



(22) Date de dépôt/Filing Date: 2002/05/21
(41) Mise à la disp. pub./Open to Public Insp.: 2002/09/04
(45) Date de délivrance/Issue Date: 2008/09/23
(62) Demande originale/Original Application: 2 387 401
(30) Priorités/Priorities: 2001/05/21 (USUSSN 60/291,959);
2001/12/03 (USUSSN 60/334,604)

(51) Cl.Int./Int.Cl. *C12N 15/55* (2006.01),
C12N 15/54 (2006.01), *C12N 9/10* (2006.01),
C12N 9/16 (2006.01), *C12P 1/00* (2006.01),
C12P 21/02 (2006.01), *C12Q 1/04* (2006.01),
C12Q 1/68 (2006.01), *C40B 30/04* (2006.01),
C40B 40/02 (2006.01)

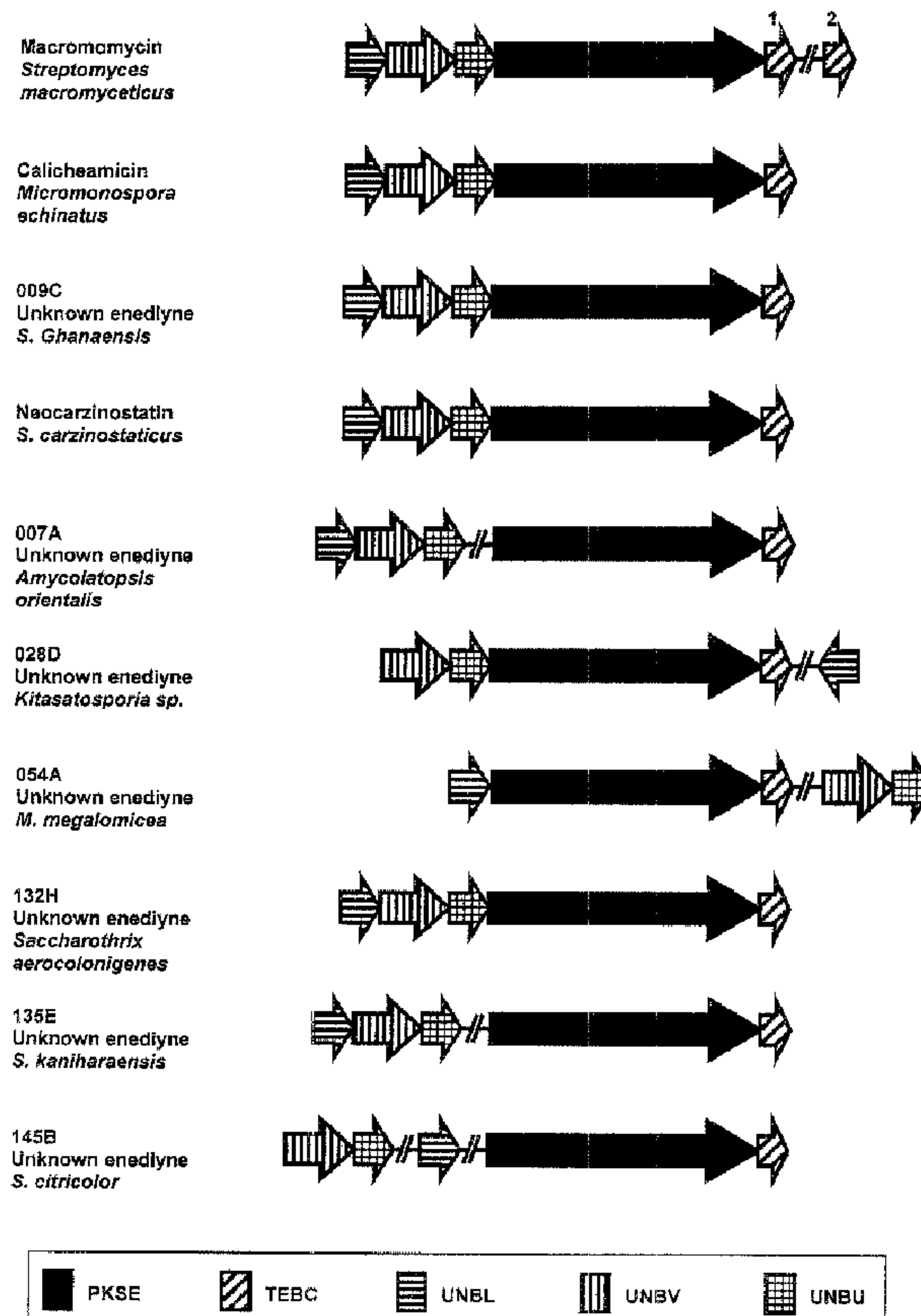
(72) Inventeurs/Inventors:
FARNET, CHRIS M., CA;
ZAZOPOULOS, EMMANUEL, CA;
STAFFA, ALFREDO, CA

(73) Propriétaire/Owner:
THALLION PHARMACEUTICALS INC./THALLION
PHARMACEUTIQUES INC., CA

(74) Agent: ROBIC

(54) Titre : COMPOSITIONS, METHODES ET DISPOSITIFS UTILISES POUR DECOUVRIR DES PRODUITS
NATURELS A BASE D'ENEDIYNE

(54) Title: COMPOSITIONS, METHODS AND SYSTEMS FOR THE DISCOVERY OF ENEDIYNE NATURAL PRODUCTS



(57) Abrégé/Abstract:

Five protein families cooperate to form the warhead structure that characterizes enediyne compounds, both chromoprotein enediynes and non-chromoprotein enediynes. The protein family of the present divisional application is a thioesterase protein family



(57) **Abrégé(suite)/Abstract(continued):**

members of which form, together with the enediynes polyketide synthase protein family of parent application CA 2,387,401, a polyketide synthase catalytic complex involved in warhead formation in enediynes. Genes encoding a member of each of the five protein families are found in all enediyne biosynthetic loci. The genes and proteins may be used in genetic engineering applications to design new enediyne compounds and in methods to identify new enediyne biosynthetic loci.

Abstract:

Five protein families cooperate to form the warhead structure that characterizes enediynes compounds, both chromoprotein enediynes and non-chromoprotein enediynes. The protein family of the present divisional application is a thioesterase protein family members of which form, together with the enediyne polyketide synthase protein family of parent application CA 2,387,401, a polyketide synthase catalytic complex involved in warhead formation in enediynes. Genes encoding a member of each of the five protein families are found in all enediyne biosynthetic loci. The genes and proteins may be used in genetic engineering applications to design new enediyne compounds and in methods to identify new enediyne biosynthetic loci.

-1-

TITLE OF THE INVENTION: Compositions, methods and systems for the discovery of enediyne natural products.

FIELD OF INVENTION

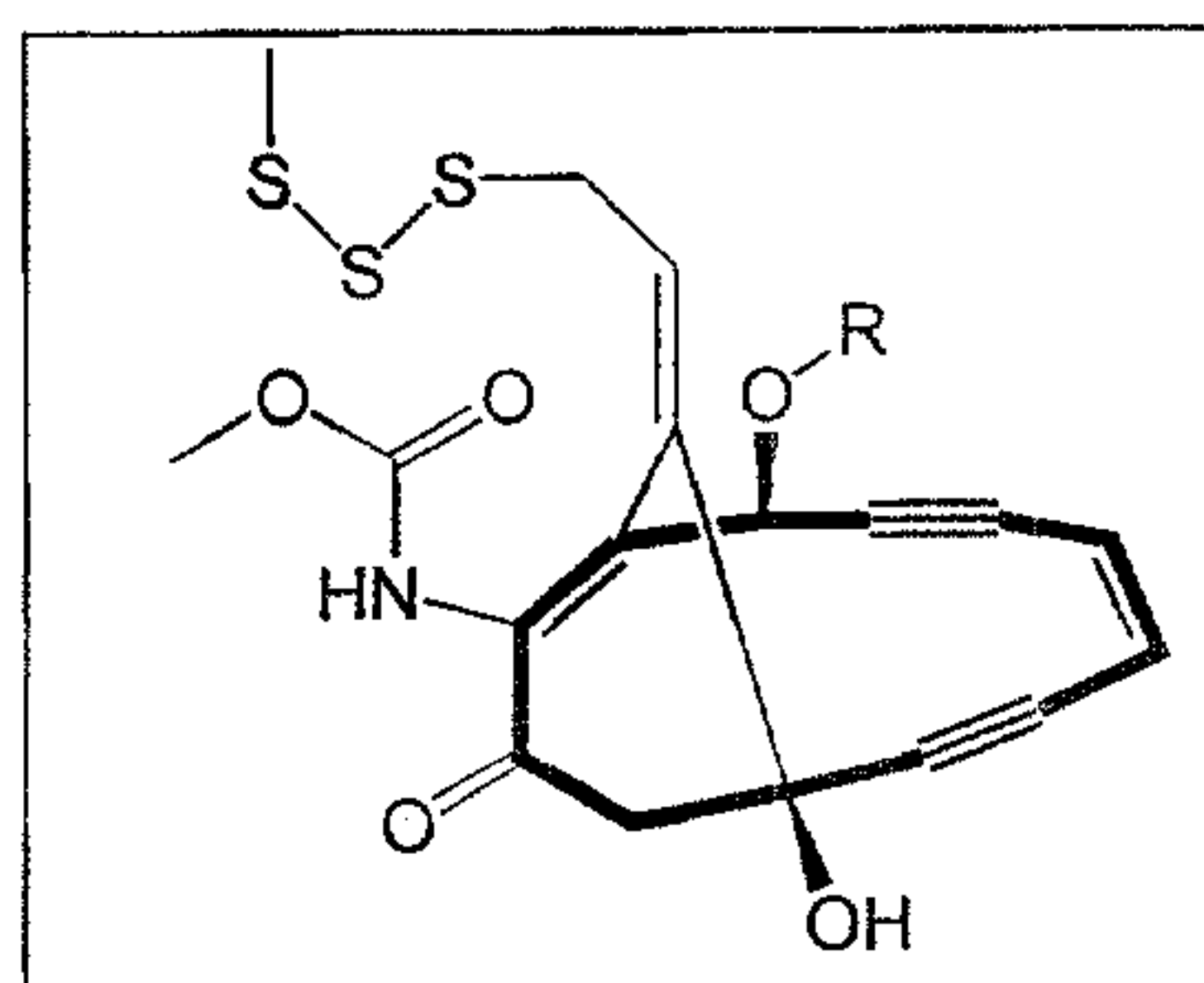
The present invention relates to the field of microbiology, and more specifically to genes and proteins involved in the production of enediynes.

BACKGROUND

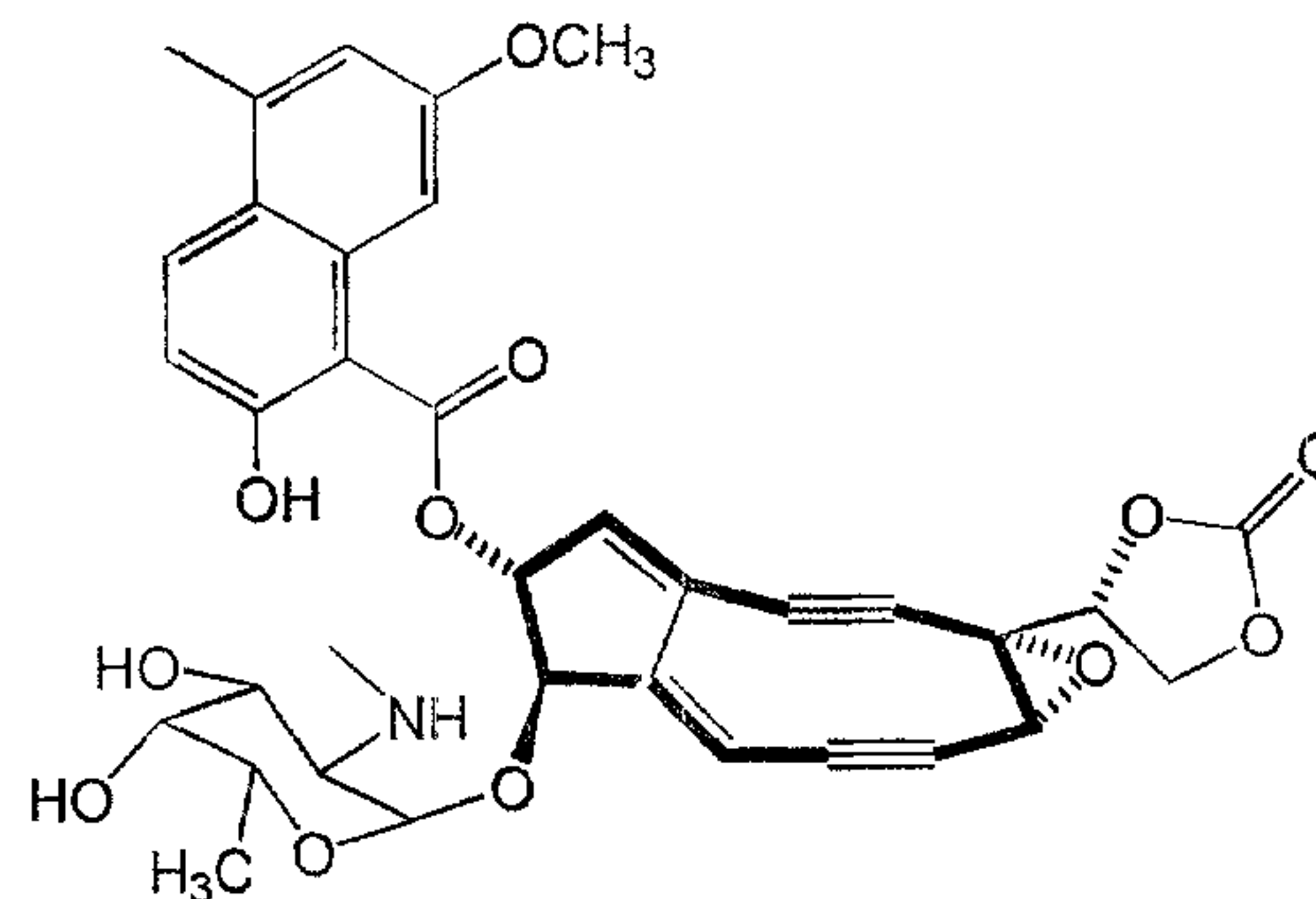
10 Enediyne natural products are characterized by the presence of the enediyne ring structure also referred to as the warhead. The labile enediyne ring structure undergoes a thermodynamically favorable Bergman cyclization resulting in transient formation of a biradical species. The biradical species is capable of inducing irreversible DNA damage in the cell. This reactivity gives rise to potential biological activity against both bacterial and tumor cell lines. Enediynes have potential as anticancer agents because of their ability to cleave DNA. Calicheamicin is currently in clinical trials as an anticancer agent for acute myeloid leukemia (Nabhan C. and Tallman MS, Clin Lymphoma (2002) Mar; 2 Suppl 1:S19-23). Enediynes also have utility as anti-infective agents. Accordingly, processes for improving production of
20 existing enediynes or producing novel modified enediynes are of great interest to the pharmaceutical industry.

Enediynes are a structurally diverse group of compounds. Chromoprotein enediynes refer to enediynes associated with a protein conferring stability to the complex under physiological conditions. Non-chromoprotein enediynes refer to enediynes that require no additional stabilization factors. The structure of the chromoprotein enediynes neocarzinostatin and C-1027, and the non-chromoprotein enediynes calicheamicin and dynemicin are shown below with the dodecapolyene backbone forming the warhead structure in each enediyne highlighted in bold.

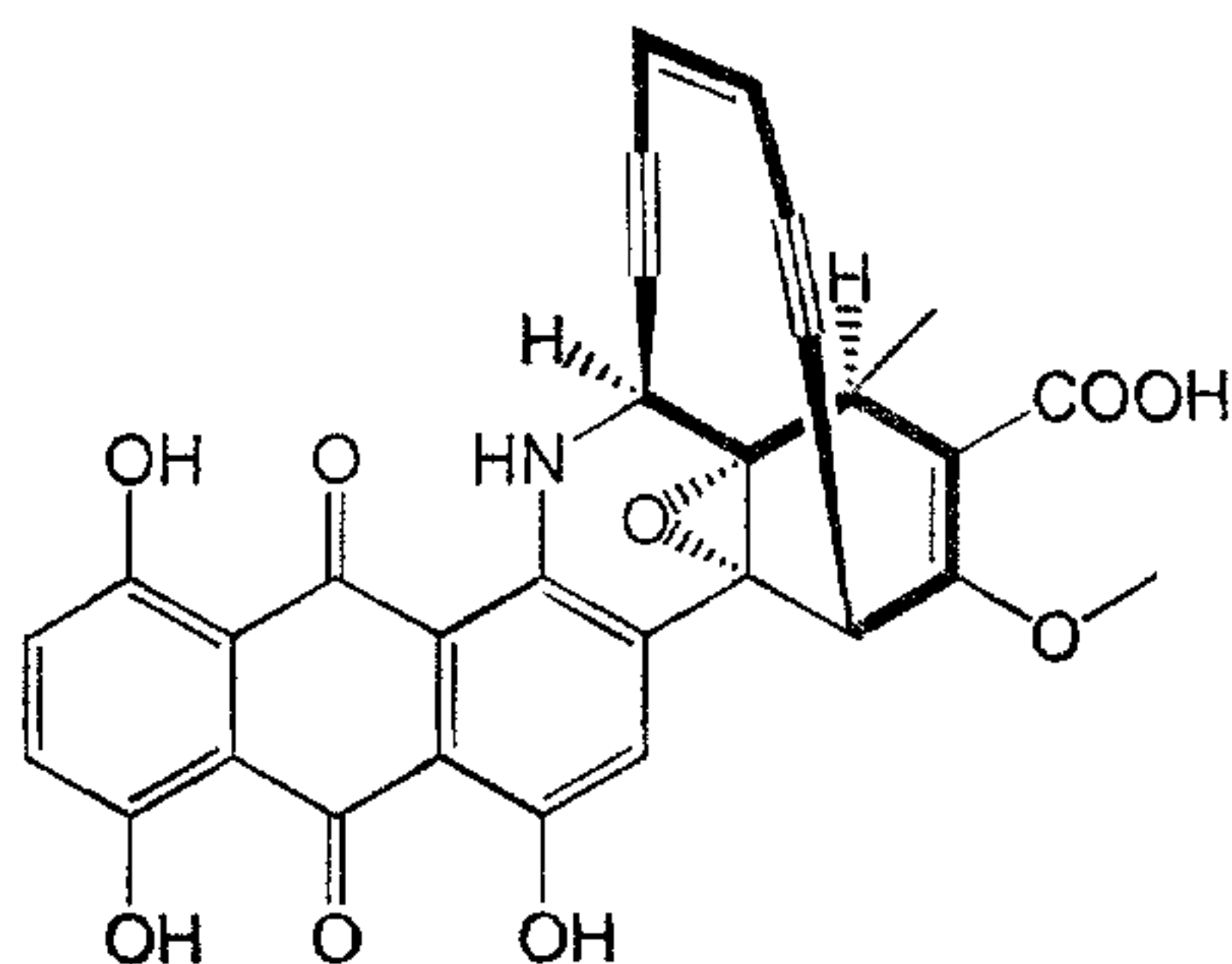
-2-



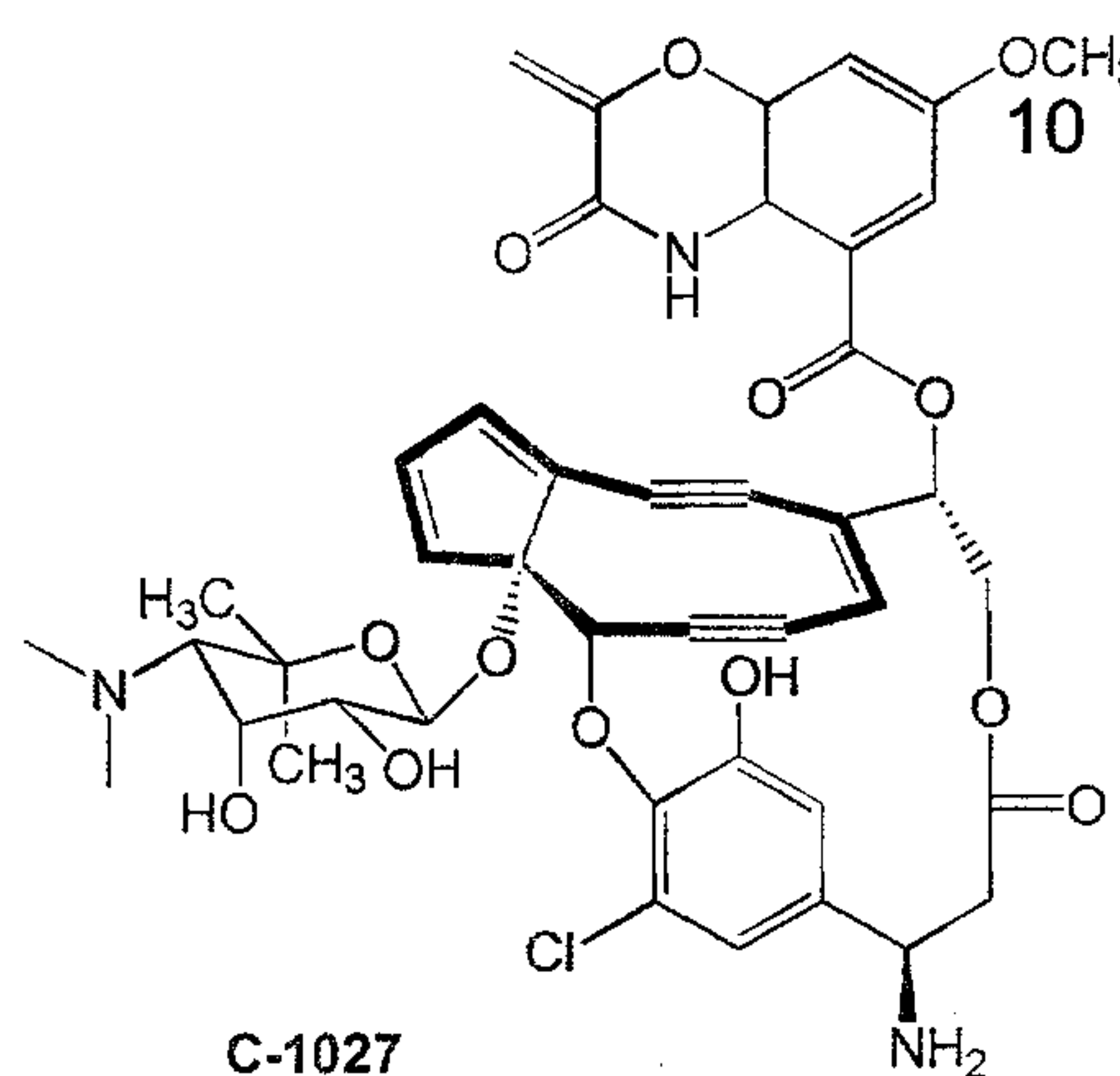
Calicheamicin



Neocarzinostatin



Dynemicin A



C-1027

Efforts at discovering the genes responsible for synthesis of the warhead structure that characterizes enediynes have been unsuccessful. Genes encoding biosynthetic enzymes for the aryltetrasaccharide of calicheamicin, and for calicheamicin resistance are described in WO 00/37608. Additional genes involved in the biosynthesis of the chromoprotein enediyne C-1027 have been isolated (Liu, *et al. Antimicrobial Agents and Chemotherapy*, vol. 44, pp 382-292 (2000); WO 00/40596). Isotopic incorporation experiments have indicated that the enediyne backbones of esperamicin, dynemicin, and neocarzinostatin are acetate derived (Hansens, O.D. *et al. J. Am. Chem Soc.* 11, vol 111 pp. 3295-3299 (1989); Lam, K. *et al. J. Am. Chem. Soc.* vol. 115, pp 12340-12345 (1993); Tokiwa, Y *et al. J. Am. Chem Soc.* vol. 113 pp. 4107-4110). However, both PCR and DNA probes homologous to type I and type II

PKSs have failed to identify the presence of PKS genes associated with biosynthesis of enediynes in known enediyne producing microorganisms (WO 00/40596; W. Liu & B. Shen, *Antimicrobial Agents Chemotherapy*, vol. 44 No. 2 pp.382-392 (2000)).

-3-

Elucidation of the genes involved in biosynthesis of enediynes, particularly the warhead structure, would provide access to rational engineering of enediyne biosynthesis for novel drug leads and makes it possible to construct overproducing strains by de-regulating the biosynthetic machinery. Elucidation of PKS genes involved in the biosynthesis of enediynes would contribute to the field of combinatorial biosynthesis by expanding the repertoire of PKS genes available for making novel enediynes via combinatorial biosynthesis.

Existing screening methods for identifying enediyne-producing microbes are laborious, time-consuming and have not provided sufficient discrimination to date to
10 detect organisms producing enediyne natural products at low levels. There is a need for improved tools to detect enediyne-producing organisms. There is also a need for tools capable of detecting organisms that produce enediynes at levels that are not detected by traditional culture tests.

SUMMARY OF THE INVENTION:

One embodiment of the present invention is an isolated, purified or enriched nucleic acid comprising a sequence selected from the group consisting of: (a) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 23; sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 23; fragments comprising at least 150, preferably at least
20 200, more preferably at least 250, still more preferably at least 300, still more preferably at least 350 and most preferably at least 400 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 23; and fragments comprising at least 150, preferably at least 200, more preferably at least 250, still more preferably at least 300, still more preferably at least 350 and most preferably at least 400 consecutive nucleotides of the sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 23; (b) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401; sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401; fragments comprising at least 2000, preferably at least 3000, more preferably at least 4000, still more preferably at least 5000, still more preferably at least 5600 and most preferably at
30 least 5750 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401; and fragments comprising at least 2000, preferably at least 3000, more preferably at least 4000, still more preferably at least 5000, still more

-4-

preferably at least 5600 and most preferably at least 5750 consecutive nucleotides of the sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401; (c) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, a divisional application of CA 2,387,401; sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692; fragments comprising at least 700, preferably at least 750, more preferably at least 800, still more preferably at least 850, still more preferably at least 900 and most preferably at least 950 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692; and fragments comprising at least 700, preferably at least 750, more preferably at least 800, still more preferably at least 850, still more preferably at least 900 and most preferably at least 950 consecutive nucleotides of the sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692; (d) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, a divisional application of CA 2,387,401; sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812; fragments comprising at least 600, preferably at least 700, more preferably at least 750, still more preferably at least 800, still more preferably at least 850 and most preferably at least 900 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812; and fragments comprising at least 600, preferably at least 700, more preferably at least 750, still more preferably at least 800, still more preferably at least 850 and most preferably at least 900 consecutive nucleotides of the sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812; and (e) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802, a divisional application of CA 2,387,401; sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; fragments comprising at least 700, preferably at least 750, more preferably at least 800, still more preferably at least 850, still more preferably at least 900 and most preferably at least 950 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; and fragments comprising at least 700, preferably at least 750, more preferably at least 800, still more preferably at least 850, still more preferably at least 900 and most preferably at least 950 consecutive nucleotides of the sequences complementary to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802. One aspect of the present invention is an isolated, purified or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of high stringency. Another aspect of the present

-5-

invention is an isolated, purified or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of moderate stringency. Another aspect of the present invention is an isolated, purified or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under low stringency. Another aspect of the present invention is an isolated, purified or enriched nucleic acid having at least 70% identity to the nucleic acid of this embodiment by analysis with BLASTN version 2.0 with the default parameters. Another aspect of the present invention is an isolated, purified or enriched nucleic acid having at least 99% identity to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

In one aspect, the invention of the present divisional application provides an isolated, purified or enriched nucleic acid coding for a polypeptide that produces an alignment of at least 49 percent identity to the consensus sequence of SEQ ID NO: 1, as determined using the BLASTP algorithm with the default parameters.

In another aspect, the present divisional application provides an isolated, purified or enriched nucleic acid that encodes an enediyne thioesterase protein comprising a polypeptide selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; (b) polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 or 22 during synthesis a warhead structure in an enediyne compound; and (c) fragments of the polypeptides of (a) and (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the nucleic acid encoding the enediyne protein may be used in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

One aspect of the parent application of the present divisional application provides an isolated, purified or enriched nucleic acid that encodes an enediyne polyketide synthase protein comprising a polypeptide selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; (b) polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as determined by BLASTP algorithm with the

-6-

default parameters, and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 during synthesis of a warhead structure in an enediyne compound; and (c) fragments of the polypeptides of (a) and (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the nucleic acid encoding the enediyne polyketide synthase protein may be used in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

10 In another aspect, the present divisional application provides an isolated, purified or enriched nucleic acid that encodes an enediyne polyketide synthase catalytic complex, said polyketide synthase catalytic complex comprising a polyketide synthase having at least 45 percent identity to SEQ ID NO: 24, and thioesterase having at least 49 percent identity to SEQ ID NO: 1, wherein the percent identity is determined using the BLASTP algorithm with the default parameters.

Another embodiment of the invention is an isolated, purified or enriched nucleic acid that encodes an enediyne polyketide synthase catalytic complex comprising (a) a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as determined using
20 the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 of CA 2,387,401 during synthesis a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 in the synthesis of the warhead structure in an enediyne compound; and (b) a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 application as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
30 during synthesis of a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of the present application in the synthesis of the warhead structure in an enediyne compound. In

-7-

one aspect of this embodiment, the nucleic acid encoding an enediyne polyketide synthase catalytic complex may be used in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

Another embodiment is a gene cassette comprising: (a) a nucleic acid encoding an enediyne polyketide synthase catalytic complex as described above; and (b) at least one nucleic acid encoding a polypeptide selected from the group consisting of (i) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 as determined using BLASTP algorithm with the default parameters and
10 having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 during synthesis of a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 in the synthesis of a warhead structure in an enediyne compound; (ii) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 during synthesis of a warhead structure in an enediyne compound; and
20 fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 in the synthesis of the warhead structure in an enediyne compound; and (iii) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 during synthesis of a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 in the synthesis of the
30 warhead structure in an enediyne compound. In one aspect of this embodiment, the nucleic acid encoding the gene cassette may be used in genetic engineering application to synthesize the warhead structure of an enediyne compound.

-8-

Another embodiment is an isolated, purified or enriched nucleic acid encoding a gene cassette comprising: (a) a nucleic acid encoding a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; a polypeptide having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 of CA 2,387,401 during synthesis a warhead structure in an enediyne compound; or a fragment thereof, which fragment has the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 in the synthesis of the warhead structure in an enediyne compound; (b) at least one nucleic acid encoding a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 of the present divisional application; a polypeptide having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 of the present divisional application as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 during synthesis of a warhead structure in an enediyne compound; or a fragment thereof, which fragment has the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 in the synthesis of the warhead structure in an enediyne compound; (c) at least one nucleic acid encoding a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 which is a divisional application of CA 2,387,401; a polypeptide having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 during synthesis of a warhead structure in an enediyne compound; and a fragment thereof, which fragment has the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 in the synthesis of the warhead structure in an enediyne compound; (d) at least one nucleic acid encoding a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 which is a divisional application of CA 2,387,401; a polypeptide having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10,

12, 14, 16, 18, 20 of CA 2,444,812 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 during synthesis a warhead structure in an enediyne compound; and a fragment thereof, which fragment has the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 in the synthesis of the warhead structure in an enediyne compound; and (e) at least one nucleic acid encoding a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; a polypeptide having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 during synthesis a warhead structure in an enediyne compound; and a fragment thereof, which fragment has the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the nucleic acid encoding the gene cassette may be used in genetic engineering application to synthesize the warhead structure of an enediyne compound.

Another embodiment of the present invention is an isolated or purified polypeptides comprising a sequence selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; and fragments comprising at least 1300, preferably at least 1450, more preferably at least 1550, still more preferably at least 1650, still more preferably at least 1750 and most preferably at least 1850 consecutive amino acids of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; (b) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; and fragments comprising at least 40, preferably at least 60, more preferably at least 80, still more preferably at least 100, still more preferably at least 120 and most preferably at least 130 consecutive amino acids of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; (c) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692; and fragments comprising at least 220, preferably at least 240, more preferably at least 260, still more preferably at least 280, still more preferably at least 300 and most preferably at least 310

-10-

consecutive amino acids of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692; (d) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812; and fragments comprising at least 520, preferably at least 540, more preferably at least 560, still more preferably at least 580, still more preferably at least 600 and most preferably at least 620 consecutive amino acids of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812; and (e) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; and fragments comprising at least 220, preferably at least 240, more preferably at least 260, still more preferably at least 280, still more preferably at least 300 and most preferably at least 320 consecutive amino acids of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802. One aspect of the present invention is an isolated or purified polypeptide having at least 70% identity to the polypeptide of this embodiment by analysis with BLASTP algorithm with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 99% identity to the polypeptides of this embodiment as determined by analysis with BLASTP algorithm with the default parameters.

An embodiment of the parent application CA 2,387,401 is an isolated or purified enediyne polyketide synthase comprising a polypeptide selected from the group consisting of (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; (b) polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 during synthesis a warhead structure in an enediyne compound; and (c) fragments of the polypeptides of (a) and (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the enediyne polyketide synthase protein may be used in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

Another embodiment is an isolated, purified enediyne polyketide synthase catalytic complex comprising (a) a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of

-11-

CA 2,387,401 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 during synthesis a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 in the synthesis of the warhead structure in an enediyne compound; and (b) a polypeptide selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 of the present divisional application as determined using
10 the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 during synthesis of a warhead structure in an enediyne compound; and fragments thereof, which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the enediyne polyketide synthase catalytic complex may be used in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

One aspect of CA 2,445,692 is a polypeptide selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692; (b) polypeptides
20 having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 during synthesis of a warhead structure in an enediyne compound; and (c) fragments of (a) or (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 in the synthesis of the warhead structure in an enediyne compound. In one aspect, the polypeptide of this embodiment may be used with an enediyne polyketide synthase catalytic complex of the invention in genetic engineering applications to synthesize the warhead of an enediyne compound.

30 One aspect of CA 2,444,812 is a polypeptide selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12,

-12-

14, 16, 18, 20 of CA 2,444,812; (b) polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 during synthesis of a warhead structure in an enediyne compound; and (c) fragments of (a) or (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 in the synthesis of the warhead structure in an enediyne compound. In one aspect, the polypeptide of this embodiment may be used with an enediyne polyketide synthase catalytic complex of the invention in genetic engineering applications to synthesize the warhead structure of an enediyne compound.

One aspect of CA 2,444,802 is a polypeptide selected from the group consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; (b) polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 as determined using the BLASTP algorithm with the default parameters and having the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 during synthesis of a warhead structure in an enediyne compound; and (c) fragments of (a) or (b), which fragments have the ability to substitute for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 in the synthesis of the warhead structure in an enediyne compound. In one aspect of this embodiment, the polypeptide of this embodiment may be used with an enediyne polyketide synthase catalytic complex of the invention in genetic engineering applications to synthesize the warhead structure of enediyne compound.

An enediyne gene cluster may be identified using compositions of the invention such as hybridization probes or PCR primers. Hybridization probes or PCR primers according to the invention are derived from protein families associated with the warhead structure characteristic of enediynes. To identify enediyne gene clusters, the hybridization probes or PCR primers are derived from any one or more nucleic acid sequences corresponding to the five protein families designated herein as PKSE, TEBC, UNBL, UNBV and UNBU. The compositions of the invention are used as probes to identify enediyne biosynthetic genes, enediyne gene fragments, enediyne gene clusters, or enediyne producing organisms from samples including potential

-13-

enediyne producing microorganisms. The samples may be in the form of environmental biomass, pure or mixed microbial culture, isolated genomic DNA from pure or mixed microbial culture, genomic DNA libraries from pure or mixed microbial culture. The compositions are used in polymerase chain reaction, and nucleic acid hybridization techniques well known to those skilled in the art.

Environmental samples that harbour microorganisms with the potential to produce enediynes are identified by PCR methods. Nucleic acids contained within the environmental sample are contacted with primers derived from the invention so as to amplify target enediyne biosynthetic gene sequences. Environmental samples deemed
10 to be positive by PCR are then pursued to identify and isolate the enediyne gene cluster and the microorganism that contains the target gene sequences. The enediyne gene cluster may be identified by generating genomic DNA libraries (for example, cosmid, BAC, etc.) representative of genomic DNA from the population of various microorganisms contained within the environmental sample, locating genomic DNA clones that contain the target sequences and possibly overlapping clones (for example, by hybridization techniques or PCR), determining the sequence of the desired genomic DNA clones and deducing the ORFs of the enediyne biosynthetic locus. The microorganism that contains the enediyne biosynthetic locus may be identified and isolated, for example, by colony hybridization using nucleic acid probes derived from
20 either the invention or the newly identified enediyne biosynthetic locus. The isolated enediyne biosynthetic locus may be introduced into an appropriate surrogate host to achieve heterologous production of the enediyne compound(s); alternatively, if the microorganism containing the enediyne biosynthetic locus is identified and isolated it may be subjected to fermentation to produce the enediyne compound(s).

A microorganism that harbours an enediyne gene cluster is first identified and isolated as a pure culture, for example, by colony hybridization using nucleic acid probes derived from the invention. Beginning with a pure culture, a genomic DNA library (for example, cosmid, BAC, etc.) representative of genomic DNA from this single species is prepared, genomic DNA clones that contain the target sequences and
30 possibly overlapping clones are located using probes derived from the invention (for example, by hybridization techniques or PCR), the sequence of the desired genomic DNA clones is determined and the ORFs of the enediyne biosynthetic locus are deduced. The microorganism containing the enediyne biosynthetic locus may be

-14-

subjected to fermentation to produce the enediynes compound(s) or the enediyne biosynthetic locus may be introduced into an appropriate surrogate host to achieve heterologous production of the enediynes compound(s).

An enediyne gene cluster may also be identified *in silico* using one or more sequences selected from enediyne-specific nucleic acid code, and enediyne-specific polypeptide code as taught by the invention. A query from a set of query sequences stored on computer readable medium is read and compared to a subject selected from the reference sequences of the invention. The level of similarity between said subject and query is determined and queries sequences representing enediyne genes are
10 identified.

Thus another embodiment of the invention is a method of identifying an enediyne biosynthetic gene or gene fragment comprising providing a sample containing genomic DNA, and detecting the presence of a nucleic acid sequence coding for a polypeptide from at least one or the groups consisting of: (a) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; and polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as determined using the BLASTP algorithm with the default parameters; (b) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; and polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 as determined using
20 the BLASTP algorithm with the default parameters; (c) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 4,445,692; and polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 as determined using the BLASTP algorithm with the default parameters; (d) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812; and polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 as determined using the BLASTP algorithm with the default parameters; and (e) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; and polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 as determined using the BLASTP algorithm with the default
30 parameters. One aspect of this embodiment provides detecting a nucleic acid sequence coding a polypeptide from at least two of the above groups (a), (b), (c), (d) and (e). Another aspect of this embodiment provides detecting a nucleic

-15-

acid sequence coding a polypeptide from at least three of the groups (a), (b), (c), (d) and (e). Another aspect of this embodiment provides detecting a nucleic acid sequence coding a polypeptide from at least four of the groups (a), (b), (c), (d) and (e). Another aspect of this embodiment provides detecting a nucleic acid sequence coding a polypeptide from each of the groups (a), (b), (c), (d) and (e). Another aspect of this embodiment of the invention provide the further step of using the nucleic acid detected to isolate an enediyne gene cluster from the sample containing genomic DNA. Another aspect of this embodiment of the invention comprises identifying an organism containing the nucleic acid sequence detected from the genomic DNA in the sample.

10 It is understood that the invention, having provided, compositions and methods to identify enediyne biosynthetic gene cluster, further provides enediynes produced by the biosynthetic gene clusters identified.

BRIEF DESCRIPTION OF THE DRAWINGS:

Figure 1 is a block diagram of a computer system which implements and executes software tools for the purpose of comparing a query to a subject, wherein the subject is selected from the reference sequences of the invention

20 Figures 2A, 2B, 2C and 2D are flow diagrams of a sequence comparison software that can be employed for the purpose of comparing a query to a subject, wherein the subject is selected from the reference sequences of the invention, wherein Figure 2A is the query initialization subprocess of the sequence comparison software, Figure 2B is the subject datasource initialization subprocess of the sequence comparison software, Figure 2C illustrates the comparison subprocess and the analysis subprocess of the sequence comparison software, and Figure 2D is the Display/Report subprocess of the sequence comparison software.

Figure 3 is a flow diagram of the comparator algorithm (238) of Figure 2C which is one embodiment of a comparator algorithm that can be used for pairwise determination of similarity between a query/subject pair.

30 Figure 4 is a flow diagram of the analyzer algorithm (244) of Figure 2C which is one embodiment of an analyzer algorithm that can be used to assign identity to a query sequence, based on similarity to a subject sequence, where the subject sequence is a reference sequence of the invention.

-16-

Figure 5 is a schematic representation comparing the calicheamicin enediyne biosynthetic locus from *Micromonospora echinospora* subsp. *calichensis* (CALI), the macromomycin (auromomycin) enediyne biosynthetic locus from *Streptomyces macromycetius* (MACR), and a chromoprotein enediyne biosynthetic locus from *Streptomyces ghanaensis* (009C). Open reading frames in each locus are identified by boxes; gray boxes indicate ORFs that are not common to the three enediyne loci, black boxes indicate ORFs that are common to the three enediyne loci and are labeled using a four-letter protein family designation. The scale is in kilobases.

Figure 6 illustrates the 5 genes conserved throughout ten enediyne biosynthetic loci from diverse genera, including both chromoprotein and non-chromoprotein enediyne loci.

Figure 7 is a graphical depiction of the domain architecture typical of enediyne polyketide synthases (PKSE).

