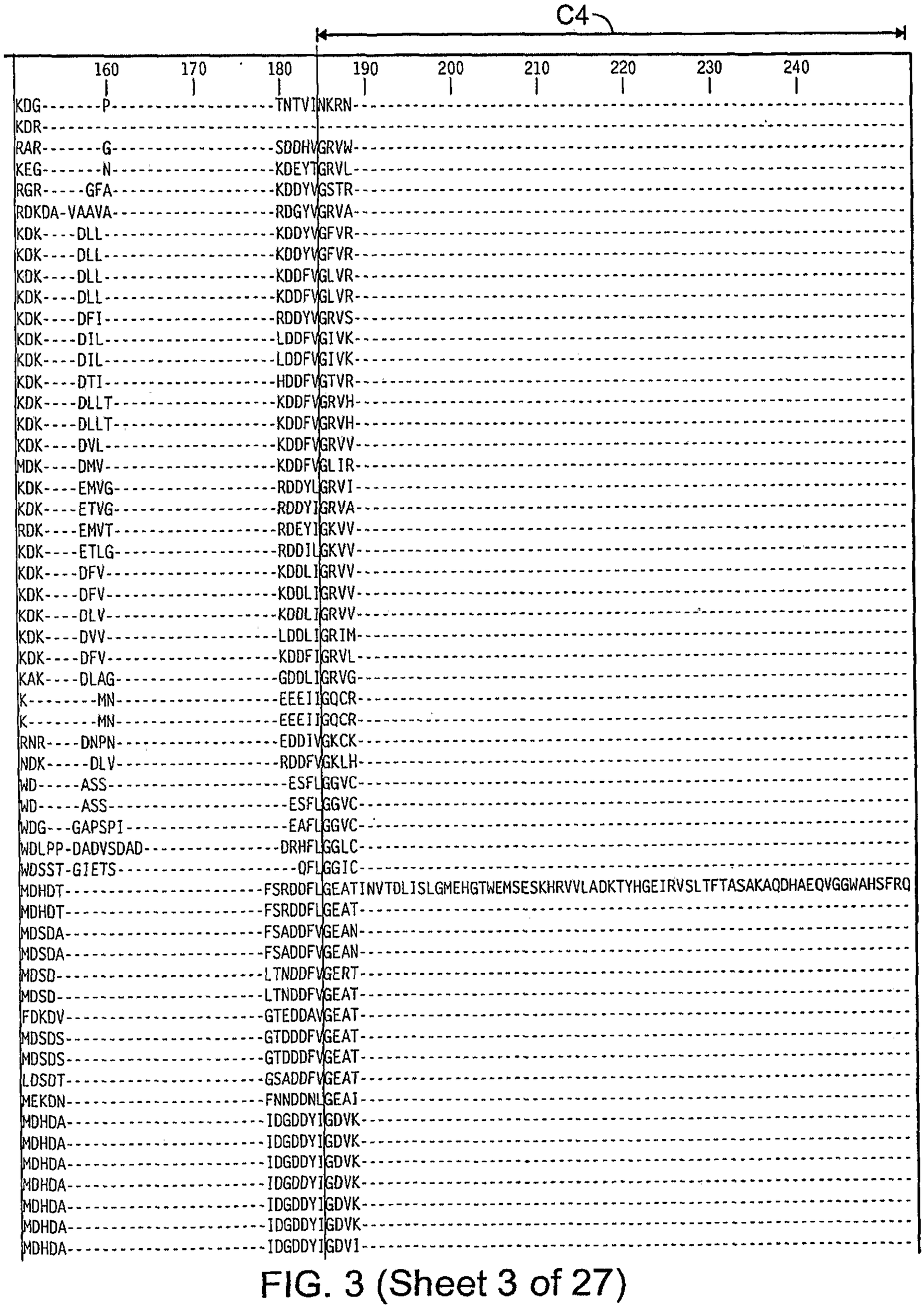


FIG. 3 (Sheet 3 of 27)

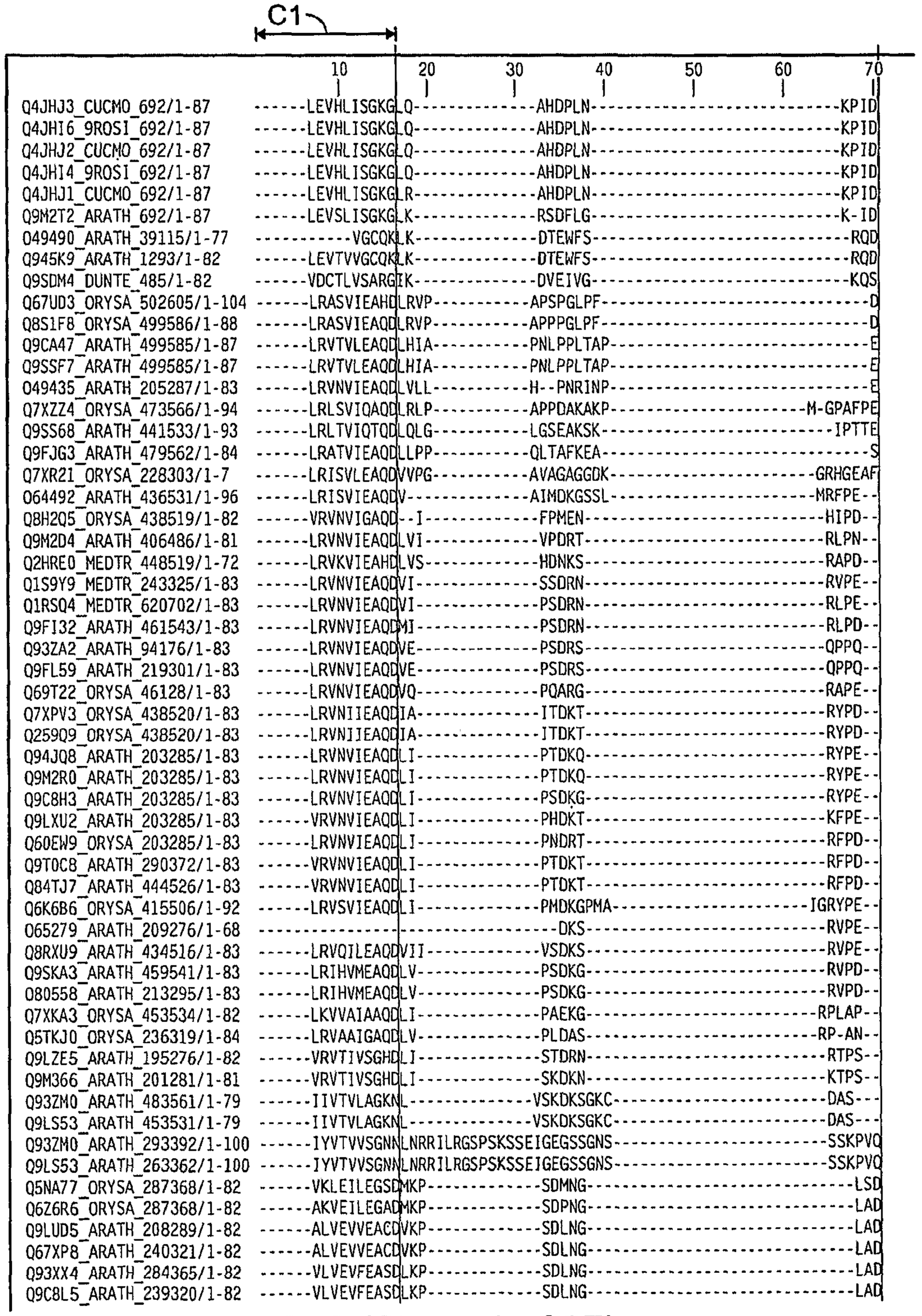


FIG. 3 (Sheet 4 of 27)

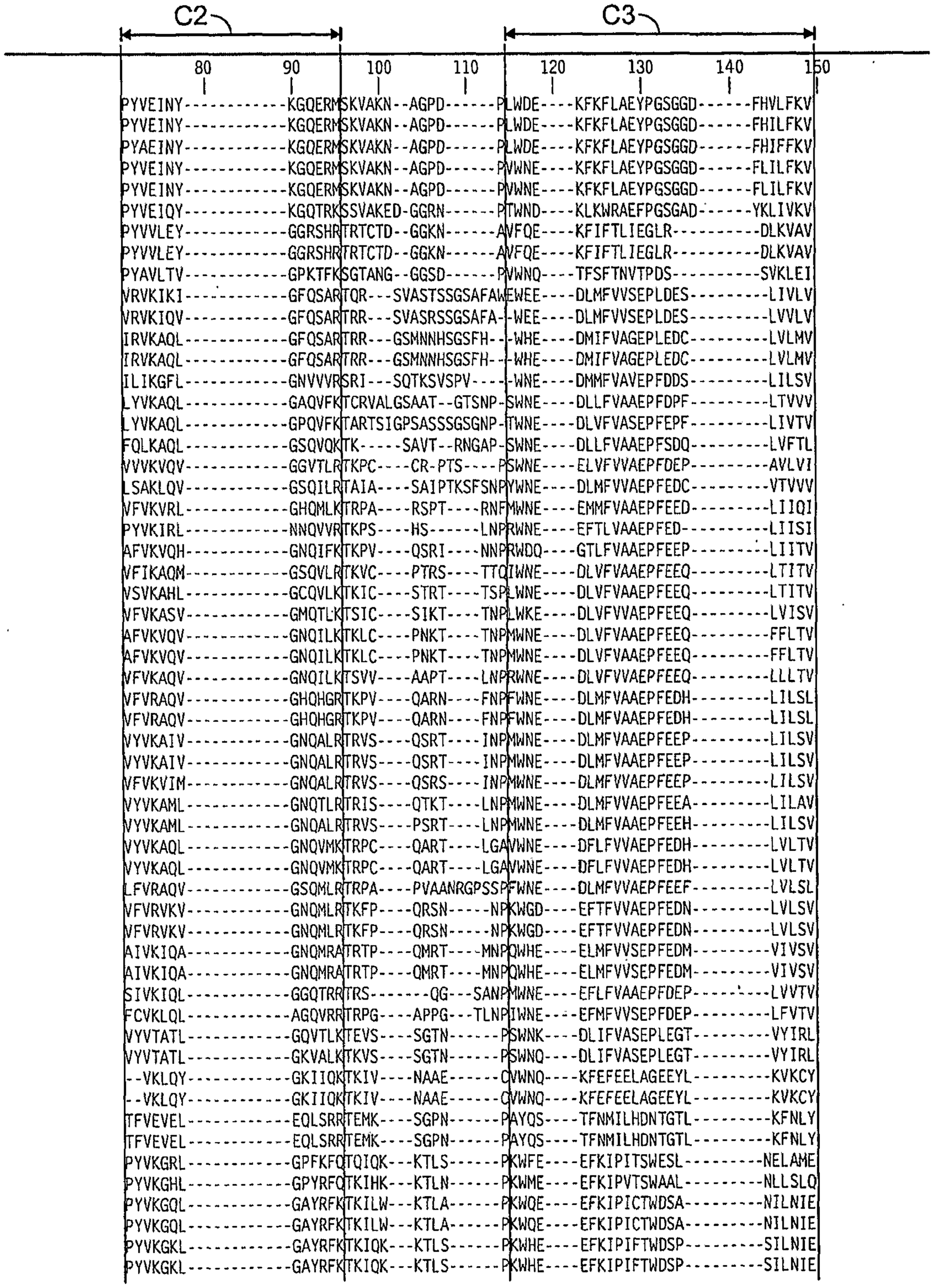


FIG. 3 (Sheet 5 of 27)

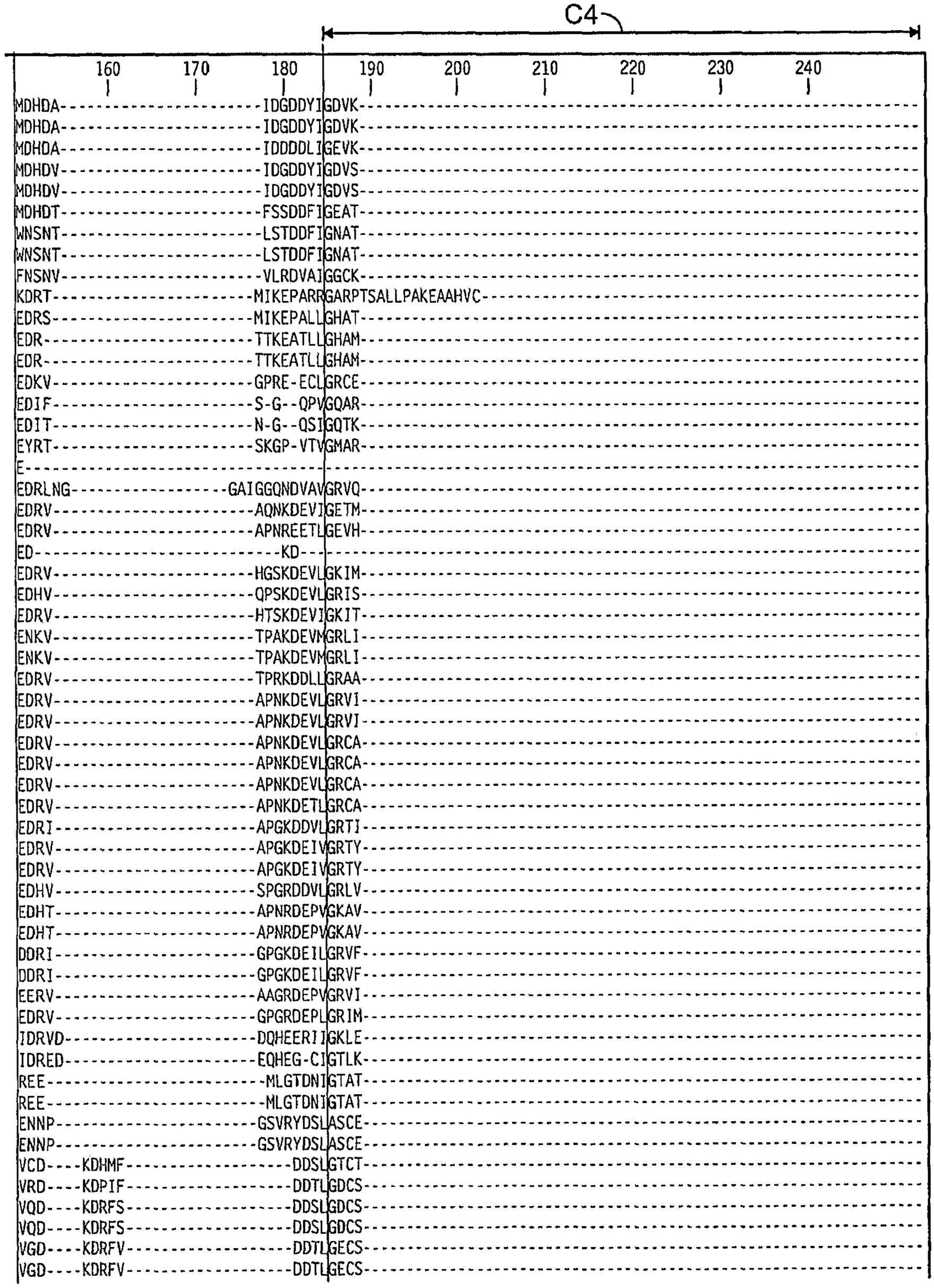


FIG. 3 (Sheet 6 of 27)