Figure 8 is an amino acid clustal alignment of full length enediyne polyketide synthase (PKSE) proteins from ten enediyne biosynthetic loci. Approximate domain boundaries are indicated above the alignment. Conserved residues or motifs important for the function of each domain are highlighted in black.

Figure 9A is an amino acid clustal alignment comparing the acyl carrier protein (ACP) domain of the PKSEs from three known enediynes, macromomycin (MACR), calicheamicin (CALI), and neocarzinostatin (NEOC), and the ACP domain of the actinorhodin Type II PKS system (1AF8; Protein DataBank database, maintained by Research Collaboratory for Structural Bioinformatics (RCSB)). Figure 9B depicts the space-filling side-chains of the conserved residues on the three dimensional structure of the ACP of the actinorhodin Type II PKS system (1AF8).

Figure 10A is an amino acid Clustal™ alignment comparing the 4'-phosphopantetheinyl transferase (PPTE) domain of the PKSEs from three known enediynes, macromomycin (MACR), calicheamicin (CALI), and neocarzinostatin (NEOC), and the 4'-phosphopantetheinyl transferase, Sfp, of *Bacillus subtilis* (sfp). Conserved residues are boxed. The known secondary structure of Sfp is shown below the aligned sequences and the predicted secondary structure of the PPTE domain of the PKSE is shown above the aligned sequences wherein the boxes indicate α -helices and the arrows indicate β -sheets. Figure 10B shows how the conserved residues of the 4'-

-17-

phosphopantetheinyl transferase Sfp co-ordinate a magnesium ion and coenzyme A; corresponding residues in the neocarzinostatin PPTE domain are shown in bold.

Figure 11 is an amino acid clustal alignment of eleven TEBC proteins and 4-hydroxybenzoyl-CoA thioesterase (1BVQ; Protein DataBank database, maintained by Research Collaboratory for Structural Bioinformatics (RCSB)) superimposed with the secondary structure of 1BVQ. Alpha-helices (α) and beta-sheets (β) are depicted by arrows.

Figure 12 is an amino acid clustal alignment of ten UNBL proteins.

Figure 13 is an amino acid clustal alignment of ten UNBV proteins highlighting
10 the putative N-terminal signal sequence that likely targets these proteins for secretion.

Figure 14 is an amino acid clustal alignment of ten UNBU proteins highlighting the putative transmembrane domains 2 to 6 composed of membrane spanning alpha-helices as defined by PSORT Analysis (Nakai, Kenta; PSORT: Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequences, National Institute for Basic Biology (Japan), <http://www.nibb.ac.jp>) as shown in dashed lines labelled 1 to 7 that anchor this family of proteins within the cell membrane.

Figure 15 shows restriction site and functional maps of plasmids pECO1202-CALI-1 and pECO1202-CALI-4 of the invention. The open reading frames of the genes forming an expression cassette according to the invention are shown as arrows
20 pointing in the direction of transcription.

Figure 16 shows restriction site and functional maps of plasmids pECO1202-CALI-5, pECO1202-CALI-2, pECO1202-CALI-3, pECO1202-CALI-6 and pECO1202-CALI-7. The open reading frames of the genes forming the expression cassette according to the invention are shown as arrows pointing in the direction of transcription.

Figure 17 is an immunoblot analysis of His-tagged TEBC protein in total protein extracts from recombinant *S. lividans* TK24 clones harboring the pECO1202-CALI-2 or the pECO1202-CALI-4 expression vector. Lane M provides molecular weight markers; lanes 1 to 6 represent crude extracts of independent transformants of *S. lividans* TK24 harboring pECO1202-CALI-2; lane 7 represents a crude extract of *S. lividans* TK24
30 harboring pECO1202-CALI-4; and lane 8 represents a crude extract of *S. lividans* TK24 harboring pECO1202 (control).

Figure 18 is an immunoblot analysis of His-tagged TEBC protein in fractionated extracts from recombinant *S. lividans* TK24 clones harboring the pECO1202-CALI-2

-17a-

expression vector. Lane M provides molecular weight markers; lanes 1 to 6 represent soluble (S) and pellet (P) protein fractions of independent transformants of *S. lividans*

-18-

TK24 harboring pECO1202-CALI-2; lane C represents protein fractions of *S. lividans* TK24 harboring pECO1202 (control).

DETAILED DESCRIPTION OF THE INVENTION:

The invention provides enediynes related compositions. The compositions can be used to produce enediyne-related compounds. The compositions can also be used to identify enediyne natural products, enediyne genes, enediyne gene clusters and enediyne producing organisms. The invention rests on the surprising discovery that all enediynes, including chromoprotein enediynes and non-chromoprotein enediynes, use
10 a conserved set of genes for formation of the warhead structure.

To provide the compositions and methods of the invention, a sample of the microorganism *Streptomyces macromyceticus* was obtained and the biosynthetic locus for the chromoprotein enediyne macromomycin was identified. The gene cluster was identified as the biosynthetic locus for macromomycin from *Streptomyces macromyceticus* NRRL B-5335 (sometimes referred to herein as MACR), firstly by confirming the sequence encoding the apoprotein associated with the chromoprotein, which sequence is disclosed in Samy TS *et al.*, *J. Biol. Chem* (1983) Jan 10; 258(1) pp.183-91, and secondly using the genome scanning procedure of Canadian Patent CA 2,352,451.

20 A sample of the microorganism *Micromonospora echinospora* subsp. *calichensis* was then obtained and the full biosynthetic locus for the non-chromoprotein enediyne calicheamicin was identified. The gene cluster was identified as the biosynthetic locus for calicheamicin from *Micromonospora echinospora* subsp. *calichensis* NRRL 15839 (sometimes referred to herein as CALI) by comparing the sequence with the partial locus for CALI which was disclosed in WO 00/40596. We were able to overcome the problems encountered in prior attempts to isolate and clone the entire biosynthetic locus by using a shotgun-based approach as described in Canadian Patent CA 2,352,451.

30 We identified two further enediyne natural products biosynthetic loci from organisms not previously reported to produce enediyne compounds, namely a chromoprotein enediyne from *Streptomyces ghanaensis* NRRL B-12104 (sometimes referred to herein as 009C), and a chromoprotein enediyne from *Amycolatopsis orientalis* ATCC 43491 (sometimes referred to herin as 007A). The presence of an

-19-

apoprotein encoding gene in 009C and 007A confirms that 009C and 007A produce chromoprotein enediyne compounds.

Comparison of the MACR, CALI, 009C and 007A loci revealed that all loci contain at least one a member of five (5) protein families. The five protein families are referred to throughout the description and figures by reference to a four-letter designation as indicated Table 1.

Table 1: Protein Family Descriptions

Families	Function
PKSE	unusual polyketide synthase, found only in enediyne biosynthetic loci and involved in warhead formation; believed to act iteratively.
TEBC	thioesterase unique to enediyne biosynthetic loci; significant similarity to small (130-150 aa) proteins of the 4-hydroxybenzoyl-CoA thioesterase family in a number of bacteria.
UNBL	unique to enediyne biosynthetic loci; these proteins are rich in basic amino acids and contain several conserved or invariant histidine residues.
UNBV	unique to enediyne biosynthetic loci; secreted proteins; contain putative cleavable N-terminal signal sequence; believed to be associated with stabilization and/or export of the enediyne chromophore and/or late modifications in the biosynthesis of enediyne chromophores.
UNBU	unique to enediyne biosynthetic loci; C-terminal domain homology to bacterial putative ABC transporters and permease transport systems; integral membrane proteins with seven or eight putative membrane-spanning alpha helices; believed to be involved in transport of enediynes and/or intermediates across the cell membrane.

10 A member of each of the five protein families was found in each of the more than ten biosynthetic loci for chromoprotein and non-chromoprotein enediynes studied. Two of the five protein families, PKSE and TEBC, form a polyketide synthase catalytic complex involved in formation of the warhead structure that distinguishes enediyne compounds. The other three protein families conserved throughout chromoprotein and non-chromoprotein enediyne biosynthetic loci are also associated with the warhead structure that characterizes enediyne compounds. Nucleic acid sequences and polypeptide sequences related to these five protein families form the basis for the compositions and methods of the invention.

20 We have discovered at least one member of each of the protein families PKSE, TEBC, UNBL, UNBV and UNBU in all of the 10 enediyne biosynthetic loci studied, including MACR, CALI, 009C, 007A, an enediyne biosynthetic locus from *Kitasatosporia* sp. (sometimes referred to herein as 028D), an enediyne biosynthetic

-20-

locus from *Micromonospora megalomicea* (sometimes referred to herein as 054A), an enediyne biosynthetic locus from *Saccharothrix aerocolonigenes* (sometimes referred to herein as 132H), an enediyne biosynthetic locus from *Streptomyces kaniharaensis* (sometimes referred to herein as 135E), an enediyne biosynthetic locus from *Streptomyces citricolor* (sometimes referred to herein as 145B), and the biosynthetic locus for the chromoprotein enediyne neocarzinostatin from *Streptomyces carzinostaticus* (sometimes referred to herein as NEOC).

10 The protein families PKSE, TEBC, UNBL, UNBV and UNBU of the present invention are associated with warhead formation in enediyne compounds and are found in both chromoprotein and non-chromoprotein enediyne biosynthetic loci. Members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU found within an enediyne biosynthetic loci are not necessarily present in a single operon and are therefore not necessarily transcriptionally linked to one another. However, the members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU that are found within a single enediyne biosynthetic locus are functionally linked to one another in that they act in a concerted fashion in the production of an enediyne product. Although expression of functionally linked enediyne specific genes encoding members of the PKSE, TEBC, UNBL, UNBV and UNBU protein families may be under control of distinct transcriptional promoters, they may nonetheless be expressed in a concerted fashion.

20 Due to high overall sequence conservation between members of the PKSE, TEBC, UNBL, UNBV and UNBU protein families, it is expected that members of the PKSE, TEBC, UNBL, UNBV and UNBU protein families may be exchanged for another member of the same protein family while retaining the ability of the new enediyne biosynthetic system to synthesize the warhead structure of an enediyne compound. Thus, it is contemplated that genes encoding a polypeptide from protein families PKSE, TEBC, UNBL, UNBV and UNBU from two or more different enediyne biosynthetic systems may be combined so as to obtain a full complement of the five-gene enediyne cassette of the invention, wherein one or more genes in the enediyne cassette has inherent or engineered optimal properties.

30 Representative nucleic acid sequences and polypeptide sequences drawn from each of the ten enediyne loci described herein are provided in the sequence listing accompanying parent application CA 2,387,401 and the sequence listings accompanying divisional applications CA 2,445,687, CA 2,445,692, CA 2,444,812

-21-

and CA 2,444,802. Unless indicated otherwise, references to SEQ ID NOS in the present divisional application CA 2,445,687 refer to the sequence listing accompanying CA 2,445,687 as examples of the compositions of the invention. Referring to the sequence listings, a nucleic acid sequence encoding a member of the PKSE protein family of the invention from the biosynthetic locus for macromomycin from *Streptomyces macromyceticus* (MACR) is provided in SEQ ID NO: 3 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 2 of CA 2,387,401. Nucleic acid sequences encoding two members of the TEBC protein family from MACR are provided in SEQ ID NOS: 3 and 5 with the corresponding deduced polypeptide sequences provided in SEQ ID NOS: 2 and 4 respectively. A nucleic acid sequence encoding a member of the UNBL protein family from MACR is provided in SEQ ID NO: 3 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 2 of CA 2,445,692. A nucleic acid sequence encoding a member of the protein family UNBV from MACR is provided in SEQ ID NO: 3 of CA 2,444,812 with the corresponding deduced polypeptide provided in SEQ ID NO: 2 of CA 2,444,812. A nucleic acid sequence encoding a member of the protein family UNBU from MACR is provided in SEQ ID NO: 3 of CA 2,444,802 with the corresponding deduced polypeptide provided in SEQ ID NO: 2 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the biosynthetic locus for calicheamicin from *Micromonospora echinospora* subsp. *calichensis* (CALI) is provided in SEQ ID NO: 5 of CA 2,387,401, with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 4 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from CALI is provided in SEQ ID NO: 7, with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 6. A nucleic acid sequence encoding a member of the UNBL protein family from CALI is provided in SEQ ID NO: 5

-22-

of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 4 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from CALI is provided in SEQ ID NO: 5 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 4 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from CALI is provided in SEQ ID NO: 5 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 4 of CA 2,444,802.

10 A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Streptomyces ghanaensis* (009C) is provided in SEQ ID NO: 7 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 6 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 009C is provided in SEQ ID NO: 9 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 8. A nucleic acid sequence encoding a member of the UNBL protein family from 009C is provided in SEQ ID NO: 7 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 6 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 009C is provided in SEQ ID NO: 7 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 6 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from 009C is provided in SEQ ID NO: 7 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 6 of CA 2,444,802.

20 A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the biosynthetic locus for neocazinoastatin from *Streptomyces carzinostaticus* subsp. *neocarzinostaticus* (NEOC) is provided in SEQ ID NO: 9 of CA 2,387,401 with the corresponding

-23-

deduced polypeptide sequence provided in SEQ ID NO: 8 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from NEOC is provided in SEQ ID NO: 11 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 10. A nucleic acid sequence encoding a member of the UNBL protein family from NEOC is provided in SEQ ID NO: 9 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 8 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from NEOC is provided in SEQ ID NO: 9 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 8 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from NEOC is provided in SEQ ID NO: 9 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 8 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Amycolatopsis orientalis* (007A) is provided in SEQ ID NO: 11 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 10 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 007A is provided in SEQ ID NO: 13 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 12. A nucleic acid sequence encoding a member of the UNBL protein family from 007A is provided in SEQ ID NO: 11 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 10 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 007A is provided in SEQ ID NO: 11 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 10 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from 007A is provided in SEQ ID NO: 11 of CA 2,444,802

-24-

with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 10 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Kitasatospora sp.* (028D) is provided in SEQ ID NO: 13 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 12 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 028D is provided in SEQ ID NO: 15 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 14. A nucleic acid sequence encoding a member of the UNBL protein
10 family from 028D is provided in SEQ ID NO: 13 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 12 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 028D is provided in SEQ ID NO: 13 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 12 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from 028D is provided in SEQ ID NO: 13 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 12 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Micromonospora megalomicea*
20 (054A) is provided in SEQ ID NO: 15 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 14 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 054A is provided in SEQ ID NO: 17 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 16. A nucleic acid sequence encoding a member of the UNBL protein family from 054A is provided in SEQ ID NO: 15 of CA

-25-

2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 14 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 054A is provided in SEQ ID NO: 15 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 14 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from 054A is provided in SEQ ID NO: 15 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 14 of CA 2,444,802.

10 A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Saccharothrix aerocolonigenes* (132H) is provided in SEQ ID NO: 17 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 16 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 132H is provided in SEQ ID NO: 19 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 18. A nucleic acid sequence encoding a member of the UNBL protein family from 132H is provided in SEQ ID NO: 17 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 16 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 132H is provided in SEQ ID NO: 17 of CA 2,444,812 with the corresponding deduced polypeptide provided in SEQ ID NO: 16 of CA 2,444,812. A nucleic acid sequence encoding a member of
20 the UNBU protein family from 132H is provided in SEQ ID NO: 17 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 16 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Streptomyces kaniharaensis* (135E) is provided in SEQ ID NO: 19 of CA 2,387,401

-26-

with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 18 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 135E is provided in SEQ ID NO: 21, with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 20. A nucleic acid sequence encoding a member of the UNBL protein family from 135E is provided in SEQ ID NO: 19 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 18 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 135E is provided in SEQ ID NO: 19 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 18 of CA 2,444,812. A nucleic acid sequence encoding a member of the UNBU protein family from 135E is provided in SEQ ID NO: 19 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 18 of CA 2,444,802.

A nucleic acid sequence encoding a member of the PKSE protein family of the invention from the enediyne biosynthetic locus from *Streptomyces citricolor* (145B) is provided in SEQ ID NO: 21 of CA 2,387,401 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 20 of CA 2,387,401. A nucleic acid sequence encoding a member of the TEBC protein family from 145B is provided in SEQ ID NO: 23 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 22. A nucleic acid sequence encoding a member of the UNBL protein family from 145B is provided in SEQ ID NO: 21 of CA 2,445,692 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 20 of CA 2,445,692. A nucleic acid sequence encoding a member of the UNBV protein family from 145B is provided in SEQ ID NO: 21 of CA 2,444,812 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 20 of CA 2,444,812. A nucleic acid

-27-

sequence encoding a member of the UNBU protein family from 145B is provided in SEQ ID NO: 21 of CA 2,444,802 with the corresponding deduced polypeptide sequence provided in SEQ ID NO: 20 of CA 2,444,802.

TEBC refers to a family of thioesterase proteins unique to enediyne biosynthesis which together with a protein from the protein family PKSE forms an enediyne polyketide catalytic complex and is involved in synthesis of a warhead structure that characterizes enediyne compounds. The TEBC protein family is defined structurally as a peptide sequence that produces an alignment of at least 49 percent identity to the following TEBC consensus sequence using BLASTP 2.0.11 algorithm with the filter
 10 option -F set to false, the gap opening penalty -G set to 11, the gap extension penalty -E set to 1, and all remaining options set to default values:

```
vtmadYfEYRHtVgFEETNLVGNVYYVNYLRWQGRCRElFLkekAPeVLadlrddLkLFTLkvd
CEFFaEitAfDeLsiRMRLaeltgTQleftFDYvrlggdgvvetLVARGrQRiACMRGPntaTvP
arVPeaLrrALaPYaagtrvlaGrga
```

where the TEBC consensus sequence is based on SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22. It is contemplated that the BLASTP 2.0.11 algorithm may be replaced with newer versions thereof, in which case more recent versions of BLASTP
 20 2.0.11 algorithm may be used with parameters selected to be substantially equivalent to those described above. Representative members of the protein family TEBC include the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22. Other members of protein family TEBC include polypeptides having at least 75%, preferably 80%, more preferably, 85% still more preferably 90% and most preferably 95% or more identity to a polypeptide having the sequence of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 as determined using the BLASTP algorithm with the default parameters and retaining the ability to act in a concerted fashion with a protein from the protein family PKSE during synthesis of a warhead structure in an enediyne compound. Other
 30 members of the protein family TEBC include fragments, analogs and derivatives of the above polypeptides, which fragments, analogs and derivatives have the ability to substitute for another TEBC protein and retain the ability to act in a concerted fashion with a PKSE protein during formation of a warhead structure in an enediyne compound.

The above consensus sequence was generated as follows. First, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22 were aligned with ClustalX 1.81 program using

-28-

default settings. Then a profile hidden Markov model (HMM), *i.e.* a statistical description of a sequence family's consensus was made from the alignment file with the hmmbuild program of the HMMER 2.2 package (Sean Eddy: Sequence analysis using profile hidden Markov models, School of Medicine, Washington University (St-Louis, USA), <http://hmmer.wustl.edu/>) and was calibrated with the calibrate program of the HMMER package, both using the default setting. HMMER is a freely distributable implementation of profile HMM software for protein sequence analysis and is available from the above web-site. Finally, the consensus sequences were generated from the HMM with the hmemit program of the HMMER package using the -C option to predict
 10 a single majority rule consensus sequence from the HMM's probability distribution. Highly conserved amino acid residues ($p \geq 0.5$) are shown in upper case in the consensus sequence, others are shown in lower case.

As used herein, PKSE refers to a family of polyketide synthase proteins that are uniquely associated with enediynes biosynthetic loci and that are involved in synthesis of the warhead structure that characterizes enediyne compounds. The PKSE protein family is defined structurally as a polypeptide sequence that produces an alignment with at least 45% identity to the following consensus sequence using BLASTP 2.0.11 algorithm with the filter option -F set to false, the gap opening penalty -G set to 11, the gap extension penalty -E set to 1, and all remaining options set to default value. PKSE
 20 protein family consensus sequence:

gghgmsmtRIAIVGmAcRYPDAtsPeeLWeNvLAGRRAFRRLPDeRMrl eDYWdAD
 PaAPDRFYArnAAViEGYEFDRiayrvAGSTyRSTDlTHWLALDtAArALADAGFP
 gGeGLPrerTGVVvGNsLTGEFSRANvMRLRWPyVRRvvAAALaeqgWdddrLaaf
 LddleAaYKaPFPaIdEDTLAGGLsNTIAGRICNHFDLkGGGYTVDGACSSSLLSV
 vTAaraLv dGdlDVAVAGGVDLSIDPFvIGFAKTGALAkGEMRVYDrgSNGFWPG
 EGCGMVLMREeDAIAaGrRIYAtiaGWGvSSDGkGGITRPEasGyRLALrRAYrr
 AGFGvetVgLFEGHGTGTAVGDaTELeALseaRraAdPaAepAAiGSIKGNIGHTK
 AAAGVAGLIKAaLAVhhqVlPPatGcvdPHplLtgdSaaLrVlrkAEIWPadaPvR
 AGVsAMGFGGINTHVvldepvgaRRraldrtrrrLaasrQDaELLLLDGadaaeLr
 30 arLtrladlvarLSyAELaDLAatLqreLrglpyRAAVVatSPedAerrLrqLarl
 LesGetellsadgGvFLGratrapRIGfLFPGQGSGrGgggGALrRRFaeadevYr
 raglpaGgDqVaTdVAQPRIVTGSlaGLRVLdaLGieAsvAVGHSLGELtALHWAG
 ALdEdtllrlArvRGrvMAehssggGaMAgLAAtPeaaeaLlaGlpvVvAGYNGPr
 QTVVaGpadaVdeVcrRAaraGVtatrLnVSHAFHSPLVApAAeafaeLasvdFg
 pparrvVSTVTGallpadtDLreLLrrQitaPVRfteAlgaaaadvDLfiEVGPGR
 VLsgLaaeiaPdvPAvalDTDaeSLrpLLavVGAAfvlGApvalerLFedRLiRPL
 pidrefsFLAsPCEqAPeikapavrpapvpvapaadaaaaaaaageapgesaLev
 LrrLaAERaELPvesVdpdsrllDDLHLSSITVGQivNqaaraLGipaaavptnFA
 40 TatlAElaEaLdeLaqtaapgdaslvAGVAPWvRpFaVdldevplPapapaaar
 GrWevFAtadhPlAepLraaLagAgvGdGVLLcLPadCaaehvglaLaAaraALaa

-29-

prgtRlVvVqhgrGAaGLAKTLrLEaPhlrtTVVhlPdpqpldeaaddAVarVvAe
 VAATtgFtEVhYdadGvRrVPvLRpLpvspaeasPLderDVLLVTGGGKGITAEC
 ALAlArdSGAaLALlGRSDPAaDeeLAdNLarMaAAGlrvrYaRADVTdpaqVaaA
 VaeLtaeLGPVTAvLHGAGRNEPaaLasLdeedFRrtlAPKvDGLrAVLaAVdper
 LkLLVTFGSIIGRAGLRGEAHYATANdWLaeLTerfarehPqcRalcLEWSVWSGv
 GMGErLgVVEsLsReGItPIspdeGVevLrrLlaDPdaptvvVVsGRtgGleTlrl
 drreLPLlRFleRplVhYpGVELVtEaeLnaGtDpYLaDHlLDGdLLfPAVlGMEA
 MaQVAaAltGrpgvPviEdveFlRPIvVpPdGsTtiRvAAlvtdpdTVdVVLrSee
 TgFaADHFRARLrytraavpdgtPaqvdddlPaVPLdPatdLYGgvLFQGkRFqRL
 10 rrYrraAARHvdAeVatsapadWFAaflPgelLLADPGtRDAlMHgiQvCVPDATL
 LPsGiERlhlaeaaeqdpeavrldArERSrDGDtYVYDvaVRDadGrvVERWeGLr
 LrAVRkrdGsGPWvpaLLGpYLERsLeevlGssiAVvVePaGddpdgsvaeRRarT
 aeAasRALGaPveVRhRPDGRPEldggrevSasHgaGlTLaVvaagrtvACDvEaV
 aeRtaeeWagLLGerhealaeLLaaEaGEppdvAATRVSAvECLrKaGvraGapL
 tLlpvtpdGWVVLsaGdvRiATfVTavrgatdPVVFAVLtgaer

where the above consensus sequences based on SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14,
 16, 18 and 20 of CA 2,387,401, the parent application of the present divisional
 20 application and was generated as described above. It is contemplated that the
 BLASTP 2.0.11 algorithm may be replaced with newer versions thereof, in which case
 more recent versions of the BLASTP 2.0.11 algorithm may be used with parameters
 selected to be substantially equivalent to those described above. Representative
 members of the protein family PKSE include the polypeptides of SEQ ID NOS: 2, 4, 6,
 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401. Other members of protein family PKSE
 include polypeptides having at least 75%, preferably 80%, more preferably, 85% still
 more preferably 90% and most preferably 95% or more identity to a polypeptide having
 the sequence of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 as
 determined using the BLASTP algorithm with the default parameters and having the
 30 ability to substitute for another PKSE protein and retaining the ability to act in a
 concerted fashion with a TEBC protein during synthesis of a warhead structure of an
 enediyne compound. Other members of the protein family PKSE include fragments,
 analogs and derivatives of the above polypeptides, which fragments, analogs and
 derivatives have the ability to substitute for another PKSE protein and retain the ability
 to act in concert fashion with TEBC during synthesis of a warhead structure of an
 enediyne compound.

UNBL refers to a family of proteins indicative of enediyne biosynthetic loci and
 which are rich in basic amino acids and contain several conserved or invariant histidine
 residues. Representative members of the protein family UNBL include the polypeptides

-30-

of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692. Other members of protein family UNBL include polypeptides having at least 75%, preferably 80%, more preferably, 85% still more preferably 90% and most preferably 95% or more identity to a polypeptide having the sequence of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 as determined using the BLASTP algorithm with the default parameters and that are present in a gene cluster associated with the biosynthesis of an enediyne compound. Other members of the protein family UNBL include fragments, analogs and derivatives of the above polypeptides, which fragments, analogs and derivatives have the ability to substitute for another UNBL protein and retain the ability to act in a

10 concerted fashion with genes in an enediyne biosynthetic locus to form a warhead structure of an enediyne compound.

UNBV refers to a family of proteins indicative of enediyne biosynthetic loci and which may contain a cleavable N-terminal signal sequence. Representative members of the protein family UNBV include the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812. Other members of protein family UNBV include polypeptides having at least 75%, preferably 80%, more preferably, 85% still more preferably 90% and most preferably 95% or more identity to a polypeptide having the sequence of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 as determined using the BLASTP algorithm with the default parameters and that are

20 present in a gene cluster associated with the biosynthesis of an enediyne compound. Other members of the protein family UNBV include fragments, analogs and derivatives of the above polypeptides, which fragments, analogs and derivatives have the ability to substitute for another UNBV protein and retain the ability to act in a concerted fashion with genes in an enediyne biosynthetic locus to form a warhead structure in an enediyne compound.

UNBU refers to a family of membrane proteins indicative of enediyne biosynthetic loci. Representative members of the protein family UNBU include the polypeptides of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802. Other members of protein family UNBU include polypeptides having at least 75%, preferably

30 80%, more preferably, 85% still more preferably 90% and most preferably 95% or more identity to

-31-

a polypeptide having the sequence of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 as determined using the BLASTP algorithm with the default parameters and that are present in a gene cluster associated with the biosynthesis of an enediyne compound. Other members of the protein family UNBU include fragments, analogs and derivatives of the above polypeptides, which fragments, analogs and derivatives have the ability to substitute for another UNBU protein and retain the ability to act in a concerted fashion with genes in an enediyne biosynthetic locus to form the warhead structure in an enediyne compound.

“Enediyne producer” or “enediyne-producing organism” refers to a
10 microorganism which carries the genetic information necessary to produce an enediyne compound, whether or not the organism is known to produce an enediyne product. The terms apply equally to organisms in which the genetic information to produce an enediyne compound is found in the organism as it exists in its natural environment, and to organisms in which the genetic information is introduced by recombinant techniques. For the sake of particularity, specific organisms contemplated herein include organisms of the family *Micromonosporaceae*, of which preferred genera include *Micromonospora*, *Actinoplanes* and *Dactylosporangium*; the family *Streptomycetaceae*, of which preferred genera include *Streptomyces* and *Kitasatospora*; the family
20 *Pseudonocardiaceae*, of which preferred genera are *Amycolatopsis* and
Saccharopolyspora; and the family *Actinosynnemataceae*, of which preferred genera include *Saccharothrix* and *Actinosynnema*; however the terms are intended to encompass all organisms containing genetic information necessary to produce an enediyne compound.

“Enediyne biosynthetic gene product” refers to any enzyme involved in the biosynthesis of an enediyne, whether a chromoprotein enediyne or a non-chromoprotein enediyne. These gene products are located in any enediyne biosynthetic locus in an organism of the family *Micromonosporaceae*, of which preferred genera include *Micromonospora*, *Actinoplanes* and *Dactylosporangium*; the family *Streptomycetaceae*, of which preferred genera include *Streptomyces* and
30 *Kitasatospora*; the family *Pseudonocardiaceae*, of which preferred genera are *Amycolatopsis* and *Saccharopolyspora*. For the sake of particularity, the enediyne biosynthetic loci described herein are associated with *Streptomyces macromyceticus*,

-32-

Micromonospora echinospora subsp. *calichensis*, *Streptomyces ghanaensis*, *Streptomyces carzinostaticus* subsp. *neocarzinostaticus*, *Amycolatopsis orientalis*, *Kitasatospora* sp., *Micromonospora megalomicea*, *Saccharothrix aerocolonigenes*, *Streptomyces kaniharaensis*, and *Streptomyces citricolor*; however, it should be understood that this term encompasses enediyne biosynthetic enzymes (and genes encoding such enzymes) isolated from any microorganism of the genus *Streptomyces*, *Micromonospora*, *Amycolatopsis*, *Kitasatospora*, or *Saccharithrix* and furthermore that these genes may have novel homologues in any microorganism, actinomycete or non-actinomycete, that falls within the scope of the present invention. Specific

10 embodiments include the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 of the present application, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802.

The term "isolated" means that the material is removed from its original environment, e.g. the natural environment if it is naturally occurring. For example, a naturally-occurring polynucleotide or polypeptide present in a living organism is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be

20 part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

The term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual nucleic acids obtained from a library have been conventionally purified to electrophoretic homogeneity. The purified nucleic acids of the present invention have been purified from the remainder of the genomic DNA in the organism by at least 10^4 to 10^6 fold. However, the term "purified" also includes nucleic acids which have been purified from the remainder of the genomic DNA or from other sequences in a library or other environment by at least one order of magnitude,

30 preferably two or three orders of magnitude, and more preferably four or five orders of magnitude.

"Recombinant" means that the nucleic acid is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. "Enriched" nucleic acids

-33-

represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. "Backbone" molecules include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid of interest. Preferably, the enriched nucleic acids represent 15% or more, more preferably 50% or more, and most preferably 90% or more, of the number of nucleic acid inserts in the population of recombinant backbone molecules.

"Recombinant polypeptides" or "recombinant proteins" refers to polypeptides or proteins produced by recombinant DNA techniques, *i.e.* produced from cells transformed by an exogenous DNA construct encoding the desired polypeptide or protein. "Synthetic" polypeptides or proteins are those prepared by chemical synthesis.

The term "gene" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as, where applicable, intervening regions (introns) between individual coding segments (exons).

The term "operon" means a transcriptional gene cassette under the control of a single transcriptional promoter, which gene cassette encodes polypeptides that may act in a concerted fashion to carry out a biochemical pathway and/or cellular process.

A DNA or nucleotide "coding sequence" or "sequence encoding" a particular polypeptide or protein, is a DNA sequence which is transcribed and translated into a polypeptide or protein when placed under the control of appropriate regulatory sequences.

"Oligonucleotide" refers to a nucleic acid, generally of at least 10, preferably 15 and more preferably at least 20 nucleotides, preferably no more than 100 nucleotides, that are hybridizable to a genomic DNA molecule, a cDNA molecule, or an mRNA molecule encoding a gene, mRNA, cDNA or other nucleic acid of interest.

A promoter sequence is "operably linked to" a coding sequence recognized by RNA polymerase which initiates transcription at the promoter and transcribes the coding sequence into mRNA.

"Plasmids" are designated herein by a lower case p followed by capital letters and/or numbers. The starting plasmids herein are commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in

-34-

accord with published procedures. In addition, equivalent plasmids to those described herein are known in the art and will be apparent to the skilled artisan.

“Digestion” of DNA refers to enzymatic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinary skilled artisan. For analytical purposes, typically 1 µg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 µl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 µg of DNA are digested with
10 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier’s instructions. After digestion, gel electrophoresis may be performed to isolate the desired fragment.

Two deposits have been made with the International Depositary Authority of Canada, Bureau of Microbiology, Health Canada, 1015 Arlington Street, Winnipeg, Manitoba, Canada R3E 3R2 on April 3, 2002. The first deposit is an *E. coli* DH10B strain harbouring a cosmid clone (020CN) of a partial biosynthetic locus for macromomycin from *Streptomyces macromyceticus*, including open reading frames
20 coding for the polypeptides of SEQ ID NOS: 2 and 4 of the present application, SEQ ID NO: 2 of CA 2,387,401, SEQ ID NO: 2 of CA 2,445,692, SEQ ID NO: 2 of CA 2,444,812, and SEQ ID NO: 2 of CA 2,444,802, which deposit was assigned deposit accession number IDAC030402-1. The second deposit is an *E. coli* DH10B strain harbouring a cosmid clone (061CR) of a partial biosynthetic locus for calicheamicin from *Micromonospora echinospora* subsp. *calichensis*, including open reading frames coding for the polypeptides of SEQ ID NO: 6 of the present application, SEQ ID NO: 4 of CA 2,387,401, SEQ ID NO: 4 of CA 2,445,692, SEQ ID NO: 4 of CA 2,444,812, and SEQ ID NO: 4 of CA 2,444,802, which deposit was assigned accession number IDAC 030402-2. The *E. coli* strain deposits are referred to herein as “the deposited strains”.

30 The deposited strains comprise a member from each of the protein families PKSE, TEBC, UNBL, UNBV and UNBU drawn from a chromoprotein enediyne biosynthetic locus (macromomycin) and a member from each of the protein families PKSE, TEBC, UNBL, UNBV and UNBU drawn from a non-chromoprotein enediyne

-35-

biosynthetic locus (calicheamicin). The sequence of the polynucleotides comprised in the deposited strains, as well as the amino acid sequence of any polypeptide encoded thereby are controlling in the event of any conflict with any description of sequences herein.

The deposit of the deposited strains has been made under the terms of the Budapest Treaty on the International Recognition of the Deposit of Micro-organisms for Purposes of Patent Procedure. The deposited strains will be irrevocably and without restriction or condition released to the public upon the issuance of a patent. The deposited strains are provided merely as convenience to those skilled in the art and are not an admission that a deposit is required for enablement. A license may be required to make, use or sell the deposited strains or nucleic acids therein, and compounds derived therefrom, and no such license is hereby granted.

Representative nucleic acid sequences encoding members of protein family TEBC are provided in SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23.

Representative nucleic acid sequences encoding members of protein family PKSE are provided in SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 of CA 2,387,401.

Representative nucleic acid sequences encoding members of protein family UNBL are provided in SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 of CA 2,445,692.

Representative nucleic acid sequences encoding members of protein family UNBV are provided in SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 of CA 2,444,812.

Representative nucleic acid sequences encoding members of protein family UNBU are provided in SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 of CA 2,444,802.

Representative polypeptides of protein family TEBC are provided in SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22. Representative polypeptides of protein family PKSE are provided in SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 of CA 2,387,401. Representative polypeptides of protein family UNBL are provided in SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 of CA 2,445,692. Representative polypeptides of protein family UNBV are provided in SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 of CA 2,444,812. Representative polypeptides of protein family UNBU are provided in SEQ ID NOS: 2, 4,

-36-

6, 8, 10, 12, 14, 16, 18 and 20 of CA 2,444,802.

One aspect of the present divisional application is an isolated, purified, or enriched nucleic acid comprising one of the sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive bases of one of the sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 or the sequences complementary thereto. The isolated, purified or enriched nucleic acids may comprise DNA, including cDNA, genomic DNA, and synthetic DNA. The DNA may be double stranded or single stranded, and if single stranded may be the coding or
10 non-coding (anti-sense) strand. Alternatively, the isolated, purified or enriched nucleic acids may comprise RNA.

As discussed in more detail below, the isolated, purified or enriched nucleic acids of one of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 may be used to prepare one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100 or 100 consecutive amino acids of one of the polypeptides of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22.

Accordingly, another aspect of the present application is an isolated, purified or enriched nucleic acid which encodes one of the polypeptides of SEQ ID NOS: 2, 4, 6,
20 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100 or 150 consecutive amino acids of one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22. The coding sequences of these nucleic acids may be identical to one of the coding sequences of one of the nucleic acids of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, or a fragment thereof or may be different coding sequences which encode one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100 or 150 consecutive amino acids of one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 as a result of the redundancy or degeneracy of the genetic code. The genetic code is well known to those of skill in the
30 art and can be obtained, for example, from Stryer, *Biochemistry*, 3rd edition, W. H. Freeman & Co., New York.

-37-

The isolated, purified or enriched nucleic acid which encodes one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, may include, but is not limited to: (1) only the coding sequences of one of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23; (2) the coding sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 and additional coding sequences, such as leader sequences or proprotein sequences; or (3) the coding sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 and non-coding sequences, such as introns or non-coding sequences 5' and/or 3' of the coding sequence. Thus, as used herein, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

The invention of the present divisional application relates to polynucleotides based on SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 but having polynucleotide changes that are "silent", for example changes which do not alter the amino acid sequence encoded by the polynucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23. The invention of the present divisional application also relates to polynucleotides which have nucleotide changes which result in amino acid substitutions, additions, deletions, fusions and truncations of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22. Such nucleotide changes may be introduced using techniques such as site directed mutagenesis, random chemical mutagenesis, exonuclease III deletion, and other recombinant DNA techniques.

The isolated, purified or enriched nucleic acids of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive bases of one of the sequence of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, or the sequences complementary thereto may be used as probes to identify and isolate DNAs encoding the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 respectively.

For example, a genomic DNA library may be constructed from a sample microorganism or a sample containing a microorganism capable of producing an enediynes. The genomic DNA library is then contacted with a probe comprising a coding sequence or a fragment of the coding sequence, encoding one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or a fragment

-38-

thereof under conditions which permit the probe to specifically hybridize to sequences complementary thereto. In one embodiment, the probe is an oligonucleotide of about 10 to about 30 nucleotides in length designed based on a nucleic acid of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23. Genomic DNA clones which hybridize to the probe are then detected and isolated. Procedures for preparing and identifying DNA clones of interest are disclosed in Ausubel *et al.*, Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997; and Sambrook *et al.*, Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989. In another embodiment, the probe is a restriction fragments or a PCR amplified nucleic acid derived from SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23.

The isolated, purified or enriched nucleic acids of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive bases of one of the sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, or the sequences complementary thereto may be used as probes to identify and isolate related nucleic acids. In some embodiments, the related nucleic acids may be genomic DNAs (or cDNAs) from potential enediyne producers. In one embodiment of the present divisional application, isolated, purified or enriched nucleic acids of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive bases of one of the sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 or the sequences complementary thereto may be used as probes to identify and isolate related nucleic acids. In such procedures, a nucleic acid sample containing nucleic acids from a potential enediyne-producer is contacted with the probe under conditions which permit the probe to specifically hybridize to related sequences. The nucleic acid sample may be a genomic DNA (or cDNA) library from the potential enediyne-producer. Hybridization of the probe to nucleic acids is then detected using any of the methods known in the art, including those referred to herein.

Hybridization may be carried out under conditions of low stringency, moderate stringency or high stringency. As an example of nucleic acid hybridization, a polymer membrane containing immobilized denatured nucleic acids is first prehybridized for 30 minutes at 45 °C in a solution consisting of 0.9 M NaCl, 50 mM NaH₂PO₄, pH 7.0, 5.0 mM Na₂EDTA, 0.5% SDS, 10X Denhardt's, and 0.5 mg/ml polyriboadenylic acid.

-39-

Approximately 2×10^7 cpm (specific activity $4-9 \times 10^8$ cpm/ μ g) of ^{32}P end-labeled oligonucleotide probe are then added to the solution. After 12-16 hours of incubation, the membrane is washed for 30 minutes at room temperature in 1X SET (150 mM NaCl, 20 mM Tris hydrochloride, pH 7.8, 1 mM Na_2EDTA) containing 0.5% SDS, followed by a 30 minute wash in fresh 1X SET at $T_m-10^\circ\text{C}$ for the oligonucleotide probe where T_m is the melting temperature. The membrane is then exposed to autoradiographic film for detection of hybridization signals.

By varying the stringency of the hybridization conditions used to identify nucleic acids, such as genomic DNAs or cDNAs, which hybridize to the detectable probe, nucleic acids having different levels of homology to the probe can be identified and isolated. Stringency may be varied by conducting the hybridization at varying temperatures below the melting temperatures of the probes. The melting temperature of the probe may be calculated using the following formulas:

For oligonucleotide probes between 14 and 70 nucleotides in length the melting temperature (T_m) in degrees Celcius may be calculated using the formula:
 $T_m = 81.5 + 16.6(\log [\text{Na}^+]) + 0.41(\text{fraction G+C}) - (600/N)$ where N is the length of the oligonucleotide.

If the hybridization is carried out in a solution containing formamide, the melting temperature may be calculated using the equation $T_m = 81.5 + 16.6(\log [\text{Na}^+]) + 0.41(\text{fraction G + C}) - (0.63\% \text{ formamide}) - (600/N)$ where N is the length of the probe.

Prehybridization may be carried out in 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 0.1 mg/ml denatured fragmented salmon sperm DNA or 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 0.1 mg/ml denatured fragmented salmon sperm DNA, 50% formamide. The composition of the SSC and Denhardt's solutions are listed in Sambrook et al., *supra*.

Hybridization is conducted by adding the detectable probe to the hybridization solutions listed above. Where the probe comprises double stranded DNA, it is denatured by incubating at elevated temperatures and quickly cooling before addition to the hybridization solution. It may also be desirable to similarly denature single stranded probes to eliminate or diminish formation of secondary structures or oligomerization. The filter is contacted with the hybridization solution for a sufficient period of time to allow the probe to hybridize to cDNAs or genomic DNAs containing sequences complementary thereto or homologous thereto. For probes over 200

-40-

nucleotides in length, the hybridization may be carried out at 15-25 °C below the T_m. For shorter probes, such as oligonucleotide probes, the hybridization may be conducted at 5-10 °C below the T_m. Preferably, the hybridization is conducted in 6X SSC, for shorter probes. Preferably, the hybridization is conducted in 50% formamide containing solutions, for longer probes.

All the foregoing hybridizations would be considered to be examples of hybridization performed under conditions of high stringency.

Following hybridization, the filter is washed for at least 15 minutes in 2X SSC, 0.1% SDS at room temperature or higher, depending on the desired stringency. The filter is then washed with 0.1X SSC, 0.5% SDS at room temperature (again) for 30 minutes to 1 hour.

Nucleic acids which have hybridized to the probe are identified by autoradiography or other conventional techniques.

The above procedure may be modified to identify nucleic acids having decreasing levels of homology to the probe sequence. For example, to obtain nucleic acids of decreasing homology to the detectable probe, less stringent conditions may be used. For example, the hybridization temperature may be decreased in increments of 5 °C from 68 °C to 42 °C in a hybridization buffer having a Na⁺ concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate stringency" conditions above 50°C and "low stringency" conditions below 50°C. A specific example of "moderate stringency" hybridization conditions is when the above hybridization is conducted at 55°C. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 45°C.

Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42 °C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of homology to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50 °C. These conditions are considered to be "moderate stringency" conditions above 25% formamide and "low stringency" conditions below 25% formamide. A specific example of "moderate stringency" hybridization conditions is when the above hybridization is

-41-

conducted at 30% formamide. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 10% formamide.

Nucleic acids which have hybridized to the probe are identified by autoradiography or other conventional techniques.

For example, the preceding methods may be used to isolate nucleic acids having a sequence with at least 97%, at least 95%, at least 90%, at least 85%, at least 80%, or at least 70% identity to a nucleic acid sequence selected from the group consisting of the sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, fragments comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 10 400, or 500 consecutive bases thereof, and the sequences complementary thereto. Identity may be measured using BLASTN version 2.0 with the default parameters. For example, the homologous polynucleotides may have a coding sequence which is a naturally occurring allelic variant of one of the coding sequences described herein. Such allelic variant may have a substitution, deletion or addition of one or more nucleotides when compared to the nucleic acids of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, or the sequences complementary thereto.

Additionally, the above procedures may be used to isolate nucleic acids which encode polypeptides having at least 99%, 95%, at least 90%, at least 85%, at least 80%, or at least 70% identity to a polypeptide having the sequence of one of SEQ ID 20 NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof as determined using the BLASTP version 2.2.2 algorithm with default parameters.

Structural features common to the biosynthesis of all enediynes compounds require one or more proteins selected from a group of five specific protein families, namely PKSE, TEBC, UNBL, UNBV and UNBU. Thus, a polypeptide representing a member of any one of these five protein families or a polynucleotide encoding a polypeptide representing a member of any one of these five protein families is considered indicative of an enediyne gene cluster, a enediyne natural product or an enediyne producing organism. It is not necessary that a member of each of the five 30 protein families considered indicative of an enediyne compound be detected to identify an enediyne biosynthetic locus and an enediyne-producing organism. Rather, the presence of at least one, preferably two, more preferably three, still more preferably four, and most preferably five of the protein families PKSE, TEBC, UNBV and UNBU

-42-

indicates the presence of an enediyne natural product, an enediyne biosynthetic locus or an enediyne producing organism.

To identify an enediyne natural product, an enediyne gene cluster or an enediyne-producing organism, nucleic acids from cultivated microorganisms or from an environmental sample, e.g. soil, potentially harboring an organism having the genetic capacity to produce an enediyne compound may be contacted with a probe based on nucleotide sequences coding a member of the five protein families PKSE, TEBC, UNBL, UNBV and UNBU.

10 In such procedures, nucleic acids are obtained from cultivated microorganisms or from an environmental sample potentially harboring an organism having the genetic capacity to produce an enediyne compound. The nucleic acids are contacted with probes designed based on the teachings and compositions of the invention under conditions which permit the probe to specifically hybridize to any complementary sequences indicative of the presence of a member of the PKSE, TEBC, UNBL, UNBV and UNBU protein families of the invention. The presence of at least one, preferably two, more preferably three, still more preferably four or five of the PKSE, TEBC, UNBL, UNBV and UNBU protein families indicates the presence of an enediyne gene cluster or an enediyne producing organism.

20 Diagnostic nucleic acid sequences encoding members of the PKSE, TEBC, UNBL, UNBV and UNBU protein families for identifying enediyne genes, biosynthetic loci, and microorganisms that harbor such genes or gene clusters may be employed on complex mixtures of microorganisms such as those from environmental samples (e.g., soil). A mixture of microorganisms refers to a heterogeneous population of microorganisms consisting of more than one species or strain. In the absence of amplification outside of its natural habitat, such a mixture of microorganisms is said to be uncultured. A cultured mixture of microorganisms may be obtained by amplification or propagation outside of its natural habitat by *in vitro* culture using various growth media that provide essential nutrients. However, depending on the growth medium used, the amplification may preferentially result in amplification of a sub-population of
30 the mixture and hence may not be always desirable. If desired, a pure culture representing a single species or strain may be obtained from either a cultured or uncultured mixture of microorganisms by established microbiological techniques such

-43-

as serial dilution followed by growth on solid media so as to isolate individual colony forming units.

Enediynes biosynthetic genes and/or enediynes biosynthetic gene clusters may be identified from either a pure culture or cultured or uncultured mixtures of microorganisms employing the diagnostic nucleic acid sequences disclosed in this invention by experimental techniques such as PCR, hybridization, or shotgun sequencing followed by bioinformatic analysis of the sequence data. The identification of one or more members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU or enediynes gene clusters including one or more members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU in a pure culture of a single organism directly distinguishes such an enediynes-producer. The identification of one or more members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU or enediynes gene clusters including one or more members of the protein families PKSE, TEBC, UNBL, UNBV and UNBU in a cultured or uncultured mixture of microorganisms requires further steps to identify and isolate the microorganism(s) that harbor(s) them so as to obtain pure cultures of such microorganisms.

By way of example, the colony lift technique (Ausubel *et al.*, Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997; and Sambrook *et al.*, Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989) may be used to to identify microorganisms that harbour enediynes genes and/or enediynes biosynthetic loci from a cultured mixture of microorganisms. In such a procedure, the mixture of microorganisms is grown on an appropriate solid medium. The resulting colony forming units are replicated on a solid matrix such as a nylon membrane. The membrane is contacted with detectable diagnostic nucleic acid sequences, the positive colony forming units are identified, and the corresponding colony forming units on the original medium are identified, purified, and amplified.

Nucleic acids encoding a member of the protein families PKSE, TEBC, UNBL, UNBV and UNBU may be used to survey a number of environmental samples for the presence of organisms that have the potential to produce enediynes compounds, *i.e.*, those organisms that contain enediynes biosynthetic genes and/or an enediynes biosynthetic locus. One protocol for use of a survey to identify polypeptides encoded by DNA isolated from uncultured mixtures of microorganisms is outlined in Seow *et al.* (1997) J. Bacteriol. Vol. 179 pp. 7360-7368.

-44-

Where necessary, conditions which permit the probe to specifically hybridize to complementary sequences from an enediyne-producer may be determined by placing a probe based on a member of the protein families PKSE , TEBC, UNBL, UNBV and UNBU in contact with complementary sequences obtained from an enediyne-producer as well as control sequences which are not from an enediyne-producer. In some analyses, the control sequences may be from organisms related to enediyne-producers. Alternatively, the control sequences are not related to enediyne-producers. Hybridization conditions, such as the salt concentration of the hybridization buffer, the formamide concentration of the hybridization buffer, or the hybridization temperature, may be varied to identify conditions which allow the probe to hybridize specifically to nucleic acids from enediyne-producers.

If the sample contains nucleic acids from enediyne-producers, specific hybridization of the probe to the nucleic acids from the enediyne-producer is then detected. Hybridization may be detected by labeling the probe with a detectable agent such as a radioactive isotope, a fluorescent dye or an enzyme capable of catalyzing the formation of a detectable product. Many methods for using the labeled probes to detect the presence of nucleic acids in a sample are familiar to those skilled in the art. These include Southern Blots, Northern Blots, colony hybridization procedures, and dot blots.

Another aspect of the present divisional application is an isolated or purified polypeptide comprising the sequence of one of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. As discussed above, such polypeptides may be obtained by inserting a nucleic acid encoding the polypeptide into a vector such that the coding sequence is operably linked to a sequence capable of driving the expression of the encoded polypeptide in a suitable host cell. For example, the expression vector may comprise a promoter, a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for modulating expression levels, an origin of replication and a selectable marker.

Promoters suitable for expressing the polypeptide or fragment thereof in bacteria include the *E.coli lac* or *trp* promoters, the *lacI* promoter, the *lacZ* promoter, the T3 promoter, the T7 promoter, the *gpt* promoter, the lambda P_R promoter, the lambda P_L promoter, promoters from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), and the acid phosphatase promoter. Fungal

-45-

promoters include the α factor promoter. Eukaryotic promoters include the CMV immediate early promoter, the HSV thymidine kinase promoter, heat shock promoters, the early and late SV40 promoter, LTRs from retroviruses, and the mouse metallothionein-I promoter. Other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses may also be used.

Mammalian expression vectors may also comprise an origin of replication, any necessary ribosome binding sites, a polyadenylation site, splice donors and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. In some embodiments, DNA sequences derived from the SV40 splice and
10 polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Vectors for expressing the polypeptide or fragment thereof in eukaryotic cells may also contain enhancers to increase expression levels. Enhancers are cis-acting elements of DNA, usually from about 10 to about 300 bp in length that act on a promoter to increase its transcription. Examples include the SV40 enhancer on the late side of the replication origin bp 100 to 270, the cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and the adenovirus enhancers.

In addition, the expression vectors preferably contain one or more selectable
20 marker genes to permit selection of host cells containing the vector. Examples of selectable markers that may be used include genes encoding dihydrofolate reductase or genes conferring neomycin resistance for eukaryotic cell culture, genes conferring tetracycline or ampicillin resistance in *E. coli*, and the *S. cerevisiae* TRP1 gene.

In some embodiments of the present divisional application, the nucleic acid encoding one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is assembled in appropriate phase with a leader sequence capable of directing secretion of the translated polypeptides or fragments thereof. Optionally, the nucleic acid can encode a fusion polypeptide in which one of
30 the polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is fused to heterologous peptides or polypeptides, such as N-terminal

-46-

identification peptides which impart desired characteristics such as increased stability or simplified purification or detection.

The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA sequence is ligated to the desired position in the vector following digestion of the insert and the vector with appropriate restriction endonucleases. Alternatively, appropriate restriction enzyme sites can be engineered into a DNA sequence by PCR. A variety of cloning techniques are disclosed in Ausbel *et al.* Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook *et al.*, Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbour Laboratory Press, 1989. Such procedures and others are deemed to be within the scope of those skilled in the art.

The vector may be, for example, in the form of a plasmid, a viral particle, or a phage. Other vectors include derivatives of chromosomal, nonchromosomal and synthetic DNA sequences, viruses, bacterial plasmids, phage DNA, baculovirus, yeast plasmids, vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies. A variety of cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989).

Particular bacterial vectors which may be used include the commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017), pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden), GEM1 (Promega Biotec, Madison, WI, USA) pQE70, pQE60, pQE-9 (Qiagen), pD10, psiX174 pBluescript™ II KS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene), ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia), pKK232-8 and pCM7. Particular eukaryotic vectors include pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). However, any other vector may be used as long as it is replicable and stable in the host cell.

The host cell may be any of the host cells familiar to those skilled in the art, including prokaryotic cells or eukaryotic cells. As representative examples of appropriate hosts, there may be mentioned: bacteria cells, such as *E. coli*, *Streptomyces lividans*, *Bacillus subtilis*, *Salmonella typhimurium* and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*, fungal cells, such

-47-

as yeast, insect cells such as *Drosophila* S2 and *Spodoptera* Sf9, animal cells such as CHO, COS or Bowes melanoma, and adenoviruses. The selection of an appropriate host is within the abilities of those skilled in the art.

The vector may be introduced into the host cells using any of a variety of techniques, including electroporation, transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer. Where appropriate, the engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter may be induced by appropriate means (e.g., temperature shift or chemical induction) and the cells may be cultured for an additional period to allow them to produce the desired polypeptide or fragment thereof.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract is retained for further purification. Microbial cells employed for expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known to those skilled in the art. The expressed polypeptide or fragment thereof can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the polypeptide. If desired, high performance liquid chromatography (HPLC) can be employed for final purification steps.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts (described by Gluzman, Cell, 23:175(1981), and other cell lines capable of expressing proteins from a compatible vector, such as the C127, 3T3, CHO, HeLa and BHK cell lines.

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Depending upon the host

-48-

employed in a recombinant production procedure, the polypeptide produced by host cells containing the vector may be glycosylated or may be non-glycosylated.

Polypeptides of the invention may or may not also include an initial methionine amino acid residue.

Alternatively, the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be synthetically produced by conventional peptide synthesizers. In other embodiments, fragments or portions of the polynucleotides may be employed for producing the corresponding full-length polypeptide by peptide
10 synthesis; therefore, the fragments may be employed as intermediates for producing the full-length polypeptides.

Cell-free translation systems can also be employed to produce one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof using mRNAs transcribed from a DNA construct comprising a promoter operably linked to a nucleic acid encoding the polypeptide or fragment thereof. In some embodiments, the DNA construct may be linearized prior to conducting an *in vitro* transcription reaction. The transcribed mRNA is then incubated with an appropriate cell-free translation extract, such as a rabbit reticulocyte extract, to produce the desired
20 polypeptide or fragment thereof.

The present divisional application also relates to variants of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. The term "variant" includes derivatives or analogs of these polypeptides. In particular, the variants may differ in amino acid sequence from the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, by one or more substitutions, additions, deletions, fusions and truncations, which may be present in any combination.

The variants may be naturally occurring or created *in vitro*. In particular, such variants may be created using genetic engineering techniques such as site directed
30 mutagenesis, random chemical mutagenesis, Exonuclease III deletion procedures, and standard cloning techniques. Alternatively, such variants, fragments, analogs, or derivatives may be created using chemical synthesis or modification procedures.

-49-

Other methods of making variants are also familiar to those skilled in the art. These include procedures in which nucleic acid sequences obtained from natural isolates are modified to generate nucleic acids which encode polypeptides having characteristics which enhance their value in industrial or laboratory applications. In such procedures, a large number of variant sequences having one or more nucleotide differences with respect to the sequence obtained from the natural isolate are generated and characterized. Preferably, these nucleotide differences result in amino acid changes with respect to the polypeptides encoded by the nucleic acids from the natural isolates.

10 For example, variants may be created using error prone PCR. In error prone PCR, DNA amplification is performed under conditions where the fidelity of the DNA polymerase is low, such that a high rate of point mutation is obtained along the entire length of the PCR product. Error prone PCR is described in Leung, D.W., *et al.*, Technique, 1:11-15 (1989) and Caldwell, R. C. & Joyce G.F., PCR Methods Applic., 2:28-33 (1992). Variants may also be created using site directed mutagenesis to generate site-specific mutations in any cloned DNA segment of interest. Oligonucleotide mutagenesis is described in Reidhaar-Olson, J.F. and Sauer, R.T., Science, 241:53-57 (1988). Variants may also be created using directed evolution strategies such as those described in US Patents Nos. 6,361,974 and 6,372,497.

20 The variants of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, may be variants in which one or more of the amino acid residues of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code.

Conservative substitutions are those that substitute a given amino acid in a polypeptide by another amino acid of like characteristics. Typically seen as conservative substitutions are the following replacements: replacements of an aliphatic amino acid such as Ala, Val, Leu and Ile with another aliphatic amino acid; replacement
30 of a Ser with a Thr or vice versa; replacement of an acidic residue such as Asp or Glu with another acidic residue; replacement of a residue bearing an amide group, such as Asn or Gln, with another residue bearing an amide group; exchange of a basic residue

-50-

such as Lys or Arg with another basic residue; and replacement of an aromatic residue such as Phe or Tyr with another aromatic residue.

Other variants are those in which one or more of the amino acid residues of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 includes a substituent group.

Still other variants are those in which the polypeptide is associated with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol).

Additional variants are those in which additional amino acids are fused to the polypeptide, such as leader sequence, a secretory sequence, a proprotein sequence or a sequence which facilitates purification, enrichment, or stabilization of the polypeptide.

In some embodiments of this divisional application, the fragments, derivatives and analogs retain the same biological function or activity as the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22. In other embodiments, the fragment, derivative or analogue includes a fused heterologous sequence which facilitates purification, enrichment, detection, stabilization or secretion of the polypeptide that can be enzymatically cleaved, in whole or in part, away from the fragment, derivative or analogue.

Another aspect of the present divisional application are polypeptides or fragments thereof which have at least 70%, at least 80%, at least 85%, at least 90%, or more than 95% identity to one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. Identity may be determined using a program, such as BLASTP version 2.2.2 with the default parameters, which aligns the polypeptides or fragments being compared and determines the extent of amino acid identity or similarity between them. It will be appreciated that amino acid "homology" includes conservative substitutions such as those described above.

The polypeptides or fragments having homology to one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof may be obtained by isolating the nucleic acids encoding them using the techniques described above.

-51-

Alternatively, the homologous polypeptides or fragments may be obtained through biochemical enrichment or purification procedures. The sequence of potentially homologous polypeptides or fragments may be determined by proteolytic digestion, gel electrophoresis and/or microsequencing. The sequence of the prospective homologous polypeptide or fragment can be compared to one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof using a program such as BLASTP version 2.2.2 with the default parameters.

10 The polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or fragments, derivatives or analogs thereof comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof invention may be used in a variety of application. For example, the polypeptides or fragments, derivatives or analogs thereof may be used to biocatalyze biochemical reactions. In particular, the polypeptides of the PKSE family, namely SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 fragments, derivatives or analogs thereof; the TEBC family, namely SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments, derivatives or analogs thereof, may be used in any combination, *in vitro* or *in vivo*, to direct the synthesis or modification of an enediyne warhead or a substructure thereof. Polypeptides of the
20 UNBL family, namely 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692 or fragments, derivatives or analogs thereof; may be used *in vitro* or *in vivo* to direct or aid the synthesis or modification of an enediyne warhead or a substructure thereof. Polypeptides of the UNBV family, namely 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 or fragments, derivatives or analogs thereof, may be used *in vitro* or *in vivo* to direct or aid the synthesis or modification of an enediyne warhead or a substructure thereof. Polypeptides of the UNBU family, namely 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 or fragments, derivatives or analogs thereof may be used *in vitro* or *in vivo* to direct or aid the synthesis or modification of an enediyne warhead or a substructure thereof.

30 The polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 of the present divisional application, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812 and SEQ ID NOS: 2, 4, 6, 8,

-52-

10, 12, 14, 16, 18, 20 of CA 2,444,802, or fragments, derivatives or analogues thereof comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof, may also be used to generate antibodies which bind specifically to the polypeptides or fragments, derivatives or analogues. The antibodies generated from SEQ ID NOS: 2 and 4, SEQ ID NO: 2 of CA 2,387,401, SEQ ID NO: 2 of CA 2,445,692, SEQ ID NO: 2 of CA 2,444,812 and SEQ ID NO: 2 of CA 2,444,802 may be used to determine whether a biological sample contains *Streptomyces macromyceticus* or a related microorganism. The antibodies generated from SEQ ID NO: 6, SEQ ID NO: 4 of CA 2,387,401, SEQ ID NO: 4 of CA 2,445,692, SEQ ID NO: 4 of CA 2,444,812 and
10 SEQ ID NO: 4 of CA 2,444,802 may be used to determine whether a biological sample contains *Micromonospora echinospora* subsp. *calichensis* or a related microorganism. The antibodies generated from SEQ ID NO: 8, SEQ ID NO: 6 of CA 2,387,401, SEQ ID NO: 6 of CA 2,445,692, SEQ ID NO: 6 of CA 2,444,812 and SEQ ID NO: 6 of CA 2,444,802 may be used to determine whether a biological sample contains *Streptomyces ghanaensis* or a related microorganism. The antibodies generated from SEQ ID NO: 10, SEQ ID NO: 8 of CA 2,387,401, SEQ ID NO: 8 of CA 2,445,692, SEQ ID NO: 8 of CA 2,444,812 and SEQ ID NO: 8 of CA 2,444,802 may be used to determine whether a biological sample contains *Streptomyces carzinostaticus* subsp. *neocarzinostaticus* or a related microorganism. The antibodies generated from SEQ ID
20 NO: 12, SEQ ID NO: 10 of CA 2,387,401, SEQ ID NO: 10 of CA 2,445,692, SEQ ID NO: 10 of CA 2,444,812 and SEQ ID NO: 10 of CA 2,444,802 may be used to determine whether a biological sample contains *Amycolatopsis orientalis* or a related microorganism. The antibodies generated from SEQ ID NO: 14, SEQ ID NO: 12 of CA 2,387,401, SEQ ID NO: 12 of CA 2,445,692, SEQ ID NO: 12 of CA 2,444,812 and SEQ ID NO: 12 of CA 2,444,802 may be used to determine whether a biological sample contains *Kitasatosporia* sp. or a related microorganism. The antibodies generated from SEQ ID NO: 16, SEQ ID NO: 14 of CA 2,387,401, SEQ ID NO: 14 of CA 2,445,692, SEQ ID NO: 14 of CA 2,444,812 and SEQ ID NO: 14 of CA 2,444,802 may be used to determine whether a biological sample contains *Micromonospora megalomicea* or a
30 related microorganism. The antibodies generated from SEQ ID NO: 18, SEQ ID NO: 16 of CA 2,387,401, SEQ ID NO: 16 of CA 2,445,692, SEQ ID NO: 16 of CA 2,444,812 and SEQ ID NO: 16 of CA 2,444,802 may be used to determine whether a biological sample contains

Saccharothrix aerocolonigenes or a related microorganism. The antibodies generated from SEQ ID NO: 20, SEQ ID NO: 18 of CA 2,387,401, SEQ ID NO: 18 of CA 2,445,692, SEQ ID NO: 18 of CA 2,444,812 and SEQ ID NO: 18 of CA 2,444,802 may be used to determine whether a biological sample contains *Streptomyces kaniharaensis* or a related microorganism. The antibodies generated from SEQ ID NO: 22, SEQ ID NO: 20 of CA 2,387,401, SEQ ID NO: 20 of CA 2,445,692, SEQ ID NO: 20 of CA 2,444,812 and SEQ ID NO: 20 of CA 2,444,802 may be used to determine whether a biological sample contains *Streptomyces citricolor* or a related microorganism.

- 10 In such procedures, a biological sample is contacted with an antibody capable of specifically binding to one of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. The ability of the biological sample to bind to the antibody is then determined. For example, binding may be determined by labeling the antibody with a detectable label such as a fluorescent agent, an enzymatic label, or a radioisotope. Alternatively, binding of the antibody to the sample may be detected using a secondary antibody having such a detectable label thereon. A variety of assay protocols may be used to detect the presence of *Micromonospora echinospora subsp. calichensis*, *Streptomyces ghanaensis*, *Streptomyces carzinostaticus* subsp.
- 20 *neocarzinostaticus*, *Amycolatopsis orientalis*, *Kitasatosporia sp.*, *Micromonospora megalomicea*, *Saccharothrix aerocolonigenes*, *Streptomyces kaniharaensis*, *Streptomyces citricolor* or the presence of polypeptides related to SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 in a sample. Particular assays include ELISA assays, sandwich assays, radioimmunoassays, and Western Blots. Alternatively, antibodies generated from SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 may be used to determine whether a biological sample contains related polypeptides that may be involved in the biosynthesis of enediyne natural products or other enediyne-like compounds.

- 30 Polyclonal antibodies generated against the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal. The antibody so obtained will then bind the polypeptide itself. In this manner,

-55-

NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; nucleotide sequences homologous to SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; or homologous to fragments of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; and sequences complementary to all of the preceding sequences. The fragments include portions of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802, comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 consecutive nucleotides of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802. Preferably, the fragments are novel fragments. Homologous sequences and fragments of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5,

-54-

even a sequence encoding only a fragment of the polypeptide can be used to generate antibodies which may bind to the whole native polypeptide. Such antibodies can then be used to isolate the polypeptide from cells expressing that polypeptide.

For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kholer and Milstein, 1975, *Nature*, 256:495-497), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., 1983, *Immunology Today* 4:72), and the EBV-hybridoma technique (Cole, et al., 1985, in *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc., pp. 77-96).

10 Techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce single chain antibodies to the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. Alternatively, transgenic mice may be used to express humanized antibodies to these polypeptides or fragments thereof.

Antibodies generated against the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof may be used in screening for similar polypeptides from a sample containing organisms or cell-free extracts thereof. In such
20 techniques, polypeptides from the sample is contacted with the antibodies and those polypeptides which specifically bind the antibody are detected. Any of the procedures described above may be used to detect antibody binding. One such screening assay is described in "Methods for measuring Cellulase Activities", *Methods in Enzymology*, Vol 160, pp. 87-116.

As used herein, the term "enediyne-specific nucleic acid codes" encompass the nucleotide sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 of the present application, the nucleic acid sequences of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID
30 NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802; fragments of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID

-56-

7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, refer to a sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 80%, 75% or 70% identity to these sequences. Homology may be determined using any of the computer programs and parameters described herein, including BLASTN and TBLASTX with the default parameters. Homologous sequences also include RNA sequences in which uridines replace the thymines in the nucleic acid codes of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and
 10 SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. It will be appreciated that the nucleic acid codes of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802, can be represented in the traditional single character format in which G, A, T and C denote the guanine, adenine, thymine and cytosine bases of the deoxyribonucleic acid (DNA) sequence respectively, or in which G, A, U and C denote the guanine, adenine, uracil
 20 and cytosine bases of the ribonucleic acid (RNA) sequence (see the inside back cover of Stryer, *Biochemistry*, 3rd edition, W. H. Freeman & Co., New York) or in any other format which records the identity of the nucleotides in a sequence.

“Enediyne-specific polypeptide codes” encompass the polypeptide sequences of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 which are encoded by the cDNAs of SEQ ID NOS:

-57-

3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,387,401, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,445,692, SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,812, and SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 of CA 2,444,802 respectively; polypeptide sequences homologous to the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802; or fragments of any of the preceding sequences. Homologous polypeptide sequences refer to a polypeptide sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75% or 70% identity to one of the polypeptide sequences of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802. Polypeptide sequence identity may be determined using any of the computer programs and parameters described herein, including BLASTP version 2.2.2 with the default parameters or with any user-specified parameters. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error.

The polypeptide fragments comprise at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100 or 150 consecutive polypeptides of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802. Preferably the fragments are novel fragments. It will be appreciated that the polypeptide codes of the SEQ ID NOS: 2, 4, 6, 8, 10, 12,

-58-

14, 16, 18, 20, 22, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802 can be represented in the traditional single character format or three letter format (see the inside back cover of Stryer, *Biochemistry*, 3rd edition, W.H. Freeman & Co., New York) or in any other format which relates the identity of the polypeptides in a sequence.

10 A single sequence selected from enediyne-specific nucleic acid codes and enediyne-specific polypeptide codes is sometimes referred to herein as a subject sequence.

It will be readily appreciated by those skilled in the art that the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, and a subject sequence can be stored, recorded and manipulated on any medium which can be read and accessed by a computer. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, and a subject sequence.

20 Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM) as well as other types of media known to those skilled in the art.

The enediyne-specific nucleic acid codes, a subset thereof and a subject sequence may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, and a subject sequence may be stored as ASCII or text in a word processing file, such as MicrosoftWORD™ or
30 WORDPERFECT™ in a variety of database programs familiar to those of skill in the art, such as DB2™ or ORACLE™. In addition, many computer programs and databases may

-59-

be used as sequence comparers, identifiers or sources of query nucleotide sequences or query polypeptide sequences to be compared to the enediyne-specific nucleic acid codes, a subset thereof, the enediyne-specific polypeptide codes, a subset thereof, and a subject sequence.

The following list is intended not to limit the invention but to provide guidance to programs and databases useful with the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, and a subject sequence. The program and databases which may be used include, but are not limited to: MacPattern™ (EMBL), DiscoveryBase™ (Molecular Applications Group), GeneMine™ (Molecular Applications Group) Look™ (Molecular Applications Group), MacLook™ (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul *et al.*, *J. Mol. Biol.* 215:403 (1990)), FASTA (Person and Lipman, *Proc. Natl. Acad. Sci. USA*, 85:2444 (1988)), FASTDB (Brutlag *et al. Comp. App. Biosci.* 6:237-245, 1990), Catalyst™ (Molecular Simulations Inc.), Catalyst/SHAPE™ (Molecular Simulations Inc.), Cerius².DBAccess™ (Molecular Simulations Inc.), HypoGen™ (Molecular Simulations Inc.), Insight II™ (Molecular Simulations Inc.), Discover™ (Molecular Simulations Inc.), CHARMM™ (Molecular Simulations Inc.), Felix™ (Molecular Simulations Inc.), DelPhi™ (Molecular Simulations Inc.), QuanteMM™ (Molecular Simulations Inc.), Homology™ (Molecular Simulations Inc.), Modeler™ (Molecular Simulations Inc.), ISIS™ (Molecular Simulations Inc.), Quanta/Protein Design™ (Molecular Simulations Inc.), WetLab™ (Molecular Simulations Inc.), WetLab Diversity Explorer™ (Molecular Simulations Inc.), Gene Explorer™ (Molecular Simulations Inc.), SeqFold™ (Molecular Simulations Inc.), the MDL Available Chemicals Directory database, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwent's World Drug Index database, the BioByteMasterFile™ database, the Genbank™ database, and the Gensyqn™ database. Many other programs and databases would be apparent to one of skill in the art given the present disclosure.

Embodiments of the present invention include systems, particularly computer systems that store and manipulate the sequence information described herein. As used herein, "a computer system", refers to the hardware components, software components, and data storage components used to analyze enediyne-specific nucleic

-60-

acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, or a subject sequence.

Preferably, the computer system is a general purpose system that comprises a processor and one or more internal data storage components for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

One example of a computer system is illustrated in Figure 1. The computer system of Figure 4 will includes a number of components connected to a central system bus 116, including a central processing unit 118 with internal 118 and/or external cache memory 120, system memory 122, display adapter 102 connected to a monitor 100, network adapter 126 which may also be referred to as a network interface, internal modem 124, sound adapter 128, IO controller 132 to which may be connected a keyboard 140 and mouse 138, or other suitable input device such as a trackball or tablet, as well as external printer 134, and/or any number of external devices such as external modems, tape storage drives, or disk drives. One skilled in the art will readily appreciate that not all components illustrated in Figure 1 are required to practice the invention and, likewise, additional components not illustrated in Figure 1 may be present in a computer system contemplated for use with the invention.

One or more host bus adapters 114 may be connected to the system bus 116. To host bus adapter 114 may optionally be connected one or more storage devices such as disk drives 112 (removable or fixed), floppy drives 110, tape drives 108, digital versatile disk DVD drives 106, and compact disk CD ROM drives 104. The storage devices may operate in read-only mode and / or in read-write mode. The computer system may optionally include multiple central processing units 118, or multiple banks of memory 122.

Arrows 142 in Figure 1 indicate the interconnection of internal components of the computer system. The arrows are illustrative only and do not specify exact connection architecture.

Software for accessing and processing the reference sequences (such as sequence comparison software, analysis software as well as search tools, annotation tools, and modeling tools etc.) may reside in main memory 122 during execution.

-61-

In one embodiment, the computer system further comprises a sequence comparison software for comparing the nucleic acid codes of a query sequence stored on a computer readable medium to a subject sequence which is also stored on a computer readable medium; or for comparing the polypeptide code of a query sequence stored on a computer readable medium to a subject sequence which is also stored on computer readable medium. A "sequence comparison software" refers to one or more programs that are implemented on the computer system to compare nucleotide sequences with other nucleotide sequences stored within the data storage means. The design of one example of a sequence comparison software is provided in
10 Figures 2A, 2B, 2C and 2D.

The sequence comparison software will typically employ one or more specialized comparator algorithms. Protein and/or nucleic acid sequence similarities may be evaluated using any of the variety of sequence comparator algorithms and programs known in the art. Such algorithms and programs include, but are no way limited to, TBLASTN, BLASTN, BLASTP, FASTA, TFASTA, CLUSTAL, HMMER, MAST, or other suitable algorithm known to those skilled in the art. (Pearson and Lipman, 1988, *Proc. Natl. Acad. Sci USA* 85(8): 2444-2448; Altschul *et al*, 1990, *J. Mol. Biol.* 215(3):403-410; Thompson *et al.*, 1994, *Nucleic Acids Res.* 22(2):4673-4680; Higgins *et al.*, 1996, *Methods Enzymol.* 266:383-402; Altschul *et al.*, 1990, *J. Mol. Biol.* 215(3):403-410;
20 Altschul *et al.*, 1993, *Nature Genetics* 3:266-272; Eddy S.R., *Bioinformatics* 14:755-763, 1998; Bailey TL *et al*, *J Steroid Biochem Mol Biol* 1997 May;62(1):29-44). One example of a comparator algorithm is illustrated in Figure 3. Sequence comparator algorithms identified in this specification are particularly contemplated for use in this aspect of the invention.

The sequence comparison software will typically employ one or more specialized analyzer algorithms. One example of an analyzer algorithm is illustrated in Figure 4. Any appropriate analyzer algorithm can be used to evaluate similarities, determined by the comparator algorithm, between a query sequence and a subject sequence (referred to herein as a query/subject pair). Based on context specific rules, the annotation of a
30 subject sequence may be assigned to the query sequence. A skilled artisan can readily determine the selection of an appropriate analyzer algorithm and appropriate context specific rules. Analyzer algorithms identified elsewhere in this specification are particularly contemplated for use in this aspect of the invention.

-62-

Figures 2A, 2B, 2C and 2D together provide a flowchart of one example of a sequence comparison software for comparing query sequences to a subject sequence. The software determines if a gene or set of genes represented by their nucleotide sequence, polypeptide sequence or other representation (the query sequence) is significantly similar to the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, of the invention (the subject sequence). The software may be implemented in the C or C++™ programming language, Java™, Perl™ or other suitable programming language known to a person skilled in the art.

10 Referring to Figure 2A, the query sequence(s) may be accessed by the program by means of input from the user 210, accessing a database 208 or opening a text file 206. The "query initialization process" allows a query sequence to be accessed and loaded into computer memory 122, or under control of the program stored on a diskdrive 112 or other storage device in the form of a query sequence array 216. The query array 216 is one or more query nucleotide or polypeptide sequences accompanied by some appropriate identifiers.

A dataset is accessed by the program by means of input from the user 228, accessing a database 226, or opening a text file 224. The "subject data source initialization process" of Figure 2B refers to the method by which a reference dataset
20 containing one or more sequence selected from the enediyne-specific nucleic acid codes, a subset thereof, enediyne-specific polypeptide codes, a subset thereof, or a subject sequence is loaded into computer memory 122, or under control of the program stored on a disk drive 112 or other storage device in the form of a subject array 234. The subject array 234 comprises one or more subject nucleotide or polypeptide sequences accompanied by some appropriate identifiers.

The "comparison subprocess" of Figure 2C is the process by which the comparator algorithm 238 is invoked by the software for pairwise comparisons between query elements in the query sequence array 216, and subject elements in the subject array 234. The "comparator algorithm" of Figure 2C refers to the pairwise comparisons
30 between a query sequence and subject sequence, i.e. a query/subject pair from their respective arrays 216, 234. Comparator algorithm 238 may be any algorithm that acts on a query/subject pair, including but not limited to homology algorithms such as BLAST, Smith Waterman™, Fasta™, or statistical representation/probabilistic

-63-

algorithms such as Markov models exemplified by HMMER, or other suitable algorithm known to one skilled in the art. Suitable algorithms would generally require a query/subject pair as input and return a score (an indication of likeness between the query and subject), usually through the use of appropriate statistical methods such as Karlin Altschul statistics used in BLAST™, Forward™ or Viterbi™ algorithms used in Markov models, or other suitable statistics known to those skilled in the art.

The sequence comparison software of Figure 2C also comprises a means of analysis of the results of the pairwise comparisons performed by the comparator algorithm 238. The “analysis subprocess” of Figure 2C is a process by which the
 10 analyzer algorithm 244 is invoked by the software. The “analyzer algorithm” refers to a process by which annotation of a subject is assigned to the query based on query/subject similarity as determined by the comparator algorithm 238 according to context-specific rules coded into the program or dynamically loaded at runtime. Context-specific rules are what the program uses to determine if the annotation of the subject can be assigned to the query given the context of the comparison. These rules allow the software to qualify the overall meaning of the results of the comparator algorithm 238.

In one embodiment of the present divisional application, context-specific rules may state that for a set of query sequences to be considered representative of an
 20 enediyne locus the comparator algorithm 238 must determine that the set of query sequences contain at least one query sequence that shows a statistical similarity to reference sequences corresponding to a nucleic acid sequence code for a polypeptide from two of the groups consisting of: (1) SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 and polypeptides having at least 75% identity to a polypeptide sequence of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; (2) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 and polypeptides having at least 75% identity to a polypeptide sequence of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401; (3) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692, and polypeptides having at least
 30 75% identity to the polypeptide sequence of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,445,692; (4) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812, and polypeptides having at least 75% identity to the polypeptide sequence SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,812; (5) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,387,401 CA 2,444,802, and polypeptides having at

-64-

least 75% identity to the polypeptide sequence SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 of CA 2,444,802. Of course preferred context specific rules may specify a wide variety of thresholds for identifying enediynes-biosynthetic genes or enediynes-producing organisms without departing from the scope of the invention. Some thresholds contemplate that at least one query sequence in the set of query sequences show a statistical similarity to the nucleic acid code corresponding to 2 or 3 or 4 or 5 of the above 5 groups polypeptides diagnostic of enediynes biosynthetic genes. Other context specific rules set the level of identity required in each of the group may be set at 70%, 80%, 85%, 90%, 95% or 98% in regards to any one or more of the subject sequences.

10 In another embodiment of the present divisional application, context-specific rules may state that for a query sequence to be considered an enediynes polyketide synthase, the comparator algorithm 238 must determine that the query sequence shows a statistical similarity to subject sequences corresponding to a nucleic acid sequence code for a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, polypeptides having at least 75% identity to a polypeptide of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 and fragment comprising at least 500 consecutive amino acids of the polypeptides of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22. Of course preferred context specific rules may specify a wide variety of thresholds for identifying enediynes polyketide synthase proteins without departing from the scope of the invention. Some context specific rules set level of identity required of the query sequence at 70%, 80%, 85%, 90%, 95% or 98% in regards to the reference sequences.

20 Thus, the analysis subprocess may be employed in conjunction with any other context specific rules and may be adapted to suit different embodiments. The principal function of the analyzer algorithm 244 is to assign meaning or a diagnosis to a query or set of queries based on context specific rules that are application specific and may be changed without altering the overall role of the analyzer algorithm 244.

30 Finally the sequence comparison software of Figure 2 comprises a means of returning of the results of the comparisons by the comparator algorithm 238 and analyzed by the analyzer algorithm 244 to the user or process that requested the comparison or comparisons. The "display / report subprocess" of Figure 2D is the process by which the results of the comparisons by the comparator algorithm 238 and

-65-

analyses by the analyzer algorithm 244 are returned to the user or process that requested the comparison or comparisons. The results 240, 246 may be written to a file 252, displayed in some user interface such as a console, custom graphical interface, web interface, or other suitable implementation specific interface, or uploaded to some database such as a relational database, or other suitable implementation specific database.

Once the results have been returned to the user or process that requested the comparison or comparisons the program exits.

10 The principle of the sequence comparison software of Figure 2 is to receive or load a query or queries, receive or load a reference dataset, then run a pairwise comparison by means of the comparator algorithm 238, then evaluate the results using an analyzer algorithm 244 to arrive at a determination if the query or queries bear significant similarity to the reference sequences, and finally return the results to the user or calling program or process.

Figure 3 is a flow diagram illustrating one embodiment of comparator algorithm 238 process in a computer for determining whether two sequences are homologous. The comparator algorithm receives a query/subject pair for comparison, performs an appropriate comparison, and returns the pair along with a calculated degree of similarity.

20 Referring to Figure 3, the comparison is initiated at the beginning of sequences 304. A match of (x) characters is attempted 306 where (x) is a user specified number. If a match is not found the query sequence is advanced 316 by one polypeptide with respect to the subject, and if the end of the query has not been reached 318 another match of (x) characters is attempted 306. Thus if no match has been found the query is incrementally advanced in entirety past the initial position of the subject, once the end of the query is reached 318, the subject pointer is advanced by 1 polypeptide and the query pointer is set to the beginning of the query 318. If the end of the subject has been reached and still no matches have been found a null homology result score is assigned 324 and the algorithm returns the pair of sequences along with a null score to
30 the calling process or program. The algorithm then exits 326. If instead a match is found 308, an extension of the matched region is attempted 310 and the match is analyzed statistically 312. The extension may be unidirectional or bidirectional. The algorithm continues in a loop extending the matched region and computing the

-66-

homology score, giving penalties for mismatches taking into consideration that given the chemical properties of the polypeptide side chains not all mismatches are equal. For example a mismatch of a lysine with an arginine both of which have basic side chains receive a lesser penalty than a mismatch between lysine and glutamate which has an acidic side chain. The extension loop stops once the accumulated penalty exceeds some user specified value, or of the end of either sequence is reached 312. The maximal score is stored 314, and the query sequence is advanced 316 by one polypeptide with respect to the subject, and if the end of the query has not been reached 318 another match of (x) characters is attempted 306. The process continues
 10 until the entire length of the subject has been evaluated for matches to the entire length of the query. All individual scores and alignments are stored 314 by the algorithm and an overall score is computed 324 and stored. The algorithm returns the pair of sequences along with local and global scores to the calling process or program. The algorithm then exits 326.