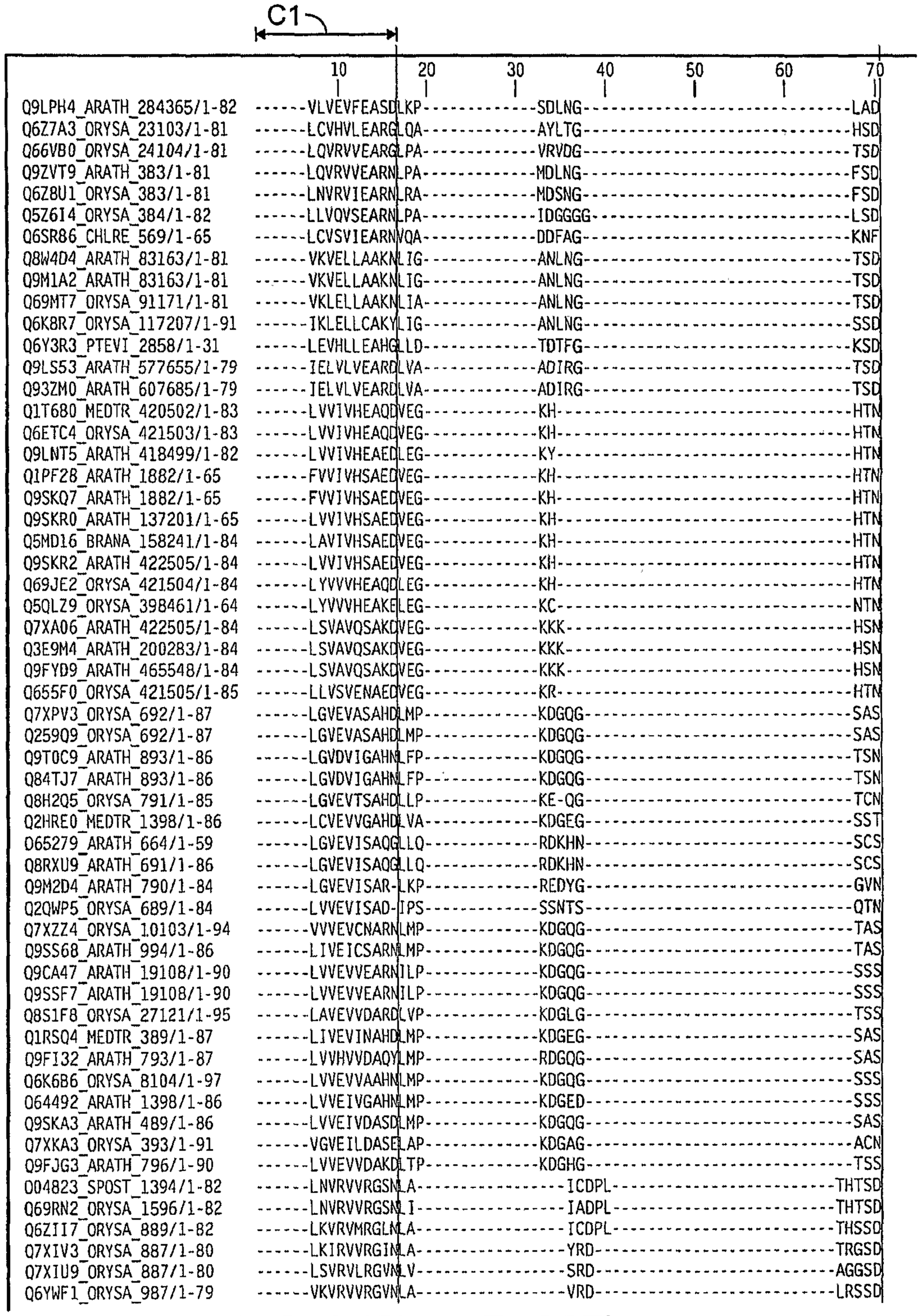


FIG. 3 (Sheet 7 of 27)

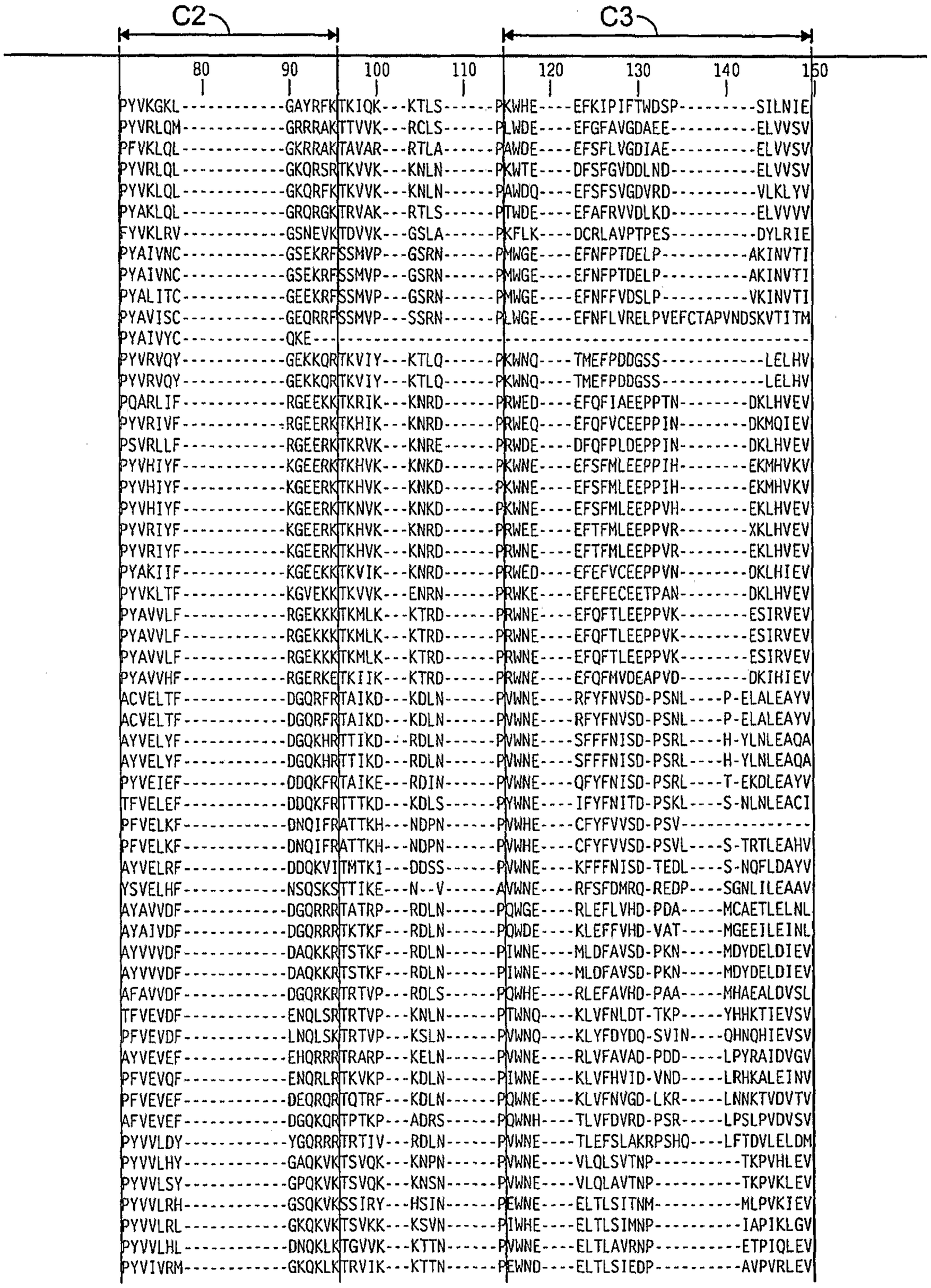


FIG. 3 (Sheet 8 of 27)

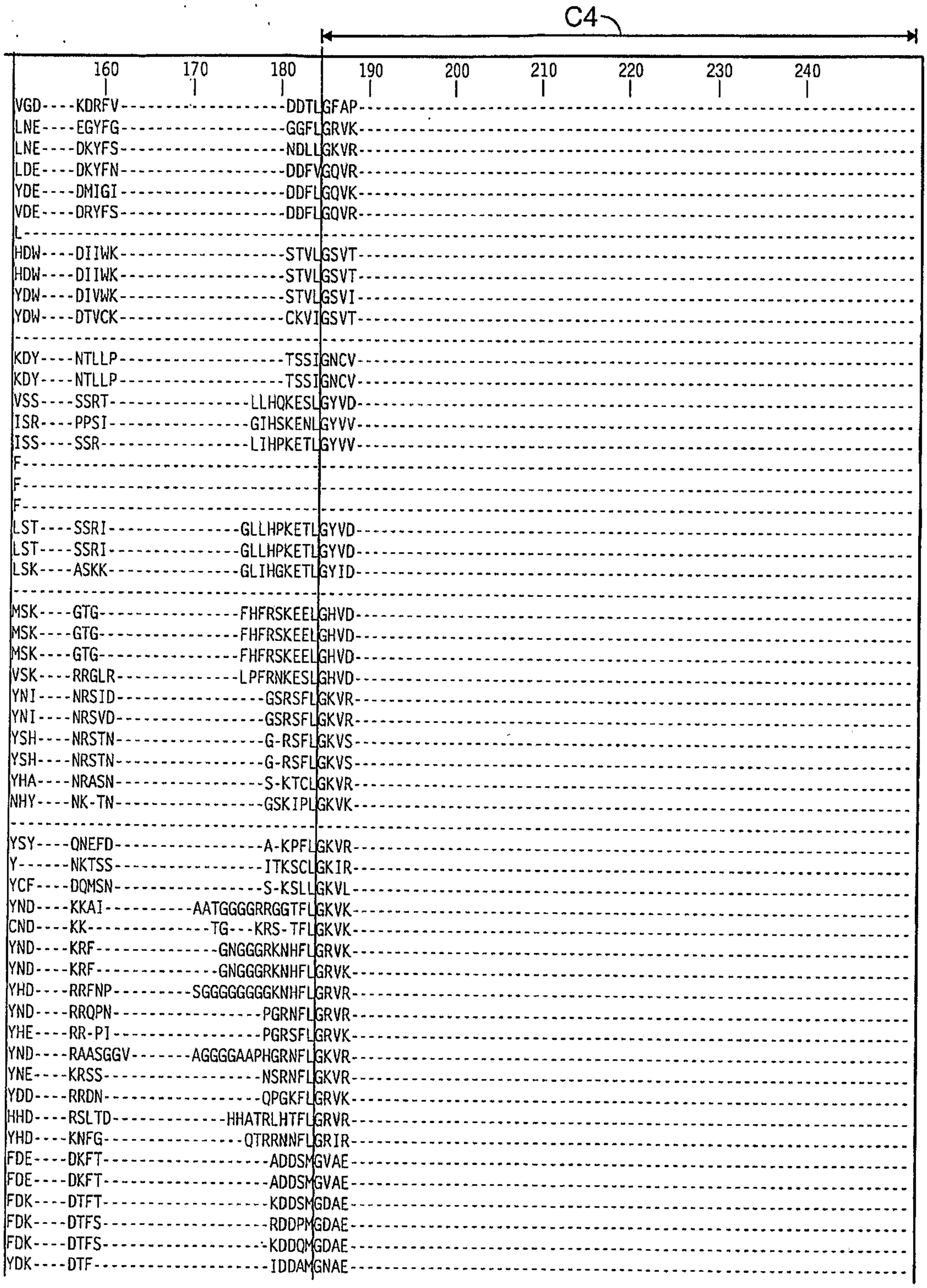


FIG. 3 (Sheet 9 of 27)