Comparator algorithm 238 algorithm may be represented in pseudocode as follows:

```

20  INPUT:      Q[m]: query, m is the length
                S[n]: subject, n is the length
                x:   x is the size of a segment

    START:
    for each i in [1,n] do
        for each j in [1,m] do
            if ( j + x - 1 ) <= m and ( i + x - 1 ) <= n then
                if Q(j, j+x-1) = S(i, i+x-1) then
                    k=1;
                    while Q(j, j+x-1+k ) = S(i, i+x-1+ k) do
                        k++;
                    Store highest local homology
30  Compute overall homology score
    Return local and overall homology scores

    END.
```

The comparator algorithm 238 may be written for use on nucleotide sequences, in which case the scoring scheme would be implemented so as to calculate scores and apply penalties based on the chemical nature of nucleotides. The comparator algorithm 238 may also provide for the presence of gaps in the scoring method for nucleotide or polypeptide sequences.

-67-

BLAST is one implementation of the comparator algorithm 238. HMMER is another implementation of the comparator algorithm 238 based on Markov model analysis. In a HMMER implementation a query sequence would be compared to a mathematical model representative of a subject sequence or sequences rather than using sequence homology.

Figure 4 is a flow diagram illustrating an analyzer algorithm 244 process for detecting the presence of an enediynes biosynthetic locus. The analyzer algorithm of Figure 4 may be used in the process by which the annotation of a subject is assigned to the query based on their similarity as determined by the comparator algorithm 238 and according to context-specific rules coded into the program or dynamically loaded at runtime. Context sensitive rules are what determines if the annotation of the subject can be assigned to the query given the context of the comparison. Context specific rules set the thresholds for determining the level and quality of similarity that would be accepted in the process of evaluating matched pairs.

The analyzer algorithm 244 receives as its input an array of pairs that had been matched by the comparator algorithm 238. The array consists of at least a query identifier, a subject identifier and the associated value of the measure of their similarity. To determine if a group of query sequences includes sequences diagnostic of an enediynes biosynthetic gene cluster, a reference or diagnostic array 406 is generated by accessing a data source and retrieving enediynes specific information 404 relating to enediynes-specific nucleic acid codes and enediynes-specific polypeptide codes. Diagnostic array 406 consists at least of subject identifiers and their associated annotation. Annotation may include reference to the five protein families diagnostic of enediynes biosynthetic genes clusters, i.e. PKSE, TEBC, UNBL, UNBV and UNBU. Annotation may also include information regarding exclusive presence in loci of a specific structural class or may include previously computed matches to other databases, for example databases of motifs.

Once the algorithm has successfully generated or received the two necessary arrays 402, 406, and holds in memory any context specific rules, each matched pair as determined by the comparator algorithm 238 can be evaluated. The algorithm will perform an evaluation 408 of each matched pair and based on the context specific rules confirm or fail to confirm the match as valid 410. In cases of successful confirmation of the match 410 the annotation of the subject is assigned to the query.

-68-

Results of each comparison are stored 412. The loop ends when the end of the query / subject array is reached. Once all query / subject pairs have been evaluated against enediynes-specific nucleic acid codes and enediynes-specific polypeptide codes, a final determination can be made if the query set of ORFs represents an enediynes locus 416.

The algorithm then returns the overall diagnosis and an array of characterized query / subject pairs along with supporting evidence to the calling program or process and then terminates 418.

The analyzer algorithm 244 may be configured to dynamically load different diagnostic arrays and context specific rules. It may be used for example in the comparison of query/subject pairs with diagnostic subjects for other biosynthetic pathways, such as chromoprotein enediynes-specific nucleic acid codes or non-chromoprotein enediynes-specific polypeptide codes, or other sets of annotated subjects.

The present invention will be further described with reference to the following examples; however, it is to be understood that the present invention is not limited to such examples.

EXAMPLES

Example 1: Identification and sequencing of the macromomycin (auromomycin)

20 biosynthetic locus

Macromomycin is a chromoprotein enediynes produced by *Streptomyces macromyceticus* (NRRL B-5335). Macromomycin is believed to be a derivative of a larger chromoprotein enediynes compound referred to as auromomycin (Vandre and Montgomery (1982) Biochemistry Vol 21 pp. 3343-3352; Yamashita et al. (1979) J. Antibiot. Vol. 32 pp. 330-339). Thus, throughout the specification, reference to macromomycin is intended to encompass the molecules referred to by some authors as auromomycin. Likewise, reference to the biosynthetic locus for macromomycin is intended to encompass the biosynthetic locus that directs the synthesis of the molecules some authors have referred to as macromomycin and auromomycin.

30 *Streptomyces macromyceticus* (NRRL B-5335) was obtained from the Agricultural Research Service collection (National Center for Agricultural Utilization Research, 1815 N. University Street, Peoria, Illinois 61604) and cultured using standard microbiological techniques (Kieser et al., supra). The organism was propagated on

-69-

oatmeal agar medium at 28 degrees Celsius for several days. For isolation of high molecular weight genomic DNA, cell mass from three freshly grown, near confluent 100 mm petri dishes was used. The cell mass was collected by gentle scraping with a plastic spatula. Residual agar medium was removed by repeated washes with STE buffer (75 mM NaCl; 20 mM Tris-HCl, pH 8.0; 25 mM EDTA). High molecular weight DNA was isolated by established protocols (Kieser *et al.* *supra*) and its integrity was verified by field inversion gel electrophoresis (FIGE) using the preset program number 6 of the FIGE MAPPERTM power supply (BIORAD). This high molecular weight genomic DNA serves for the preparation of a small size fragment genomic sampling library (GSL), *i.e.*, the small insert library, as well as a large size fragment cluster identification library (CIL), *i.e.*, the large insert library. Both libraries contained randomly generated *S. macromyceticus* genomic DNA fragments and, therefore, are representative of the entire genome of this organism.

For the generation of the *S. macromyceticus* GSL library, genomic DNA was randomly sheared by sonication. DNA fragments having a size range between 1.5 and 3 kb were fractionated on a agarose gel and isolated using standard molecular biology techniques (Sambrook *et al.*, *supra*). The ends of the obtained DNA fragments were repaired using T4 DNA polymerase (Roche) as described by the supplier. This enzyme creates DNA fragments with blunt ends that can be subsequently cloned into an appropriate vector. The repaired DNA fragments were subcloned into a derivative of pBluescript SK+ vector (Stratagene) which does not allow transcription of cloned DNA fragments. This vector was selected as it contains a convenient polylinker region surrounded by sequences corresponding to universal sequencing primers such as T3, T7, SK, and KS (Stratagene). The unique *EcoRV* restriction site found in the polylinker region was used as it allows insertion of blunt-end DNA fragments. Ligation of the inserts, use of the ligation products to transform *E. coli* DH10B (Invitrogen) host and selection for recombinant clones were performed as previously described (Sambrook *et al.*, *supra*). Plasmid DNA carrying the *S. macromyceticus* genomic DNA fragments was extracted by the alkaline lysis method (Sambrook *et al.*, *supra*) and the insert size of 1.5 to 3 kb was confirmed by electrophoresis on agarose gels. Using this procedure, a library of small size random genomic DNA fragments is generated that covers the entire genome of the studied microorganism. The number of individual clones that can be

-70-

generated is infinite but only a small number is further analyzed to sample the microorganism's genome.

A CIL library was constructed from the *S. macromyceticus* high molecular weight genomic DNA using the SuperCos-1™ cosmid vector (Stratagene™). The cosmid arms were prepared as specified by the manufacturer. The high molecular weight DNA was subjected to partial digestion at 37 degrees Celsius with approximately one unit of *Sau3AI* restriction enzyme (New England Biolabs) per 100 micrograms of DNA in the buffer supplied by the manufacturer. This enzyme generates random fragments of DNA ranging from the initial undigested size of the DNA to short fragments of which the length is dependent upon the frequency of the enzyme DNA recognition site in the genome and the extent of the DNA digestion. At various timepoints, aliquots of the digestion were transferred to new microfuge tubes and the enzyme was inactivated by adding a final concentration of 10 mM EDTA and 0.1% SDS. Aliquots judged by FIGE analysis to contain a significant fraction of DNA in the desired size range (30-50kb) were pooled, extracted with phenol/chloroform (1:1 vol:vol), and pelleted by ethanol precipitation.

The 5' ends of *Sau3AI* DNA fragments were dephosphorylated using alkaline phosphatase (Roche) according to the manufacturer's specifications at 37 degrees Celsius for 30 min. The phosphatase was heat inactivated at 70 degrees Celsius for 10 min and the DNA was extracted with phenol/chloroform (1:1 vol:vol), pelleted by ethanol precipitation, and resuspended in sterile water. The dephosphorylated *Sau3AI* DNA fragments were then ligated overnight at room temperature to the SuperCos-1 cosmid arms in a reaction containing approximately four-fold molar excess SuperCos-1 cosmid arms.

The ligation products were packaged using Gigapack® III XL packaging extracts (Stratagene™) according to the manufacturer's specifications. The CIL library consisted of 864 isolated cosmid clones in *E. coli* DH10B (Invitrogen). These clones were picked and inoculated into nine 96-well microtiter plates containing LB broth (per liter of water: 10.0 g NaCl; 10.0 g tryptone™; 5.0 g yeast extract) which were grown overnight and then adjusted to contain a final concentration of 25% glycerol. These microtiter plates were stored at -80 degrees Celsius and served as glycerol stocks of the CIL library. Duplicate microtiter plates were arrayed onto nylon membranes as follows. Cultures grown on microtiter plates were concentrated by pelleting and

-71-

resuspending in a small volume of LB broth. A 3 X 3 96-pin-grid was spotted onto nylon membranes.

The membranes, representing the complete CIL library, were then layered onto LB agar and incubated overnight at 37 degrees Celcius to allow the colonies to grow. The membranes were layered onto filter paper pre-soaked with 0.5 N NaOH/1.5 M NaCl for 10 min to denature the DNA and then neutralized by transferring onto filter paper pre-soaked with 0.5 M Tris (pH 8)/1.5 M NaCl for 10 min. Cell debris was gently scraped off with a plastic spatula and the DNA was crosslinked onto the membranes by UV irradiation using a GS GENE LINKER™ UV Chamber (BIORAD). Considering an
10 average size of 8 Mb for an actinomycete genome and an average size of 35 kb of genomic insert in the CIL library, this library represents roughly a 4-fold coverage of the microorganism's entire genome.

The GSL library was analyzed by sequence determination of the cloned genomic DNA inserts. The universal primers KS or T7, referred to as forward (F) primers, were used to initiate polymerization of labeled DNA. Extension of at least 700 bp from the priming site can be routinely achieved using the TF, BDT v2.0 sequencing kit as specified by the supplier (Applied Biosystems). Sequence analysis of the small genomic DNA fragments (Genomic Sequence Tags, GSTs) was performed using a
20 3700 ABI capillary electrophoresis DNA sequencer (Applied Biosystems). The average length of the DNA sequence reads was ~700 bp. Further analysis of the obtained GSTs was performed by sequence homology comparison to various protein sequence databases. The DNA sequences of the obtained GSTs were translated into amino acid sequences and compared to the National Center for Biotechnology Information (NCBI) nonredundant protein database and the proprietary Ecopia natural product biosynthetic gene Decipher™ database using previously described algorithms (Altschul et al., *supra*). Sequence similarity with known proteins of defined function in the database enables one to make predictions on the function of the partial protein that is encoded by the translated GST.

A total of 479 *S. macromyceticus* GSTs obtained with the forward sequencing
30 primer were analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Sequence alignments displaying an E value of at least e-5 were considered as significantly homologous and retained for further evaluation. GSTs showing similarity to a gene of interest can be at this point selected and used to identify larger

-72-

segments of genomic DNA from the CIL library that include the gene(s) of interest. Several *S. macromyceticus* GSTs that contained genes of interest were pursued. One of these GSTs encoded a portion of an oxidoreductase based on Blast analysis of the forward read and a portion of the macromomycin apoprotein based on Blast analysis of the reverse read. Oligonucleotide probes derived from such GSTs were used to screen the CIL library and the resulting positive cosmid clones were sequenced. Overlapping cosmid clones provided in excess of 125 kb of sequence information surrounding the macromomycin apoprotein gene (Figure 5).

10 Hybridization oligonucleotide probes were radiolabeled with P^{32} using T4 polynucleotide kinase (New England Biolabs) in 15 microliter reactions containing 5 picomoles of oligonucleotide and 6.6 picomoles of $[\gamma-P^{32}]ATP$ in the kinase reaction buffer supplied by the manufacturer. After 1 hour at 37 degrees Celcius, the kinase reaction was terminated by the addition of EDTA to a final concentration of 5 mM. The specific activity of the radiolabeled oligonucleotide probes was estimated using a Model 3 Geiger counter (Ludlum Measurements Inc., Sweetwater, Texas) with a built-in integrator feature. The radiolabeled oligonucleotide probes were heat-denatured by incubation at 85 degrees Celcius for 10 minutes and quick-cooled in an ice bath immediately prior to use.

20 The *S. macromyceticus* CIL library membranes were pretreated by incubation for at least 2 hours at 42 degrees Celcius in Prehyb Solution (6X SSC; 20mM NaH_2PO_4 ; 5X Denhardt's; 0.4% SDS; 0.1 mg/ml sonicated, denatured salmon sperm DNA) using a hybridization oven with gentle rotation. The membranes were then placed in Hyb Solution (6X SSC; 20mM NaH_2PO_4 ; 0.4% SDS; 0.1 mg/ml sonicated, denatured salmon sperm DNA) containing 1×10^6 cpm/ml of radiolabeled oligonucleotide probe and incubated overnight at 42 degrees Celcius using a hybridization oven with gentle rotation. The next day, the membranes were washed with Wash Buffer (6X SSC, 0.1% SDS) for 45 minutes each at 46, 48, and 50 degrees Celcius using a hybridization oven with gentle rotation. The *S. macromyceticus* CIL membranes were then exposed to X-ray film to visualize and identify the positive cosmid clones. Positive clones were
30 identified, cosmid DNA was extracted from 30 ml cultures using the alkaline lysis method (Sambrook et al., *supra*) and the inserts were entirely sequenced using a shotgun sequencing approach (Fleischmann et al., (1995) *Science*, 269:496-512).

-73-

Sequencing reads were assembled using the Phred-Phrap™ algorithm (University of Washington, Seattle, USA) recreating the entire DNA sequence of the cosmid insert. Reiterations of hybridizations of the CIL library with probes derived from the ends of the original cosmid allow indefinite extension of sequence information on both sides of the original cosmid sequence until the complete sought-after gene cluster is obtained. The structure of macromomycin (auromomycin) has not been elucidated, however the apoprotein component has been well characterized (Van Roey and Beerman (1989) Proc Natl Acad Sci USA Vol. 86 pp. 6587-6591). An unusual polyketide synthase (PKSE) was found approximately 40 kb upstream of the macromomycin apoprotein gene (Figure 5). No other polyketide synthase or fatty acid synthase gene cluster was found in the vicinity of the macromomycin apoprotein gene, suggesting that the PKSE may be the only polyketide synthase involved in the biosynthesis of macromomycin (auromomycin).

Four other enediynes-specific genes clustered with or in close proximity to the PKSE gene were found in the macromomycin biosynthetic locus. These genes and the polypeptides that they encode have been assigned the family designations TEBC, UNBL, UNBV, and UNBU. The macromomycin locus contains two copies of the TEBC gene (Figure 6, Table 2). Table 2 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of these enediynes-specific polypeptides from the macromomycin locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 2

MACR locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1936	T37056,2082aa	6e-86	273/897 (30.43%)	372/897 (41.47%)	multi-domain beta keto-acyl synthase, Streptomyces coelicolor
		NP_485686.1,1263aa	5e-82	256/900 (28.44%)	388/900 (43.11%)	heterocyst glycolipid synthase, Nostoc sp.
		AAL01060.1,2573aa	6e-78	244/884 (27.6%)	376/884 (42.53%)	polyunsaturated fatty acid synthase, Photobacterium profundum
TEBC1	162	NP_249659.1,148aa	4e-06	38/134 (28.36%)	59/134 (44.03%)	hypothetical protein, Pseudomonas aeruginosa
		CAB50777.1,150aa	4e-06	39/145 (26.9%)	65/145 (44.83%)	hypothetical protein, Pseudomonas putida
		NP_214031.1,128aa	2e-04	33/129 (25.58%)	55/129 (42.64%)	hypothetical protein, Aquifex aeolicus
TEBC2	157	NP_242865.1,138aa	0.27	31/131 (23%)	50/131 (37%)	4-hydroxybenzoyl-CoA thioesterase, Bacillus halodurans

-74-

UNBL	327	NP_422192.1,423aa	0.095	30/86 (34.88%)	40/86 (46.51%)	peptidase, <i>Caulobacter crescentus</i>
UNBV	642	NO HOMOLOG				
UNBU	433	NP_486037.1,300aa	1e-06	49/179 (27.37%)	83/179 (46.37%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_107088.1,503aa	2e-04	72/280 (25.71%)	126/280 (45%)	hypothetical protein, <i>Mesorhizobium loti</i>
		NP_440874.1,285aa	4e-04	47/193 (24.35%)	86/193 (44.56%)	hypothetical protein, <i>Synechocystis</i> sp.

The macromomycin genes listed in Table 2 are arranged as depicted in Figure 6. The UNBL, UNBV, UNBU, PKSE, and TEBC1 genes span approximately 10.5 kb and are tandemly arranged in the order listed. Thus these five genes may constitute an operon. A second TEBC gene (TEBC2) is found approximately 6.6 kb downstream of the 5-gene enediynes-specific cassette. The macromomycin enediynes-specific cassette is composed of six functionally linked genes and polypeptides, five of which may be expressed as a single operon.

10 Example 2: Identification and sequencing of the calicheamicin biosynthetic locus

Calicheamicin is a non-chromoprotein enediyne produced by *Micromonospora echinospora* subsp. *calichensis* NRRL 15839. Both GSL and CIL genomic DNA libraries of *M. echinospora* genomic DNA were prepared as described in Example 1. A total of 288 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*) to identify those clones that contained inserts related to the macromomycin (auromomycin) biosynthetic genes, particularly the PKSE. Such GST clones were identified and were used to isolate cosmid clones from the *M. echinospora* CIL library. Overlapping cosmid clones were sequenced and assembled as described in Example 1. The resulting DNA

20 sequence information was more than 125 kb in length and included the calicheamicin genes described in WO 00/37608. The calicheamicin biosynthetic genes disclosed in WO 00/37608 span only from 37140 bp to 59774 bp in Figure 5 and do not include the unusual PKS gene (PKSE) and four other flanking genes (UNBL, UNBV, UNBU, and TEBC) that are homologous to those in the macromomycin biosynthetic locus. Table 3 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of these enediynes-specific polypeptides from the calicheamicin locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 3

-75-

CALI locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1919	AAF26923.1,2439aa	1e-60	228/876 (26.03%)	317/876 (36.19%)	polyketide synthase, <i>Polyangium cellulorum</i>
		NP_485686.1,1263aa	5e-59	148/461 (32.1%)	210/461 (45.55%)	heterocyst glycolipid synthase, <i>Nostoc</i> sp.
		T37056,2082aa	9e-58	161/466 (34.55%)	213/466 (45.71%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
TEBC	148	NP_249659.1,148aa	8e-06	41/133 (30.83%)	62/133 (46.62%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		AAD49752.1,148aa	1e-05	41/138 (29.71%)	63/138 (45.65%)	orf1, <i>Pseudomonas aeruginosa</i>
		NP_242865.1,138aa	2e-04	32/130 (24.62%)	56/130 (43.08%)	4-hydroxybenzoyl-CoA thioesterase, <i>Bacillus halodurans</i>
UNBL	322	NO HOMOLOG				
UNBV	651	NO HOMOLOG				
UNBU	321	NP_486037.1,300aa	8e-09	61/210 (29.05%)	99/210 (47.14%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_107088.1,503aa	5e-05	58/208 (27.88%)	96/208 (46.15%)	hypothetical protein, <i>Mesorhizobium loti</i>

The calicheamicin genes listed in Table 3 are arranged as depicted in Figure 6. The UNBL, UNBV, UNBU, PKSE, and TEBC genes span approximately 10.5 kb and are tandemly arranged in the order listed. Thus these five genes may constitute an operon. Therefore, the calicheamicin enediyne-specific cassette is composed of five functionally linked genes and polypeptides that may be expressed as a single operon.

Example 3: Identification and sequencing of the biosynthetic locus for an unknown chromoprotein enediyne in *Streptomyces ghanaensis*

10 The genomic sampling method described in Example 1 was applied to genomic DNA from *Streptomyces ghanaensis* NRRL B-12104. *S. ghanaensis* has not previously been described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *S. ghanaensis* genomic DNA were prepared as described in Example 1. A total of 435 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*).

Surprisingly, two GSTs from *S. ghanaensis* were identified as encoding portions of genes in the 5-gene cassette common to both the macromomycin and calicheamicin enediyne biosynthetic loci. One of these GSTs encoded a portion of a TEBC homologue and the other encoded a portion of a UNBV homologue. These *S. ghanaensis* GSTs were subsequently found in a genetic locus referred to herein as 009C (Figure 5). As in the macromomycin and calicheamicin enediyne biosynthetic

20

-76-

loci, the UNBV and TEBC genes in 009C were found to flank a PKSE gene and adjacent to UNBL and UNBU genes. The 009C locus included a gene encoding a homologue of the macromomycin apoprotein approximately 50 kb downstream of the UNBV-UNBU-UNBL-PKSE-TEBC cassette. The presence of the 5-gene cassette in the vicinity of an apoprotein suggests that 009C represents a biosynthetic locus for an unknown chromoprotein enediynes that was not previously described to be produced by *S. ghanaensis* NRRL B-12104.

Table 4 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of these enediynes-specific polypeptides from the 009C locus. Homology was determined using the BLASTP algorithm with the default parameters.

-77-

Table 4

009C locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1956	T37056,2082aa	1e-101	298/902 (33.04%)	395/902 (43.79%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		NP_485686.1,1263aa	2e-99	274/900 (30.44%)	407/900 (45.22%)	heterocyst glycolipid synthase, <i>Nostoc</i> sp.
		BAB69208.1,2365aa	3e-89	282/880 (32.05%)	366/880 (41.59%)	polyketide synthase, <i>Streptomyces avermitilis</i>
TEBC	152	NP_249659.1,148aa	5e-07	39/131 (29.77%)	59/131 (45.04%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		NP_231474.1,155aa	2e-04	30/129 (23.26%)	62/129 (48.06%)	hypothetical protein, <i>Vibrio cholerae</i>
		NP_214031.1,128aa	2e-04	31/128 (24.22%)	55/128 (42.97%)	hypothetical protein, <i>Aquifex aeolicus</i>
UNBL	329	NO HOMOLOG				
UNBV	636	NP_615809.1,2275aa	6e-05	72/314 (22.93%)	114/314 (36.31%)	cell surface protein, <i>Methanosarcina acetivorans</i>
UNBU	382	NP_486037.1,300aa	4e-07	46/175 (26.29%)	81/175 (46.29%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_107088.1,503aa	6e-06	68/255 (26.67%)	118/255 (46.27%)	hypothetical protein, <i>Mesorhizobium loti</i>

The 009C genes listed in Table 4 are arranged as depicted in Figure 6. The UNBL, UNBV, UNBU, PKSE, and TEBC genes span approximately 10.5 kb and are tandemly arranged in the order listed. These five genes may constitute an operon. Therefore, the 009C enediynes-specific cassette is composed of five functionally linked genes and polypeptides that may be expressed as a single operon.

Example 4: The 5-gene enediynes cassette is present in the neocarzinostatin

10 biosynthetic locus

Neocarzinostatin is a chromoprotein enediyne produced by *Streptomyces carzinostaticus* subsp. *neocarzinostaticus* ATCC 15944. The neocarzinostatin biosynthetic locus was sequenced and was shown to contain, in addition to the neocarzinostatin apoprotein gene, the 5-gene cassette that is present in the macromomycin and calicheamicin enediyne biosynthetic loci. The genes and proteins involved in the biosynthesis of neocarzinostatin are disclosed in co-pending application USSN 60/354,474. The presence of the 5-gene cassette in the neocarzinostatin biosynthetic locus reconfirms that it is present in all enediyne biosynthetic loci.

20 Table 5 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of these enediyne-specific polypeptides from the

-78-

neocarzinostatin locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 5

NEOC locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1977	T37056,2082aa	7e-93	285/891 (31.99%)	384/891 (43.1%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		NP_485686.1,1263aa	8e-88	261/890 (29.33%)	397/890 (44.61%)	heterocyst glycolipid synthase, <i>Nostoc</i> sp.
		BAB69208.1,12365aa	2e-85	276/876 (31.51%)	370/876 (42.24%)	polyketide synthase, <i>Streptomyces avermitilis</i>
TEBC	153	NP_249659.1,148aa	3e-06	37/129 (28.68%)	56/129 (43.41%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		CAB50777.1,150aa	1e-04	32/114 (28.07%)	53/114 (46.49%)	hypothetical protein, <i>Pseudomonas putida</i>
		NP_214031.1,128aa	2e-04	34/129 (26.36%)	55/129 (42.64%)	hypothetical protein, <i>Aquifex aeolicus</i>
UNBL	328					
UNBV	636	NP_618575.1,1881aa	2e-05	77/317 (24.29%)	117/317 (36.91%)	cell surface protein, <i>Methanosarcina acetivorans</i>
UNBU	364	NP_107088.1,503aa	2e-05	49/158 (31.01%)	79/158 (50%)	hypothetical protein, <i>Mesorhizobium loti</i>
		NP_486037.1,300aa	8e-05	33/126 (26.19%)	60/126 (47.62%)	hypothetical protein, <i>Nostoc</i> sp.

The neocarzinostatin genes listed in Table 5 are arranged as depicted in Figure 6. The UNBL, UNBV, UNBU, PKSE, and TEBC genes span approximately 10.5 kb and are tandemly arranged in the order listed. Thus these five genes may constitute an operon. Therefore, the neocarzinostatin enediyne-specific cassette is composed of five functionally linked genes and polypeptides that may be expressed as a single operon.

10

Example 5: The 5-gene enediyne cassette is present in the biosynthetic locus of an unknown chromoprotein enediyne in *Amycolatopsis orientalis*

The genomic sampling method described in Example 1 was applied to genomic DNA from *Amycolatopsis orientalis* ATCC 43491. *A. orientalis* has not previously been described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *A. orientalis* genomic DNA were prepared as described in Example 1.

A total of 1025 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Several secondary metabolism loci were identified and sequenced as described in Example 1. One of these loci (herein referred to as 007A) includes a 5-gene cassette

20

-79-

common to all enediyne biosynthetic loci. The arrangement of the five genes of the cassette in 007A is shown in Figure 6. Interestingly, the *A. orientalis* genome also contains an enediyne apoprotein gene that is similar to that from the macromomycin and 009C loci as well as other chromoprotein enediynes (data not shown). Therefore, *A. orientalis*, the producer of the well-known glycopeptide antibiotic vancomycin, has the genomic potential to produce a chromoprotein enediyne.

Table 6 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of the enediyne-specific polypeptides from the 007A locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 6

007A locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1939	T37056,2082aa	5e-96	291/906 (32.12%)	399/906 (44.04%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		NP_485686.1,1263aa	9e-87	255/897 (28.43%)	395/897 (44.04%)	heterocyst glycolipid synthase, <i>Nostoc</i> sp.
		BAB69208.1,2365aa	8e-86	285/926 (30.78%)	393/926 (42.44%)	modular polyketide synthase, <i>Streptomyces avermitilis</i>
TEBC	146	NP_214031.1,128aa	0.052	28/124 (22.58%)	51/124 (41.13%)	hypothetical protein, <i>Aquifex aeolicus</i>
UNBL	324	NO HOMOLOG				
UNBV	654	NP_618575.1,1881aa	0.001	80/332 (24.1%)	117/332 (35.24%)	cell surface protein, <i>Methanosarcina acetivorans</i>
UNBU	329	NP_486037.1,300aa	0.005	56/245 (22.86%)	96/245 (39.18%)	hypothetical protein, <i>Nostoc</i> sp.

The 007A genes listed in Table 6 are arranged as depicted in Figure 6. The UNBL, UNBV, and UNBU genes span approximately 4 kb and are tandemly arranged in the order listed. The PKSE and TEBC genes span approximately 6.5 kb and are tandemly arranged in the order listed. Thus these five genes may constitute two operons. The two putative operons are separated by approximately 5 kb. Although these two clusters of genes may not be transcriptionally linked to one another, they are still functionally linked. Therefore, the 007A enediyne-specific cassette is composed of five functionally linked genes and polypeptides, three of which may be expressed as a one operon and two of which may be expressed as a second operon.

Example 6: The 5-gene enediyne cassette is present in the biosynthetic locus of an unknown enediyne in *Kitasatosporia* sp. CECT 4991

-80-

The genomic sampling method described in Example 1 was applied to genomic DNA from *Kitasatosporia* sp. CECT 4991. This organism was not previously described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *Kitasatosporia* sp. genomic DNA were prepared as described in Example 1.

A total of 1390 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Surprisingly, two GSTs from *Kitasatosporia* sp. were identified as encoding portions of genes in the 5-gene cassette common to enediyne biosynthetic loci. One of these GSTs encoded a portion of a PKSE homologue and the other encoded a portion of a UNBV homologue. These *Kitasatosporia* sp. GSTs were subsequently found in a genetic locus referred to herein as 028D which includes a 5-gene cassette common to all enediyne biosynthetic loci. The arrangement of the five genes of the cassette in 028D is shown in Figure 6. Therefore, *Kitasatosporia* sp. CECT 4991 has the genomic potential to produce enediyne compound(s).

Table 7 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of the enediyne-specific polypeptides from the 028D locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 7

028D locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1958	BAB69208.1,2365aa	1e-81	273/926 (29.48%)	354/926 (38.23%)	polyketide synthase, <i>Streptomyces avermitilis</i>
		T37056,2082aa	3e-78	263/895 (29.39%)	356/895 (39.78%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		NP_485686.1,1263aa	7e-71	231/875 (26.4%)	345/875 (39.43%)	heterocyst glycolipid synthase, <i>Nostoc</i> sp.
TEBC	158	NP_249659.1,148aa	1e-04	38/133 (28.57%)	61/133 (45.86%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		AAD49752.1,148aa	3e-04	38/138 (27.54%)	62/138 (44.93%)	orf1, <i>Pseudomonas aeruginosa</i>
		NP_231474.1,155aa	7e-04	31/127 (24.41%)	61/127 (48.03%)	hypothetical protein, <i>Vibrio cholerae</i>
UNBL	327	NO HOMOLOG				
UNBV	676	NO HOMOLOG				
UNBU	338	NP_486037.1,300aa	5e-08	66/240 (27.5%)	105/240 (43.75%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_440874.1,285aa	2e-04	51/190 (26.84%)	98/190 (51.58%)	hypothetical protein, <i>Synechocystis</i> sp.

The 028D genes listed in Table 7 are arranged as depicted in Figure 6. The UNBV, UNBU, PKSE, and TEBC genes span approximately 9.5 kb and are tandemly

-81-

arranged in the order listed. Thus these four genes may constitute an operon. This putative operon is separated from the UNBL gene, which is oriented in the opposite direction relative to the putative operon, by approximately 10.5 kb. Although the UNBL gene cannot be transcriptionally linked to the other genes, it is still functionally linked to the former. Therefore, the 028D enediyne-specific cassette is composed of five functionally linked genes and polypeptides, four of which may be expressed as a single operon. Although expression of functionally linked enediyne-specific genes may be under control of distinct transcriptional promoters they may, nonetheless, be expressed in a concerted fashion. As depicted in Figure 6, the 028D biosynthetic locus is unique
10 in that it is the only example whose enediyne-specific genes are not all oriented in the same direction.

Example 7: The 5-gene enediyne cassette is present in the biosynthetic locus of an unknown enediyne in *Micromonospora megalomicea*

The genomic sampling method described in Example 1 was applied to genomic DNA from *Micromonospora megalomicea* NRRL 3275. This organism was not previously described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *M. megalomicea* genomic DNA were prepared as described in
Example 1.

20 A total of 1390 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Surprisingly, one GST from *M. megalomicea* was identified as encoding a portion of the PKSE gene present in the 5-gene cassette common to biosynthetic loci. The forward read of this GST encoded the C-terminal portion of the KS domain and the N-terminal portion of the AT domain of a PKSE gene. The complement of the reverse read of this GST encoded the C-terminal portion of the AT domain of a PKSE gene. This *M. megalomicea* GST was subsequently found in a genetic locus referred to herein as 054A which includes a 5-gene cassette common to all enediyne biosynthetic loci. The arrangement of the five genes of the cassette in 054A is shown in Figure 6. Therefore,
30 *M. megalomicea* has the genomic potential to produce enediyne compound(s).

Table 8 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of the enediyne-specific polypeptides from the 054A

-82-

locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 8

054A locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1927	NP_485686.1,1263aa	3e-76	247/886 (27.88%)	365/886 (41.2%)	heterocyst glycolipid synthase, Nostoc sp.
		T37056,2082aa	3e-75	269/903 (29.79%)	354/903 (39.2%)	multi-domain beta keto-acyl synthase, Streptomyces coelicolor
		BAB69208.1,2365aa	9e-74	277/923 (30.01%)	359/923 (38.89%)	polyketide synthase, Streptomyces avermitilis
TEBC	154	NP_249659.1,148aa	2e-06	43/147 (29.25%)	66/147 (44.9%)	hypothetical protein, Pseudomonas aeruginosa
		AAD49752.1,148aa	2e-05	42/147 (28.57%)	65/147 (44.22%)	orf1, Pseudomonas aeruginosa
		CAB50777.1,150aa	1e-04	40/139 (28.78%)	61/139 (43.88%)	hypothetical protein, Pseudomonas putida
UNBL	322	NO HOMOLOG				
UNBV	659	CAC44518.1,706aa	0.048	50/166 (30.12%)	67/166 (40.36%)	putative secreted esterase, Streptomyces coelicolor
UNBU	354	NP_486037.1,300aa	5e-06	66/268 (24.63%)	118/268 (44.03%)	hypothetical protein, Nostoc sp.

The 054A genes listed in Table 8 are arranged as depicted in Figure 6. The UNBL, PKSE, and TEBC genes span approximately 7.5 kb and are tandemly arranged in the order listed. The UNBV and UNBU genes span approximately 3 kb and are tandemly arranged in the order listed. Thus these five genes may constitute two operons. The two putative operons are separated by approximately 2 kb. Therefore, the 054A enediyne-specific cassette is composed of five functionally linked genes and polypeptides, three of which may be expressed as a one operon and two of which may be expressed as another operon.

Example 8: The 5-gene enediyne cassette is present in the biosynthetic locus of an unknown enediyne in *Saccharothrix aerocolonigenes*

The genomic sampling method described in Example 1 was applied to genomic DNA from *Saccharothrix aerocolonigenes* ATCC 39243. This organism was not previously described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *Saccharothrix aerocolonigenes* genomic DNA were prepared as described in Example 1.

A total of 513 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Several

-83-

secondary metabolism loci were identified and sequenced as described in Example 1. One of these loci (herein referred to as 132H) includes a 5-gene cassette common to all enediynes biosynthetic loci. The arrangement of the five genes of the cassette in 132H is shown in Figure 6. Therefore, *Saccharothrix aerocolonigenes* has the genomic potential to produce enediyne compound(s).

Table 9 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of these enediynes-specific polypeptides from the 132H locus. Homology was determined using the BLASTP algorithm with the default parameters.

10 **Table 9**

132H locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1892	BAB69208.1,2365aa	1e-108	312/872 (35.78%)	404/872 (46.33%)	polyketide synthase, <i>Streptomyces avermitilis</i>
		T37056,2082aa	1e-101	290/886 (32.73%)	407/886 (45.94%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		T30183,2756aa	4e-94	271/886 (30.59%)	398/886 (44.92%)	hypothetical protein, <i>Shewanella</i> sp.
TEBC	143	NP_442358.1,138aa	0.001	32/127 (25.2%)	48/127 (37.8%)	hypothetical protein, <i>Synechocystis</i> sp.
UNBL	313	NO HOMOLOG				
UNBV	647	AAD34550.1,1529aa	0.012	76/304 (25%)	105/304 (34.54%)	esterase, <i>Aspergillus terreus</i>
UNBU	336	NP_486037.1,300aa	1e-04	42/172 (24.42%)	79/172 (45.93%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_440874.1,285aa	1e-04	48/181 (26.52%)	90/181 (49.72%)	hypothetical protein, <i>Synechocystis</i> sp.

The 132H genes listed in Table 9 are arranged as depicted in Figure 6. The UNBL, UNBV, UNBU, PKSE, and TEBC genes span approximately 10.5 kb and are tandemly arranged in the order listed. Thus, these five genes may constitute an operon. Therefore, the 132H enediynes-specific cassette is composed of five functionally linked genes and polypeptides that may be expressed as a single operon.

Example 9: The 5-gene enediynes cassette is present in the biosynthetic locus of an unknown enediynes in *Streptomyces kaniharaensis*

20 The genomic sampling method described in Example 1 was applied to genomic DNA from *Streptomyces kaniharaensis* ATCC 21070. This organism was not previously described to produce enediynes compounds. Both GSL and CIL genomic

-84-

DNA libraries of *S. kaniharaensis* genomic DNA were prepared as described in Example 1.

A total of 1020 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Surprisingly, one GST from *S. kaniharaensis* was identified as encoding a portion of the PKSE gene present in the 5-gene cassette common to biosynthetic loci. The forward read of this GST encoded the N-terminal portion of the KS domain of a PKSE gene. The complement of the reverse read of this GST encoded the C-terminal portion of the AT domain of a PKSE gene. This *S. kaniharaensis* GST was subsequently found in a genetic locus referred to herein as 135E which includes a 5-gene cassette common to all enediyne biosynthetic loci. The arrangement of the five genes of the cassette in 135E is shown in Figure 6. Therefore, *S. kaniharaensis* has the genomic potential to produce enediyne compound(s).

Table 10 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of the enediyne-specific polypeptides from the 135E locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 10

135E locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1933	T37056,2082aa	1e-85	282/909 (31.02%)	365/909 (40.15%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		BAB69208.1,2365aa	3e-84	285/925 (30.81%)	366/925 (39.57%)	polyketide synthase, <i>Streptomyces avermitilis</i>
		T30937,1053aa	2e-69	246/907 (27.12%)	356/907 (39.25%)	glycolipid synthase, <i>Nostoc punctiforme</i>
TEBC	154	NP_249659.1,148aa	2e-07	41/132 (31.06%)	63/132 (47.73%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		AAD49752.1,148aa	2e-06	40/132 (30.3%)	62/132 (46.97%)	orf1, <i>Pseudomonas aeruginosa</i>
		NP_214031.1,128aa	5e-04	35/127 (27.56%)	60/127 (47.24%)	hypothetical protein, <i>Aquifex aeolicus</i>
UNBL	323	NO HOMOLOG				
UNBV	655	CAC44518.1,706aa	9e-04	41/135 (30.37%)	59/135 (43.7%)	putative secreted esterase, <i>Streptomyces coelicolor</i>
UNBU	346	NP_486037.1,300aa	4e-09	52/191 (27.23%)	87/191 (45.55%)	hypothetical protein, <i>Nostoc</i> sp.
		NP_440874.1,285aa	9e-06	47/197 (23.86%)	89/197 (45.18%)	hypothetical protein, <i>Synechocystis</i> sp.

The 135E genes listed in Table 10 are arranged as depicted in Figure 6. The UNBL, UNBV, and UNBU genes span approximately 4 kb and are tandemly arranged

-85-

in the order listed. The PKSE and TEBC genes span approximately 6.5 kb and are tandemly arranged in the order listed. Thus these five genes may constitute two operons. The two putative operons are separated by approximately 6 kb. Although these two clusters of genes may not be transcriptionally linked to one another, they are still functionally linked. Therefore, the 135E enediyne-specific cassette is composed of five functionally linked genes and polypeptides, three of which may be expressed as a one operon and two of which may be expressed as another operon.

Example 10: The 5-gene enediyne cassette is present in the biosynthetic locus of an unknown enediyne in *Streptomyces citricolor*

The genomic sampling method described in Example 1 was applied to genomic DNA from *Streptomyces citricolor* IFO 13005. This organism was not previously described to produce enediyne compounds. Both GSL and CIL genomic DNA libraries of *S. citricolor* genomic DNA were prepared as described in Example 1.

A total of 1245 GSL clones were sequenced with the forward primer and analyzed by sequence comparison using the Blast algorithm (Altschul et al., *supra*). Several secondary metabolism loci were identified and sequenced as described in Example 1. One of these loci (herein referred to as 145B) includes a 5-gene cassette common to all enediyne biosynthetic loci. The arrangement of the five genes of the cassette in 145B is shown in Figure 6. Therefore, *S. citricolor* has the genomic potential to produce enediyne compound(s).

Table 11 lists the results of sequence comparison using the Blast algorithm (Altschul et al., *supra*) for each of the enediyne-specific polypeptides from the 145B locus. Homology was determined using the BLASTP algorithm with the default parameters.

Table 11

145B locus						
Family	#aa	GenBank homology Accession, #aa	probability	identity	similarity	proposed function of GenBank match
PKSE	1958	T37056,2082aa	4e-88	285/929 (30.68%)	378/929 (40.69%)	multi-domain beta keto-acyl synthase, <i>Streptomyces coelicolor</i>
		BAB69208.1,2365aa	3e-82	284/923 (30.77%)	375/923 (40.63%)	polyketide synthase, <i>Streptomyces avermittis</i>
		AAL01060.1,2573aa	5e-78	240/855 (28.07%)	354/855 (41.4%)	polyunsaturated fatty acid synthase, <i>Photobacterium profundum</i>
TEBC	165	NP_249659.1,148aa	2e-07	39/133 (29.32%)	60/133 (45.11%)	hypothetical protein, <i>Pseudomonas aeruginosa</i>
		NP_231474.1,155aa	3e-04	30/127 (23.62%)	60/127 (47.24%)	hypothetical protein, <i>Vibrio cholerae</i>
		CAB50777.1,150aa	4e-04	37/135 (27.41%)	58/135 (42.96%)	hypothetical protein, <i>Pseudomonas putida</i>
UNBL	324	NO HOMOLOG				
UNBV	659	NP_618575.1,1881aa	0.003	57/245 (23.27%)	85/245 (34.69%)	cell surface protein, <i>Methanosarcina acetivorans</i>
UNBU	337	NP_486037.1,300aa	0.002	62/267 (23.22%)	109/267 (40.82%)	hypothetical protein, <i>Nostoc sp.</i>

The 145B genes listed in Table 11 are arranged as depicted in Figure 6. The UNBV, and UNBU genes span approximately 3 kb and are tandemly arranged in the order listed. The PKSE and TEBC genes span approximately 6.5 kb and are tandemly arranged in the order listed. Thus these four genes may constitute two operons. The two putative operons are separated by approximately 9.5 kb that includes the UNBL gene. Although these genes may not be transcriptionally linked to one another, they are still functionally linked. Therefore, the 145B enediyne-specific cassette is composed of five functionally linked genes and polypeptides, four of which may be expressed as two operons each containing two genes.

Example 11: Analysis of the polypeptides encoded by the 5-gene enediyne-specific cassette

The amino acid sequences of the PKSE, TEBC, UNBL, UNBV, and UNBU protein families from the ten enediyne biosynthetic loci described above were compared to one another by multiple sequence alignment using the Clustal algorithm (Thompson *et al.*, 1994, *Nucleic Acids Res.* 22(2):4673-4680; Higgins *et al.*, 1996, *Methods Enzymol.* 266:383-402; Higgins and Sharp (1988) *Gene Vol.* 73 pp.237-244).

The alignments are shown in Figures 8, 11, 12, 13, and 14, respectively. Where

-87-

applicable, conserved residues or motifs important for the function are highlighted in black and additional features are indicated.

The PKSE family is a family of polyketide synthases that are involved in formation of enediyne warhead structures. Figure 7 summarizes schematically the domain organization of a typical PKSE, showing the position and relative size of the putative domains based on Markov modeling of PKS domains: ketosynthase (KS), acyltransferase (AT), acyl carrier protein (ACP), ketoreductase (KR), dehydratase (DH), and 4'-phosphopantetheinyl transferase (PPTE) activities. Using the calicheamicin PKSE as an example, the full-length PKSE protein is 1919 amino acids in length. As indicated in Figure 8 for the calicheamicin PKSE, the KS domain spans positions 3 to 467 of the PKSE; the AT domain spans positions 482 to 905 of the PKSE; the ACP domain spans positions 939 to 1009 of the PKSE; a small domain of unknown function of approximately 130 amino acids (spanning positions 1025 to 1144 of the PKSE) is present between the ACP and the KR domains; the KR domain spans positions 1153 to 1414 of the PKSE; the DH domain spans positions 1421 to 1563 of the PKSE; a C-terminal 4'-phosphopantetheinyl transferase (PPTE) domain spans positions 1708 to 1914 of the PKSE; a small domain of about 110 amino acids (spanning positions 1591 to 1701 of the PKSE) is present between the DH and the PPTE domains.

The PKSE contains a conserved unusual ACP domain (Figure 9A). This ACP domain contains several conserved residues that are also present in the well-characterized ACP of the actinorhodin type II PKS (PDBid:1AF8 in Figure 9B). The most important conserved residue is the serine residue to which a 4'-phosphopantetheine prosthetic group is covalently attached (corresponding to Ser-42 of 1AF8). In addition to Ser-42, several surface-exposed charged residues are conserved, namely Glu-20, Asp-37, and Glu-84 (highlighted in the alignment of Figure 9A and highlighted and labeled in the three dimensional structure shown in Figure 9B). Several buried uncharged or non-polar residues that may be important in stabilizing the overall fold of the ACP domain are also conserved, namely Leu-14, Val-15, Gly-57, Pro-71, Ala-83, and Ala-85 (highlighted in the alignment and three dimensional structure shown in Figure 9). Interestingly, the conserved serine (Ser-42) is almost always immediately preceded by another serine in the ACP domains of PKSEs. As shown in Figure 8, nine of the ten PKSE members contain this double serine arrangement, the only exception being that from the 132H locus in which the first of the

-88-

serine is replaced by a threonine. Therefore, PKSEs contain ACP domains with two potential hydroxyl-containing residues in close proximity to one another. These ACPs may carry two 4'-phosphopantetheine prosthetic groups. The positioning of the KR and DH domains after the ACP is unusual among PKSs, but is described in one of the three PKS-like components of the eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) biosynthetic machinery (Metz et al. (2001) *Science* Vol. 293 pp. 290-293). The unusual domain organization shared by the PKSE genes of the invention and the PKS-like synthetase involved in synthesis of polyunsaturated fatty acids suggests that enediynes warhead formation involves intermediates similar to those generated during assembly of polyunsaturated fatty acids.

The presence of an unusual ACP domain in the PKSE, and the absence of any obvious 4'-phosphopantetheinyl transferase or holo-ACP synthase (involved in phosphopantetheinyl transfer onto the conserved serine of the ACP) common to enediyne biosynthetic loci led us to search for the presence of a 4'-phosphopantetheinyl transferase. We examined the conserved domains of the PKSE whose functions were unaccounted for as well as the UNBL, UNBV, and UNBU polypeptides in more detail and determined that the PPTE domain was a 4'-phosphopantetheinyl transferase.

The C-terminal domains of the PKSEs from the biosynthetic loci of three known enediynes, namely neocarzinostatin (NEOC, aa 1620-1977), calicheamicin (CALI, aa 1562-1919) and macromomycin (MACR, aa 1582-1936), were analyzed for their folding using secondary structure predictions and solvation potential information (Kelley *et al.* (2000) *J. Mol. Biol.* Vol. 299 pp. 499-520). Comparison searches using a database of known 3-D structures of proteins revealed similarities between the C-terminal domains of the PKSEs and Sfp, the 4'-phosphopantetheinyl transferase from the *Bacillus subtilis* surfactin biosynthetic locus (Reuter *et al.* (1999) *EMBO* Vol. 18 pp. 6823-6831). The alignment shown in Figure 10A indicates the predicted secondary structures of all three C-terminal PKSE domains (PPTE domains) along with the X-ray crystallography-determined secondary structure of Sfp (PDB id: 1QR0). Alpha-helices are indicated by rectangles and β -sheets by arrows.

An overall conservation of secondary structure over the entire length of the proteins is evident. All major structural constituents of Sfp, namely α -helices α 1- α 5 and

-89-

β -sheets β 2- β 4 and β 8 are also present in PPTE domains. Similar to Sfp, the PPTE domains are predicted to have an intramolecular 2-fold pseudosymmetry.

The loop formed between α 5 and β 7 in Sfp is not present in the PPTE domains. It is believed that this region of Sfp is in part responsible for ACP recognition and contributes to the broad substrate specificity observed for this enzyme. The size of this loop appears to vary among phosphopantetheinyl transferases, as the EntD enzyme, which exhibits a greater ACP substrate specificity than Sfp, has a region between α 5 and β 7 structures shorter than that of Sfp but longer than that found in the PPTE domains. The short α 5/ β 7 loop region found in the PPTE domains may reflect the need
10 for a specific interaction with the rather unusual ACP domain found in the PKSE enzymes. Residues conserved in all phosphopantetheinyl transferases and shown in Sfp to make contacts with the CoA substrate and Mg^{++} cofactor are also conserved in the PPTE domains (highlighted in Figure 10A).

Referring to Figure 10B, Sfp residues Lys-28 and Lys-31 make salt bridges with the 3'-phosphate of CoA and are not found in the PPTE domains; however, a similar interaction could be provided by the corresponding conserved residue Arg-26. Sfp Thr-44 makes a hydrogen bond and His-90 a salt bridge with the 3'-phosphate of CoA; similar hydrogen bonding potential is provided by the conserved serine found at the corresponding position 44 of the PPTE domains, while the histidine 90 residue is
20 absolutely conserved in all three PPTE domains.

Sfp amino acid residues 73-76 hold in place the adenine base of CoA. The main chain carbonyl of Tyr-73 forms a hydrogen bond with the adenine amino group and residues Gly-74, Lys-75 and Pro-76 hold firmly in place the adenine ring. In the PPTE domains, a conserved aspartic acid that may form a salt bridge with the adenine amino group is substituted for Tyr-73 and a conserved arginine residue is substituted for Lys-75. The remaining two residues, Gly-74 and Pro-76, are also found in the PPTE domains.

Sfp residues Ser-89 and His-90 interact via hydrogen bonding and salt bridging with the α -phosphate of the CoA substrate. Similarly, Lys-155 in helix α 5 interacts with
30 the CoA α -phosphate. The His-90 and Lys-155 residues are highly conserved in the PPTE domains whereas Ser-89 is found only in the neocarzinostatin PPTE domain.

-90-

Sfp residues Asp-107, Glu-109 in the β 4 sheet and Glu-151 in the α 5 helix participate in the complexation of a metal ion (presumably Mg^{++}) together with the α and β phosphates of the CoA pyrophosphate and a water molecule. All three residues are also conserved in PPTE domains. Importantly, Asp-107 was altered by mutagenesis in Sfp and shown to be critical for catalytic activity but not for CoA binding of the protein suggesting the Mg^{++} ion is important for catalysis (Quadri *et al.*, 1998, Biochemistry, Vol. 37, 1585-1595).

In the Sfp protein, residue Glu-127 salt-bridges the amino group of Lys-150. In the PPTE domains, a Glu/Asp residue is found at the corresponding position 127, 10 whereas Lys-150 is not conserved. Since Glu-127 is highly conserved in the PPTE domains, it is conceivable that the role of Lys-150 is served by other basic residues in the vicinity, namely the conserved arginine at the corresponding position 145. Residue Trp-147, conserved in all phosphopantetheinyl transferases and shown to be critical for catalytic activity, is also present in all three PPTE domains (Quadri *et al.*, 1998, Biochemistry, Vol. 37, 1585-1595).

The presence of a phosphopantetheinyl domain (PPTE) in the C-terminal part of the PKSE enediynes warhead PKS is reminiscent of the 4'-phosphopantetheinyl domain found in the yeast fatty acid synthase (FAS) complex, where it resides in the C-terminal region of the FAS α subunit. FAS is capable of auto-pantetheinylation resulting in a 20 post-translational autoactivation of this enzyme (Fichtlscherer *et al.*, 2000, Eur. J. Biochem., Vol. 267, 2666-2671). In a similar manner, the PKSE warhead PKSs are likely to be capable of auto-pantetheinylation and activation of their ACP domains before proceeding to the iterative synthesis of the polyunsaturated polyketide intermediate forming the enediyne core.

The ACP and KR domains of the PKSEs are separated by approximately 130 amino acids. The presence of a considerable number of invariable residues within this stretch of amino acids suggests that the putative domain formed by these 130 amino acids has a functional role. The putative domain may serve a structural role, for example as a protein-protein interaction domain or it may form a cleft adjacent to the 30 ACP that acts as a "chain length factor" for the growing polyketide chain. A search of NCBI's Conserved Domain Database with Reverse Position Specific BLAST revealed several short stretches of homology to proteins that bind substrates such as ATP, AMP, NAD(P), as well as folates and double stranded RNA (adenosine deaminase). Thus,

-91-

the putative domain may adopt a structure accommodating an adenosine or adenosine-like structure and serve as a cofactor-binding site. Alternatively, the domain might interact with the adenosine moiety of coenzyme A (CoA). As such, the physical proximity of the CoA to the ACP domain may facilitate the phosphopantetheinylation of the ACP. Yet another possibility is that a molecule of CoA is noncovalently-bound to the putative domain downstream of the ACP via its adenosine moiety and its phosphopantetheinyl tail protrudes out from the enzyme, as would the phosphopantetheinyl tail on the holo-ACP. Alternatively, the PPTe domain can carry a molecule of noncovalently-bound CoA. Thus, it is expected that KS carries out several iterations of condensation reactions involving the transfer of an acetyl group from an acetyl-ACP-thioester to a growing acyl-CoA chain that is non-covalently bound to the enzyme. The proposed scenario explains the presence of the TEBC, an acyl-CoA thioesterase rather than a "conventional" PKS-type thioesterase: the full-length polyketide chain generated by the PKSE is not tethered to the holo-ACP, but rather to a non-covalently bound CoA and the TEBC hydrolyzes the thioester bond of a polyketide-CoA to release the full-length polyketide and CoA. A CoA-activated thioester may render the polyketide more accessible to auxiliary enzymes involved in cyclization and acetylation prior to or concomitant to hydrolytic release by TEBC.

Figure 11 is a Clustal amino acid alignment showing the relationship between the TEBC family of proteins and the enzyme 4-hydroxybenzoyl-CoA thioesterase (1BVQ) of *Pseudomonas* sp. Strain CBS-3 for which the crystal structure has been previously determined (Benning *et al.* (1998) *J. Biol. Chem.* Vol. 273 pp. 33572-33579). The black bars highlight the three regions of conservation believed to play important roles in the catalysis for 4-hydroxybenzoyl-CoA thioesterase. Homology between the TEBC family of proteins and 1BVQ is concentrated in these three highlighted regions.

Figure 12 is a Clustal amino acid alignment of the UNBL family of proteins. The UNBL family of proteins represents a novel group of conserved proteins that are unique to enediyne biosynthetic loci. The UNBL proteins are rich in basic residues and contain several conserved or invariant histidine residues. Besides the PKSE and TEBC proteins, the UNBL proteins are the only other proteins predicted by the PSORT program (Nakai *et al.* (1999) *Trends Biochem. Sci.* Vol. 24 pp. 34-36) to be cytosolic that are encoded by the enediyne warhead gene cassette and thus represent the best

-92-

candidates for the acetylenase activity that is required to introduce triple bonds into the warhead structure.

Figure 13 is a Clustal amino acid alignment of the UNBV family of proteins. PSORT analysis of the UNBV family of proteins predicts that they are secreted proteins. The approximate position of the putative cleavable N-terminal signal sequence is indicated above the alignment. The UNBV proteins display considerable amino acid conservation but do not have any known homologue. Thus, the UNBV family of proteins represents a novel group of conserved proteins of unknown function that are unique to enediynes biosynthetic loci.

10 Figure 14 is a Clustal amino acid alignment of the UNBU family of proteins. PSORT analysis of the UNBU family of proteins predicts that they are integral membrane proteins with seven or eight putative membrane-spanning alpha helices (indicated by dashes in Figure 14). The UNBU proteins display considerable amino acid conservation but do not have any known homologue. The UNBU family of proteins represents a novel group of conserved proteins that are unique to enediynes biosynthetic loci.

UNBU is likely involved in transport of the enediynes across the cell membrane. UNBU may also contribute, in part, to the biochemistry involved in the completion of the warhead. In the case of chromoprotein enediynes, the apoprotein carries its own
20 cleavable N-terminal signal sequence and is probably exported independently of the chromoprotein by the general protein secretion machinery. Formation of the bioactive warhead, export, and binding of the chromophore and protein component must occur in and around the cell membrane to minimize damage to the producer and to maximize the stability of the natural product. UNBV is predicted to be an extracellular protein. UNBV may finalize or stabilize the warhead structure. UNBV may act in close association with the extracellularly exposed portion(s) of UNBU.

To date, we have sequenced over ten enediynes biosynthetic loci that contain the 5-gene cassette made up of PKSE, TEBC, UNBL, UNBV, and UNBU genes. In all cases, the PKSE and TEBC genes are adjacent to one another and the TEBC gene is
30 always downstream of the PKSE gene. Moreover, these two genes are usually, if not always, translationally coupled. These observations suggest that the expression of the PKSE and TEBC genes is tightly coordinated and that their gene products, *i.e.*, polypeptides, act together. Likewise, the UNBV and UNBU genes are always adjacent

-93-

to one another and the UNBU gene is always downstream of the UNBV gene. Moreover, these two genes are usually, if not always, translationally coupled. These observations suggest that the expression of the UNBV and UNBU genes is tightly coordinated and that their gene products, *i.e.*, polypeptides, act together.

Example 12: Common mechanism for the biosynthesis of enediyne warheads

Without intending to be limited to any particular biosynthetic scheme or mechanism of action, the genes and proteins of the present invention can explain formation of enediyne warheads in both chromoprotein enediynes and non-
10 chromoprotein enediynes.

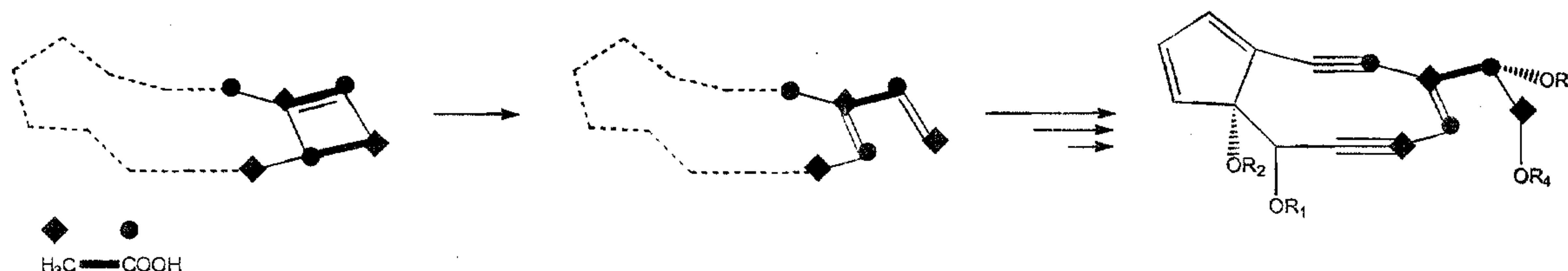
The PKSE is proposed to generate a highly conjugated polyunsaturated hepta/octaketide intermediate in a manner analogous to the action of polyunsaturated fatty acid synthases (PUFAs). The polyunsaturated fatty acyl intermediate is then modified by tailoring enzymes involving one or more of UNBL, UNBU and UNBV to introduce the acetylene bonds and form the ring structure(s). The conserved auxiliary proteins UNBL, UNBU and UNBV are expected to be involved in modulating iterations performed by the PKSE, or in subsequent transformations to produce the enediyne core in a manner analogous to action of lovastatin monaketide synthase, a fungal
20 iterative type I polyketide synthase that is able to perform different oxidative/reductive chemistry at each iteration with the aid of at least one auxiliary protein (Kennedy et al., 1999, Science Vol. 284 pp. 1368-1372).

The acetate enrichment pattern of the enediyne moiety of esperamicin and dynemicin suggest that both are derived from an intact heptaketide/octaketide. There has been suggestion that esperamicin and dynemicin may share a common precursor (Lam *et. al.*, J. Am. Chem. Soc. 1993, Vol. 115 pp. 12340). However, in the case of neocarzinostatin, representative of other chromoprotein enediynes, incorporation studies investigating carbon-carbon connectivities revealing that the final enediyne core contains uncoupled acetate atoms (Hensens et al., 1989 JACS, Vol. 111, pp. 3295-3299), and other studies regarding polyacetylene biosynthesis (Hensens *et. al.*, *supra*),
30 suggest that the chromoprotein enediyne precursors are distinct from those of the non-chromoprotein enediynes. Thus, prior art studies regarding formation of the enediyne core teach away from the present invention that genes and proteins common to both

-94-

chromoprotein enediynes and non-chromoprotein enediynes are responsible for formation of the warhead in both classes of enediynes.

We propose that skeletal rearrangements may account for the distinct chromoprotein/nonchromoprotein enediyne labeling patterns. For instance, thermal electrocyclic rearrangement of an intermediate cyclobutene to a 1,3 diene could result in an isotopic labeling pattern consistent with that which has been reported.



Accordingly, the warhead precursor in the formation of neocarzinostatin could be a heptaketide, similar to that proposed for the other classes of enediynes. Since calicheamicin and esperimicin do not contain any uncoupled acetates, the common unsaturated polyketidic precursor must rearrange differently from the chromoprotein class. However, the proposed biosynthetic scheme is consistent with one aspect of the present invention, namely that warhead formation in all enediynes involves common genes, proteins and common precursors.

Example 13: Heterologous expression of genes and proteins of the calicheamicin enediyne cassette

Escherichia coli was used as a general host for routine subcloning.

Streptomyces lividans TK24 was used as a heterologous expression host. The plasmid pECO1202 was derived from plasmid pANT1202 (Desanti, C. L. 2000. The molecular biology of the *Streptomyces snp* Locus, 262 pp., Ph.D dissertation, Ohio State Univ., Columbus, OH) by deleting the *KpnI* site in the multi-cloning site (MCS). pECO1202RBS contains a DNA sequence encoding a putative ribosome-binding site (AGGAG) introduced just upstream of the *ClaI* site located in the MCS of pECO1202.

E. coli strains carrying plasmids were grown in Luria-Bertani (LB) medium and were selected with appropriate antibiotics. *S. lividans* TK24 strains were grown on R2YE medium. (Kieser, T. *et al.*, Practical *Streptomyces* Genetics, The John Innes Foundation, Norwich, United Kingdom, 2000).

-95-

Preparation of *S. lividans* TK24 protoplasts was carried out using the standard protocols. (Kieser et al., *supra*). Polyethylene glycol-induced protoplast transformation was carried out with 1 µg DNA per transformation. After protoplast regeneration on R5 agar medium for 16 h at 30 °C, transformants were selected by overlaying each R5 plate with 50 µg/ml apramycin solutions. Transformants were grown in 50 ml flasks containing R2YE medium plus apramycin for seven days.

SDS-PAGE and Western-blotting were carried out by standard procedures (Sambrook, J. et al. 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.). Penta-His™ antibody was obtained
10 from Qiagen. Western blots were performed using the ECL detection kit from Amersham Pharmacia biotech using the manufacturer's suggested protocols. One milliliter of seven-day *S. lividans* culture was centrifuged and mycelium resuspended in cold extraction buffer (0.1M Tris-HCl, pH 7.6, 10 mM MgCl₂ and 1 mM PMSF). The mycelium was sonicated 4 x 20 sec on ice with 1 min intervals to release soluble protein. After 10 min centrifugation at 20,000g, the supernatant and pellet fractions were diluted with sample buffer and subjected to SDS-PAGE and Western-blotting analysis.

DNA manipulations used in construction of expression plasmids were carried out using standard methods (Sambrook, J. et al., *supra*). The plasmid pECO1202 was
20 used as the parent plasmid. Cosmid 061CR, carrying the calicheamicin biosynthetic gene locus was digested with *MfeI*, and the restriction fragments were made blunt ended by treatment with the Klenow fragment of DNA polymerase I. Upon additional digestion with *BglII* after phenol extraction and ethanol precipitation, the resulting 11.5 kb blunt-ended, *BglII* fragment was gel purified and cloned into pECO1202 (previously digested with *EcoRI*, made blunt ended by treatment with Klenow fragment of polymerase I, then digested with *BamHI*), to yield pECO1202-CALI-1, as shown in Figure 15.

PCR was carried out on a PTC-100 programmable thermal controller (MJ research) with *Pfu*™ polymerase and buffer from Stratagene. A typical PCR mixture
30 consisted of 10 ng of template DNA, 20 µM dNTPs, 5% dimethyl sulfoxide, 2U of *Pfu* polymerase, 1 µM primers, and 1X buffer in a final volume of 50 µl. The PCR temperature program was the following: initial denaturation at 94 °C for 2 min, 30 cycles of 45 sec at 94 °C, 1 min at 55 °C, and 2 min at 72 °C, followed by an additional 7 min

-96-

at 72 °C. A PCR product amplified by primer 1402, 5' –
 GAGTTGTATCG**ATG**AGCAGGATCGCCGTCGTCGGC –3' [containing *Cla* I site (*italic*)
 and the start codon of PKSE gene (**bold**)], and primer 1420,
 5'GTAGCCGGCCGCCTCCGGCC (corresponding to the nucleotide sequence 940 to
 959 bp of PKSE), was digested with *Cla*I and *Nhe*I and gel purified. This fragment was
 then cloned into *Cla*I, *Nhe*I digested pECO1202-CALI-1 to yield pECO1202-CALI-5
 (Figure 16).

PCR products were amplified by primer 1421, 5'-
 GACCTGCCGTACACCGTCTCC –3' (corresponding to the nucleotide sequence 5367
 10 to 5387 bp of PKSE), and primer 1403, 5'-
 CCCAAGCT**TCAGTGGTGGTGGTGGTGGT**GGCCCTGCCCCACCGTGGCCGAC-
 3'[containing a His Tag (underlined), *Hind*III site (*italic*) and stop codon of TEBC (**bold**)],
 or primer 1500, 5'- CCCAAGCT**TC**ACCCCTGCCCCACCGTGGCCGAC- 3' (containing
*Hind*III site (*italic*) and stop codon (**bold**) of TEBC). These PCR products were digested
 with *Hind*III and *Pst*I, gel purified, and then cloned into *Hind*III, *Pst*I digested pECO1205
 to yield pECO1202-CALI-2 (with HisTag) and pECO1202-CALI-3 (without HisTag),
 respectively (Figure 16).

The *Cla*I and *Hind* III fragments from pECO1202-CALI-2 and pECO1202-CALI-3
 were cloned into pECO1202RBS to yield pECO1202-CALI-6 (with HisTag) and
 20 pECO1202-CALI-7 (without HisTag), respectively, as shown in Figure 16.

Six transformants of *S. lividans* TK24 harboring pECO1202-CALI-2 were
 analyzed for expression of the His-tagged TEBC protein. Referring to Figure 17, lane
 M provides molecular weight markers; lanes 1 to 6 represent crude extracts of
 independent transformants of *S. lividans* TK24 harboring pECO1202-CALI-2; lane 7
 represents a crude extract of *S. lividans* TK24 harboring pECO1202-CALI-4; and lane 8
 represents a crude extract of *S. lividans* TK24 harboring pECO1202 (control). TEBC
 protein expression was detected in four pECO1202-CALI-2 transformants by Western
 blotting using an antibody that recognizes the His-tag (lanes 2, 3, 5, 6). TEBC protein
 expression was also observed in the transformant of *S. lividans* TK24 harboring
 30 pECO1202-CALI-4 (lane 7).

As shown in Figure 18, the TEBC protein was expressed as a soluble protein in
S. lividans although the pellet fraction also contains TEBC protein, perhaps reflecting
 insoluble protein or incomplete lysis of *S. lividans* by the sonication procedure used.

-97-

Figure 18 provides an analysis of His-tagged TEBC protein derived from recombinant *S. lividans* TK24 by immunoblotting. The soluble and insoluble protein fractions of *S. lividans* transformants were separated by 12% SDS-polyacrylamide gel electrophoresis, blotted to PVDF membrane, and detected with the Penta-His antibody. Referring to Figure 18, lane M provides molecular weight markers; lane 1 to 6 represent soluble (S) and pellet (P) protein fractions of independent transformants of *S. lividans* TK24 harboring pECO1202-CALI-2; lane C represents protein fractions of *S. lividans* TK24 harboring pECO1202 (control).

10 Example 14: Disruption of the PKSE gene abolishes production of enediynes

To confirm that the PKSE is critical to the biosynthesis of enediynes, the PKSE gene of the calicheamicin producer, *M. echinospora*, was disrupted by introduction of an apramycin selectable marker as follows. *M. echinospora* was grown with a 1:100 fresh inoculum in 50 mL MS medium (Kieser *et al.*, *supra*) supplemented with 5 % PEG 8000 and 5 mM MgCl₂ for 24 – 36 h and 6 h prior to harvest, 0.5 % glycine was added. The digest of the cell wall was accomplished via published procedures with the exception that 5 mg mL⁻¹ lysozyme and 2000 U mutanolysin were used. Under these conditions, protoplast formation was complete within 30-60 min after which the mixture was filtered twice through cotton wool. Transformation was accomplished via typical methodology (Kieser *et al.*, *supra*) with a 1:1 mixture of T-buffer and PEG 2000 containing up to 10 µg of alkaline denatured DNA per transformation. The protoplasts were then plated on R2YE plates supplemented with 10 mg L⁻¹ CoCl₂ and submitted to antibiotic pressure (70 µg mL⁻¹ apramycin) after 3 – 4 days. To date, all attempts to use methods other than protoplast chemical transformation (e.g. phage transduction, conjugation and electroporation) have failed to introduce DNA into *M. echinospora*. Low transformation efficiencies were observed in all calicheamicin-producing *Micromonospora* strains tested, including those developed from strain improvement efforts. In comparison to other actinomycetes, *M. echinospora* protoplast regeneration was found to be slow (~ 4 weeks). Moreover, integration into the locus requires homologous fragments exceeding 3 kb in size as constructs containing PKSE fragments (or other calicheamicin gene fragments) smaller than 3 kb all failed to integrate into the chromosome (data not shown).

20

30

-98-

Nine independent apramycin-resistant PKSE disruption clones were obtained. All nine isolates mapped consistently with the expected PKSE gene disruption both by PCR fragment amplification and by Southern hybridization (data not shown). All nine PKSE disruption mutants and two parental controls were subsequently tested in parallel for calicheamicin production. Extracts from these strains were prepared as follows. Fresh *M. echinospora* cells grown in R2YE were inoculated 1:100 in 10 mL medium E (Kieser *et al.*, *supra*) in stoppered 25 ml glass tubes containing a 4 cm stainless coil spring for better aeration and incubated on an orbital shaker with 230 rpm at 28 °C for one to three weeks. A 600 µl aliquot was removed at various time points, extracted with an equal volume of EtOAc and centrifuged at 10000 xg for 5 min in a benchtop centrifuge. The supernatant was concentrated to dryness, the pellet redissolved in 200 µl acetonitrile, centrifuged again and the supernatant removed, concentrated to dryness and the residual material finally dissolved in 10 µl acetonitrile. One µl of this solution was utilized for the bioassays and the remaining 8 µl aliquot was utilized for analysis by HPLC (Ultrasphere-ODS™ chromatography, 5 µm, 4.6 mm x 250 mm, 55:45 CH₃CN-0.2 NH₄OAc, pH 6.0, 1.0 mL min⁻¹, 280 nm detection). A typical *M. echinospora* fermentation contains a mixture of calicheamicins that are resolved by HPLC - γ_1^I (retention time - 7 min, ~60%), δ_1^I (retention time - 5.7 min, ~30%), and α_3^I (retention time - 3.8 min, ~10%) – and all of these calicheamicin components contribute to bioassay activities. The best production was found to occur during late log or early stationary phase growth. The estimate of calicheamicin production by parental *M. echinospora* is 0.78-0.85 mg mL⁻¹. Extracts were analyzed by i) the biological induction assay, a modified prophage induction assay used in the original discovery of the calicheamicins (Greenstein *et al.* (1986) *Antimicrob. Agents Chemotherap.* Vol. 29, 861); ii) the molecular break light assay, a DNA-cleavage assay based upon intramolecular fluorescence quenching optimized for DNA-cleavage by enediynes (in which fM calicheamicin concentrations are detectable) (Biggins *et al.* (2000) *Proc. Natl. Acad. Sci. USA* Vol. 97, 13537); and iii) high-performance liquid chromatography (HPLC) (described above). As expected, all three methods revealed that the parental *M. echinospora* fermentations produced 0.5-0.8 mg L⁻¹. In contrast, the PKSE gene disruption mutant strains were both devoid of any calicheamicin, known calicheamicin derivatives and/or enediyne activity by all three methods of detection. The elimination of calicheamicin production brought about by disruption of the PKSE gene indicates

-99-

that it provides an essential activity for biosynthesis of calicheamicin. Based on the presence of the PKSE in all enediyne biosynthetic loci sequenced to date and on their overall conservation, it is expected that PKSEs fulfill the same, essential function in the biosynthesis of all enediyne structures.

The present invention is not to be limited in scope by the specific embodiments described herein. Indeed, various modifications of the invention in addition to those described herein will become apparent to those skilled in the art from the foregoing description and the accompanying figures. Such modifications are intended to fall
10 within the scope of the appended claims.

It is further to be understood that all sizes and all molecular weight or mass values are approximate, and are provided for description.

Some open reading frames listed herein initiate with non-standard initiation codons (e.g. GTG – Valine or TTG - Leucine) rather than the standard initiation codon ATG, namely SEQ ID NOS: 3, 13, 17 and 19 of CA 2,387,401, SEQ ID NOS: 7, 15, 19 and 21 of CA 2,445,687, SEQ ID NOS: 3, 7, 9, 11, 17, 19 and 21 of CA 2,445,692, SEQ ID NOS: 7, 9, 17 and 19 of CA 2,444,802 and SEQ ID NOS: 7, 9, 15, 17 and 21 of CA 2,444,812. All ORFs are listed with M, V or L amino acids at the amino-terminal position to indicate the specificity of the first codon of the ORF. It is expected, however,
20 that in all cases the biosynthesized protein will contain a methionine residue, and more specifically a formylmethionine residue, at the amino terminal position, in keeping with the widely accepted principle that protein synthesis in bacteria initiates with methionine (formylmethionine) even when the encoding gene specifies a non-standard initiation codon (e.g. Stryer, Biochemistry 3rd edition, 1998, W.H. Freeman and Co., New York, pp. 752-754).

SEQUENCE LISTING

APPLICANT NAME: ECOPIA BIOSCIENCES INC.
 Farnet, Chris
 Staffa, Alfredo
 Zazopoulos, Emmanuel

TITLE OF INVENTION: COMPOSITIONS, METHODS AND SYSTEMS FOR THE DISCOVERY
 OF ENEDIYNE NATURAL PRODUCTS

NUMBER OF SEQUENCES: 24

CORRESPONDANCE ADDRESS: 7290 Frederick-Banting
 Saint-Laurent, Quebec, H4S 2A1

COMPUTER READABLE FORM:
 SOFTWARE: PatentIn version 3.0

CURRENT APPLICATION DATA:
 APPLICATION NUMBER: CA 2,445,687
 FILING DATE: 2002-05-21

ATTORNEY/PATENT AGENT INFORMATION
 NAME: Ywe J. Looper
 REFERENCE NUMBER: 10961

FILE REFERENCE: 3011-11CA

INFORMATION FOR SEQ ID NO: 1
 LENGTH: 154
 TYPE: PRT
 STRANDEDNESS: Unknown
 TOPOLOGY: Unknown

ORGANISM: consensus sequence

SEQUENCE: 1

Val Thr Met Ala Asp Tyr Phe Glu Tyr Arg His Thr Val Gly Phe Glu
 1 5 10 15

Glu Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp
 20 25 30

Gln Gly Arg Cys Arg Glu Leu Phe Leu Lys Glu Lys Ala Pro Glu Val
 35 40 45

Leu Ala Asp Leu Arg Asp Asp Leu Lys Leu Phe Thr Leu Lys Val Asp
 50 55 60

Cys Glu Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu Ser Ile Arg
 65 70 75 80

Met Arg Leu Ala Glu Leu Thr Gln Thr Gln Leu Glu Phe Thr Phe Asp
 85 90 95

Tyr Val Arg Leu Gly Gly Asp Gly Val Glu Thr Leu Val Ala Arg Gly
 100 105 110

Arg Gln Arg Ile Ala Cys Met Arg Gly Pro Asn Thr Ala Thr Val Pro
 115 120 125

Ala Arg Val Pro Glu Ala Leu Arg Arg Ala Leu Ala Pro Tyr Ala Ala
 130 135 140

Gly Thr Arg Val Leu Ala Gly Arg Gly Ala
 145 150

INFORMATION FOR SEQ ID NO: 2
 LENGTH: 162
 TYPE: PRT
 STRANDEDNESS:
 TOPOLOGY: Unknown

ORGANISM: Streptomyces macromyceticus

SEQUENCE: 2

Met Ser Gly Ser Ala Asp Ser Leu Gly Tyr Phe Glu Tyr Arg His Thr
 1 5 10 15

Val Ala Phe Ala Glu Thr Asp Leu Ala Gly Ser Ala Asp Tyr Val Asn
 20 25 30

Tyr Leu Gln Trp Gln Ala Arg Cys Arg Gln Leu Phe Leu Arg Gln Thr
 35 40 45

Ala Phe Gly Thr Val Leu Asp Asp Asp Leu Asp Ala Gly His Ala Asp
 50 55 60

Leu Arg Leu Phe Thr Leu Gln Val Glu Cys Glu Leu Phe Glu Ala Val
 65 70 75 80

Ser Ala Leu Asp Arg Leu Ala Ile Arg Met Arg Val Ala Glu Ile Gly
 85 90 95

His Thr Gln Phe Asp Leu Thr Phe Asp Tyr Val Lys Gly Ala Gly Glu
 100 105 110

Gly Asp Val Pro Val Ala Arg Gly Arg Gln Arg Val Val Cys Leu Arg
 115 120 125

Gly Pro Ala Gly Ala Pro Val Pro Ala Leu Ile Pro Asp Ala Leu Ala
 130 135 140

Gln Ala Leu Ala Pro Tyr Ala Ala Gly Thr Arg Pro Leu Ala Gly Arg
 145 150 155 160

His Thr

INFORMATION FOR SEQ ID NO: 3
 LENGTH: 489
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: Streptomyces macromyceticus

SEQUENCE: 3

```

atgagcggca ggcgggacag cctcgggtac ttcgagtacc ggcacacggt cgccttcgcc      60
gagaccgatc tcgcggggcag cgccgactac gtgaactacc tccagtggca ggcacgttgc      120
cggcagttgt tcctgcgcca gacggcggtc gggacgggtcc tcgacgacga cctggacgcc      180
gggcacgccc acttgaggct gttcacgctg caggtcgagt gcgagctctt cgaagcggtc      240
tcggcactcg accgcctggc catccggatg cgggtggccg agatcggaca cacacagttc      300
gacttgacgt tcgactacgt caagggggcga ggggagggcg acgtaccggt ggctcgcggc      360
aggcagcgcg tcgtgtgtct gcgcggggcg gccggcgccc ccgtcccggc cctgatcccc      420
gacgcgctgg cacaagcgct ggcgccctac gcggccggga cccggccggt ggcagggagg      480
catacatga                                     489

```

INFORMATION FOR SEQ ID NO: 4

LENGTH: 157

TYPE: PRT

STRANDEDNESS:

TOPOLOGY: Unknown

ORGANISM: Streptomyces macromyceticus

SEQUENCE: 4

```

Met Thr Thr Thr Ala Thr Thr Asp Tyr Phe Glu Tyr Arg His Thr Val
1          5          10          15
Gly Phe Glu Glu Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr
20          25          30
Leu Arg Trp Gln Gly Arg Cys Arg Glu Leu Phe Leu Lys Gln Lys Ala
35          40          45
Pro Ala Val Leu Ala Asp Val Gln Glu Asp Leu Lys Leu Phe Thr Leu
50          55          60
Lys Val Asp Cys Glu Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu
65          70          75          80
Ser Ile Arg Met Arg Leu Ala Glu Gln Ala Gln Thr Gln Leu Glu Phe
85          90          95
Thr Phe Asp Tyr Val Lys Val Thr Glu Asp Gly Thr Glu Thr Leu Val
100          105          110
Ala Arg Gly Lys Gln Arg Ile Ala Cys Met Arg Gly Pro Asn Thr Ala
115          120          125
Thr Val Pro Ser Leu Ile Pro Asp Ala Leu Ala Gln Ala Leu Ala Pro
130          135          140
Tyr Ala Thr Gln Asn Arg Ser Leu Val Gly Arg Ala Ala
145          150          155

```

INFORMATION FOR SEQ ID NO: 5
 LENGTH: 474
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: *Streptomyces macromyceticus*

SEQUENCE: 5

```

atgacgacca ccgcgacgac cgactacttc gaggaccggc acaccgttgg cttcgaggag      60
accaacctgg tgggcaacgt gtactacgtg aactacctcc ggtggcaggg acgctgccgg      120
gagctgttcc tcaagcagaa ggcacccgcg gtccctcgccg acgtccagga ggacctcaag      180
ctcttcaccc tgaaggtcga ctgcgagttc ttcgccgaga tcacggcctt cgacgagctg      240
tcgatccgga tgcggctggc cgagcaggcg cagaccagc tggagttcac cttcgactac      300
gtcaaggtga ccgaggacgg cacggagacc ctggtggccc gcggcaagca gcggatcgcc      360
tgcattcggg gtccgaacac ggccaccgtc ccctcgctga tccccgacgc cctcgcccag      420
gcgctggcgc cgtacgccac ccagaaccgc tcgctcgctg gccggggccgc ctga          474
  