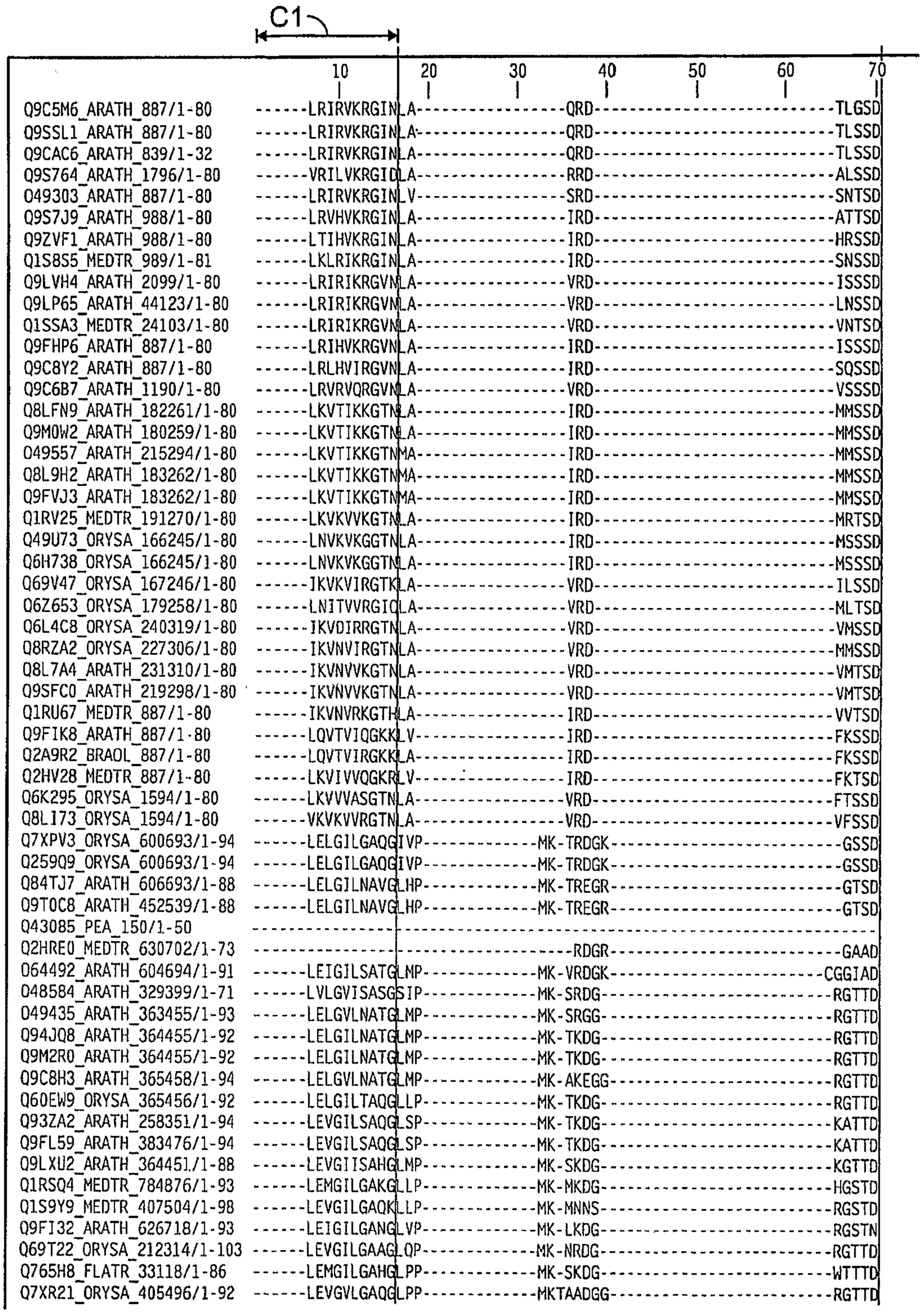


FIG. 3 (Sheet 10 of 27)

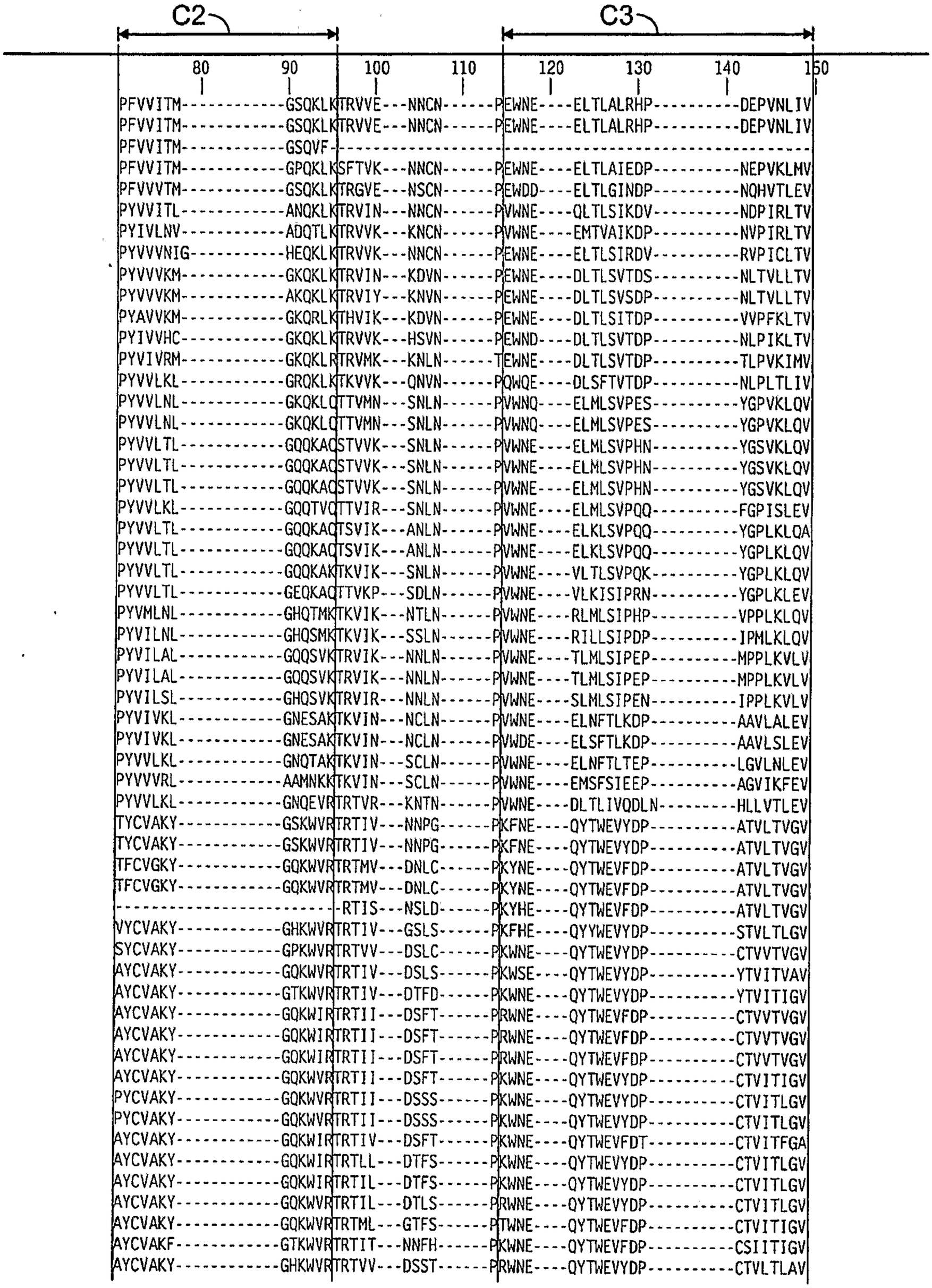


FIG. 3 (Sheet 11 of 27)

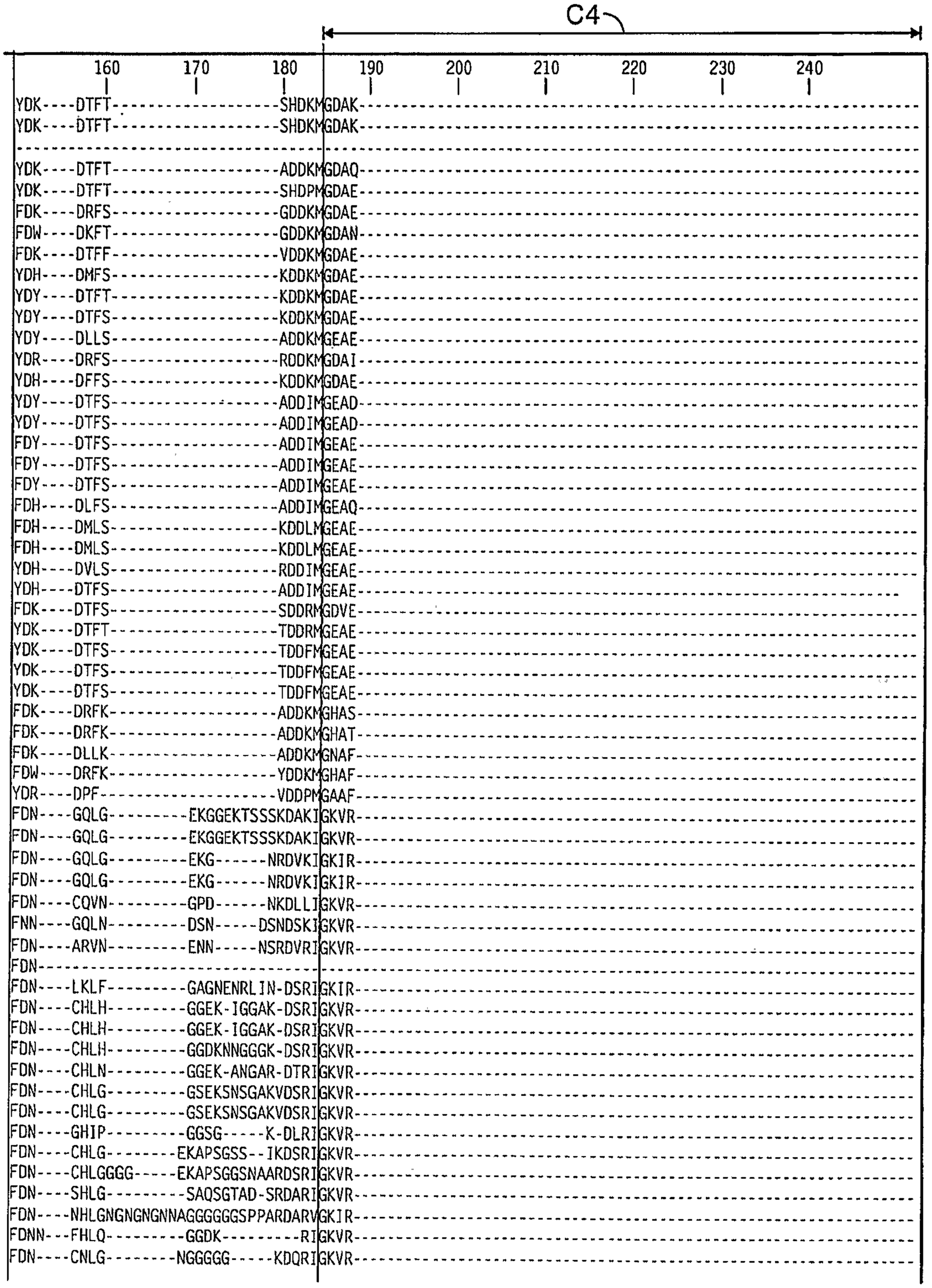


FIG. 3 (Sheet 12 of 27)

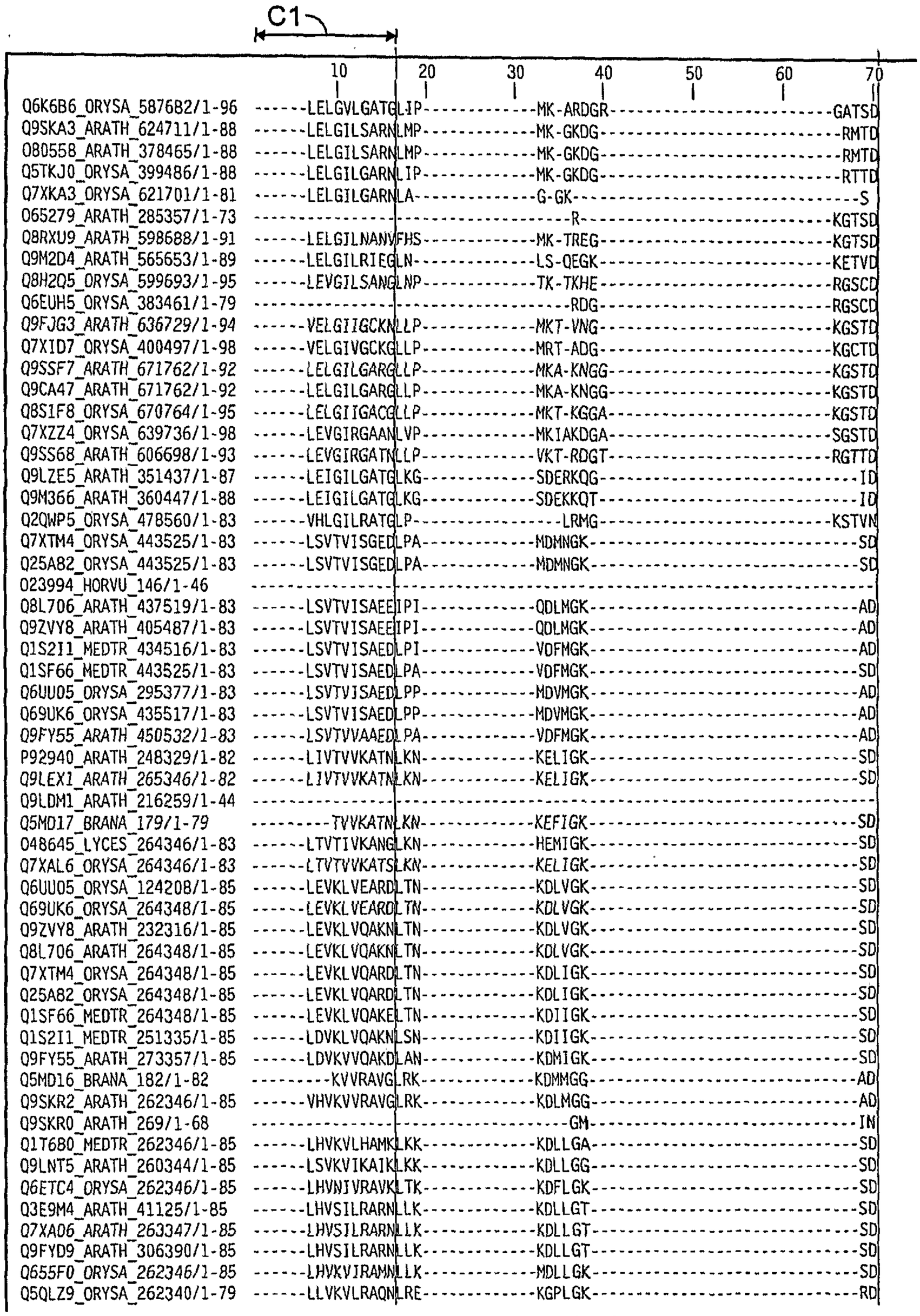


FIG. 3 (Sheet 13 of 27)