```

INFORMATION FOR SEQ ID NO: 6
 LENGTH: 148
 TYPE: PRT
 STRANDEDNESS:
 TOPOLOGY: Unknown

ORGANISM: *Micromonospora echinospora calichensis*

SEQUENCE: 6

```

Val Ser Met Pro Arg Tyr Tyr Glu Tyr Arg His Val Val Gly Phe Glu
1              5              10              15
Glu Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp
20              25              30
Gln Gly Arg Cys Arg Glu Met Phe Leu Tyr Glu His Ala Pro Glu Ile
35              40              45
Leu Asp Glu Leu Arg Ala Asp Leu Lys Leu Phe Thr Leu Lys Ala Glu
50              55              60
Cys Glu Phe Phe Ala Glu Leu Ala Pro Phe Asp Arg Leu Ala Val Arg
65              70              75              80
Met Arg Leu Val Glu Leu Thr Gln Thr Gln Met Glu Leu Gly Phe Asp
85              90              95
Tyr Leu Arg Leu Gly Gly Asp Asp Leu Leu Val Ala Arg Gly Arg Gln
100             105             110
Arg Ile Ala Cys Met Arg Gly Pro Asn Gly Arg Thr Glu Pro Val Arg
  
```

115 120 125
 Val Pro Ala Gly Leu Val Arg Ala Phe Ala Pro Phe Arg Ser Ala Thr
 130 135 140
 Val Gly Gln Gly
 145

INFORMATION FOR SEQ ID NO: 7
 LENGTH: 447
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: Micromonospora echinospora calichensis

SEQUENCE: 7

gtgagcatgc cgcgctacta cgagtaccgg cacgtcgtcg gcttcgagga gaccaacctc 60
 gtcggcaacg tgtactacgt caactacctg cgctggcagg gccggtgccg ggagatgttc 120
 ctgtacgagc acgcgccgga gatcctcgac gagctgcgcg ccgacctgaa gctgttcacc 180
 ctcaaggccg agtgcgagtt cttcgccgag ctggcgccgt tcgaccgcct cgcggtccgg 240
 atgcggctgg tcgaactcac ccagaccag atggagctgg gcttcgacta cctgcggctc 300
 ggcgggcgacg atctgctggt cgcccggggg cggcagcgga tcgcgtgcat gcgcgggccc 360
 aacgggcccga ccgagccggt ccgggtgccg gccggcctgg tgcgggcggt cgccccgttc 420
 cggtcggcca cggtggggca ggggtga 447

INFORMATION FOR SEQ ID NO: 8
 LENGTH: 152
 TYPE: PRT
 STRANDEDNESS:
 TOPOLOGY: Unknown

ORGANISM: Streptomyces ghanaensis

SEQUENCE: 8

Met Ala Glu Asp Tyr Phe Glu Tyr Arg His Thr Val Gly Phe Glu Glu
 1 5 10 15
 Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp Gln
 20 25 30
 Gly Arg Cys Arg Glu Leu Phe Leu Gln Gln Lys Ala Pro Glu Val Leu
 35 40 45
 Ala Glu Val Gln Asp Asp Leu Lys Leu Phe Thr Leu Lys Val Asp Cys
 50 55 60
 Glu Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu Ser Ile Arg Met
 65 70 75 80

Arg Leu Ser Glu Leu Gly Gln Thr Gln Leu Glu Phe Ser Phe Asp Tyr
 85 90 95

Val Lys Val Thr Gly Gly Ala Glu Leu Leu Val Ala Arg Gly Arg Gln
 100 105 110

Arg Ile Ala Cys Met Arg Gly Pro Asn Thr Asn Thr Val Pro Ser Arg
 115 120 125

Ile Pro Glu Ala Leu Ala His Ala Leu Glu Pro Tyr Thr Ala His Gly
 130 135 140

Arg Val Pro Thr Gly Arg Ala Ala
 145 150

INFORMATION FOR SEQ ID NO: 9
 LENGTH: 459
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: Streptomyces ghanaensis

SEQUENCE: 9

atggcggaag actacttcga gtaccggcac acggtcgggtt tcgaggagac caacctggtc 60
 ggcaacgtct actacgtgaa ctacctgcgc tggcagggcc ggtgccggga gctcttcctg 120
 cagcagaagg cgccggaggt actggccgag gtgcaggacg acctgaagct gttcacgctg 180
 aaggtggact gcgagttctt cgccgagatc accgccttcg acgagctgtc catccgcatg 240
 cggctgtccg aactggggca gacacagctg gagttctcct tcgactacgt caaggtgacc 300
 ggcgggggcgg agctcctcgt ggctcgcggg cgccagcgga tcgcgtgcat gcgcggaccc 360
 aacaccaaca ccgtgccctc ccgcattccc gagggccttg cccacgccct ggagccgtac 420
 accgcccacg gccgggtgcc gacggggcgt gcggcatga 459

INFORMATION FOR SEQ ID NO: 10
 LENGTH: 153
 TYPE: PRT
 STRANDEDNESS:
 TOPOLOGY: Unknown

ORGANISM: Streptomyces carzinostaticus neocarzinostaticus

SEQUENCE: 10

Met Ser Asp Asp Tyr Phe Glu Tyr Arg His Thr Val Gly Phe Glu Glu
 1 5 10 15

Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp Gln
 20 25 30

Gly Arg Cys Arg Glu Leu Phe Leu Lys Gln Lys Ala Pro Glu Val Leu
 35 40 45

Ala Asp Val Gln Asp Asp Leu Lys Leu Phe Thr Leu Lys Val Asp Cys
50 55 60

Glu Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu Ser Ile Arg Met
65 70 75 80

Arg Leu Ser Asp Phe Gly Gln Thr Gln Leu Glu Phe Thr Phe Asp Tyr
85 90 95

Val Lys Val Asp Glu Asp Gly Gly Glu Thr Leu Val Ala Arg Gly Arg
100 105 110

Gln Arg Val Ala Cys Met Arg Gly Pro Asn Thr Asn Thr Val Pro Ser
115 120 125

Leu Val Pro Glu Ala Leu Val Arg Ala Leu Glu Pro Tyr Gly Ala Gln
130 135 140

Arg Arg Val Leu Pro Gly Arg Thr Ala
145 150

INFORMATION FOR SEQ ID NO: 11
LENGTH: 462
TYPE: DNA
STRANDEDNESS: Double stranded
TOPOLOGY: Unknown

ORGANISM: Streptomyces carzinostaticus neocarzinostaticus

SEQUENCE: 11

atgtcggatg actacttcga gtaccggcac acggtcggct tcgaggaaac caatctggtc 60
ggcaacgtct actacgtgaa ctacctacgc tggcagggac gttgccggga gctgttcctc 120
aagcagaagg caccggaggt cctcgcggac gtacaggacg acctcaagct gttcacgctc 180
aaggtggact gtgagttctt cgccgagatc accgccttcg acgagttgtc catacggatg 240
cggctctccg acttcgggca gaccagttg gagttcacct tcgactacgt caaggtggac 300
gaggacggcg gcgagaccct ggtggcccgg ggccggcagc gggtcgcctg catgcgaggg 360
cccaacacca acacagtgcc ctcaactggc cccgaggcac tgggtccgagc cctcgagccg 420
tacggcgcac agaggcgggt gctgccgggg cggacggcat ga 462

INFORMATION FOR SEQ ID NO: 12
LENGTH: 146
TYPE: PRT
STRANDEDNESS:
TOPOLOGY: Unknown

ORGANISM: Amycolatopsis orientalis

SEQUENCE: 12

Met Ala Asp Tyr Tyr Glu Ile Leu His Thr Val Gly Phe Glu Glu Thr

1	5	10	15												
Asn	Leu	Val	Gly	Asn	Val	Tyr	Tyr	Val	Asn	Tyr	Val	Arg	Trp	Gln	Gly
		20					25					30			
Arg	Cys	Arg	Glu	Met	Phe	Leu	Lys	Glu	Lys	Ala	Pro	Ala	Val	Leu	Glu
	35					40					45				
Glu	Val	Arg	His	Asp	Leu	Lys	Leu	Phe	Thr	Leu	Lys	Val	Asp	Cys	Glu
	50				55						60				
Phe	Tyr	Ala	Glu	Ile	Thr	Ala	Phe	Asp	Glu	Leu	Ser	Ile	Arg	Leu	Arg
65				70					75					80	
Leu	Glu	Glu	Leu	Thr	Gln	Thr	Gln	Ile	Gln	Phe	Thr	Phe	Asp	Tyr	Val
			85				90							95	
His	Leu	Thr	Ala	Glu	Gly	Glu	Arg	Leu	Val	Ala	Arg	Gly	Arg	Gln	Arg
		100					105					110			
Ile	Ala	Cys	Met	Arg	Gly	Pro	Asn	Thr	Ala	Thr	Val	Pro	Ser	Arg	Val
	115					120					125				
Pro	Glu	Gln	Leu	Arg	Glu	Ala	Leu	Ala	Pro	Tyr	Ala	Val	Asp	Gly	Lys
	130					135					140				
Gly	Glu														
145															

INFORMATION FOR SEQ ID NO: 13
 LENGTH: 441
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: Amycolatopsis orientalis

SEQUENCE: 13

atggccgact actacgagat cctccacacg gtcggattcg aagagaccaa cctggtgggc	60
aacgtctact acgtgaacta cgtgcgctgg cagggccggt gccgcgagat gttcctgaag	120
gagaaggcgc ccgcggtgct cgaagaggtc cgccacgacc tcaagctggt cacgctcaag	180
gtggactgcg agttctacgc ggagatcacc gcgttcgacg agctgtccat ccggctgcgg	240
ctggaggagc tgaccacagac ccagatccag ttcaccttcg actacgtcca cctcaccgcg	300
gaaggcgagc ggctggtggc ccgcggacgg cagcggatcg cgtgcatgcg cggcccgaac	360
acggccacgg tgcccagccg ggtgcccga cagctgcgtg aggcgctggc cccgtacgcg	420
gtcgacggca agggggaatg a	441

INFORMATION FOR SEQ ID NO: 14
 LENGTH: 158
 TYPE: PRT
 STRANDEDNESS:

TOPOLOGY: Unknown

ORGANISM: Kitasatosporia sp.

SEQUENCE: 14

Val	Thr	Gly	Pro	Asp	Tyr	Tyr	Glu	Tyr	Arg	His	Leu	Val	Gly	Phe	Glu	
1				5					10					15		
Glu	Thr	Asn	Leu	Val	Gly	Asn	Val	Tyr	Tyr	Val	Asn	Tyr	Leu	Arg	Trp	
			20					25					30			
Gln	Gly	Arg	Cys	Arg	Glu	Met	Phe	Leu	Leu	Glu	Lys	Ala	Pro	Glu	Val	
		35					40					45				
Leu	Ala	Asp	Ile	Arg	Ala	Asp	Leu	Lys	Leu	Phe	Thr	Leu	Lys	Val	Asp	
	50					55					60					
Cys	Glu	Phe	Phe	Ala	Glu	Ile	Thr	Ala	Phe	Asp	Glu	Leu	Ser	Ile	Arg	
65					70				75						80	
Met	Arg	Leu	Ala	Asp	Leu	Thr	Gln	Thr	Gln	Val	Ala	Phe	Thr	Phe	Asp	
				85					90						95	
Tyr	Val	Lys	Leu	Gly	Pro	Asp	Gly	Thr	Glu	Tyr	Leu	Val	Ala	Arg	Gly	
			100					105						110		
Gln	Gln	Arg	Val	Ala	Cys	Met	Arg	Gly	Pro	Asn	Thr	Asp	Thr	Arg	Pro	
		115					120					125				
Thr	Arg	Val	Pro	Glu	Pro	Leu	Arg	Leu	Ala	Leu	Glu	Pro	Tyr	Ala	Val	
	130					135					140					
Pro	Ala	Thr	Ala	Pro	Ser	Leu	Thr	Gly	Thr	Thr	Thr	Val	Gly			
145					150					155						

INFORMATION FOR SEQ ID NO: 15

LENGTH: 477

TYPE: DNA

STRANDEDNESS: Double stranded

TOPOLOGY: Unknown

ORGANISM: Kitasatosporia sp.

SEQUENCE: 15

gtgaccgggc	ccgactacta	cgagtaccgc	cacctggtgg	gcttcgagga	gaccaacctg	60
gtcggcaacg	tctactacgt	caactacctg	cgctggcagg	gacgttgccg	ggagatgttc	120
ctgctggaga	aggccccga	ggtgctcgcc	gacatccgcg	ccgacctcaa	gctgttcacc	180
ctcaagggtg	actgcgagtt	cttcgccgag	atcaccgcct	tcgacgagct	gtccatccgg	240
atgcgcctcg	ccgacctcac	ccagaccag	gtcgccttca	ccttcgacta	cgtaagctc	300
ggccccgacg	gcaccgagta	cctggtcgcc	cgcgggcagc	agcgggtcgc	ctgcatgcgc	360
ggccccaaca	ccgacacccg	cccgaaccgg	gtgcccgaac	cgctgcgggt	cgccctggag	420

ccctacgccg tccccgcgac ggcaccctcc ctgaccggca ccaccaccgt ggggtga

477

INFORMATION FOR SEQ ID NO: 16

LENGTH: 154

TYPE: PRT

STRANDEDNESS:

TOPOLOGY: Unknown

ORGANISM: Micromonospora megalomicea

SEQUENCE: 16

Met	Glu	Gln	Tyr	Tyr	Glu	Tyr	Arg	His	Val	Val	Gly	Phe	Glu	Glu	Thr	
1				5					10					15		
Asn	Ile	Val	Gly	Asn	Val	Tyr	Tyr	Val	Asn	Tyr	Leu	Arg	Trp	Gln	Gly	
			20					25					30			
Arg	Cys	Arg	Glu	Met	Phe	Leu	Arg	Glu	Arg	Ala	Pro	Gln	Val	Leu	Ala	
			35				40					45				
Asp	Leu	Gln	Asp	Asp	Leu	Lys	Leu	Phe	Thr	Leu	Arg	Val	Asp	Cys	Glu	
			50			55					60					
Phe	Phe	Ala	Glu	Ile	Thr	Ala	Phe	Asp	Glu	Leu	Ala	Ile	Arg	Met	Arg	
65					70					75					80	
Leu	Leu	Glu	Leu	Ala	Gln	Thr	Gln	Val	Glu	Phe	Gly	Phe	Asp	Tyr	Val	
				85					90					95		
Arg	Leu	Gly	Val	Ala	Gly	Val	Glu	Thr	Leu	Val	Ala	Arg	Gly	Thr	Gln	
			100					105					110			
Arg	Val	Ala	Cys	Met	Arg	Gly	Pro	Asn	Asn	Arg	Thr	Val	Pro	Ala	Arg	
			115				120					125				
Val	Pro	Glu	Ala	Leu	Gly	Arg	Ala	Leu	Ala	Pro	Tyr	Ala	Thr	Gly	Ala	
			130			135					140					
Pro	Val	Thr	Val	Ala	Ala	Gly	Arg	Pro	Leu							
145					150											

INFORMATION FOR SEQ ID NO: 17

LENGTH: 465

TYPE: DNA

STRANDEDNESS: Double stranded

TOPOLOGY: Unknown

ORGANISM: Micromonospora megalomicea

SEQUENCE: 17

atggagcagt	actacgagta	ccggcatgtc	gtcgggttcg	aggagacgaa	catcgtcggc	60
aacgtctact	acgtcaacta	cctgcgatgg	cagggccgct	gccgggagat	gttcctccgg	120
gagcgggccc	cgcaggtgct	ggccgacctg	caggacgacc	tcaagttggt	cactctgcgg	180
gtcgactgcg	agttcttcgc	cgagatcacc	gccttcgacg	aactggcgat	ccggatgagg	240

```

ctgttggagc tggcccagac ccaggtcgag ttcggcttcg actacgtccg gctcggcgtc      300
gccggtgtcg agacgctcgt cgcccggggc acgcagcggg tcgcctgcat gcggggggccg      360
aacaaccgta cggtgcccgc ccgggtgccg gaggcgctcg gccgtgcact cgcgccgtac      420
gccaccggcg cacccgtcac cgtcgcggca gggaggccac tgtga                        465

```

INFORMATION FOR SEQ ID NO: 18
 LENGTH: 143
 TYPE: PRT
 STRANDEDNESS:
 TOPOLOGY: Unknown

ORGANISM: *Saccharothrix aerocolonigenes*

SEQUENCE: 18

```

Val Thr Val Ala Arg Thr Phe Asp Tyr Arg His Val Ile Thr Leu Glu
1          5          10          15
Glu Thr Asn Leu Val Gly Asn Val Tyr Phe Thr Asn Tyr Leu Arg Trp
20          25          30
Gln Gly His Cys Arg Glu Arg Phe Leu Met Glu His Ala Pro Gly Val
35          40          45
Leu Arg Ala Leu Arg Gly Ala Leu Ala Leu Val Thr Val Ser Cys Gln
50          55          60
Cys Asp Phe Phe Asp Glu Leu Phe Ala Ser Asp Thr Val Glu Leu Arg
65          70          75          80
Met Ala Leu Gln Gly Thr Ser Asp Asn Arg Val Thr Met Ala Phe Asp
85          90          95
Tyr Tyr Arg Thr Ser Gly Ser Val Ala Gln Leu Val Ala Arg Gly Ser
100         105         110
Gln Thr Ile Ala Cys Met Ser Arg Thr Glu Glu Gly Thr Val Pro Val
115         120         125
Ser Val Pro Ala Glu Leu Arg Asp Ala Leu Ser His Tyr Ala Glu
130         135         140

```

INFORMATION FOR SEQ ID NO: 19
 LENGTH: 432
 TYPE: DNA
 STRANDEDNESS: Double stranded
 TOPOLOGY: Unknown

ORGANISM: *Saccharothrix aerocolonigenes*

SEQUENCE: 19

```

gtgaccgtgg ctaggacgtt cgactaccgg cacgtgatca ccctcgagga gacgaacctg      60
gtcgggaacg tctacttcac gaactacctg cgctggcagg gacattgccg tgaacgtttc      120

```

```

ctgatggagc acgcgcccgg tgtgctccgc gcgttgcgag gggcactcgc cctgggtcacg      180
gtctcctgcc agtgcgactt cttcgacgag ctcttcgcgt cggacacggg cgaactccgc      240
atggcgttgc agggcaccag cgacaacagg gtcacgatgg cgttcgacta ctaccggacc      300
tcgggttcgg tggcgcagct ggtggccagg ggcagtcaga ccatcgcggtg catgagcagg      360
accgaggagg ggaccgtgcc ggtgagcgtg cccgccgaac tgcgggacgc gttgtcgcac      420
tacgccgagt ga                                                                432

```

INFORMATION FOR SEQ ID NO: 20

LENGTH: 154

TYPE: PRT

STRANDEDNESS:

TOPOLOGY: Unknown

ORGANISM: Streptomyces kaniharaensis

SEQUENCE: 20

```

Val Met Ala Gly Tyr Tyr Glu Ile Arg His Thr Val Gly Phe Glu Glu
1           5           10           15

```

```

Thr Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp Gln
          20           25           30

```

```

Gly Arg Cys Arg Glu Met Phe Leu Lys Glu Lys Ala Pro Gly Val Leu
          35           40           45

```

```

Ala Glu Leu Arg Asp Asp Leu Lys Leu Phe Thr Leu Arg Val Asp Cys
          50           55           60

```

```

Glu Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu Ala Val Arg Met
65           70           75           80

```

```

Arg Leu Glu Glu Ile Ala Gln Thr Gln Leu Gln Phe Ser Phe Asp Tyr
          85           90           95

```

```

Leu Arg Leu Asp Gly Ala Gly Glu His Leu Val Ala Arg Gly Arg Gln
          100          105          110

```

```

Arg Ile Ala Cys Met Arg Gly Pro Asn Thr Asp Thr Val Pro Ala Arg
          115          120          125

```

```

Val Pro Glu Glu Leu Arg Arg Ala Leu Ala Pro Tyr Ala Thr Gly Pro
          130          135          140

```

```

Val Gly Ala Ala Ala Ala Gly Arg Pro Arg
145           150

```

INFORMATION FOR SEQ ID NO: 21

LENGTH: 465

TYPE: DNA

STRANDEDNESS: Double stranded

TOPOLOGY: Unknown

ORGANISM: *Streptomyces kaniharaensis*

SEQUENCE: 21

```

gtgatggccg gctactacga gatccggcac accgtcggct tcgaggagac caacctcgtc      60
ggcaacgtct actacgtcaa ctacctacgc tggcaaggtc gttgccggga gatgttcctc      120
aaggagaagg cgcccggggg gctcgccgaa ctgcggggacg acctgaagct gttcaccctc      180
cgggtggact gcgagttctt cgccgagatc accgcgttcg acgaactcgc cgtccggatg      240
cggctggagg agatcgccca gacgcagctc cagttcagct tcgactacct gcgcctcgac      300
ggcgccggcg agcacctcgt cgcccgcggg cggcagcgga tcgcctgcat gcgcggcccc      360
aacaccgaca ccgtgccggc ccgggtgccc gaggaactgc ggcggggcct ggctccgtac      420
gcgacggggc cggtcggggc ggccgcggcc gggaggcccc ggtga                      465

```

INFORMATION FOR SEQ ID NO: 22

LENGTH: 165

TYPE: PRT

STRANDEDNESS:

TOPOLOGY: Unknown

ORGANISM: *Streptomyces citricolor*

SEQUENCE: 22

```

Met Ser Gly Tyr Tyr Glu Ile Arg His Thr Val Gly Phe Glu Glu Thr
1           5           10           15
Asn Leu Val Gly Asn Val Tyr Tyr Val Asn Tyr Leu Arg Trp Gln Gly
20           25           30
Arg Cys Arg Glu Met Phe Leu Lys Glu Lys Ala Pro Gly Val Leu Ala
35           40           45
Glu Leu Arg Asp Asp Leu Lys Leu Phe Thr Leu Lys Val Asp Cys Asp
50           55           60
Phe Phe Ala Glu Ile Thr Ala Phe Asp Glu Leu Ser Ile Arg Met Arg
65           70           75           80
Leu Glu Glu Leu Thr Gln Thr Gln Ile Gln Phe Ser Phe Asp Tyr Leu
85           90           95
Arg Leu Asp Gly Gly Gln Glu Asn Leu Val Ala Arg Gly Arg Gln Arg
100          105          110
Ile Ala Cys Met Arg Gly Pro Asn Thr Ala Thr Val Pro Ala Arg Val
115          120          125
Pro Glu Glu Leu Arg Leu Ala Leu Ala Pro Tyr Ala Glu Gly Pro Val
130          135          140
Ala Ala Arg Leu Pro Ala Ala Pro Thr Ser Pro Gly Gly Pro Val Arg
145          150          155          160

```

Thr Gly Arg Gly Arg
165

INFORMATION FOR SEQ ID NO: 23
LENGTH: 498
TYPE: DNA
STRANDEDNESS: Double stranded
TOPOLOGY: Unknown

ORGANISM: Streptomyces citricolor

SEQUENCE: 23

```

atgtcgggct actacgagat ccgccacacc gtgggttttg aggagaccaa cctcgtcggc      60
aacgtctact acgtgaacta cctgcgctgg cagggggcgtt gccgggagat gttcctcaag    120
gagaaggcgc ccgggggtgct cgccgagctg cgggacgacc tgaagctgtt caccctcaag    180
gtggactgcg acttcttcgc cgagatcacc gcgttcgacg agctgtcgat ccggatgcgg     240
ctggaggagc tgacgcagac ccagatccag ttcagcttcg actacctgcg gctcgacggc     300
gggcaggaga acctggtcgc ccgtggccgt cagcggatcg cgtgcatgcg cgggcccgaac     360
acggcgacgg tccccgccag ggtgcccag gagctgcgcc tcgccctggc gccctacgcc     420
gagggcccgg tggccgcccg actgccggcg gcgccgacgt cgcccggcgg gccggtgagg     480
acggggaggg ggcggtga                                     498

```

INFORMATION FOR SEQ ID NO: 24
LENGTH: 1948
TYPE: PRT
STRANDEDNESS:
TOPOLOGY: Unknown

ORGANISM: consensus sequence

SEQUENCE: 24

```

Gly Gly His Gly Met Ser Met Thr Arg Ile Ala Ile Val Gly Met Ala
1              5              10              15
Cys Arg Tyr Pro Asp Ala Thr Ser Pro Glu Glu Leu Trp Glu Asn Val
20              25              30
Leu Ala Gly Arg Arg Ala Phe Arg Arg Leu Pro Asp Glu Arg Met Arg
35              40              45
Leu Glu Asp Tyr Trp Asp Ala Asp Pro Ala Ala Pro Asp Arg Phe Tyr
50              55              60
Ala Arg Asn Ala Ala Val Ile Glu Gly Tyr Glu Phe Asp Arg Ile Ala
65              70              75              80
Tyr Arg Val Ala Gly Ser Thr Tyr Arg Ser Thr Asp Leu Thr His Trp
85              90              95

```

Leu Ala Leu Asp Thr Ala Ala Arg Ala Leu Ala Asp Ala Gly Phe Pro
 100 105 110
 Gly Gly Glu Gly Leu Pro Arg Glu Arg Thr Gly Val Val Val Gly Asn
 115 120 125
 Ser Leu Thr Gly Glu Phe Ser Arg Ala Asn Val Met Arg Leu Arg Trp
 130 135 140
 Pro Tyr Val Arg Arg Val Val Ala Ala Ala Leu Ala Glu Gln Gly Trp
 145 150 155 160
 Asp Asp Asp Arg Leu Ala Ala Phe Leu Asp Asp Leu Glu Ala Ala Tyr
 165 170 175
 Lys Ala Pro Phe Pro Ala Ile Asp Glu Asp Thr Leu Ala Gly Gly Leu
 180 185 190
 Ser Asn Thr Ile Ala Gly Arg Ile Cys Asn His Phe Asp Leu Lys Gly
 195 200 205
 Gly Gly Tyr Thr Val Asp Gly Ala Cys Ser Ser Ser Leu Leu Ser Val
 210 215 220
 Val Thr Ala Ala Arg Ala Leu Val Asp Gly Asp Leu Asp Val Ala Val
 225 230 235 240
 Ala Gly Gly Val Asp Leu Ser Ile Asp Pro Phe Glu Val Ile Gly Phe
 245 250 255
 Ala Lys Thr Gly Ala Leu Ala Lys Gly Glu Met Arg Val Tyr Asp Arg
 260 265 270
 Gly Ser Asn Gly Phe Trp Pro Gly Glu Gly Cys Gly Met Val Val Leu
 275 280 285
 Met Arg Glu Glu Asp Ala Leu Ala Ala Gly Arg Arg Ile Tyr Ala Thr
 290 295 300
 Ile Ala Gly Trp Gly Val Ser Ser Asp Gly Lys Gly Gly Ile Thr Arg
 305 310 315 320
 Pro Glu Ala Ser Gly Tyr Arg Leu Ala Leu Arg Arg Ala Tyr Arg Arg
 325 330 335
 Ala Gly Phe Gly Val Glu Thr Val Gly Leu Phe Glu Gly His Gly Thr
 340 345 350
 Gly Thr Ala Val Gly Asp Ala Thr Glu Leu Glu Ala Leu Ser Glu Ala
 355 360 365
 Arg Arg Ala Ala Asp Pro Ala Ala Glu Pro Ala Ala Ile Gly Ser Ile
 370 375 380
 Lys Gly Asn Ile Gly His Thr Lys Ala Ala Ala Gly Val Ala Gly Leu
 385 390 395 400
 Ile Lys Ala Ala Leu Ala Val His His Gln Val Leu Pro Pro Ala Thr
 405 410 415

Gly Cys Val Asp Pro His Pro Leu Leu Thr Gly Asp Ser Ala Ala Leu
 420 425 430
 Arg Val Leu Arg Lys Ala Glu Leu Trp Pro Ala Asp Ala Pro Val Arg
 435 440 445
 Ala Gly Val Ser Ala Met Gly Phe Gly Gly Ile Asn Thr His Val Val
 450 455 460
 Leu Asp Glu Pro Val Gly Ala Arg Arg Arg Ala Leu Asp Arg Arg Thr
 465 470 475 480
 Arg Arg Leu Ala Ala Ser Arg Gln Asp Ala Glu Leu Leu Leu Leu Asp
 485 490 495
 Gly Ala Asp Ala Ala Glu Leu Arg Ala Arg Leu Thr Arg Leu Ala Asp
 500 505 510
 Phe Val Ala Arg Leu Ser Tyr Ala Glu Leu Ala Asp Leu Ala Ala Thr
 515 520 525
 Leu Gln Arg Glu Leu Arg Gly Leu Pro Tyr Arg Ala Ala Val Val Ala
 530 535 540
 Thr Ser Pro Glu Asp Ala Glu Arg Arg Leu Arg Gln Leu Ala Arg Leu
 545 550 555 560
 Leu Glu Ser Gly Glu Thr Glu Leu Leu Ser Ala Asp Gly Gly Val Phe
 565 570 575
 Leu Gly Arg Ala Thr Arg Ala Pro Arg Ile Gly Phe Leu Phe Pro Gly
 580 585 590
 Gln Gly Ser Gly Arg Gly Gly Gly Gly Ala Leu Arg Arg Arg Phe
 595 600 605
 Ala Glu Ala Asp Glu Val Tyr Arg Arg Ala Gly Leu Pro Ala Gly Gly
 610 615 620
 Asp Gln Val Ala Thr Asp Val Ala Gln Pro Arg Ile Val Thr Gly Ser
 625 630 635 640
 Leu Ala Gly Leu Arg Val Leu Asp Ala Leu Gly Ile Glu Ala Ser Val
 645 650 655
 Ala Val Gly His Ser Leu Gly Glu Leu Thr Ala Leu His Trp Ala Gly
 660 665 670
 Ala Leu Asp Glu Asp Thr Leu Leu Arg Leu Ala Arg Val Arg Gly Arg
 675 680 685
 Val Met Ala Glu His Ser Ser Gly Gly Gly Ala Met Ala Gly Leu Ala
 690 695 700
 Ala Thr Pro Glu Ala Ala Glu Ala Leu Leu Ala Gly Leu Pro Val Val
 705 710 715 720
 Val Ala Gly Tyr Asn Gly Pro Arg Gln Thr Val Val Ala Gly Pro Ala
 725 730 735
 Asp Ala Val Asp Glu Val Cys Arg Arg Ala Ala Arg Ala Gly Val Thr

740					745					750					
Ala	Thr	Arg	Leu	Asn	Val	Ser	His	Ala	Phe	His	Ser	Pro	Leu	Val	Ala
		755					760					765			
Pro	Ala	Ala	Glu	Ala	Phe	Ala	Glu	Glu	Leu	Ala	Ser	Val	Asp	Phe	Gly
	770					775					780				
Pro	Pro	Ala	Arg	Arg	Val	Val	Ser	Thr	Val	Thr	Gly	Ala	Leu	Leu	Pro
785					790					795					800
Ala	Asp	Thr	Asp	Leu	Arg	Glu	Leu	Leu	Arg	Arg	Gln	Ile	Thr	Ala	Pro
				805					810					815	
Val	Arg	Phe	Thr	Glu	Ala	Leu	Gly	Ala	Ala	Ala	Ala	Asp	Val	Asp	Leu
			820					825					830		
Phe	Ile	Glu	Val	Gly	Pro	Gly	Arg	Val	Leu	Ser	Gly	Leu	Ala	Ala	Glu
		835					840					845			
Ile	Ala	Pro	Asp	Val	Pro	Ala	Val	Ala	Leu	Asp	Thr	Asp	Ala	Glu	Ser
	850					855					860				
Leu	Arg	Pro	Leu	Leu	Ala	Val	Val	Gly	Ala	Ala	Phe	Val	Leu	Gly	Ala
865					870					875					880
Pro	Val	Ala	Leu	Glu	Arg	Leu	Phe	Glu	Asp	Arg	Leu	Ile	Arg	Pro	Leu
				885					890					895	
Pro	Ile	Asp	Arg	Glu	Phe	Ser	Phe	Leu	Ala	Ser	Pro	Cys	Glu	Gln	Ala
		900						905					910		
Pro	Glu	Ile	Lys	Ala	Pro	Ala	Val	Arg	Pro	Ala	Arg	Pro	Val	Val	Ala
	915						920					925			
Pro	Ala	Glu	Ala	Asp	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Gly	Glu	Ala
	930					935						940			
Pro	Gly	Glu	Ser	Ala	Leu	Glu	Val	Leu	Arg	Arg	Leu	Ala	Ala	Glu	Arg
945					950					955					960
Ala	Glu	Leu	Pro	Val	Glu	Ser	Val	Asp	Pro	Asp	Ser	Arg	Leu	Leu	Asp
				965					970					975	
Asp	Leu	His	Leu	Ser	Ser	Ile	Thr	Val	Gly	Gln	Ile	Val	Asn	Gln	Ala
			980					985					990		
Ala	Arg	Ala	Leu	Gly	Ile	Pro	Ala	Ala	Ala	Val	Pro	Thr	Asn	Phe	Ala
			995				1000					1005			
Thr	Ala	Thr	Leu	Ala	Glu	Leu	Ala	Glu	Ala	Leu	Asp	Glu	Leu	Ala	
	1010					1015					1020				
Gln	Thr	Ala	Ala	Pro	Gly	Asp	Ala	Ala	Ala	Ser	Leu	Val	Ala	Gly	
	1025					1030					1035				
Val	Ala	Pro	Trp	Val	Arg	Pro	Phe	Ala	Val	Asp	Leu	Asp	Glu	Val	
	1040					1045					1050				
Pro	Leu	Pro	Ala	Pro	Ala	Pro	Ala	Ala	Ala	Arg	Gly	Arg	Trp	Glu	
	1055					1060					1065				

Val	Phe	Ala	Thr	Ala	Asp	His	Pro	Leu	Ala	Glu	Pro	Leu	Arg	Ala
1070						1075					1080			
Ala	Leu	Ala	Gly	Ala	Gly	Val	Gly	Asp	Gly	Val	Leu	Leu	Cys	Leu
1085						1090					1095			
Pro	Ala	Asp	Cys	Ala	Ala	Glu	His	Val	Gly	Leu	Ala	Leu	Ala	Ala
1100						1105					1110			
Ala	Arg	Ala	Ala	Leu	Ala	Ala	Pro	Arg	Gly	Thr	Arg	Leu	Val	Val
1115						1120					1125			
Val	Gln	His	Gly	Arg	Gly	Ala	Ala	Gly	Leu	Ala	Lys	Thr	Leu	Arg
1130						1135					1140			
Leu	Glu	Ala	Pro	His	Leu	Arg	Thr	Thr	Val	Val	His	Leu	Pro	Asp
1145						1150					1155			
Pro	Gln	Pro	Leu	Asp	Glu	Ala	Ala	Asp	Asp	Ala	Val	Ala	Arg	Val
1160						1165					1170			
Val	Ala	Glu	Val	Ala	Ala	Thr	Thr	Gly	Phe	Thr	Glu	Val	His	Tyr
1175						1180					1185			
Asp	Ala	Asp	Gly	Val	Arg	Arg	Val	Pro	Val	Leu	Arg	Pro	Leu	Pro
1190						1195					1200			
Val	Ser	Pro	Ala	Glu	Glu	Ala	Ser	Pro	Leu	Asp	Glu	Arg	Asp	Val
1205						1210					1215			
Leu	Leu	Val	Thr	Gly	Gly	Gly	Lys	Gly	Ile	Thr	Ala	Glu	Cys	Ala
1220						1225					1230			
Leu	Ala	Leu	Ala	Arg	Asp	Ser	Gly	Ala	Ala	Leu	Ala	Leu	Leu	Gly
1235						1240					1245			
Arg	Ser	Asp	Pro	Ala	Ala	Asp	Glu	Glu	Leu	Ala	Asp	Asn	Leu	Ala
1250						1255					1260			
Arg	Met	Ala	Ala	Ala	Gly	Leu	Arg	Val	Arg	Tyr	Ala	Arg	Ala	Asp
1265						1270					1275			
Val	Thr	Asp	Pro	Ala	Gln	Val	Ala	Ala	Ala	Val	Ala	Glu	Leu	Thr
1280						1285					1290			
Ala	Glu	Leu	Gly	Pro	Val	Thr	Ala	Val	Leu	His	Gly	Ala	Gly	Arg
1295						1300					1305			
Asn	Glu	Pro	Ala	Ala	Leu	Ala	Ser	Leu	Asp	Glu	Glu	Asp	Phe	Arg
1310						1315					1320			
Arg	Thr	Leu	Ala	Pro	Lys	Val	Asp	Gly	Leu	Arg	Ala	Val	Leu	Ala
1325						1330					1335			
Ala	Val	Asp	Pro	Glu	Arg	Leu	Lys	Leu	Leu	Val	Thr	Phe	Gly	Ser
1340						1345					1350			
Ile	Ile	Gly	Arg	Ala	Gly	Leu	Arg	Gly	Glu	Ala	His	Tyr	Ala	Thr
1355						1360					1365			

Ala	Asn	Asp	Trp	Leu	Ala	Glu	Leu	Thr	Glu	Arg	Phe	Ala	Arg	Glu
1370						1375					1380			
His	Pro	Gln	Cys	Arg	Ala	Leu	Cys	Leu	Glu	Trp	Ser	Val	Trp	Ser
1385						1390					1395			
Gly	Val	Gly	Met	Gly	Glu	Arg	Leu	Gly	Val	Val	Glu	Ser	Leu	Ser
1400						1405					1410			
Arg	Glu	Gly	Ile	Thr	Pro	Ile	Ser	Pro	Asp	Glu	Gly	Val	Glu	Val
1415						1420					1425			
Leu	Arg	Arg	Leu	Leu	Ala	Asp	Pro	Asp	Ala	Pro	Thr	Val	Val	Val
1430						1435					1440			
Val	Ser	Gly	Arg	Thr	Gly	Gly	Leu	Glu	Thr	Leu	Arg	Leu	Asp	Arg
1445						1450					1455			
Arg	Glu	Leu	Pro	Leu	Leu	Arg	Phe	Leu	Glu	Arg	Pro	Leu	Val	His
1460						1465					1470			
Tyr	Pro	Gly	Val	Glu	Leu	Val	Thr	Glu	Ala	Glu	Leu	Asn	Ala	Gly
1475						1480					1485			
Thr	Asp	Pro	Tyr	Leu	Ala	Asp	His	Leu	Leu	Asp	Gly	Asp	Leu	Leu
1490						1495					1500			
Phe	Pro	Ala	Val	Leu	Gly	Met	Glu	Ala	Met	Ala	Gln	Val	Ala	Ala
1505						1510					1515			
Ala	Leu	Thr	Gly	Arg	Pro	Gly	Val	Pro	Val	Ile	Glu	Asp	Val	Glu
1520						1525					1530			
Phe	Leu	Arg	Pro	Ile	Val	Val	Pro	Pro	Asp	Gly	Ser	Thr	Thr	Ile
1535						1540					1545			
Arg	Val	Ala	Ala	Leu	Val	Thr	Asp	Pro	Asp	Thr	Val	Asp	Val	Val
1550						1555					1560			
Leu	Arg	Ser	Glu	Glu	Thr	Gly	Phe	Ala	Ala	Asp	His	Phe	Arg	Ala
1565						1570					1575			
Arg	Leu	Arg	Tyr	Thr	Arg	Ala	Ala	Val	Pro	Asp	Gly	Thr	Pro	Ala
1580						1585					1590			
Gln	Val	Asp	Asp	Asp	Leu	Pro	Ala	Val	Pro	Leu	Asp	Pro	Ala	Thr
1595						1600					1605			
Asp	Leu	Tyr	Gly	Gly	Val	Leu	Phe	Gln	Gly	Lys	Arg	Phe	Gln	Arg
1610						1615					1620			
Leu	Arg	Arg	Tyr	Arg	Arg	Ala	Ala	Ala	Arg	His	Val	Asp	Ala	Glu
1625						1630					1635			
Val	Ala	Thr	Ser	Ala	Pro	Ala	Asp	Trp	Phe	Ala	Ala	Phe	Leu	Pro
1640						1645					1650			
Gly	Glu	Leu	Leu	Leu	Ala	Asp	Pro	Gly	Thr	Arg	Asp	Ala	Leu	Met
1655						1660					1665			
His	Gly	Ile	Gln	Val	Cys	Val	Pro	Asp	Ala	Thr	Leu	Leu	Pro	Ser

1670		1675		1680
Gly Ile Glu Arg Leu His Leu Ala Glu Ala Ala Glu Gln Asp Pro				
1685		1690		1695
Glu Ala Val Arg Leu Asp Ala Arg Glu Arg Ser Arg Asp Gly Asp				
1700		1705		1710
Thr Tyr Val Tyr Asp Val Ala Val Arg Asp Ala Asp Gly Arg Val				
1715		1720		1725
Val Glu Arg Trp Glu Gly Leu Arg Leu Arg Ala Val Arg Lys Arg				
1730		1735		1740
Asp Gly Ser Gly Pro Trp Val Pro Ala Leu Leu Gly Pro Tyr Leu				
1745		1750		1755
Glu Arg Ser Leu Glu Glu Val Leu Gly Ser Ser Ile Ala Val Val				
1760		1765		1770
Val Glu Pro Ala Gly Asp Asp Pro Asp Gly Ser Val Ala Glu Arg				
1775		1780		1785
Arg Ala Arg Thr Ala Glu Ala Ala Ser Arg Ala Leu Gly Ala Pro				
1790		1795		1800
Val Glu Val Arg His Arg Pro Asp Gly Arg Pro Glu Leu Asp Gly				
1805		1810		1815
Gly Arg Glu Val Ser Ala Ser His Gly Ala Gly Leu Thr Leu Ala				
1820		1825		1830
Val Val Ala Ala Gly Arg Thr Val Ala Cys Asp Val Glu Ala Val				
1835		1840		1845
Ala Glu Arg Thr Ala Glu Glu Trp Ala Gly Leu Leu Gly Glu Arg				
1850		1855		1860
His Glu Ala Leu Ala Glu Leu Leu Ala Ala Glu Ala Gly Glu Pro				
1865		1870		1875
Pro Asp Val Ala Ala Thr Arg Val Trp Ser Ala Val Glu Cys Leu				
1880		1885		1890
Arg Lys Ala Gly Val Arg Ala Gly Ala Pro Leu Thr Leu Leu Pro				
1895		1900		1905
Val Thr Pro Asp Gly Trp Val Val Leu Ser Ala Gly Asp Val Arg				
1910		1915		1920
Ile Ala Thr Phe Val Thr Ala Val Arg Gly Ala Thr Asp Pro Val				
1925		1930		1935
Val Phe Ala Val Leu Thr Gly Ala Glu Arg				
1940		1945		

-100-

CLAIMS:

1. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23;
- b) a nucleic acid encoding a polypeptide of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22; and
- c) a nucleic acid with at least 75% identity to a nucleic acid of a) as determined by analysis with BLASTN version 2.0 with the default parameters, and which encodes a polypeptide having thioesterase activity.