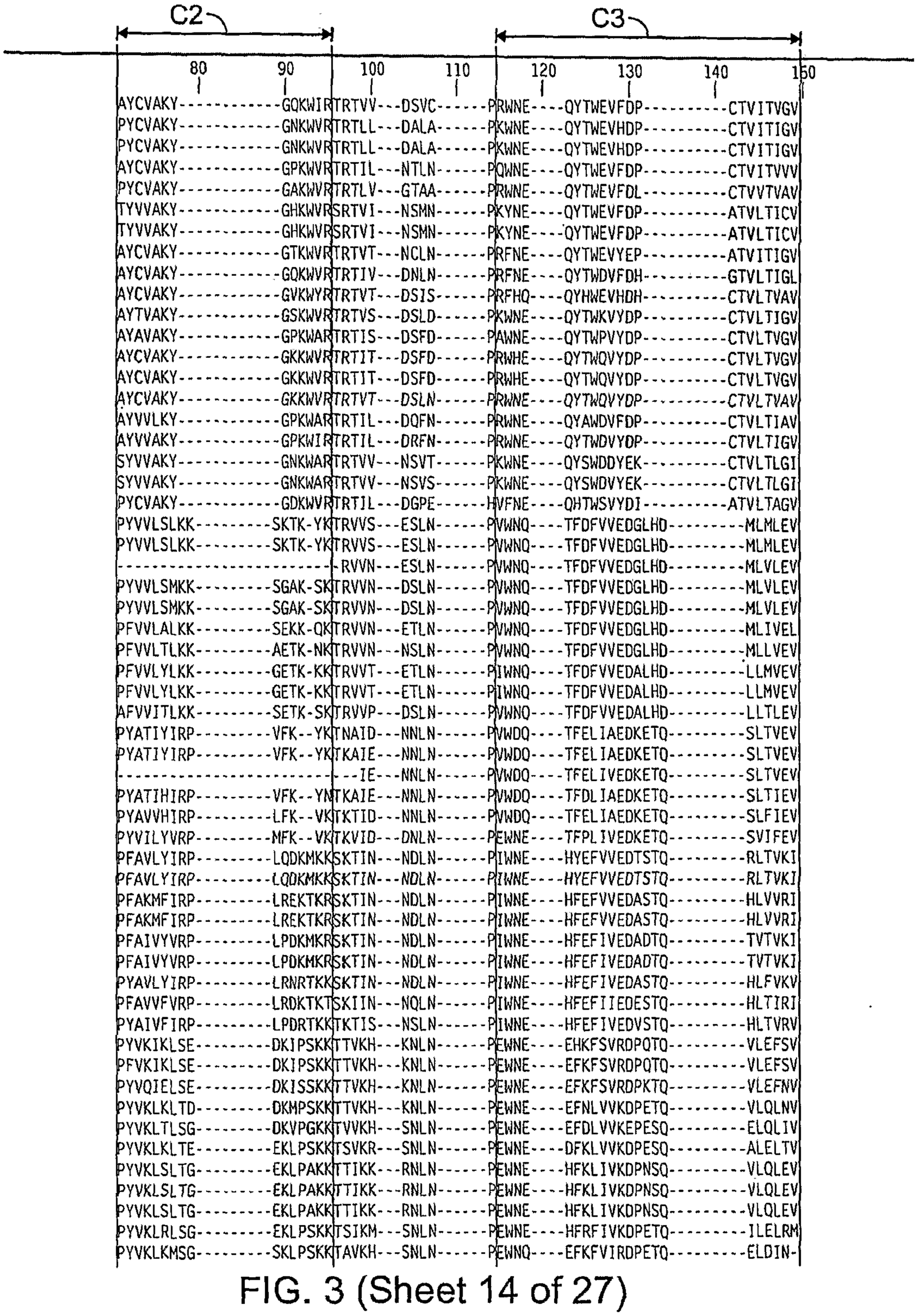


FIG. 3 (Sheet 14 of 27)

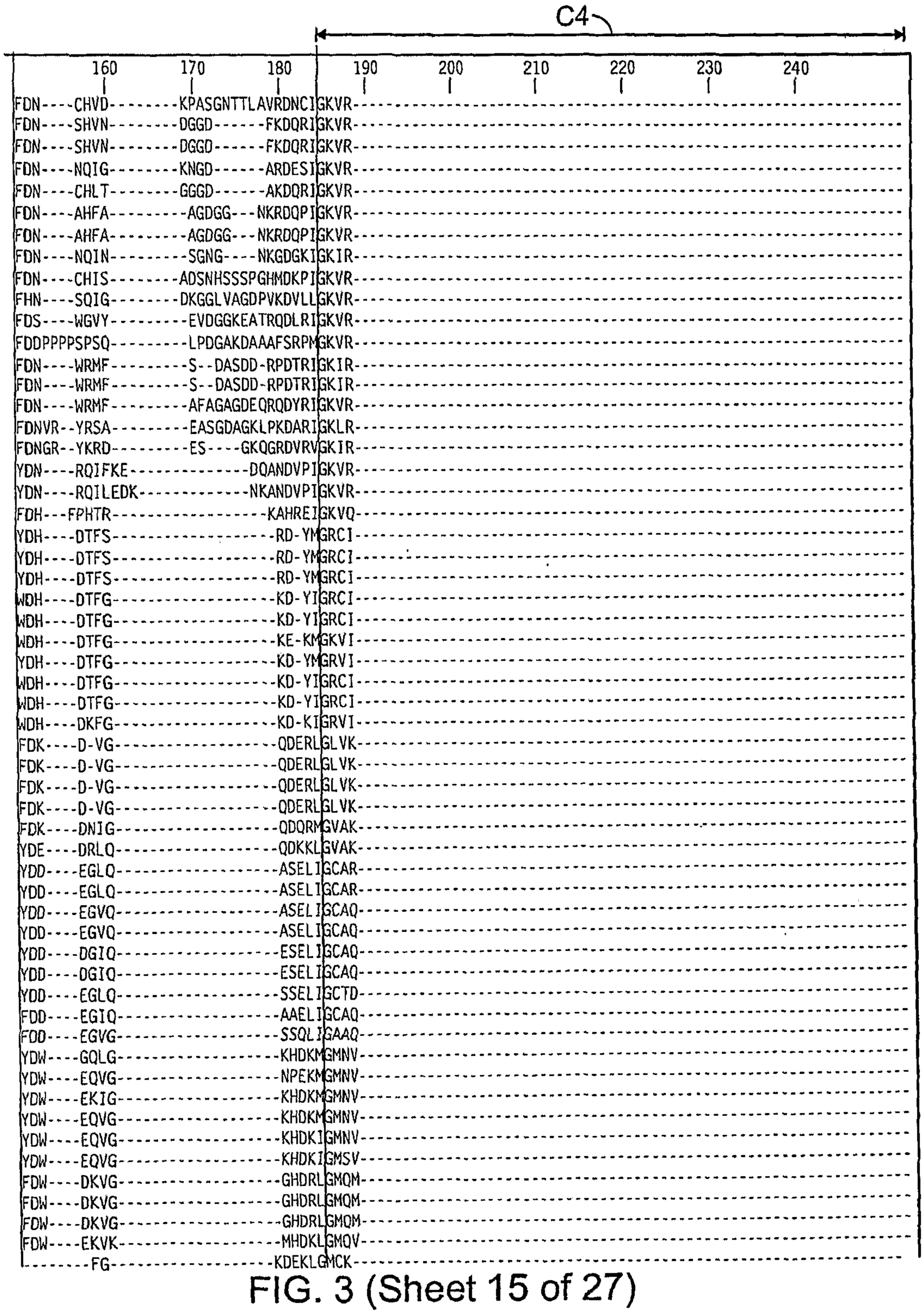


FIG. 3 (Sheet 15 of 27)

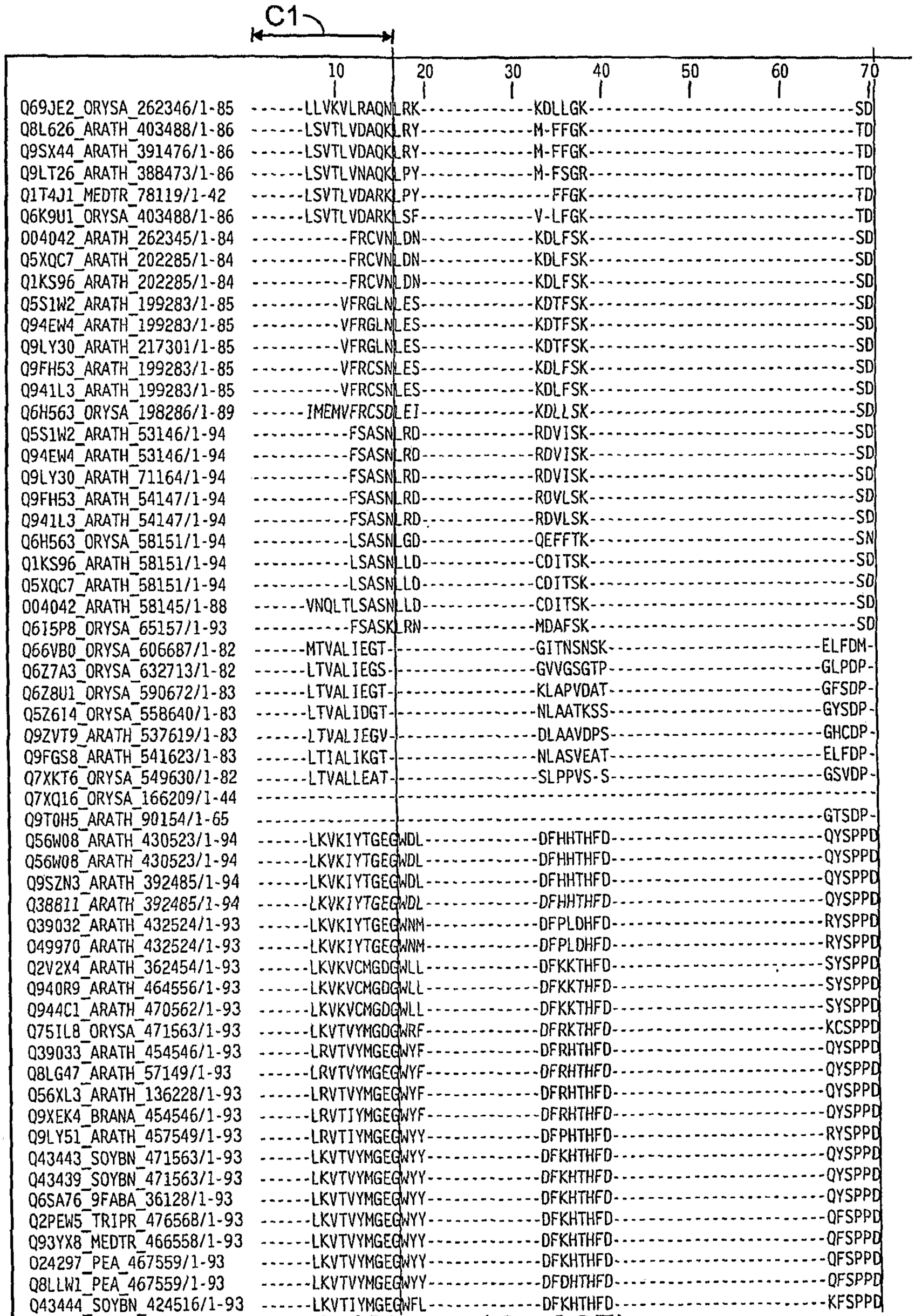


FIG. 3 (Sheet 16 of 27)

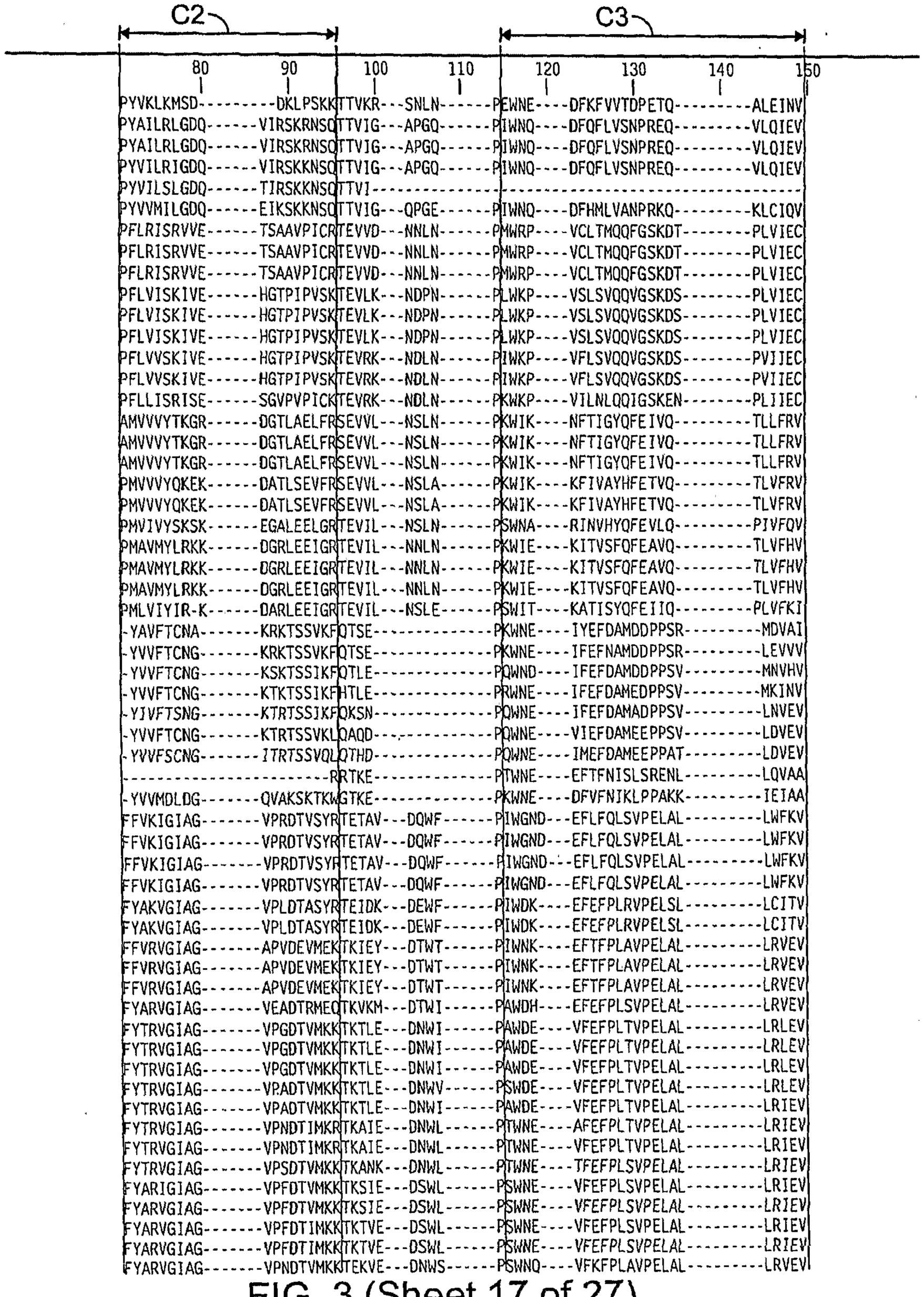


FIG. 3 (Sheet 17 of 27)

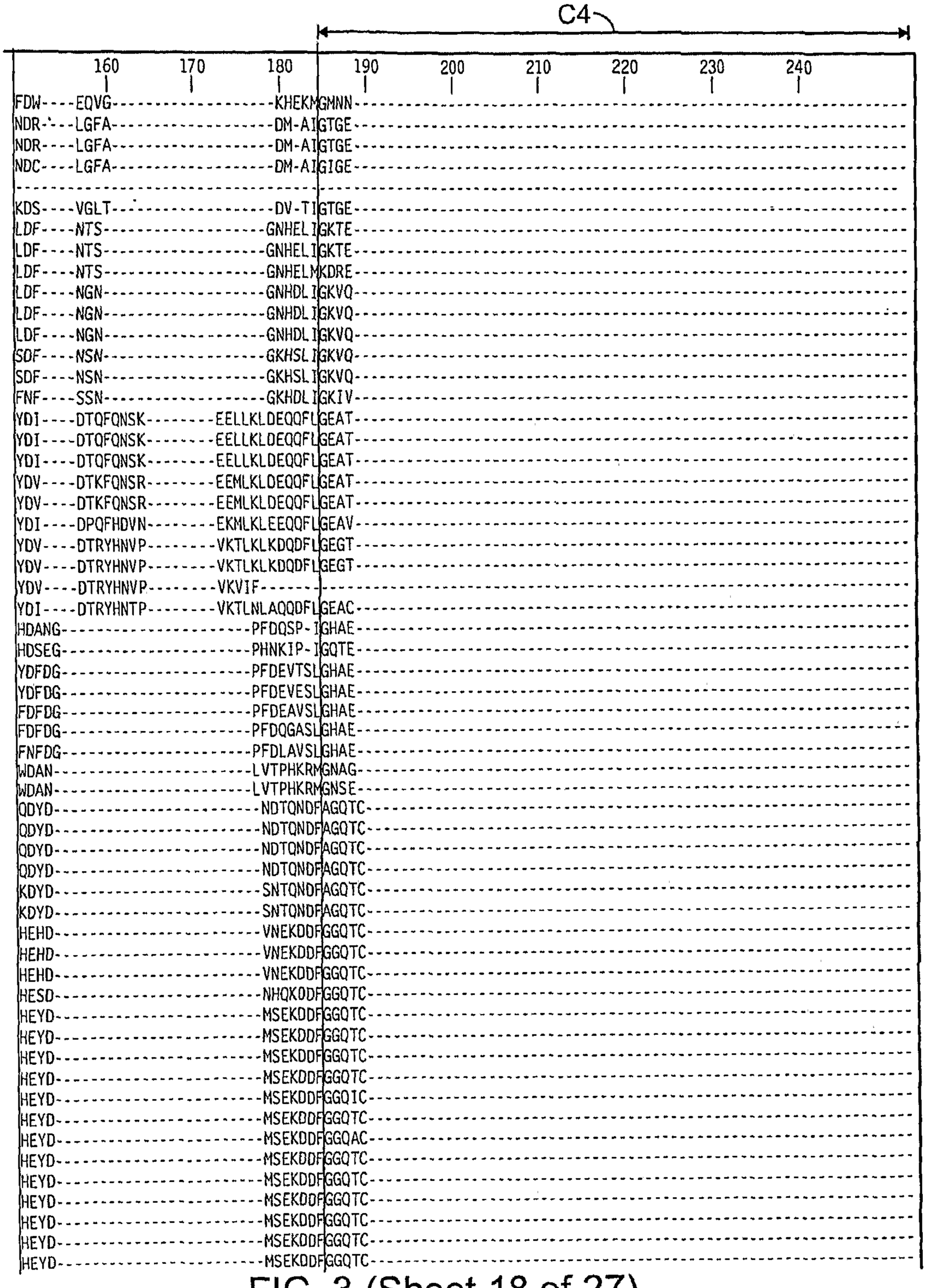


FIG. 3 (Sheet 18 of 27)

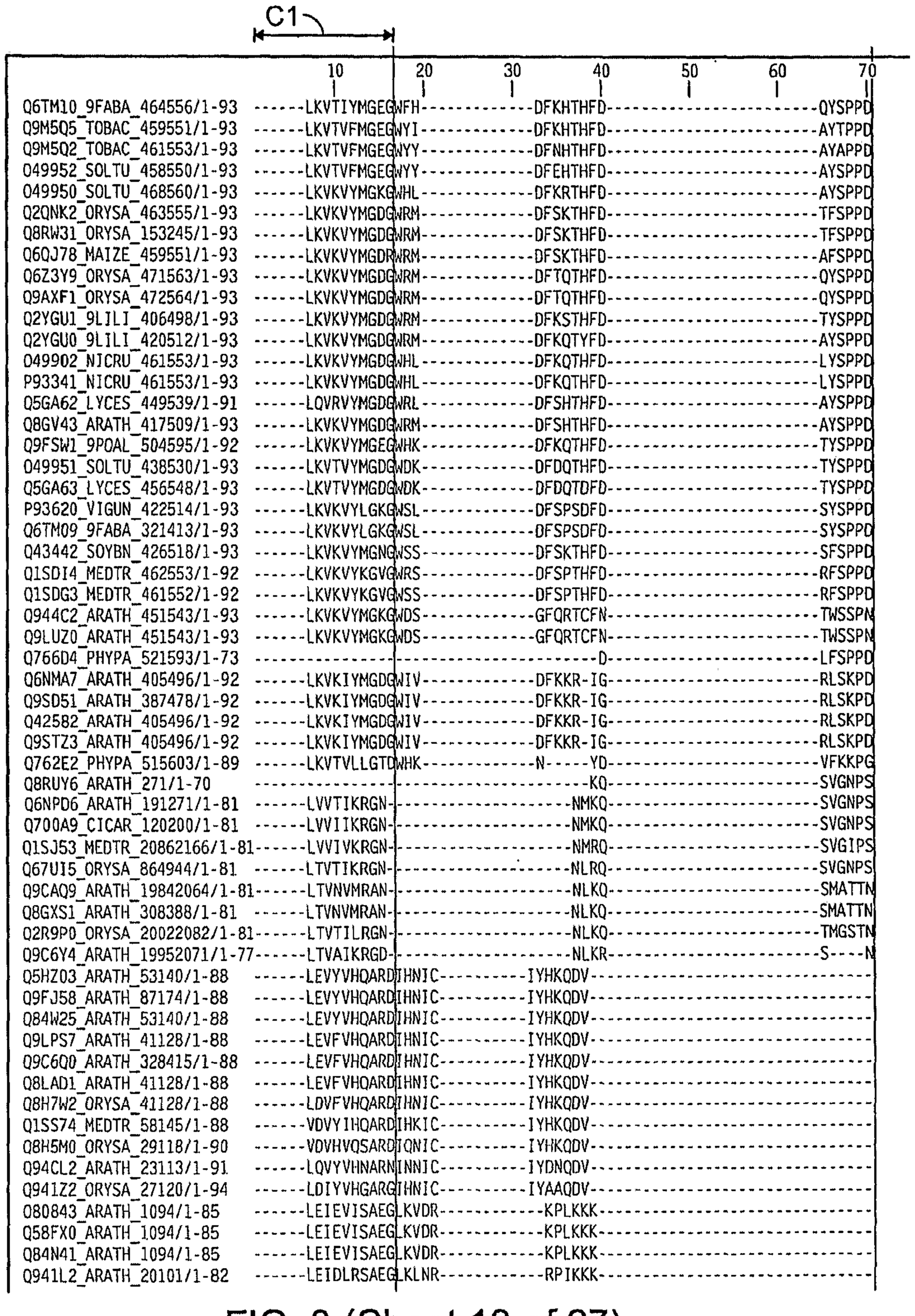


FIG. 3 (Sheet 19 of 27)

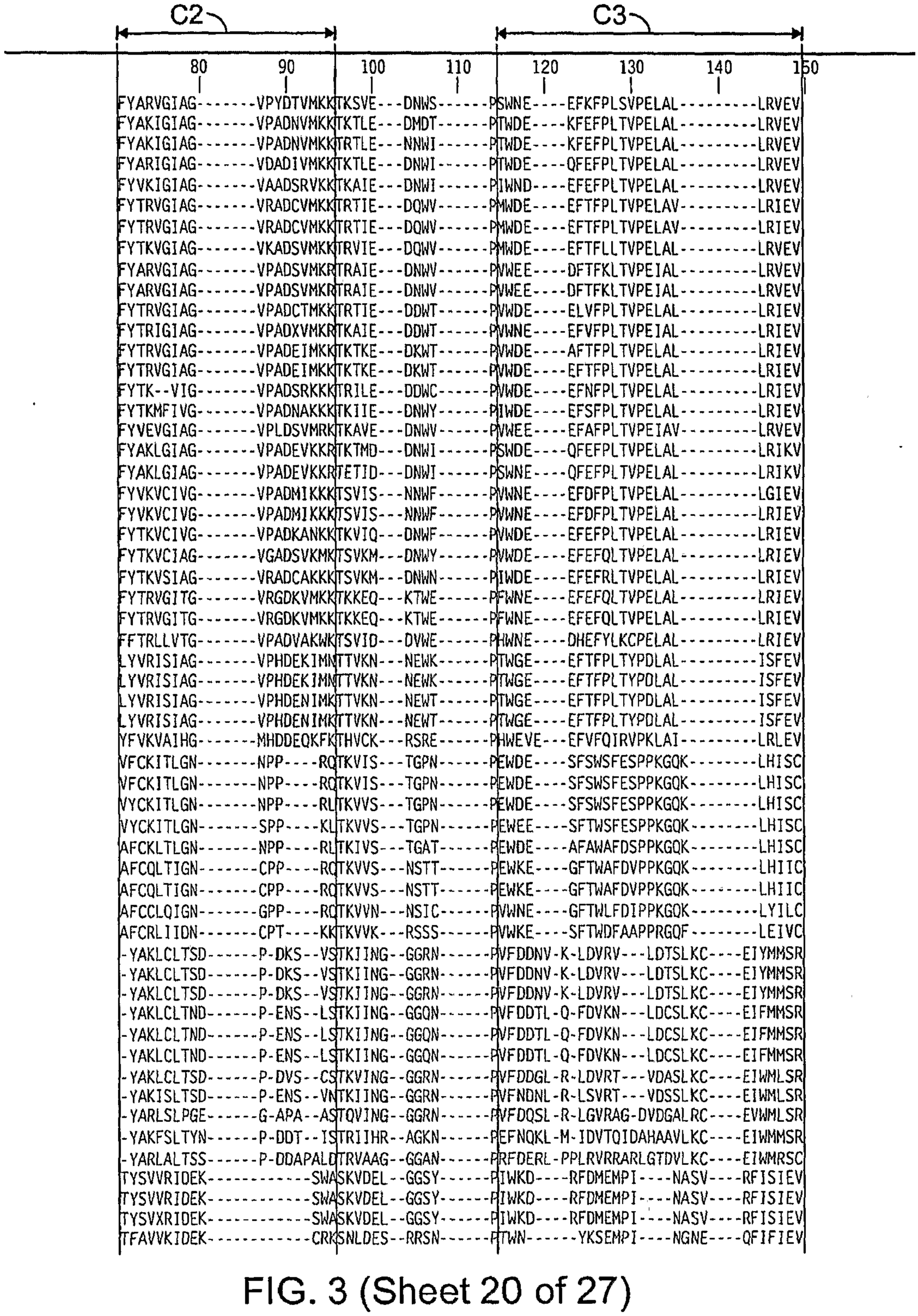


FIG. 3 (Sheet 20 of 27)

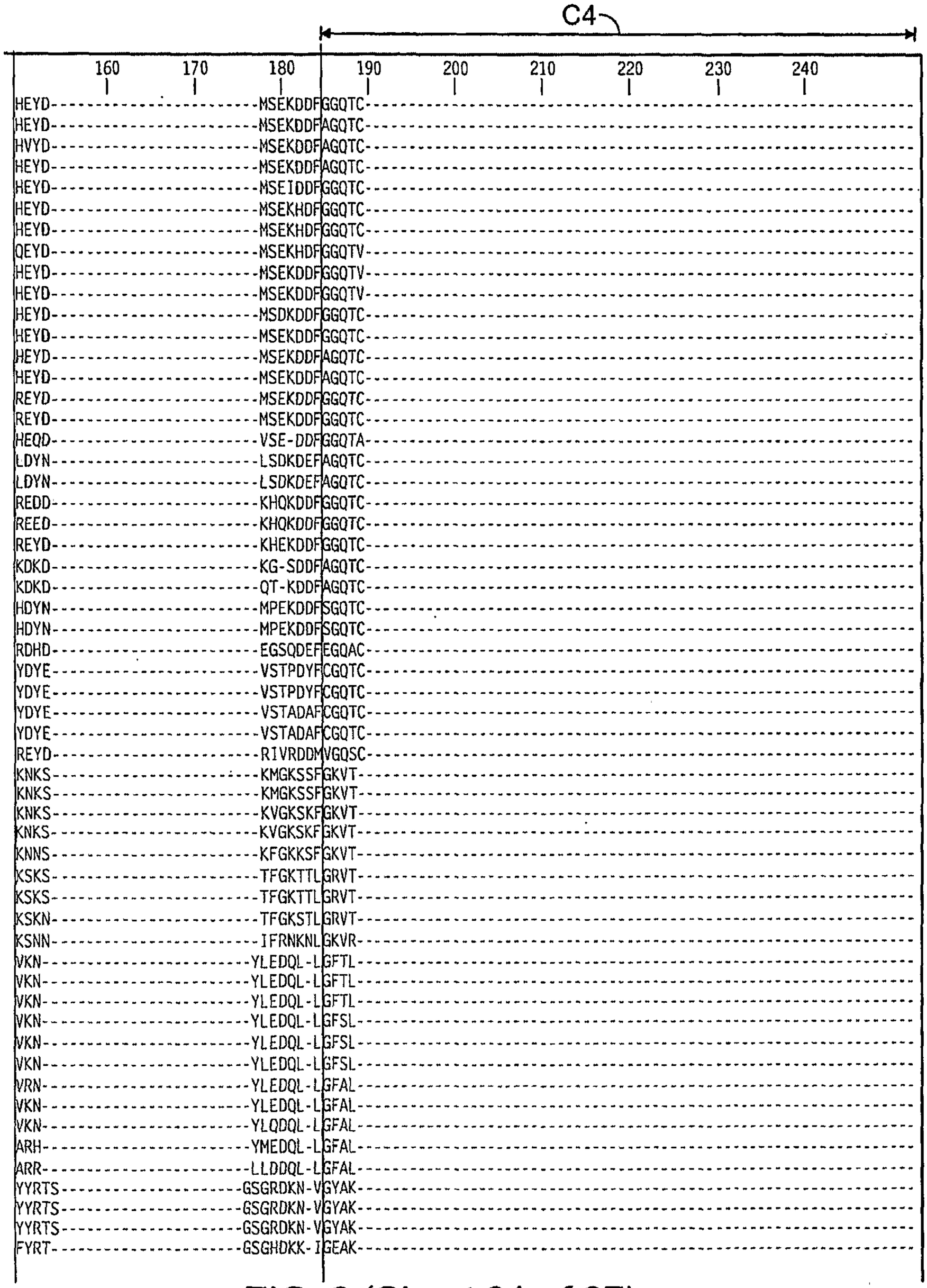


FIG. 3 (Sheet 21 of 27)