10

2. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 3;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 3 and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 3 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

20

3. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 5;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 5 and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 85% identity to the nucleic acid of SEQ ID NO: 5 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

30

-101-

4. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 7;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 7, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 7 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

10

5. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 9;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 9, and encoding a polypeptide having thioesterase activity;
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 9 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

20

6. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead cassette, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 11;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 11, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 11 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

30

7. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

-102-

- a) SEQ ID NO: 13;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 13, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 13 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

8. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence
10 selected from the group consisting of:

- a) SEQ ID NO: 15;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 15, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 15 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

9. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence
20 selected from the group consisting of:

- a) SEQ ID NO: 17;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 17, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 17 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

10. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence
30 selected from the group consisting of:

- a) SEQ ID NO: 19;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 19, and encoding a polypeptide having thioesterase activity; and

-103-

- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 19 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

11. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 21;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 21, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 21 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

12. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NO: 23;
- b) a fragment comprising at least 150 consecutive nucleotides of SEQ ID NO: 23, and encoding a polypeptide having thioesterase activity; and
- c) a nucleic acid with at least 75% identity to the nucleic acid of SEQ ID NO: 23 as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

13. An isolated, purified or enriched nucleic acid encoding a thioesterase suitable for production of an enediyne warhead structure, said nucleic acid comprising a sequence selected from the group consisting of:

- a) SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23;
- b) a nucleic acid encoding a polypeptide of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 or 22; and
- c) a nucleic acid with at least 90% identity to a nucleic acid of a) as determined by analysis with BLASTN version 2.0 with the default parameters, and encoding a polypeptide having thioesterase activity.

-104-

14. An isolated, purified or enriched nucleic acid encoding an enediyne polyketide synthase catalytic complex, said nucleic acid comprising:

- a) a nucleic acid encoding an enediyne polyketide synthase; and
- b) a nucleic acid encoding a thioesterase, of any one of claims 1-13.

15. The nucleic acid of claim 14 encoding a enediyne polyketide synthase catalytic complex, wherein the nucleic acid of paragraph b) is a nucleic acid of claim 13.

10 16. An expression vector comprising a nucleic acid of any one of claims 1 to 15.

17. An isolated host cell transformed with an expression vector of claim 16.

18. A microbial host cell transformed with an expression vector of claim 16.

19. The host cell of claim 17 or 18 wherein the host cell is selected from a species of *Pseudomonas* and *Streptomyces*.

20. The host cell of claim 17 or 18 wherein the host cell is *E.coli*.

20

21. A nucleic acid encoding a thioesterase obtainable from cosmid 020CN deposited with the International Depository Authority of Canada (IDAC) having accession no. IDAC 030402-1, said nucleic acid having at least 90% sequence identity with SEQ ID NO: 3 or SEQ ID NO: 5, as determined using BLASTN version 2.0 with the default parameters.

30

22. A nucleic acid encoding a thioesterase obtainable from cosmid 061CR deposited with the International Depository Authority of Canada (IDAC) having accession no. IDAC 030402-2, said nucleic acid having at least 90% sequence identity to SEQ ID NO: 7, as determined using BLASTN version 2.0 with the default parameters.

-105-

23. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22; and
- b) a sequence with at least 85% identity to a polypeptide SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 or 22 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

24. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- 10
- a) SEQ ID NO: 2;
 - b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 2, and having thioesterase activity; and
 - c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 2 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

25. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- 20
- a) SEQ ID NO: 4;
 - b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 4, and having thioesterase activity; and
 - c) a polypeptide having at least 85% identity to the polypeptide of SEQ ID NO: 4 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

26. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- 30
- a) SEQ ID NO: 6;
 - b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 6, and having thioesterase activity; and
 - c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 6 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

-106-

27. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 8;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 8, and having thioesterase activity; and
- c) a polypeptide having at least 85% identity to the polypeptide of SEQ ID NO: 8 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

10

28. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 10;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 10, and having thioesterase activity; and
- c) a polypeptide having at least 85% identity to the polypeptide of SEQ ID NO: 10 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

20 29. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 12;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 12, and having thioesterase activity; and
- c) a polypeptide having at least 85% identity to the polypeptide of SEQ ID NO: 12 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

30 30. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 14;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 14, and having thioesterase activity; and

-107-

- c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 14 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

31. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 16;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 16, and having thioesterase activity; and
- c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 16 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

32. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 18;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 18, and having thioesterase activity; and
- c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 18 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

33. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- a) SEQ ID NO: 20;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 20, and having thioesterase activity; and
- c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 20 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

34. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

-108-

- a) SEQ ID NO: 22;
- b) a fragment comprising at least 40 consecutive amino acids of SEQ ID NO: 22, and having thioesterase activity; and
- c) a polypeptide having at least 75% identity to the polypeptide of SEQ ID NO: 22 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

35. An isolated polypeptide comprising a thioesterase suitable for production of an enediyne warhead structure, said thioesterase selected from the group consisting of:

- 10
- a) SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22;
 - b) a sequence having at least 95% identity to a polypeptide of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 as determined using the BLASTP algorithm with the default parameters, and having thioesterase activity.

36. A method of making a polypeptide of any one of claims 23 to 35 comprising the steps of:

- a) introducing, into a microbial host cell, a nucleic acid encoding said polypeptide, said nucleic acid being operable linked to a promoter; and
- b) culturing the host cell in conditions allowing expression of the nucleic acid.

20

37. A method of making a polypeptide having a sequence selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22 comprising the steps of:

- a) introducing, into a microbial host cell, a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter; and
- b) culturing the host cell in conditions allowing expression of the nucleic acid.

38. A method of making a polypeptide of any one of claims 23 to 35 comprising the steps of:

- 30
- a) introducing, into a host cell *in vitro*, a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter; and
 - b) culturing the host cell in conditions allowing expression of the nucleic acid.

-109-

39. A method of making a polypeptide having a sequence selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22 comprising the steps of:

- a) introducing, into a host cell *in vitro*, a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter; and
- b) culturing the host cell in conditions allowing expression of the nucleic acid.

40. A method of making a polypeptide of any one of claims 23 to 35 comprising the steps of:

- 10
- a) introducing, into an isolated host cell, a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter; and
 - b) culturing the host cell in conditions allowing expression of the nucleic acid.

41. A method of making a polypeptide having a sequence selected from the group consisting of SEQ ID NOS: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22 comprising the steps of:

- a) introducing, into an isolated host cell, a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter; and
- b) culturing the host cell in conditions allowing expression of the nucleic acid.

20

42. A method of identifying an enediyne biosynthetic gene cluster comprising the steps of:

- a) providing a sample containing genomic DNA; and
- b) detecting the presence of a nucleic acid coding for a polypeptide of any one of claims 23 to 35.

43. The method of claim 42 further comprising the step of using the nucleic acid detected to isolate an enediyne gene cluster from the sample containing genomic DNA.

- 30
44. A method of identifying an enediyne-producing organism comprising the steps of:
- a) providing a sample containing genomic DNA from a microorganism; and
 - b) detecting the presence of a nucleic acid coding for a polypeptide of any one of claims 23 to 35.

-110-

45. The method of any one of claims 42, 43 and 44 wherein the sample is biomass from environmental sources.
46. The method of claim 45 wherein the biomass is a mixed microbial culture.
47. The method of any one of claims 42, 43 and 44 wherein the sample is a mixed population of organisms.
- 10 48. The method of any one of claims 42, 43 and 44 wherein the sample containing genomic DNA is a genomic library obtained from a mixed population of organisms.
49. The method of any one of claims 42, 43 and 44 wherein the sample containing genomic DNA is obtained from a pure culture.
50. The method of any one of claims 42, 43 and 44 wherein the sample containing genomic DNA is a genomic library containing a plurality of clones, wherein the DNA for generating the clones is obtained from a pure culture.
- 20 51. An isolated polypeptide representing an enediyne polyketide synthase catalytic complex comprising an enediyne polyketide synthase and a thioesterase according to any one of claims 23-35.
52. An isolated polypeptide forming an enediyne polyketide synthase catalytic complex, said enediyne polyketide synthase catalytic complex comprising an enediyne polyketide synthase and a thioesterase of claim 35.
53. Use of a nucleic acid of any one of SEQ ID NOS: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23 according to any one of claims 2 to 12 to detect an enediyne biosynthetic gene
30 cluster.
54. Use of a polypeptide according to any one of claims 23 to 35, 51 and 52 to detect an enediyne biosynthetic gene cluster.

-111-

55. Use of a nucleic acid according to any one of claims 1 to 13 to produce an enediyne warhead structure.

56. Use of a polypeptide according to any one of claims 23 to 35, 51 and 52 to produce an enediyne warhead structure.

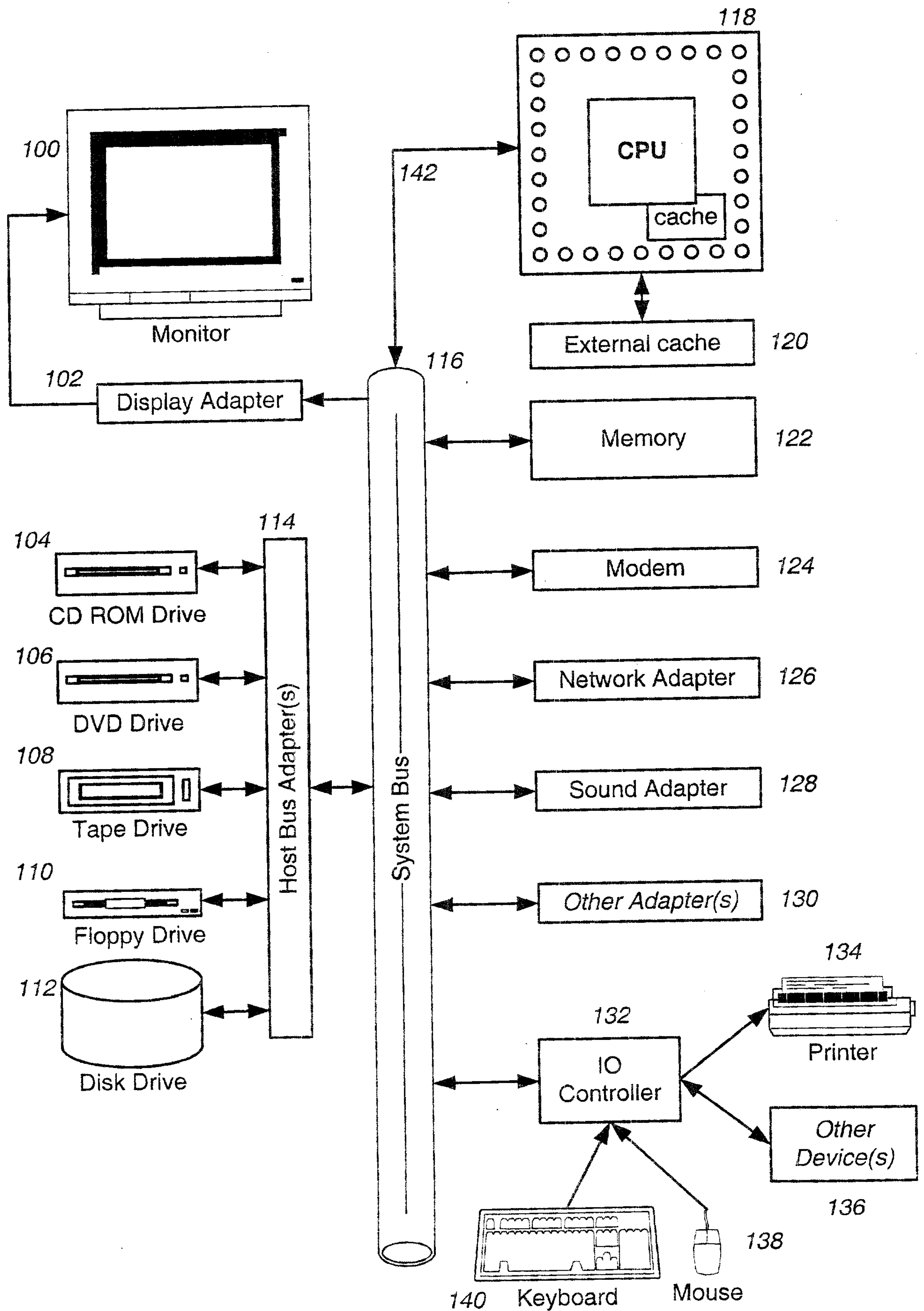


Figure 1

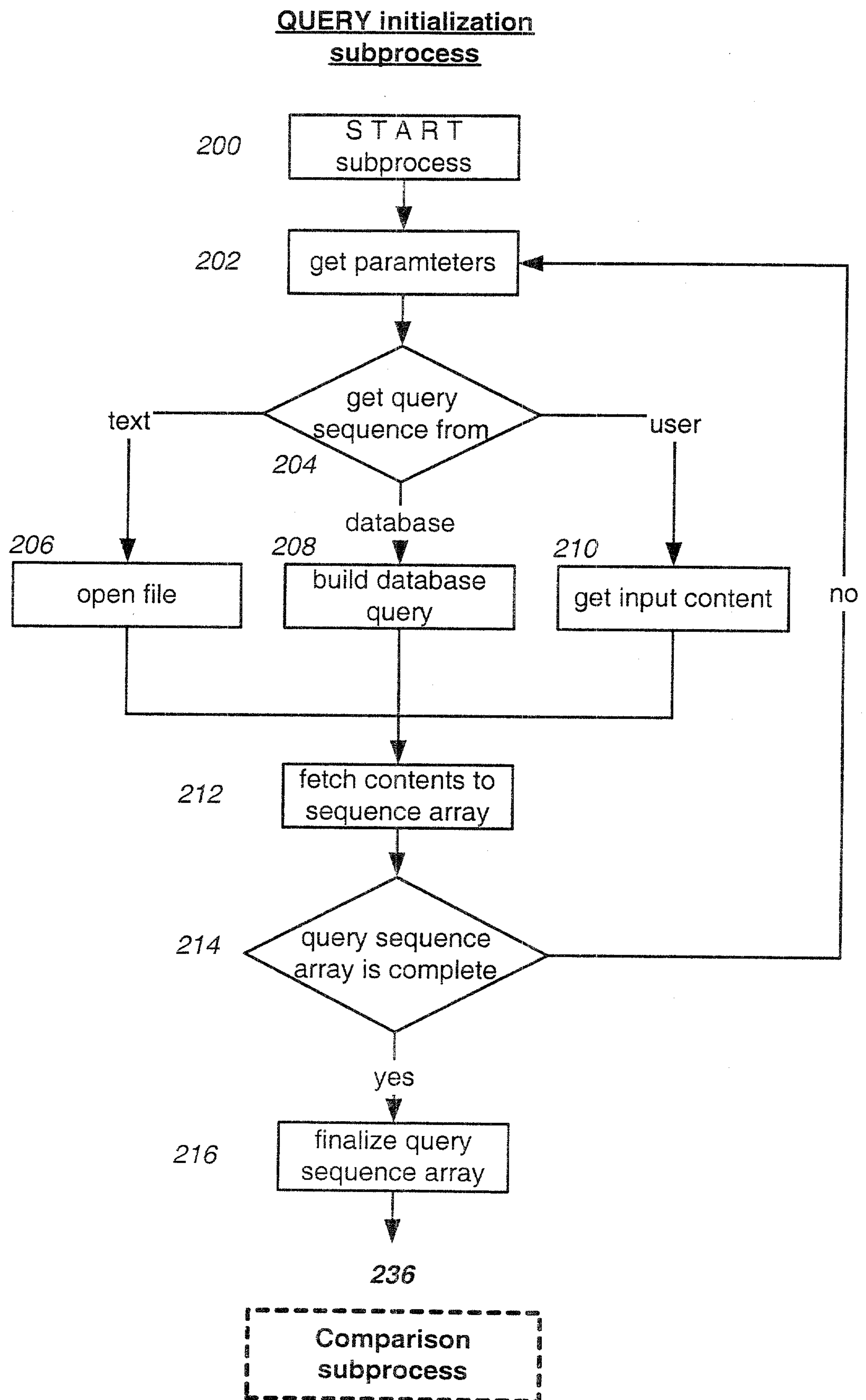


Figure 2-A

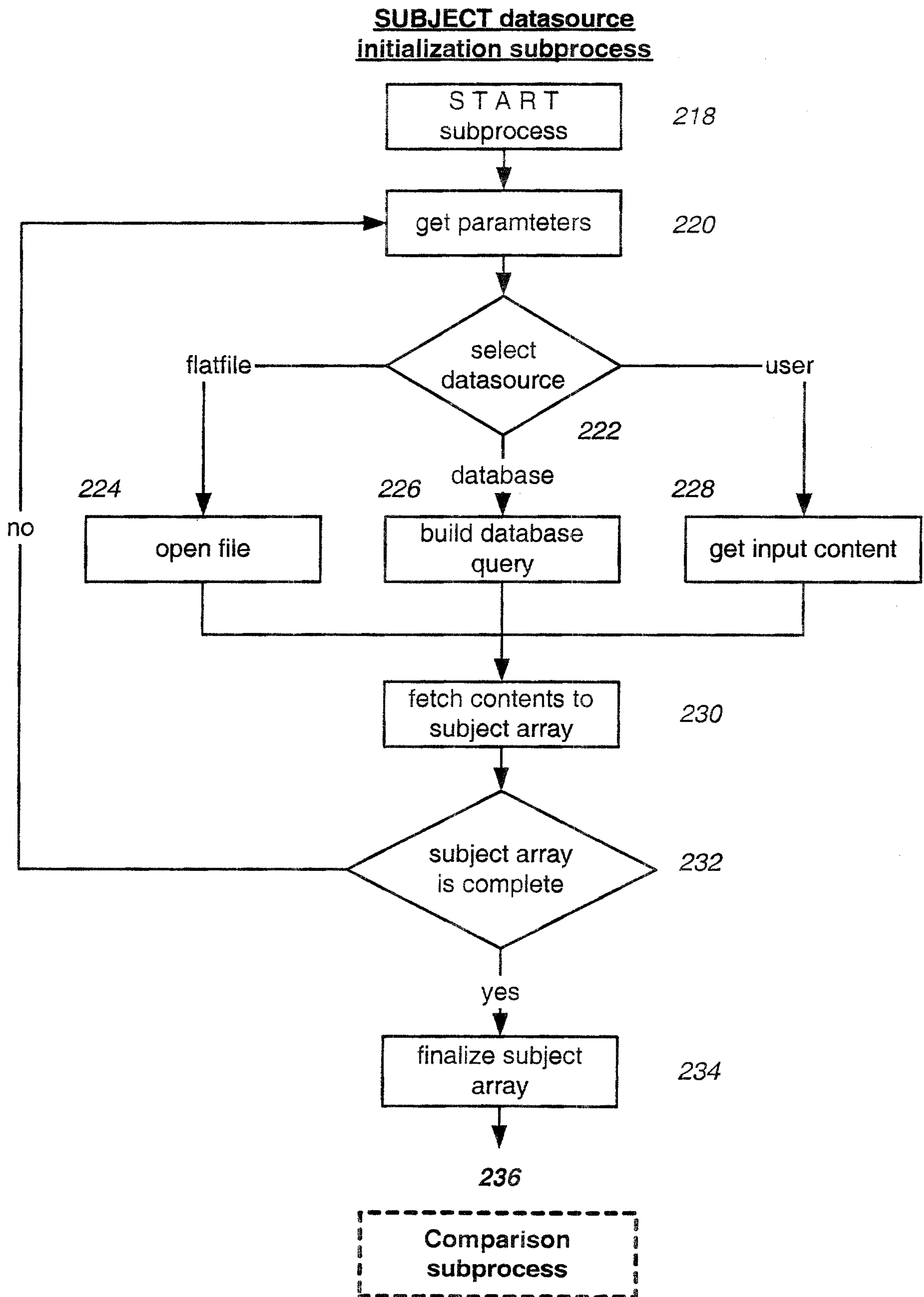


Figure 2-B

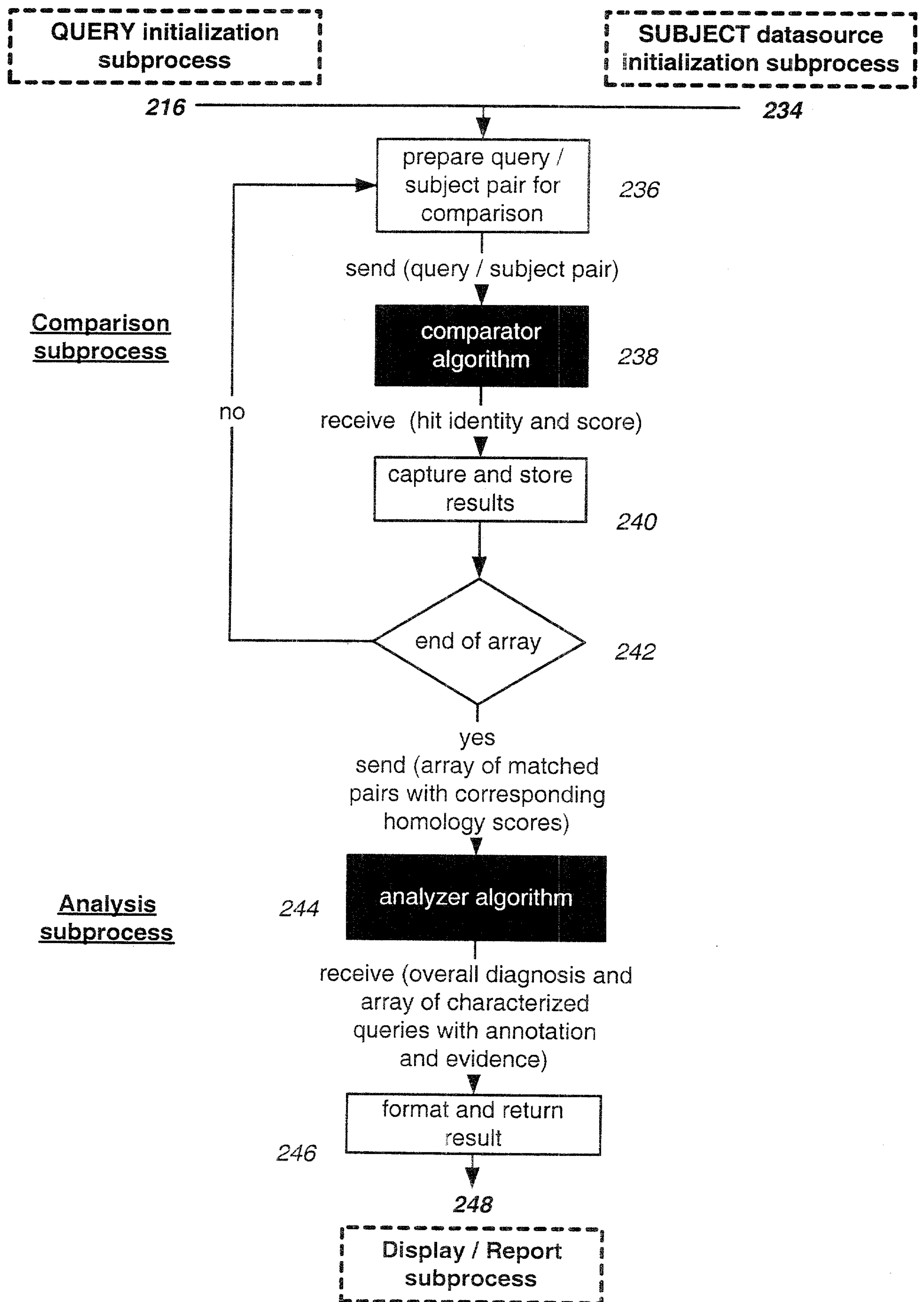


Figure 2-C

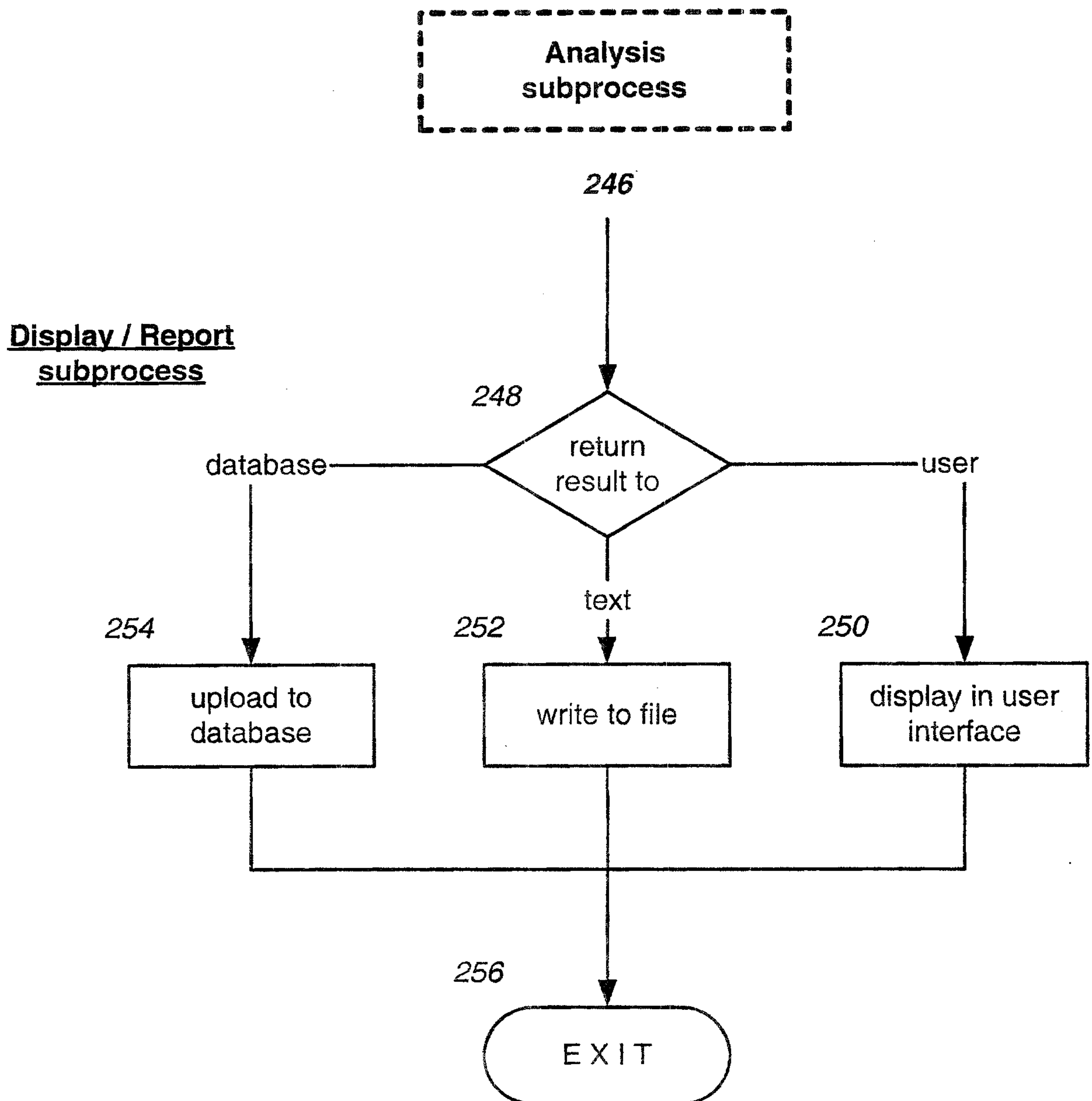


Figure 2-D

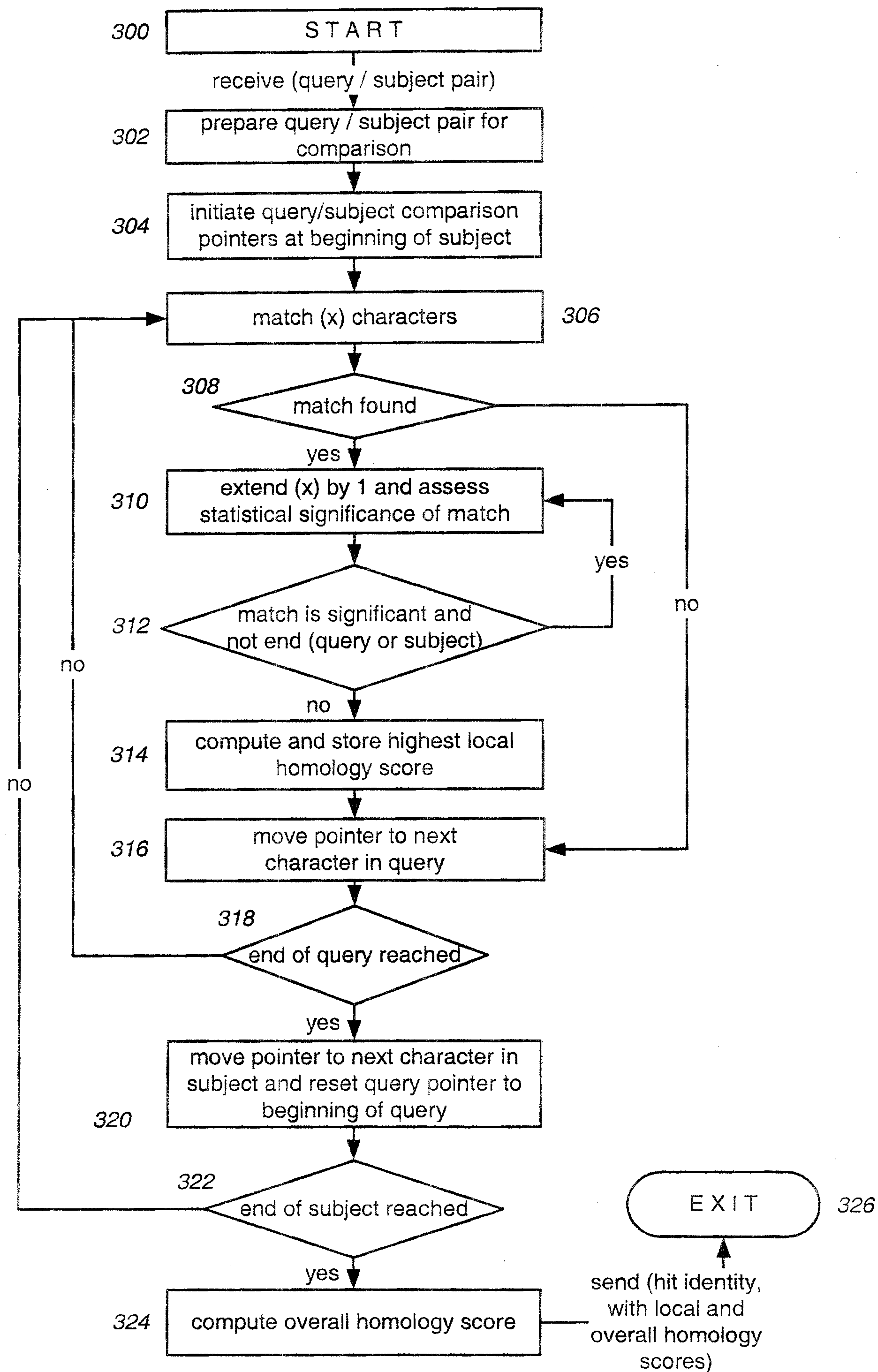


Figure 3
6/38

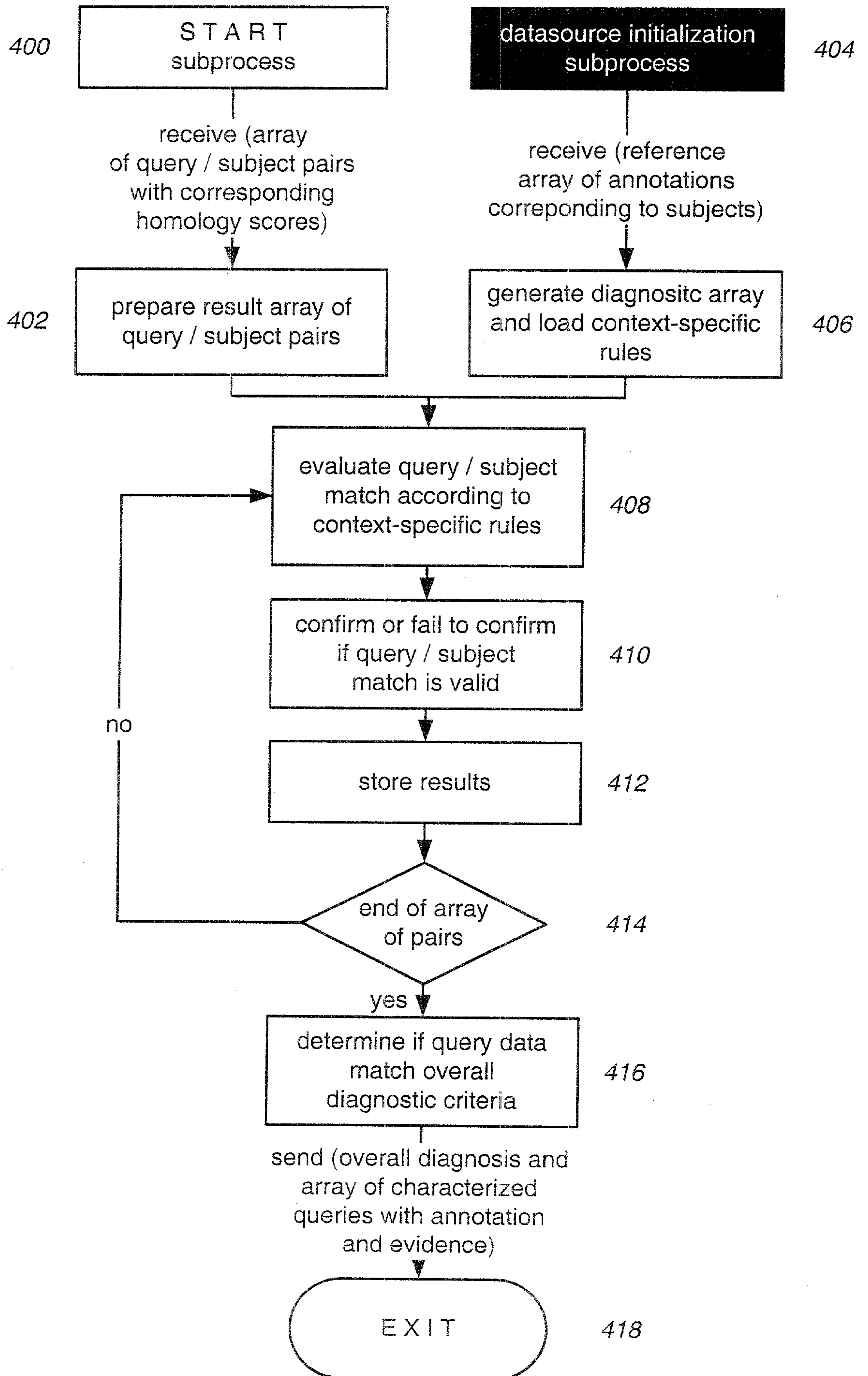


Figure 4

Figure 5

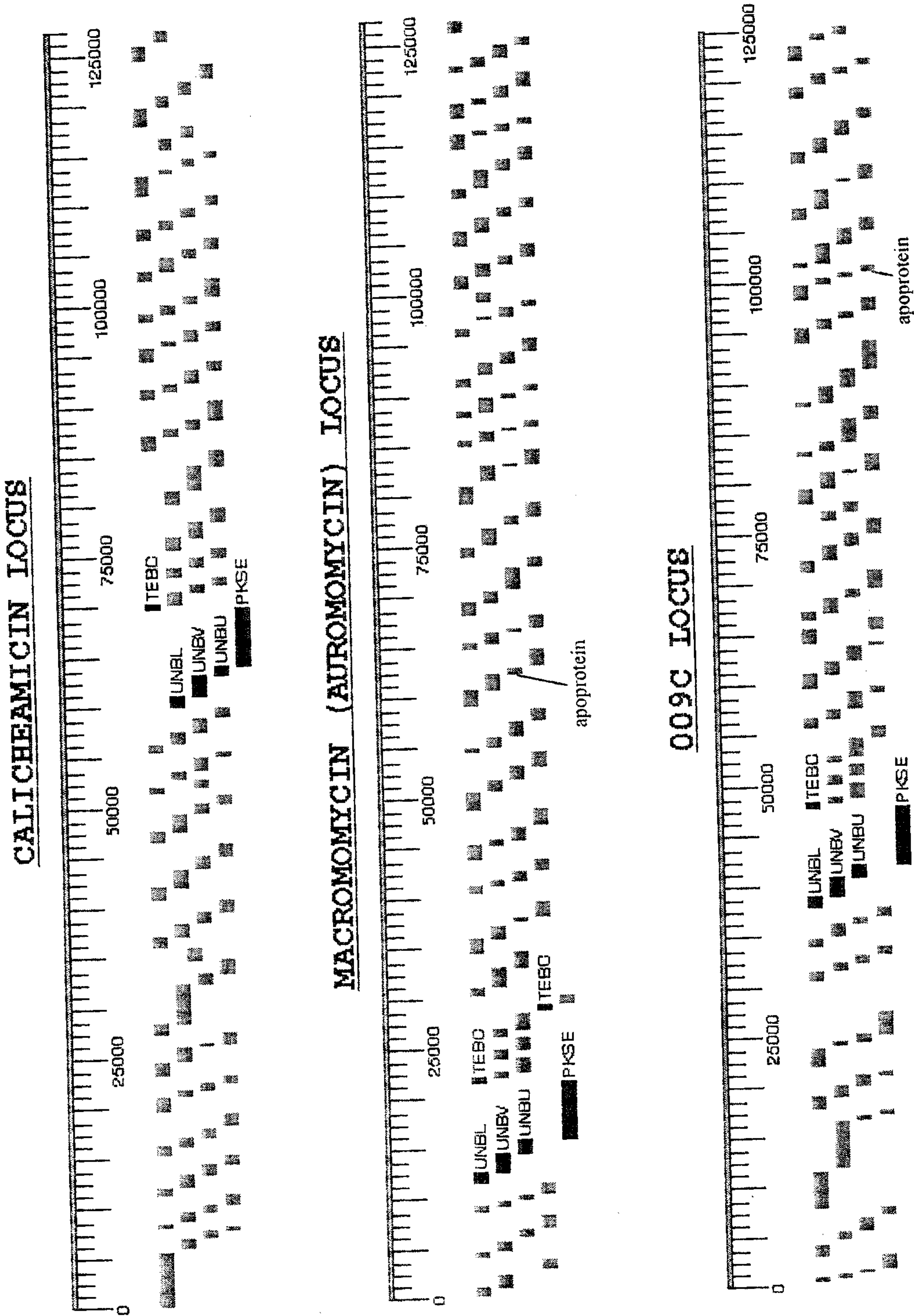


Figure 6

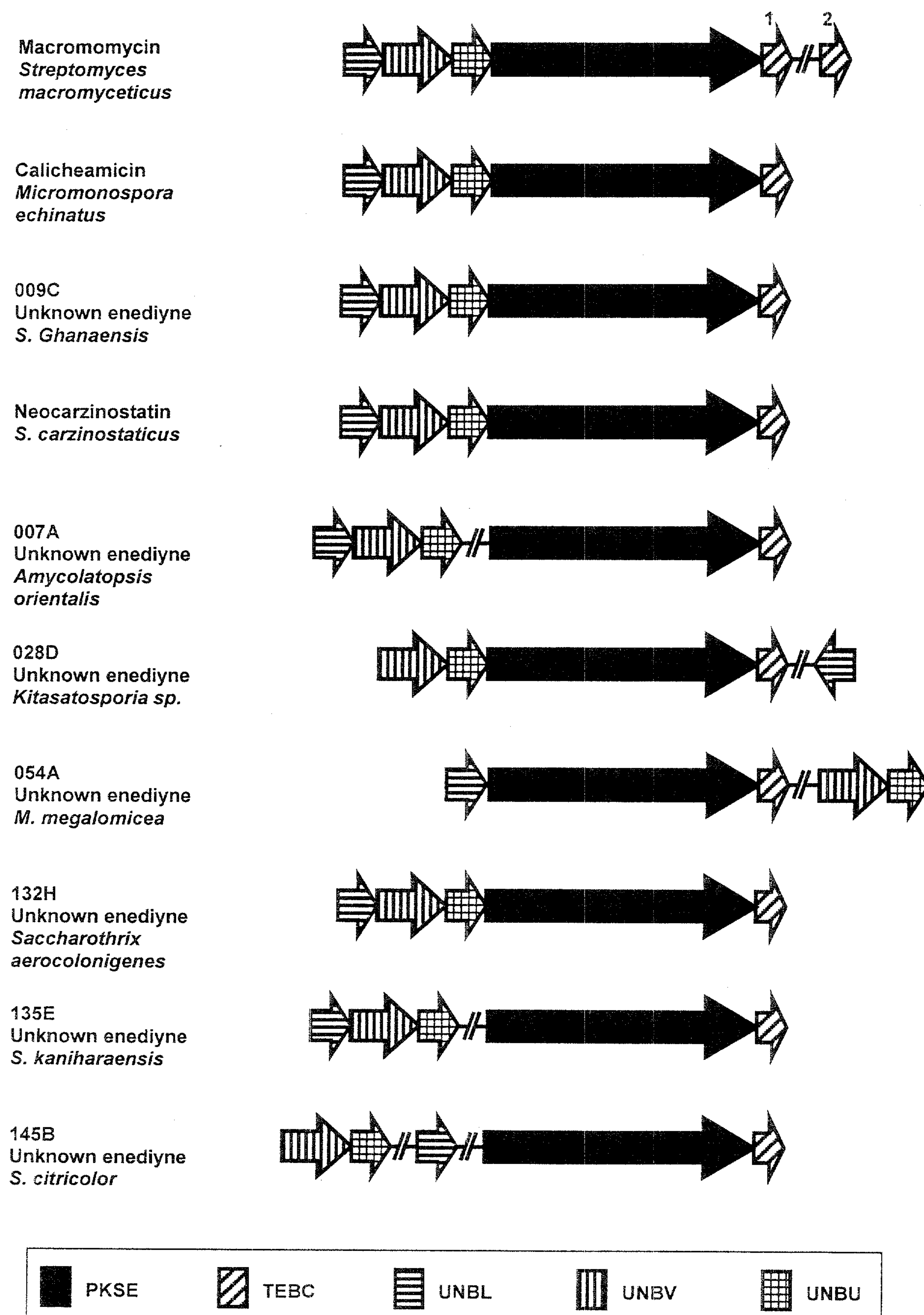
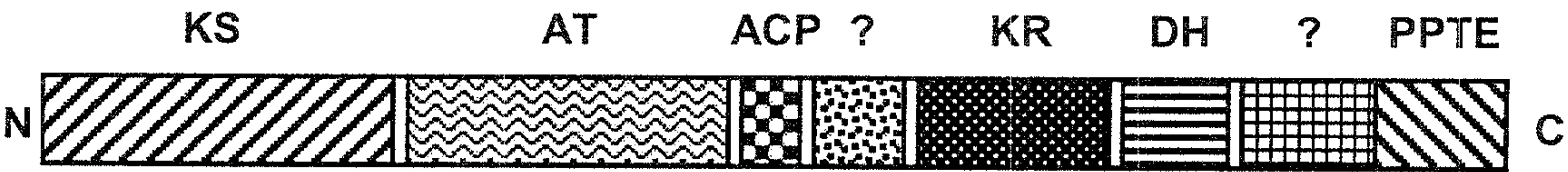