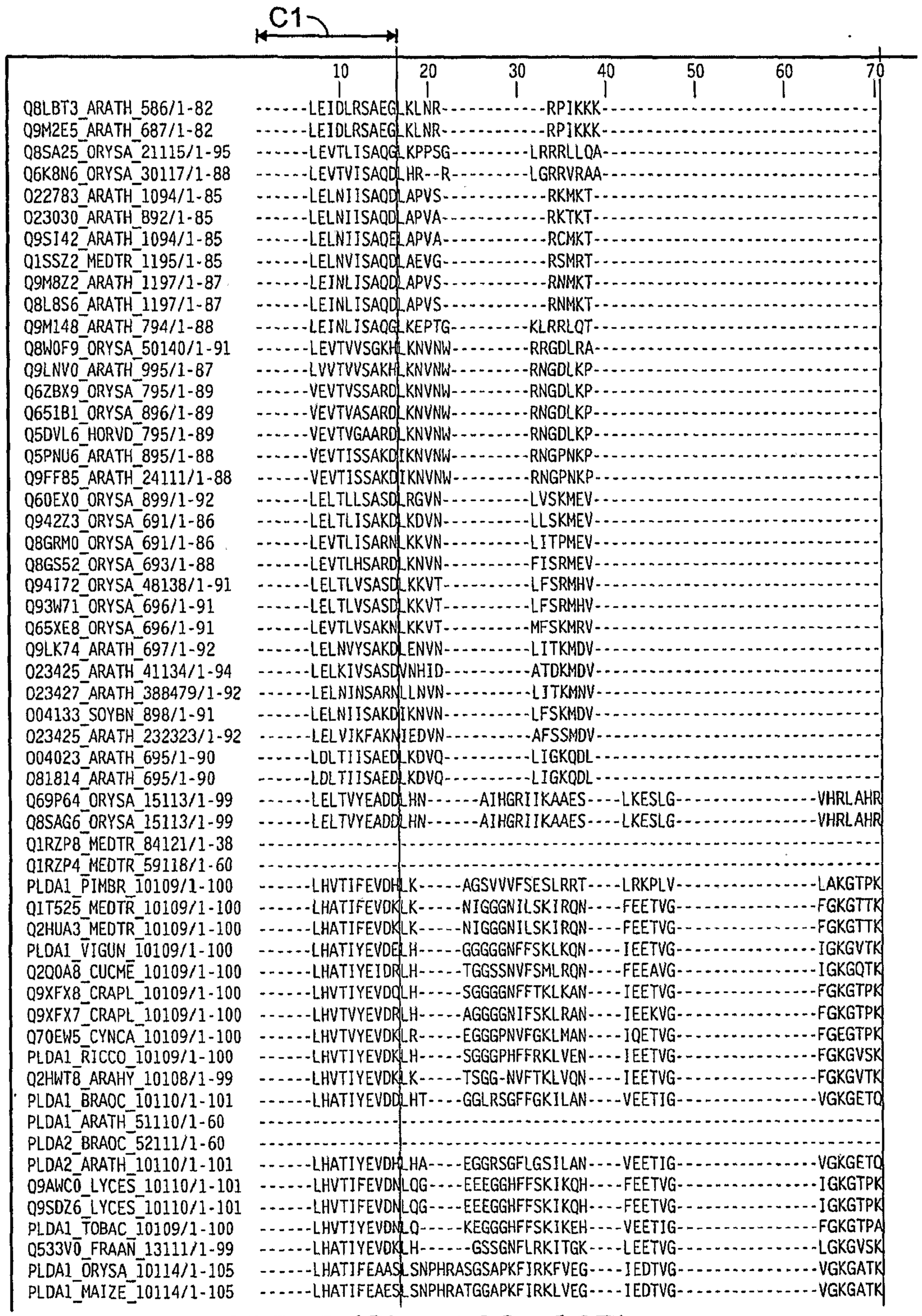


FIG. 3 (Sheet 22 of 27)

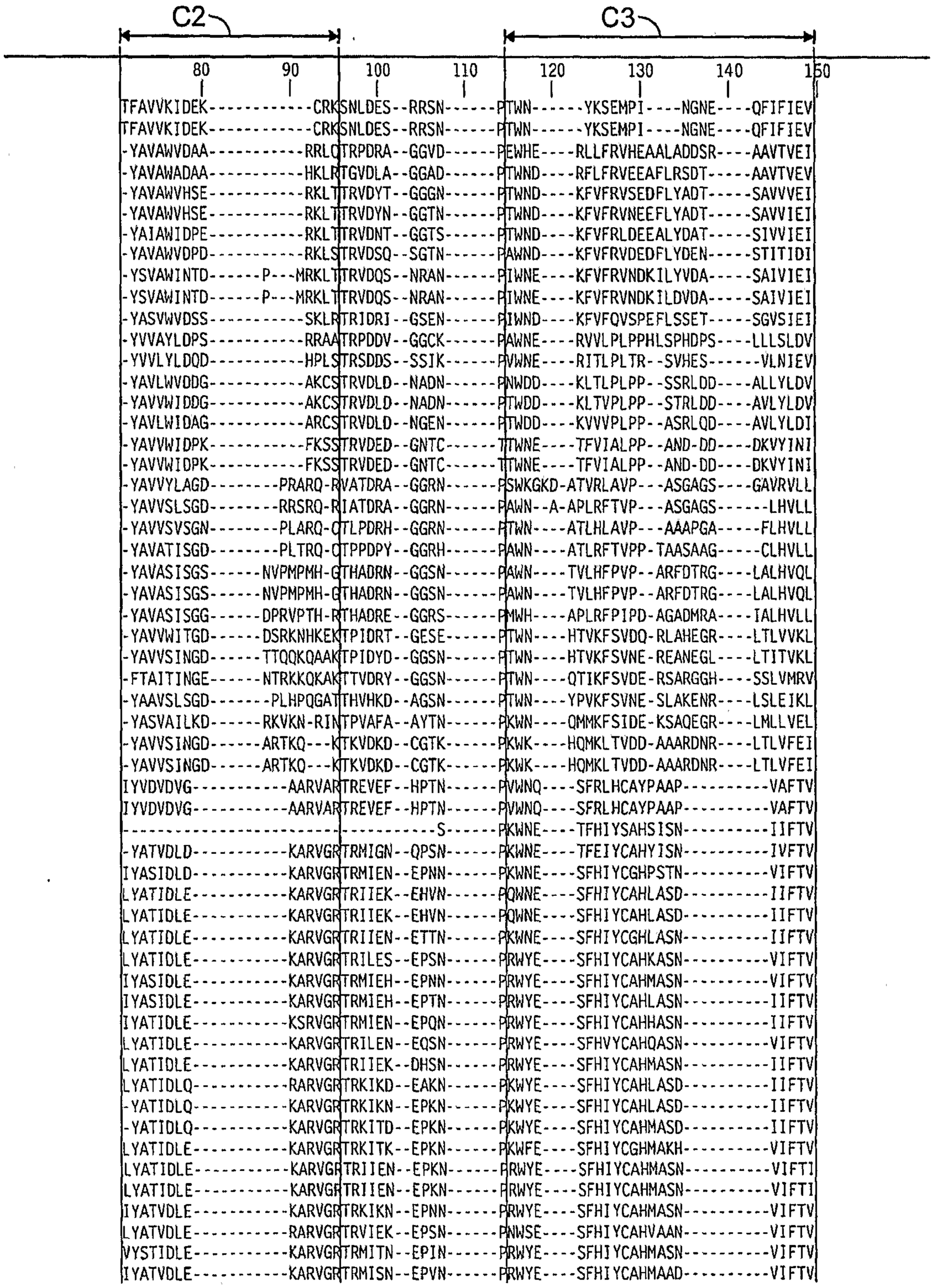


FIG. 3 (Sheet 23 of 27)

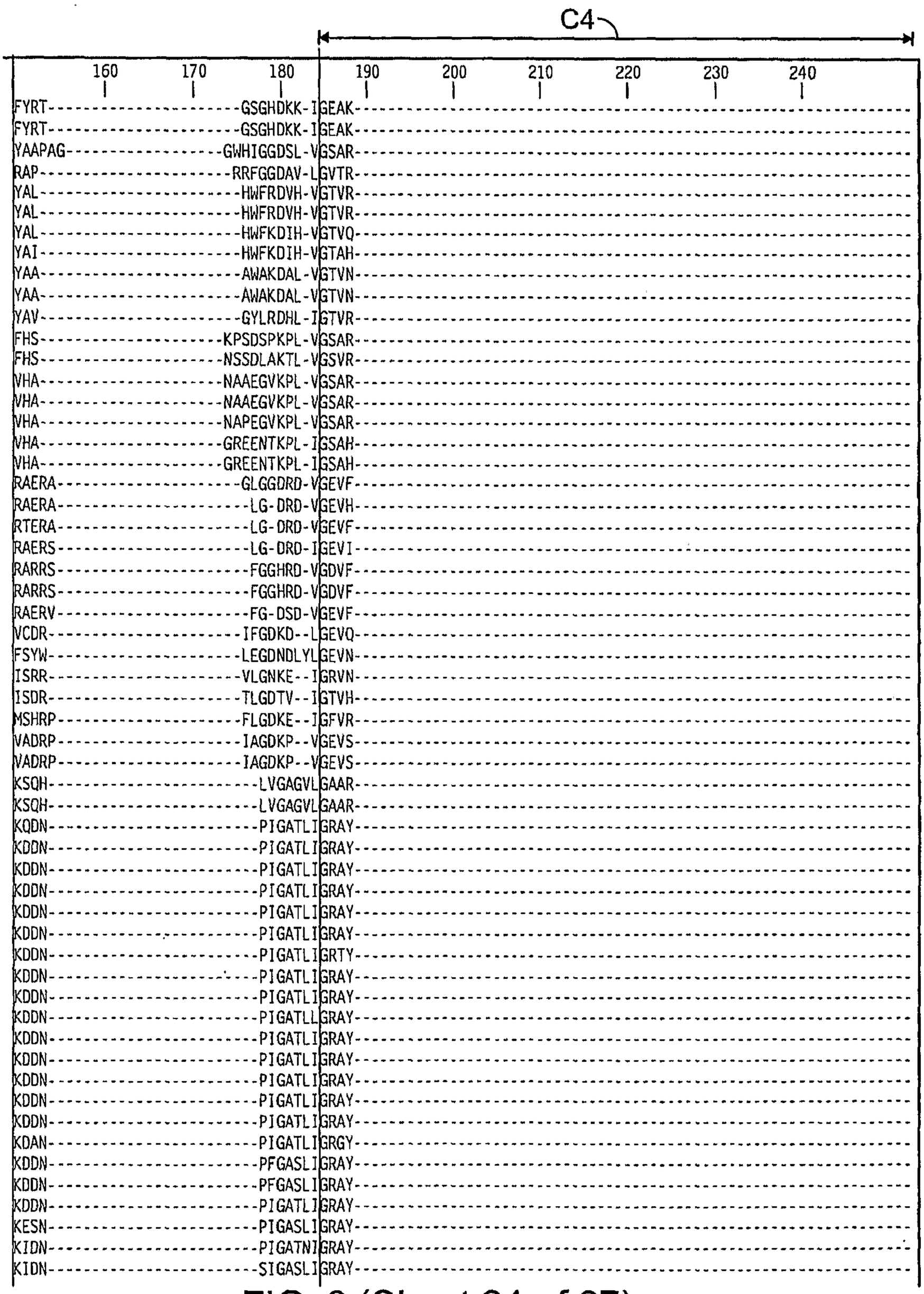


FIG. 3 (Sheet 24 of 27)

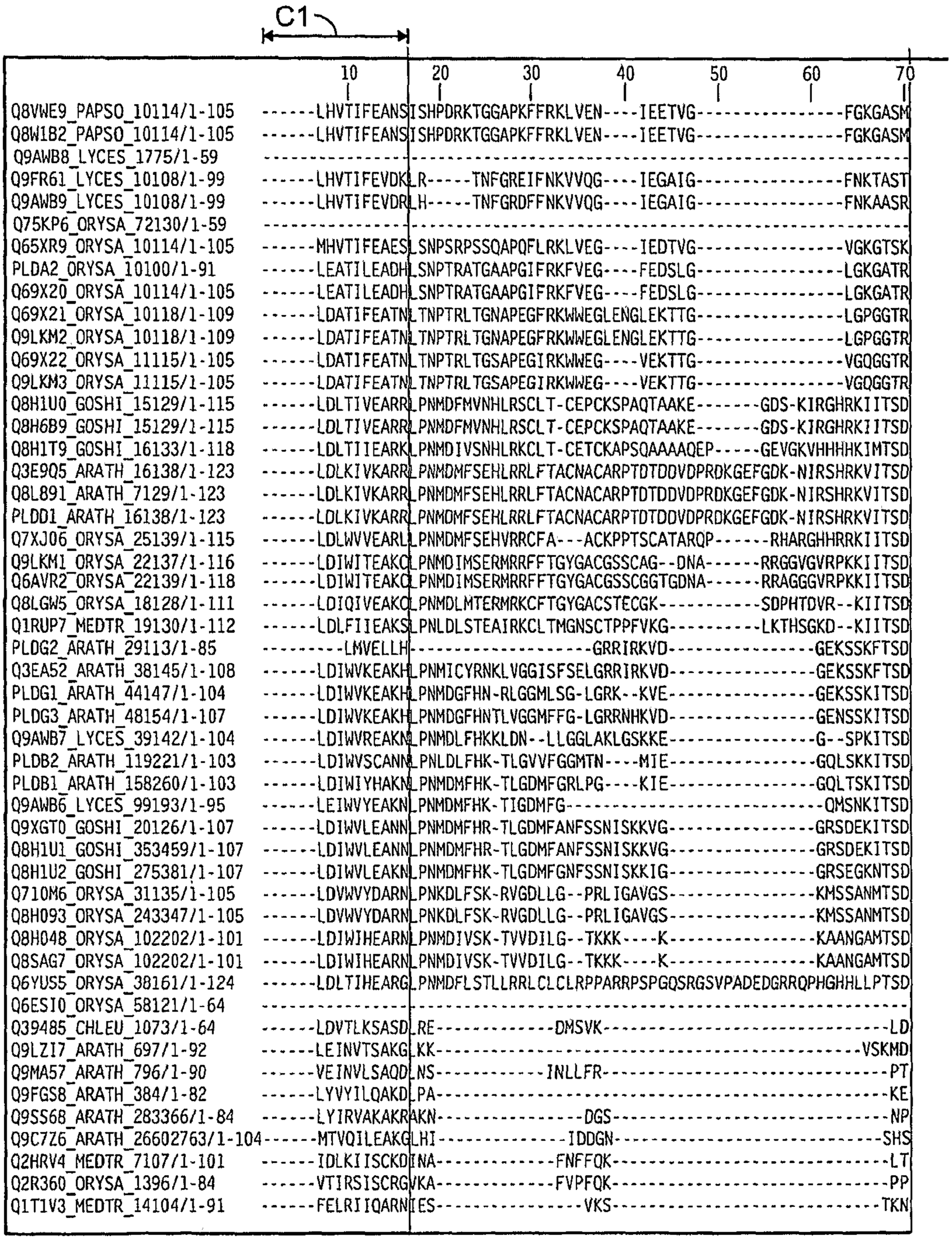


FIG. 3 (Sheet 25 of 27)

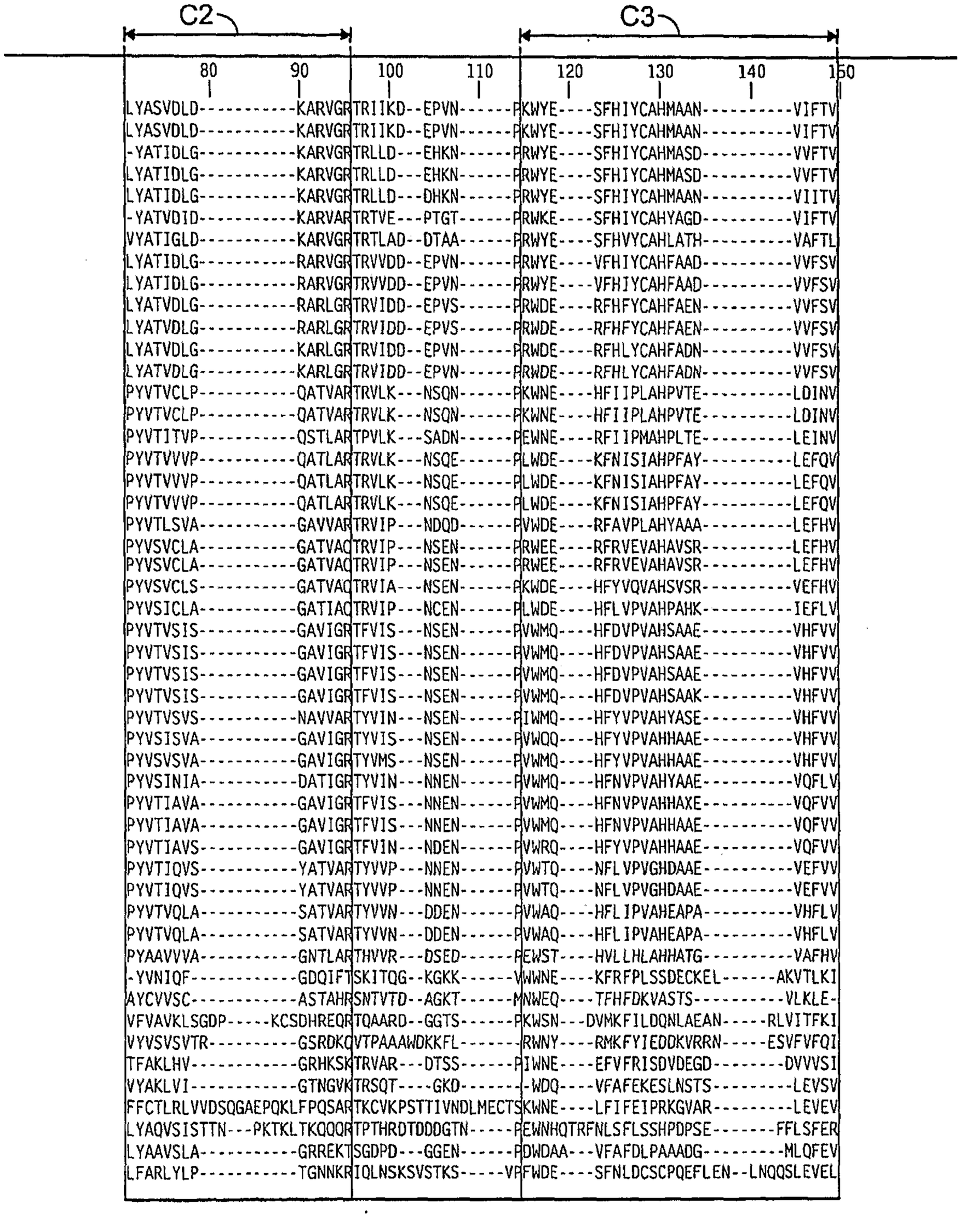


FIG. 3 (Sheet 26 of 27)

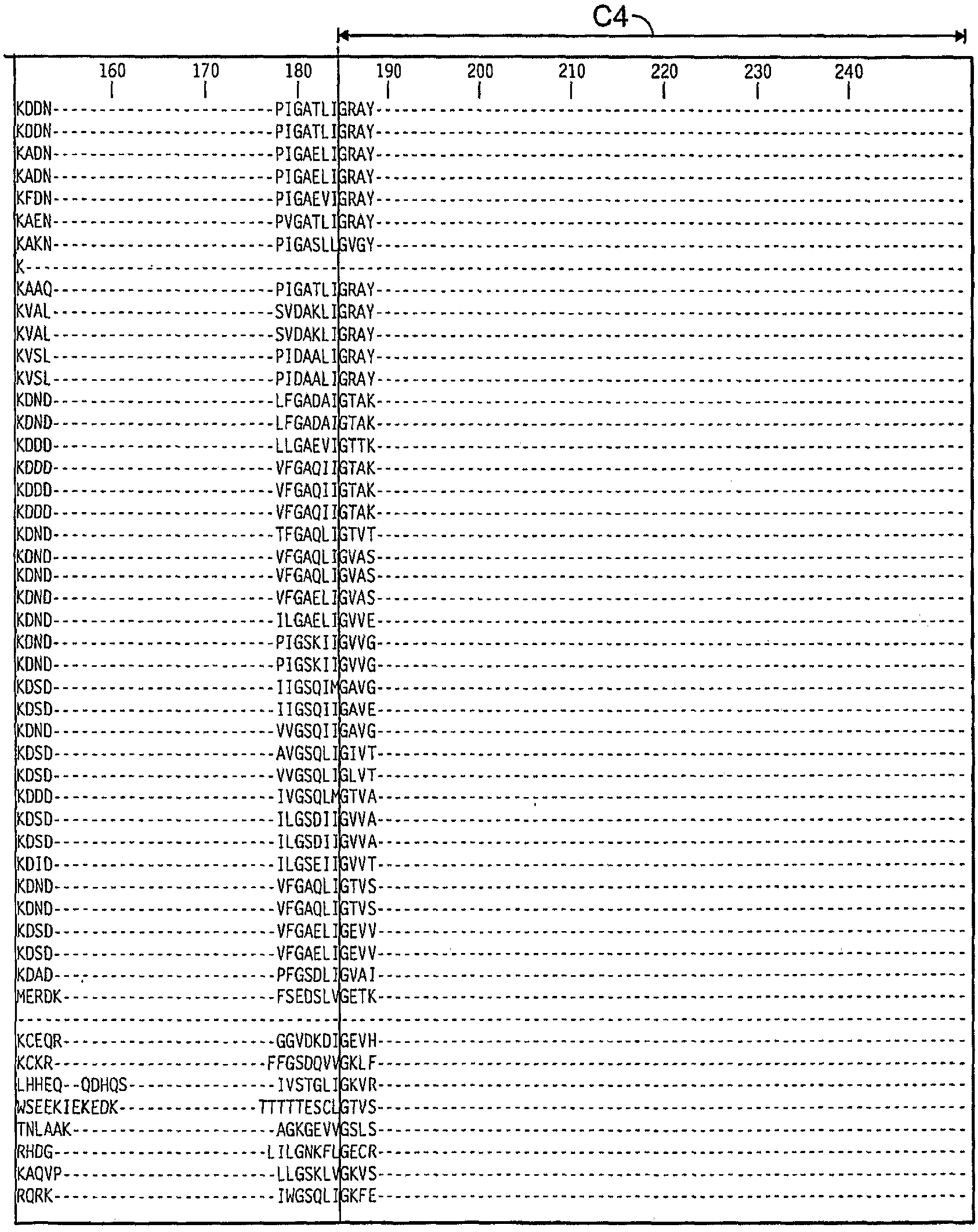


FIG. 3 (Sheet 27 of 27)

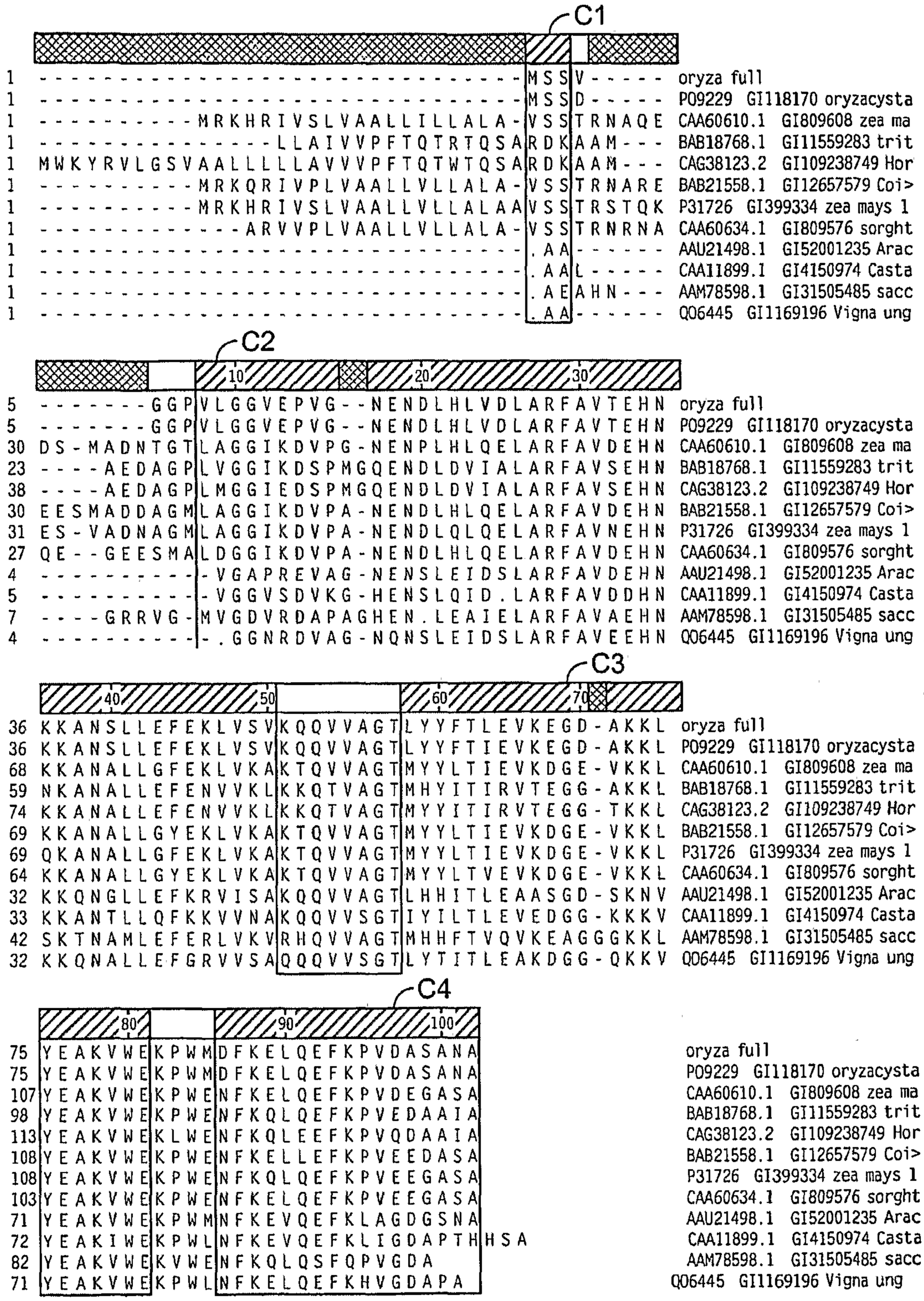

C1
C2
C3
C4
10
20
30
40
50
60
70
80
90
100
1 - M S S V - - - - - oryza full
1 - M S S D - - - - - P09229 GI118170 oryzacysta
1 - - - - - - - - - - M R K H R I V S L V A A L L I L L A L A - V S S T R N A Q E CAA60610.1 GI809608 zea ma
1 - - - - - - - - - - - - - - - - L L A I V V P F T Q T R T Q S A R D K A A M - - - BAB18768.1 GI11559283 trit
1 M W K Y R V L G S V A A L L L L L A V V V P F T Q T W T Q S A R D K A A M - - - CAG38123.2 GI109238749 Hor
1 - - - - - - - - - - M R K Q R I V P L V A A L L V L L A L A - V S S T R N A R E BAB21558.1 GI12657579 Coi>
1 - - - - - - - - - - M R K H R I V S L V A A L L V L L A L A A V S S T R S T Q K P31726 GI399334 zea mays 1
1 - - - - - - - - - - - - - A R V V P L V A A L L V L L A L A - V S S T R N R N A CAA60634.1 GI809576 sorght
1 - . A A - - - - - - AAU21498.1 GI52001235 Arac
1 - . A A L - - - - - CAA11899.1 GI4150974 Casta
1 - . A E A H N - - - AAM78598.1 GI31505485 sacc
1 - . A A - - - - - - Q06445 GI1169196 Vigna ung
5 - - - - - - - G G P V L G G V E P V G - - N E N D L H L V D L A R F A V T E H N oryza full
5 - - - - - - - G G P V L G G V E P V G - - N E N D L H L V D L A R F A V T E H N P09229 GI118170 oryzacysta
30 D S - M A D N T G T L A G G I K D V P G - N E N P L H L Q E L A R F A V D E H N CAA60610.1 GI809608 zea ma
23 - - - - A E D A G P L V G G I K D S P M G Q E N D L D V I A L A R F A V S E H N BAB18768.1 GI11559283 trit
38 - - - - A E D A G P L M G G I E D S P M G Q E N D L D V I A L A R F A V S E H N CAG38123.2 GI109238749 Hor
30 E E S M A D D A G M L A G G I K D V P A - N E N D L H L Q E L A R F A V D E H N BAB21558.1 GI12657579 Coi>
31 E S - V A D N A G M L A G G I K D V P A - N E N D L Q L Q E L A R F A V N E H N P31726 GI399334 zea mays 1
27 Q E - - G E E S M A L D G G I K D V P A - N E N D L H L Q E L A R F A V D E H N CAA60634.1 GI809576 sorght
4 - - - - - - - - - - - V G A P R E V A G - N E N S L E I D S L A R F A V D E H N AAU21498.1 GI52001235 Arac
5 - - - - - - - - - - - V G G V S D V K G - H E N S L Q I D . L A R F A V D D H N CAA11899.1 GI4150974 Casta
7 - - - - G R R V G - M V G D V R D A P A G H E N . L E A I E L A R F A V A E H N AAM78598.1 GI31505485 sacc
4 - - - - - - - - - - - - . G G N R D V A G - N Q N S L E I D S L A R F A V E E H N Q06445 GI1169196 Vigna ung
36 K K A N S L L E F E K L V S V K Q Q V V A G T L Y Y F T L E V K E G D - A K K L oryza full
36 K K A N S L L E F E K L V S V K Q Q V V A G T L Y Y F T I E V K E G D - A K K L P09229 GI118170 oryzacysta
68 K K A N A L L G F E K L V K A K T Q V V A G T M Y Y L T I E V K D G E - V K K L CAA60610.1 GI809608 zea ma
59 N K A N A L L E F E N V V K L K K Q T V A G T M H Y I T I R V T E G G - A K K L BAB18768.1 GI11559283 trit
74 K K A N A L L E F E N V V K L K K Q T V A G T M Y Y I T I R V T E G G - T K K L CAG38123.2 GI109238749 Hor
69 K K A N A L L G Y E K L V K A K T Q V V A G T M Y Y L T I E V K D G E - V K K L BAB21558.1 GI12657579 Coi>
69 Q K A N A L L G F E K L V K A K T Q V V A G T M Y Y L T I E V K D G E - V K K L P31726 GI399334 zea mays 1
64 K K A N A L L G Y E K L V K A K T Q V V A G T M Y Y L T V E V K D G E - V K K L CAA60634.1 GI809576 sorght
32 K K Q N G L L E F K R V I S A K Q Q V V A G T L H H I T L E A A S G D - S K N V AAU21498.1 GI52001235 Arac
33 K K A N T L L Q F K K V V N A K Q Q V V S G T I Y I L T L E V E D G G - K K K V CAA11899.1 GI4150974 Casta
42 S K T N A M L E F E R L V K V R H Q V V A G T M H H F T V Q V K E A G G G K K L AAM78598.1 GI31505485 sacc
32 K K Q N A L L E F G R V V S A Q Q Q V V S G T L Y T I T L E A K D G G - Q K K V Q06445 GI1169196 Vigna ung
75 Y E A K V W E K P W M D F K E L Q E F K P V D A S A N A oryza full
75 Y E A K V W E K P W M D F K E L Q E F K P V D A S A N A P09229 GI118170 oryzacysta
107 Y E A K V W E K P W E N F K E L Q E F K P V D E G A S A CAA60610.1 GI809608 zea ma
98 Y E A K V W E K P W E N F K Q L Q E F K P V E D A A I A BAB18768.1 GI11559283 trit
113 Y E A K V W E K L W E N F K Q L E E F K P V Q D A A I A CAG38123.2 GI109238749 Hor
108 Y E A K V W E K P W E N F K E L L E F K P V E E D A S A BAB21558.1 GI12657579 Coi>
108 Y E A K V W E K P W E N F K Q L Q E F K P V E E G A S A P31726 GI399334 zea mays 1
103 Y E A K V W E K P W E N F K E L Q E F K P V E E G A S A CAA60634.1 GI809576 sorght
71 Y E A K V W E K P W M N F K E V Q E F K L A G D G S N A AAU21498.1 GI52001235 Arac
72 Y E A K I W E K P W L N F K E V Q E F K L I G D A P T H H S A CAA11899.1 GI4150974 Casta
82 Y E A K V W E K V W E N F K Q L Q S F Q P V G D A AAM78598.1 GI31505485 sacc
71 Y E A K V W E K P W L N F K E L Q E F K H V G D A P A Q06445 GI1169196 Vigna ung

FIG. 4

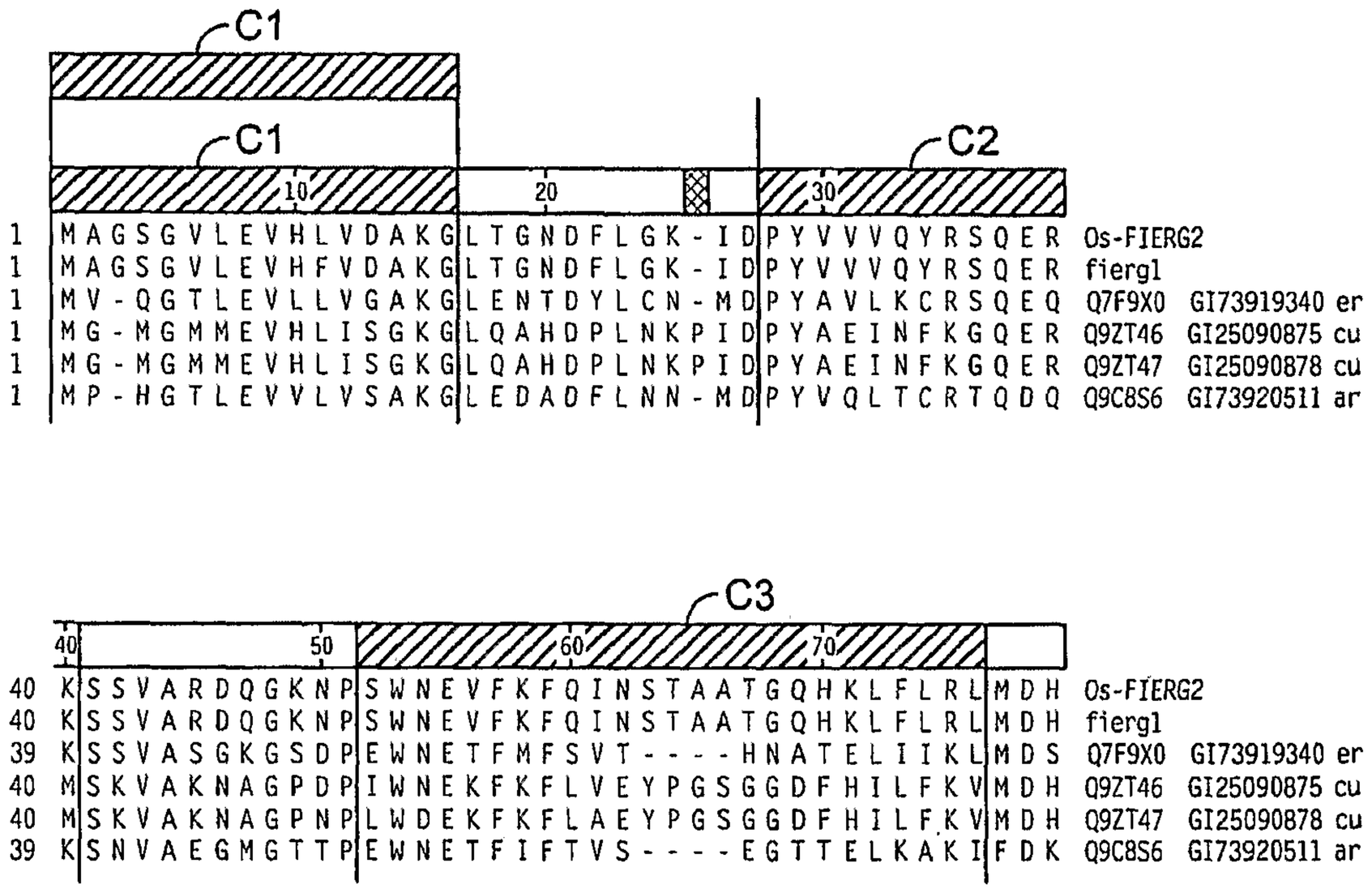

C1
C1
C2
C3
C4
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
1 MAGSGVLEVHLVDAKGLTGNDFLGK-IDPYVVVQYRSQER Os-FIERG2
1 MAGSGVLEVHFVDAKGLTGNDFLGK-IDPYVVVQYRSQER fiergl
1 MV-QGTLEVLLVGAKGLENTDYLCN-MDPYAVLKCRSQEQ Q7F9X0 GI73919340 er
1 MG-MGMMEVHLISGKGLQAHDPLNKPIDPYAEINFKGQER Q9ZT46 GI25090875 cu
1 MG-MGMMEVHLISGKGLQAHDPLNKPIDPYAEINFKGQER Q9ZT47 GI25090878 cu
1 MP-HGTLEVVLVSAKGLEDADFLNN-MDPYVQLTCRTQDQ Q9C8S6 GI73920511 ar
40 KSSVARDQGKNPSWNEVFKFQINSTAATGQHKLFLRLMDH Os-FIERG2
40 KSSVARDQGKNPSWNEVFKFQINSTAATGQHKLFLRLMDH fiergl
39 KSSVASGKGSDPEWNETFMFSVT----HNATELIIKLMDS Q7F9X0 GI73919340 er
40 MSKVAKNAGPDPIWNEKFKFLVEYPGSGGDFHILFKVMDH Q9ZT46 GI25090875 cu
40 MSKVAKNAGPNPLWDEKFKFLAEYPGSGGDFHILFKVMDH Q9ZT47 GI25090878 cu
39 KSNVAEGMGTTPEWNETFIFTVS----EGTTELKAKIFDK Q9C8S6 GI73920511 ar
80 DTFSRDDFLGEATINVTDLISLGMEHGTWEMSESKHRVVL Os-FIERG2
80 DTFSRDDFLGEATINVTDLISLGMEHGTWEMSESKHRVVL fiergl
75 DSGTDDDFVGEATISLEAIYTEG------SIPPTVYNVVK Q7F9X0 GI73919340 er
80 DAIDGDDYIGDVKIDVQNLLAEGVRKGWSELPPRMYQVLA Q9ZT46 GI25090875 cu
80 DAIDGDDYIGDVKIDVKNLLAEGVRKGKSEMPPRMYHVLA Q9ZT47 GI25090878 cu
75 DVGTEDDAVGEATIPLEPVFVEG------SIPPTAYNVVK Q9C8S6 GI73920511 ar
120 ADKTYHGEIRVSLTFTASAKAQD--HAEQVGGWAHSFRQ Os-FIERG2
120 ADKTYHGEIRVSLTFTASAKAQD--HAQQVGGWAHSFRQ fiergl
109 EEE-YRGEIKVGLTFTPEDDRDRGLSEEDIGGWKQSS Q7F9X0 GI73919340 er
120 HKIYFKGEIEVGVFFQRQG Q9ZT46 GI25090875 cu
120 HKIHFKGEIEVGVSFKLQGGGGCG----GCYPWEN Q9ZT47 GI25090878 cu
109 DEE-YKGEIWVALSFKP.ENRSRGMDEESYGGWKNSEASY Q9C8S6 GI73920511 ar

FIG. 5

Encoded Scaffold Polypeptide Sequence

C_1 X_1 C_2 X_2 C_3 X_3 C_4

PQQVHITQGDHVGKAVIVSWVTMDEPGSSVVVYWSENSKYKKSAEGTVTTYRFYNYTSGYIHHCYIKGLEYDTKYYYVVGIGNTSREFWFR

Sequence Diversification of SPS at X_1, X_2, X_3

Library of Encoded Chimeric Binding Polypeptides

e.g.

PQQVHITQGDHVGKAVIVSWVT*HF*EPASKVVVYWSENSKYKKSAEGTVTTYRF*K*NYN_GYIHHCYIKGLEYDTKYYYV*V*F*I*GNT*RA*SREFWFR

PQQVHITQGDHVGKAVIVSWVTM*FSPH*SA*K*VVVYWSENSKYKKSAEGTVTTY*QSI*FAN_NS*R*YIHHCYIKGLEYDTKYYY*VDVKFG*PIG*LI*NTRASREFWFR

PQQVHITQGDHVGKAVIVSWVTM*ARKLSP*VVVYWSENSKYKKSAEGTVTTYR*D*YN*K*TSGYIHHCYIKGLEYDTKYYYVVGIGNTSREFWFR