Figure 7



11/38

12/38

Figure 8C

```

PKSE|MACR      GHGTGTEVGDATELTAIMGARAEADPKAPLAAISSIKMIGHTKAAAGVAGLIKAAMAVD 402
PKSE|CALI      GHGTGTAVGDGTELRLALGEERRAADPDADPAAIGSIKMGIGHTKAAAGVAGLIKAVLAVH 402
PKSE|009C      GHGTGTAVGDDTELTALSDARRRADPDAPAAAITSIKMGIGHTKAAAGVAGLIKAAMAVN 403
PKSE|NEOC      GHGTGTAVGDATELAALIKARSAADPQAPVAAIGSIKMGIGHTKAAAGVAGLIKAALAVD 403
PKSE|007A      GHGTGTALGDATEIEALSTARRDADPLAERAALSTIKGNIGHTKAAAGVAGLIKATLAVY 404
PKSE|028D      GHGTGTAVGDATELRAFSEARRAAG-ATAPALSTVKGNFGHTKAAAGVAGLLKAILAVR 418
PKSE|054A      GHGTGTSVGDDTELTALSAARTAAGGAGLPAAVGSIKAMIGHTKAAAGVAGLIKAAMAVH 402
PKSE|132H      GHGTGTAVGDETELSALAE LRGGAR---EQAVVGSIKANIGHTKAASGVAGLIKAVLSIA 405
PKSE|135E      GHGTGTALGDATEIEALSSARRAADPVARPAALGTVKGNFGHTKAAAGVAGLIKAALAVH 404
PKSE|145B      GHGTGTALGDATEIEALSSARRAADPTARPAALGTVKGNFGHTKAAAGVAGLIKAALAVH 409
                ***** : ** ** : * : * : * : : * : ***** : ***** : :

PKSE|MACR      AAMLPPAIGCVDPHDLTGEQSNLRVLRKAEAWPKDAPLRAAVTAMGFGGINTHVVVDKA 462
PKSE|CALI      HQVVPPTVGCVEPHPELAADRPALRAVRRAEPWPAGAAQRAGVTAMGFGGINTHLVVDGP 462
PKSE|009C      HQVLPPSIGTIDPHALLTDDNATLKLARKAEPWPTGAPRRAGVTAMGFGGINTHVVLDEP 463
PKSE|NEOC      NQTLPPSIGTSDPHELLTEPGANLKLARKAETWPRELPRRAGITAMGFGGINTHVVLDEP 463
PKSE|007A      HQVIPPATGHFEPHESLVGDSARMYVPAEAGLWPSDQPVVAGVSAMGFGGINSHVTVTEA 464
PKSE|028D      HQVIPPATSHVDPHPELTGPAPALRVPDRALWPA GAP I RAGISSMGFGGINAHVVVEHA 478
PKSE|054A      QEVLPPAVGCVDPHEVLTGTSPALRVLRKAEAWPTDVPVVRAGVTAMGFGGINTHIVLENP 462
PKSE|132H      SGVLPPVTGCERPHHVLTA RETPLRVLT E AQWPAG - PRLAGVSSFGFGGINAHVALRDP 464
PKSE|135E      HQVIPPGTGHHDPHPGLLGDDAALYVPGRAELWPADSPVRAGVSAMGFGGINTHVAVTAA 464
PKSE|145B      HQVIPPATGHYDHPHPGLLGETAAMYVPRQAGLWPADQPVVAGVSAMGFGGINTHIALTQA 469
                : ** . ** * . : . * ** . * : : : ***** : * : :

                KS << >> AT

PKSE|MACR      VP-KRRPAPSRRATTLAASLQDAELLLLDGES PQALAA RLTEVA AF AAQVS YAQVGD LAA 521
PKSE|CALI      TR-PRRRSLDRRTQQ LARSVQDAELLLVDADTRDELRLDLDLRTVVAGLAF AELGDLAT 521
PKSE|009C      AG-RRRTAPSRRSATLAHTPQDCELLVLDGES PKALHARL TEVA AF VAQVS YGQVAD LAA 522
PKSE|NEOC      SG-RRRPASVRRLTPLADSMQDSELLLLFEGASARELSHRLSEVADYTVRLSYGEIADLAA 522
PKSE|007A      PGAARRKELDERTRSLVAGRQDSELLLLDADDAASLRGKVTGLLEVVPKLSFAELADLAG 524
PKSE|028D      DG-VRRTAVPAVAHRLVASRQDAELLLLDGADPAELHAKATRLAAFAAQLSYAEIGDLAA 537
PKSE|054A      RP-RRRVPLDTRTRALAASIQDAELLA VDAASAPELVQRLTRLVDFVGSVSYAQLADLGA 521
PKSE|132H      VT-ALPRTVSTPIRPHHRPAPRTDAFVLAGSDAAELRATLERIAELAPRLSEAEHLDLAC 523
PKSE|135E      PAAPRRTALDADTGR LVAGRQDAELLLLEARDRDGLRAEAARLLDLVPRLAQAE LADLAA 524
PKSE|145B      PGTARREALDERITRLVAGRQDAELLLLDGADQAALRAELVRLLDLVPRLAQAE LADLAG 529
                : : . . * : . : : : ** .

PKSE|MACR      ----TLQRELRLDPYRAAAVVTSPEDADLRLRQLAGTVE-AGTTS LFAPDGRTFLGRTGD 576
PKSE|CALI      ----NLHRSQRGRAYRAAVVARSPEEADRALGLAARALAP EGAGTLVDPARGVFVGRVT- 576
PKSE|009C      ----TLQRELRLGLSHRAAVVVTSPEDAERRLTHLADLLQ-TGETSYTAADGRGFLGRAT- 576
PKSE|NEOC      ----TLQRELRLGLPHRAAAVVTSPDDAENRLRHLADLLD-RGETEHWAADGRTLLGKAT- 576
PKSE|007A      ----TLSAELSGKPVRAAVVAAGPDDAERKLAKLLDLLG-EGEPEVFSAKEGIFAGSRS- 578
PKSE|028D      ----ALQRDLADRPLRAAVLADSPEQAAQRFTGLAQLLD-SGARSLLSPAQGVFLGSAG- 591
PKSE|054A      ----TLHGELRLDPYRAAVVATSPEDAERRLRQLRTAVE-AGETRQFSSDGRSLLGHVN- 575
PKSE|132H      ----QWGRDVAPGEHRVALVASTPRQLAERAVVAARSLE-SAPRGRLVVEDGVFLGTAV- 577
PKSE|135E      GLAAGLADGLTGAPVRAAVVATSPDDAARALERLLGLLA-SGATRALCSGEGVFLGTGR- 582
PKSE|145B      ----TLADRLSDGPVRAAVVASSPDDAVRALERLVALLD-SGTREAFSAGEGIFLGRAR- 583
                * . * : . * : : : *

PKSE|MACR      GDARIGFLFPGOGSGKGTTGGGALRRRFTEAAEVYDKAGLPTDGMV-ATDVAQPRIVTGS 635
PKSE|CALI      RPARVGFLFPGOGSGRGWGGGALRRRFTEIDDVYRAAGEPPGDEAAGSTVFAQPRIVTGS 636
PKSE|009C      RPARIGFLFPGOGSGHGTG G G A L C R R F P E A A E V F A R A A L P A T G D M T - A T N V A Q P R I A T G S 635
PKSE|NEOC      GRKRIGLLFPGOGSGRGTTGGGALSRRFPEVAEVLARAGSAAGSDTV-ATEVAQPRIVTGS 635
PKSE|007A      QSPKIGFLFPGOGSGQ-GRVGALRKRF AHADDIYRAANLSTGADQV-ATDVAQPRIVTGS 636
PKSE|028D      RAPRIGFLFPGOGAGRRGDGGALRRRF TAVRDL Y A H L D L P A D G D Q V - A T D V A Q P R I V A A S 650
PKSE|054A      GPGRIGLLFPGOGSGRGTTGGGALRRRFTEVEETYLRADLPTGGDVV-ATEVAQPRIVTGS 634
PKSE|132H      -AGRVTVLLFPGOGAPVRAELGALGRDLALTGGELRLDEELAGTR---GTATAQPSIFRAS 633
PKSE|135E      TEPTIAYLFPGOGSGR-GAVGAIRRRFAQADEVYRRAGLPTGADQV-DTRVAQPRIVTGS 640
PKSE|145B      SAPRIAYLFPGOGSGR-GGVGAIRRRFATAERVFHDFGPPTGDDQV-ATQVAQPRIVTGS 641
                : * : ** * : ** : : : * * * * *

```

Figure 8D

PKSE MACR	TAGLRVLDALGIEADVAVGHSLGELSALHWAGALDGPTVLEAARVRGAAMAEHSAS-GTM	694
PKSE CALI	LAGLRALAALDIDATVVVGHSLGELTTLHWAGCLDEDELRELVTLRGEAMARHAPP-GAM	695
PKSE 009C	AAGLRVLDALRLEASVAVGHSLGELSALHWAGALDEETLLQAARVRGRAMAEHSAT-GTM	694
PKSE NEOC	AAGLRVLDDEL RVEASVGIGHSLGELSALCWAGALDEEDVLEAAGVRGRAMAEHGSS-GTM	694
PKSE 007A	LAGLRVLKSLGIEAATVTGHSLGELTALHWGGALTEREVLKLAKIRGKVMATASDGDGAM	696
PKSE 028D	VAGLRVLDLLGVQADLATGHSLGELTALHWAGAMDEDTVLRAAAARGRIMAAAGDGGGTM	710
PKSE 054A	TAGLRVLHTLGIESSIAVGHSLGELSALHWAGVLDEADLLRIASLRGAAMARHSSS-GMM	693
PKSE 132H	LAALRWLDRLGVVAGAAVGHSLGEIAALVWAGCLSTEDADRLVREGRVMEDFGPRATGM	693
PKSE 135E	LAALRVLDGLGIRAAAAGVGHSLGELTALHWAGALTEDQVLRRLATVRGQVMAEVGSGGGAM	700
PKSE 145B	LAALRVLDGLGIRADTAVGHSLGELTALHWAGAMSEEQLVRLATIRGRVMARASHGGGAM	701
	*.*** * * : : *****:::* *. * : . . ** * . *	

Figure 8E

```

PKSE|MACR      ASLAAAPDAVAPLIDGLPVVISGYNGPQQTVVAGPVDVESVAQRAGQAGVKCTRLAVSH 754
PKSE|CALI      LGVTAGPEETVALLAGTNAVIAGYNGPRQTVVAGADDIVAEVARRAATAGVNCTRLPVPH 755
PKSE|009C      ASLAAAPERAEELLADLDAVIAGYNGPEQTVIAGSPADIEELQRRRAERAETVCTRLNVSH 754
PKSE|NEOC      ASLGAAPEQAEELIGALS VVVAGYNGPQQTVVSGPVHEVEEVRRAARSGVTCTPLAVSH 754
PKSE|007A      AAIATPSVAEGLAEGEEVVIAGYNAPEQTVLSGPAAEIDRVVARARAEGVTAARINVSH 756
PKSE|028D      AALATTPALAEALIVGEPVVVAGLNSPTQTVVSGPVDVDRVCALAARQGIGVGRNVSH 770
PKSE|054A      ASLAADPEALDPLLADLPVVVAAYNGPGNTVVAGTDEAVRAACQQRAGDAGFTATVLPVSH 753
PKSE|132H      VGIVADVPTAHGLCEGTGMVVACYNGPRSQVLGARTAIDEVVARAARLGVQTVVLPVTH 753
PKSE|135E      AGLAATPEDGTRLGAGLDVVIAGYNGPRQTVVSGPAAAVDEVCRRAVAEGVTATRLNVSH 760
PKSE|145B      AGLAATPERTRLSAGQDVVAGYNGPRQTVVSGPAEAVDEVCRRAAEGVTATRLNVSH 761
               . : .      *      * : . * . * : . :      *      .      : * .

PKSE|MACR      AFHSPLVAPAAESFGEWLAGADFGSVDRRIVSTVTGADLEQDGLAKLLRQQITDPVLFT 814
PKSE|CALI      AFHSPLMASAAAAFAERLRSRRFGPLLRRVASTVTGAVLPSDTDLPRHLHRQIEAPVRFA 815
PKSE|009C      AFHSPLVAHSAEVFGAWLAEARLGSPSGRVVSTVTGEELTAGTDLAALLTEQITGPVRFT 814
PKSE|NEOC      AFHSPLVASAAESFGNWLKSVDFREPAGRVVSTVTGAELTPGTDLAALLREQITAARVFT 814
PKSE|007A      AFHSPAVVPAAEAMTGEAAIDFARLDRPVVSTVTGDLVHAAEDLRDLLRDQVVLPRFR 816
PKSE|028D      AFHSPAVAPAAAGLAHLAGERFGPVGPGLVSTVTGAPLPADTDVVDLLTRQVVPVRFT 830
PKSE|054A      AFHSPLVAPAADEFCAALADRQWHPLAGRVISTVTGDLLEPHTDVPALLRRQITDPVLFT 813
PKSE|132H      GFHSPLVAPAAHAMAERLGEFDFARPVRPVASTVTGALLDPAADLRLLRDQVARPVRFH 820
PKSE|135E      AFHSPLVAPAAHAMAERLGEFDFARPVRPVASTVTGALLDPAADLRLLRDQVARPVRFH 820
PKSE|145B      AFHSPLVEPAAVAMAAELAGFDFREPVRPIASTVTGELLDPADLRLLRDQVLRPVRFH 821
               . * * * : . * :      *      : * * * *      * :      * * . * *

PKSE|MACR      QALEAAAADVDFIEVGPGRVLSLTLAEAGV-DVPAVALNTDDESLRALLQVVGAAYVVG 873
PKSE|CALI      AALGRAAAEVDLFLEVGPGRVLTGLAREQAPDVPALAVDTDAESLSGLLAAGVAVYALGG 875
PKSE|009C      RAVSEAAARHVDLFVEVGPGRVLSGLARATT-GVPAVALNTDDESLRSALAVAGAAAFVGA 873
PKSE|NEOC      EAVRAAAQDVDFIEVGPGRVLSGLAGTAT-NIPAVSLDTDDDESLRSLLQVVGAAYVVG 873
PKSE|007A      EAAAKVAERSDLVIEVGPGRVLTGLLGTIAPGTPVLSIDTDSLTLAPVLKVAGAAAFVGA 876
PKSE|028D      DALRAMDGQVDLLIEVGPQILRTLAAEVLPAVPAVATEADALSLAGLLATVATAWTMGA 890
PKSE|054A      QAVGLAAKSVDLFVEVGPGRVLAGLAGRAT-DVPAVSLDTDDDESIASLLTVVGSAYVVG 872
PKSE|132H      QAMDEVLPDLDLFCEAGPGRVLSLVAAGC-PVPVVGVDAGALDDRPLAETVAALFAAGA 872
PKSE|135E      EAAAAATADADLVVEVGPGRVLSGLLAEIAPDRPALAVDTSSSLGPLLRVAAAFAVLGS 880
PKSE|145B      EAAGVAAVGADLVVEVGPGRVLSGLLAEIAPDPTVLALDTSASLGALLRVVGAAYVLGA 881
               *      * . * . * . : *      . . . : . . . : . * .

               AT <<

PKSE|MACR      PLIHERLFRDRLTRPLEIGAEFSLTSPCEQAPESLPAGR----- 914
PKSE|CALI      PAAYPVLFEFDRLTRPFDPHRARTFFASPCEAAPELAGPAPAA----- 917
PKSE|009C      PVALERFFEDRLIRPLRVGQEFSLANPCEQAPREKAPAGR-----PRPVTP 921
PKSE|NEOC      PVAPERLFRDRLIRPLRIGQELSFLASPCEQAPATTLPVSRRSAQPPAVPADR-EQEPQP 932
PKSE|007A      QLETSTLFDGRVVRLPADGEFSFLASPCEAAPSIGAVLTRD----- 918
PKSE|028D      PVRHERLFADRFRTRPLPLDKEFRFFASPCEGGEFVLEHAG----- 932
PKSE|054A      ARIEAAALFHGRLIRPLAVGAEFSSFFANPCEQAPSVDLPV-R----- 912
PKSE|132H      LHDLSPVFTGRPARPIDIWRDRRFLANPCSSVPNAKPIEVVP----- 914
PKSE|135E      PVRAAGLFEGRLVRPLPADGAMTFLASPCEAPADAARLTP----- 922
PKSE|145B      PVRTGALFGDRLIRPLPADGVMSFLANPCEAAPPIGAGLVPQD----- 924
               . * . *      * :      * : . * .

               >> ACP

PKSE|MACR      -----APRTEGAGDGDGGEQAGQAO-----GESALEVLRALVAERAE 951
PKSE|CALI      -----VAPVPAPARADDTALPAATG-----ALELVRHLVAERAE 951
PKSE|009C      AAETPHDAHAPATPADATTAAEAPTT--EEAPED-----GAGALEVLRVLVAERAE 970
PKSE|NEOC      AAVSPPAQNSPASNDTSTASTASTAGSERTPQEEESI-----GAKALDVLSALVVERAE 987
PKSE|007A      -----R--VAEPAEAAAGTASESG-----GSSTLDLLRKLASERVE 952
PKSE|028D      -----ATPATAAAPRPAAAAAPAAG-----EATSLEVLIRLAAARAE 969
PKSE|054A      -----AAGVPVVAEADAQSAAVVP-----GETTVDLLRRIAAERAE 948
PKSE|132H      -----IEVVTPEVAPPAAEIRD-----PRTVVLELLAEASE 946
PKSE|135E      -----ARP-AVEAATGTATAPAEAG-----GESTLDLLRRLAAERVE 958
PKSE|145B      -----GGDRG-DGAAGRDEGTTTARIADSGAC-----PDSTLELLRRLAAERVE 967
               . : . :      *

```

16/38

Figure 8G

```

>> domain?
PKSE|MACR      LAGTGGGAAG-SGPLVTGSVWARPFVAVDLDEVPLAVA---APGGENGWELFTAGSDPF 1066
PKSE|CALI      REREAPHERA---DLVAGVAPWVRFTFVVDLVDEPLPAT---DPTEPCGRWQVFAGADHPL 1064
PKSE|009C      LVDGTGTTGEP-TTSAVVGAAPWARPFVAVDLDEVARPHA---AADGADGDWELFAPQDHPY 1085
PKSE|NEOC      LASTGGPADAGAASFIAGAAPWARPFVAVDLDAVARPPA---RPAAVRGTWELFAPAGYGI 1104
PKSE|007A      LAQTAKPADS-NQAEVAGVGPWVRPFVEYVVPKPS-PLATGISTAEWTAFAFAGHPL 1069
PKSE|028D      LAQOS-PQDT-RPGAAAGVAPWVRPFRIIDLTTETPPPAP---AAGPGGRWEVVFATDRHPL 1082
PKSE|054A      LAATASADDAVAAPVVGAAAPWARAWRIDLDVAEPPT-ADAPEDGTWQLFAADDHPL 1064
PKSE|132H      LPAAEAIGEN---EPVAGVASWIRCFAEVPGPVVEPGP---PGGTRRWRIHHSQGRPDV 1059
PKSE|135E      LAQTAEDGPA-PASEVPGVAPWVRPFVEHVEAALPAR-TAAPAAATGSWTVYSTPGHPL 1075
PKSE|145B      LADTSQDGAA-AADEVPGVAPWVRPFVEYMEALVPRPLPGPQAAAGDWAVYSTPGHPL 1085
      * . * * :

PKSE|MACR      GQQLKAALEGAGVGAG---VVVWLPPACPAEHIAQALDGAKAALAGD---RERRFVLVQH 1120
PKSE|CALI      ADALRRALEAAGVGEG---VLVCLPDEPDEEHLVTAVRGAQAALRQP---PGGRLVVVQP 1118
PKSE|009C      AERLRRELAGAGVGAG---VVAVLPKGCARQEVDRVLAHAHSAALAGD---RTRRFVLVQD 1139
PKSE|NEOC      AATLRAALQDAQAGSG---VLVCLPPQCSADGIDLALAAAKRALAAP---KDSRFVLVQH 1158
PKSE|007A      AEPLRAALATAGVGDG---VLLCLNADSASGDVGLFLDAGRAVLAAP---NGTRFVVVQH 1123
PKSE|028D      AGPLAERLTATAPGGG---VLLALPRDCDQRHLGLMLAAARAALDPARRAAGTRLVAVGD 1139
PKSE|054A      ATELLAELHRARLGGG---VLVWLPQDCPEEALHALRGAQQAARGG---PGTRFVLVDH 1118
PKSE|132H      ADEIRLLFGGSDAGSGDVADLLYLPDPSAQEAVGTLLSAVSSALGSG-----RLVVITH 1113
PKSE|135E      AEPLRTALAEAGIGDG---VLLCLPAECGAGDTEFLAAGRAAATAP---GGTRLVVVQH 1129
PKSE|145B      AEPLRAALARAGIGDG---VLLCLPAECGAGEADFLAAGRAVLAAP---EGTRLVVVQH 1139
      . : : : * * : * : . . * : * :

domain? << >> KR
PKSE|MACR      GRGAAGLAKTLHQEG-HLRTTIVHTPR-----PDADAVRTVVAEVAATARFTEVHYDT 1172
PKSE|CALI      AARAGALAKTARLEGDRRLRTTVVTP-----LDPAAVDRVVADVAATDDFTEAVYDP 1170
PKSE|009C      GRGAAGLAKTLYLEAPHLRTTVVHTP-----AAQDVVERVVAEVAATTRFTEVHYDE 1191
PKSE|NEOC      GRAAAGLVKTLHQEASHLVTTVVDT------LTEDTVDRVVAEVSATTRFSEVHISA 1210
PKSE|007A      GLGASGLAKTLRLDPSARTTIVDLADLGPVDPEALDAVSTVTEVAATDDFSEVRYDT 1183
PKSE|028D      HRGAAGLAKTLHLEAPDIPVTVVTLPLDQELPAPAAEQAAARIAADTAATTGFSEVHYDA 1199
PKSE|054A      GRGGAGLAKTLRLLEAPHLRVTVVHLS-----DASGAVERVTAEVAGTVGFVEVDYDA 1170
PKSE|132H      GSGLSGFLRSLRMEHPRLGVTLLRVPP-----GVDGVRAAARHAVVAAGEWRELTVGA 1166
PKSE|135E      RLGATGLAKTLHLEHPSVPTTVVELPDPLAP-----EAVGLVVAEAAATTGFTEVRYGP 1183
PKSE|145B      RFGAGGMAKTLHLEHPSVLTTVVELADPAPKG-AALDEAVARVVAEAAATAGFAEVRYRQ 1198
      . : : * . * : . . : : *

PKSE|MACR      EGARRVPTLRALPVAPARK-EHVLGSSDVLLVTGGGKGITAEALAVAKETGAKLAVLGR 1231
PKSE|CALI      GGRRRVFVLRPLPASDGEPLGADVLLVTGGGKGITAESALMLARESGARLAVLGR 1230
PKSE|009C      AGVRRVPTLRALPVAPQHT-ASPLDASDVLLVTGGGKGITAEALAVAQRTGAALAVLGR 1250
PKSE|NEOC      DGVRRVPTLRALPMSPEQQ-DKPLSASDVLLVTGGGKGITAEALATAQDSGTRLAVLGR 1269
PKSE|007A      AGVRTVPKLAALTPAEAE--TPLDTGDVLLVTGGGKGITAEALALAKDSGAKLALLGR 1241
PKSE|028D      DGTRRVFVLRPVPLEPDG-RQALGPRDVLLVTGGGKGITAEALALAGGNGAAIGLIGR 1258
PKSE|054A      DGTRRVFVLRAMSVPQIS-RPALDDTDVLLVTGGGKGITAEALAMATDSGASLALLGR 1229
PKSE|132H      EGVATEPAHRPVWHLSDGE--PPLGERDVLLVTGGGKGIGYECALARRSGAALALVGR 1224
PKSE|135E      DGRRTVFPVLRPLTPTEAPAGASPLDEADVLLVTGGGKGITAEALAMARDSGAALALIGR 1243
PKSE|145B      DGRRTVFPVLRPLQLSPAPAGESPLDARDVLLVTGGGKGITAEALAIKDSGAGLALIGR 1258
      * * : . * . * * * * * * * * * * * * * * * * * * * * * *

PKSE|MACR      SDPAEDKDLGDNLMADSGVTVAYARADVTDPAVAAVAELAEKLGPTALLHGAGRN 1291
PKSE|CALI      SDPTADEALADNLKRLADAASDLRYLRVDVTDAGAVAAAVATVTADWGPVTAVLHGAGRN 1290
PKSE|009C      SDPASRELADNIERMRRAGGARVHYARADVTVPEQVTAVAELTERLGTITALLHGAGRN 1310
PKSE|NEOC      SDPATRELADNLKRMEDSGVTMRYARADVTNPEQVTAVAELRGELGPITGVVHGAGRN 1329
PKSE|007A      SDPADDAELSENLMRMAAGITYRYERADVTDGRQVADAIGRVQAEFGPVTAVLHGAGRN 1301
PKSE|028D      SDPARDTELADNLARMAAGMRVHYARADVTSADQVKAATEITRELGPVTGLLHGAGRN 1318
PKSE|054A      SAPAEDPELAANLSRMTAAGVTVRYARADVTDPDQVRRVAELTADLGPVTAVLHGAGRN 1289
PKSE|132H      ADPHADELLRSNVDNLSAAGLRVAYESVDVADPAAVEAGVRRLEQRLGPITALLMHASGVN 1284
PKSE|135E      ADPAEDAELAANLARMATAAGLRVRYERADVTSAAQTAEVERLEHALGPVTAVLHGAGRN 1303
PKSE|145B      ADPAADTELAENLARMADAAGLRVRYARADVTSADQVAAAVDLEAELGPVTAVLHGAGRN 1318
      : * * * * : : . . * . * : . . : : * : * : * : * :

```

Figure 8H

PKSE MACR	EPAALTALGIEDFRRTFAPKVDGLRAVLDAVGEKSLKLLVTFGSIIGRAGLRGEAHYATA	1351
PKSE CALI	TPAALADLDEAALRGVFVAAKVDGLRAVLAAVDPARLRLLVTFGSIIGRAGLRGEAHYATA	1350
PKSE 009C	EPNALARLAPRDFERTFAPKVDGLRTVLDVDPGNLKLVTFGSIIGRAGLRGEAHYATA	1370
PKSE NEOC	EPGELHALEPEDFRRTFAPKVDGLRTVLEAVDAEELKLLVTFGSIIGRAGLRGEAHYATA	1389
PKSE 007A	EPAALFSLTEESFRKTLAPKIGGLNAVLDVDAVDKDKIKLLVTFGSIIGRAGLRGEAHYATA	1361
PKSE 028D	EPQSLATLDEDSFRRTLATKIDGVEAVLAAVDTAALRLFTFGSIIGRAGLRGEAHYATA	1378
PKSE 054A	EPAALANLDMGAVRRTFAPKLDGLSATLAAVDPDRLRLVTLGSIIIGRAGLRGEAHYATA	1349
PKSE 132H	EPTRFDPLDDTRFTTHLAPKTIGLRNLLAALEPRRLRLVTFGSVIGRHGLTGECHYATA	1344
PKSE 135E	EPAAVTSLTTPDDFRRTLAPKTDGLAAVLDAVAPERLKLLITFGSIIGRAGLRGEAHYATA	1363
PKSE 145B	EPAALETLAEDFRRTLAPKTDGLEAVLAAVEPERLKLLITFGSIIGRAGLRGEAHYATA	1378
	* . * . :*. * *: * *: :*:*:*:*:*** ** **..** *	

Figure 8I

```

PKSE|MACR      NEWLADLTEEVARNHPGCRALCMEWSVWSGVGMGEKLSVVETLSREGIVPVSPDHGVEIL 1411
PKSE|CALI      NEALAEALTREVAATRPECRAVCLEWSVWSGVGMGERLSVVESLSGSGVTPISPDDGLRVL 1410
PKSE|009C      NEWLADLTEEVARAHPNLRARCMWSVWSGVGMGEKLSVVESLSREGITPVSPDQGV DIL 1430
PKSE|NEOC      NEWLADLTEEIIARTHPPQVRARCVESVWSGVGMGEKLSVVESLSRQGIVPVSPDQGV EIL 1449
PKSE|007A      NDWMTELTVRFCQEHPRAKAIALEWSVWSGTGMGEKLGVSALMRDGITPIPTTEEGIEIL 1421
PKSE|028D      NDWLTDLTVRFQDDHPHCRCLALEWSVWSGSGMGERLGVLEALVREGIEPIPTEDGVALL 1438
PKSE|054A      NDWLAEATTAFGRHRHCRSLCLEWSVWSGVGMGERLSVVESLHREGVAALTPDQGVAVL 1409
PKSE|132H      NGALRAEAERLAAELPDCRVLNLDWSVWSGAGMGESLGVLDLTLRLDVTPIPVPEGVELF 1404
PKSE|135E      NDWMTELTTLRFAEKHPQARVLAIEWSVWSGAGMGERLGVVEALMREGITPISTEEGIRVL 1423
PKSE|145B      NDWMTELTTLRFRRRHPQARAIALEWSVWSGAGMGERLGVVEALIREGITPISTENGIOVL 1438
*      :      :      :      :      :      :      :      :      :      :      :      :      :

```

KR << >> DH

```

PKSE|MACR      LRLISDPDAPVVTVISGRTEGIATVRRDLPQLPLLRFACTPLVRYHGVELVTEVELNSGT 1471
PKSE|CALI      REVVADDTLPPVVVVTGRTGGVETLRYHRSELPLLRFTERPLVRYDGIELVCEVDLARTT 1470
PKSE|009C      LRLIEDPDAPVVTVVSGRTEGIDTVRRDLPDLPLMLRFTGNPLVRYHGVELVTEVELNAGT 1490
PKSE|NEOC      LRLIRDPDAPVVTVVSGRTEGIETVRRDLPPLPLLRFTGTPLVRYHGVELVTEVELNAGT 1509
PKSE|007A      RQVVGDPAAPSVLVVCGRTAGLATLPVEKRELPLTRFVDRAVVHYPGVELITEADLSAGS 1481
PKSE|028D      GRLLATPGTDTALVVMGRAGGLPTLTLEQRELPLLRFLERPVHYPGIELVADAELTGGG 1498
PKSE|054A      RRLADPEATGTVVVSGRTQGIDTIGYDRPELPLRRFLEKPLVHYPGVELVAETELNVGT 1469
PKSE|132H      LRLLGTHDLPTTVAVHGRLGGLFTVG---KPLFGGRFLETVPAYCPEVELVADSRLDLDR 1461
PKSE|135E      RELADPSAGPVLVVSAGRAAGLPTLALEQRDLPLARFLERVVTHYPGVELVTEAELSEGS 1483
PKSE|145B      REVLADPSAGPVLVVSAGRVGGLPTLTTAHRELPLTRFVERVVVHHPDIELITEAELTEGS 1498
*      :      :      :      :      :      :      :      :      :      :      :      :

```

```

PKSE|MACR      DAYLADHLLDGNLLMPAVLGMEAMVQVAHAATGWE--KVPVIEGAKFLRPIVVPNGATR 1529
PKSE|CALI      DPYLEDEHRLDGDLLFPVAVLGLEAMAQVATALARHP--GVPVIEEVRFDRPVVVDPTGT 1528
PKSE|009C      DPYLSDEHLLDGNLLFPVAVIGMEAMAQVASAVTGRT--GVPVIEDAEFLRPIVVPSPGSTR 1548
PKSE|NEOC      DPYLGDEHLLDGNLLFPVAVMGMEAMVQVAAAATGWP--GTPVIEGARFLRPIVVPDGGSTT 1567
PKSE|007A      DPYLADHLLDGDQLLEFPVAVIGMEAMTQVAKAALAAETLPAPVFSDFEFLRPIIVSPGGSTT 1541
PKSE|028D      DRYLPDEHLLDGDLLFPVAVLGMEAMTQAATALTGRR--DTPVLEGMEFLRPIVVPVTGATT 1556
PKSE|054A      DLYLADHLLDGNLLFPVAVFGMEAMAQVAAAVRGTD--DVPVIERAEFLRPIIVPPYGRGR 1527
PKSE|132H      DAYLRDHRIDGLAVLFPVAVGMEAMAQVASALAGRP---LREMTDVTLERPVIVPEDGDRM 1518
PKSE|135E      DPYLTDHQLDGDLLFPVAVLGMEAMAQAAAASVGHQ--GPPMLEAVEFLRPIAVRPPGGSTT 1541
PKSE|145B      DPYLTDHRLQGDLLFPVAVLGMEAMAQAAAASVGHQ--GPPLLEDVEFRPVVVRPPGGSTT 1556
*      :      :      :      :      :      :      :      :      :      :      :

```

DH <<

```

PKSE|MACR      IRIAATVTGPDTVDVAVHAEETGFVAEHFRARLRYA-EGAI PDGAPDQVGAGVPAAPLVP 1588
PKSE|CALI      VRVAALVRSEQVIDVVLRSVAVTGFAADHFRARLRFAPDETYPAYTAAPAPAE LPAVPLDP 1588
PKSE|009C      IRIAAVVTAPDTVDVAVHAEETGFVAEHFRARLLFT-GAAVPDGPPLQVPDDTPVVPLDP 1607
PKSE|NEOC      IRVAATVTGPDTVDVAVHASDTGFAAEHFRARLVYS-VAGVPDGPPLQTGSDTPEVPLDP 1626
PKSE|007A      IRLAALARDAETVDVVLRSGETGFSAHFRARLSFS-RPDPLG-DTVARDVALPPVPVDP 1599
PKSE|028D      LRTAVLATGPDTVQAVLRSGETGFQADHFRATLRYG-AARPED-EPAPVTDEVPRVPLTP 1614
PKSE|054A      IRVGAVATDDDTVEVAVRSED TDFVADHFTARLRYG-VGPAPEGPPEQLSDDLGPVSLAP 1586
PKSE|132H      VRVCALRQD-DAVLVVLRSDETRCQVDHFRARFPLT---PVSGATPSEEDFPEGEAGLNG 1574
PKSE|135E      LRTAALVQDTGTVDVVLRTSDTGFAADHFRARLRYL-RPALPD-SPRPAALDLPAVPVDP 1599
PKSE|145B      IRIAALVRAPGTVDVVLRSADTDFAADHFRARLRYL-RPGVPT-TPVPVAFGLPTVPVDP 1614
*      :      :      :      :      :      :      :      :      :      :

```

>> domain ?

```

PKSE|MACR      ATDLYGGVLFQGRFQRLGTFHRAAARHVDADVAIG-APTGWFAPYLPATLLMADPGMRD 1647
PKSE|CALI      ARDLYGDVLFQAGRFRRIKGYRQVAARVAEAEVVTS-DAASWFS AFLPGRLVLGDPGARD 1647
PKSE|009C      ATDLYGGILFQGARFQRLHRFHRAAARHVD AEVAVQQRPEGWFAGFLPGRLLLADPGMRD 1667
PKSE|NEOC      ASDLYGGILFQGSRFQRLRRFHRMAARHVDADVTVR-RPEGWFAGFLPAEMLLADPGMRD 1685
PKSE|007A      TTELYGTVLFQGKRFQRTGYRRASARHVAE VATG-AEVDWFAFPLPQEQLLADPGTRD 1658
PKSE|028D      -AQLYGPVLFQGRFRRLAYRDLAATHCLAEIDDT-PRTDWFAGYHPGELLLADPGTRD 1672
PKSE|054A      DADLYGGLLFQGSRFQRLRGYRRAAAKCVDAEVAAL-DGVDWFAFVPDEL LLGDPGVDRD 1645
PKSE|132H      -DELYGP LFFHTCRFRLVRRFSALAAARHCRVRLHAS-EHAPDGLALLGDP SLLGD LASND 1632
PKSE|135E      VTELYGSVLFQGKRFQRLLDYRRAGARHVAE VSTT-TPAPWFAAYLPQEQLLADPGTRD 1658
PKSE|145B      VTELYGSVLFQGKRFQRLLEYRRAGARHALAEISTT-AQAPWFAAFLPQDQLLADPGTRD 1673
*      :      :      :      :      :      :      :      :      :      :

```

Figure 8J

		domain ? <<	
PKSE	MACR	ALMHGNQVCVPDATLLPSGIERLYPMAAGTDLP-AKVRYCATERYRDGDTYVYDIAVRTE	1706
PKSE	CALI	AFMHGIQVCVPDATLLPEGIDRIWSAGPKLSAT-EAVTMTAREREQHGTAYVYDVVVRDA	1706
PKSE	009C	ALMHGNQVCVPDATLLPSGVERIHGSGEHVP-DRLRYTAVERSRDGDTYVYDIAVRDE	1726
PKSE	NEOC	ALMHGNQVCVPDATLLPSGVERVHPLGNSGNVP-DQLRYCAVERSRDGDTYVYDIAVRDA	1744
PKSE	007A	AMMHAIQCCVPDATLLPQGIERLYLAEPGEQHP-EYVLLDARERSQDGDSYVYDLDVRNP	1717
PKSE	028D	ALMHAIQACVPDATLLPVSVLHLAEAAAARTGRLLFLDARERSRDGDSYLYDLDVRDA	1732
PKSE	054A	ALMHGNQVCVPDATLLPMGVDRHHPGAAALSGR-RDLRFCAVERSRDGDTYVYDVALRDP	1704
PKSE	132H	ATVHALQACVPHRLLPVGCERFAVEPDAG----AAVEVLASERHAGGGEYVWDVVALDR	1688
PKSE	135E	TMMHAIQCCVPDATLLPRGIERLHLAERADQDS-EFIVLDARERSQDGDTYVYDVDVRTP	1717
PKSE	145B	AMMHAIQCCVPDATLLPQSIERLWLADRADQDS-EYVVLDAVERSRDGDTYVYDLDVRTP	1732
		: :*. * ***. *** . :*. : * ** * *::*:	

Figure 8K

```

>> PPTE
PKSE|MACR      DGTVVERWEGTLHAVRKTGSGPWVEPLLGSYLERTLEEVLGSHIAVAVEPDAPD--AD 1764
PKSE|CALI      TGAVIEHWMGLRLRAVRPHAPRGSWPPALLGPLLQRLAEVFPGDIAVAAAPGG----- 1760
PKSE|009C      NGVVVERWDGLTLHAVRKTGSGPWVAPLLGPYLERSLLEDVTGSRIAVAVEPHGDA--PA 1784
PKSE|NEOC      EGTVVERWEGTLHAVRKTNGSGPWVAPLLGPYLERRTLEEVLGSHIAVAVEPHGDN--PA 1802
PKSE|007A      DGKLVERWEGTLRAVRKRDGSGPWVPSMLGSYLERSVRLLGSSRAIVVEPDFVGV-PV 1776
PKSE|028D      AGSPVEQWEGLLRAVRKQDGSWPVAPLLGPFLERRVEAALGHRVRCVVLPGGED--AD 1790
PKSE|054A      DGRTVERWEGTLRAVRKQDGSWPVAPLLGAYLERRTLDDLVTGPVAVTVQPDGPDGPD 1764
PKSE|132H      DGRRRASWSGLRLRDTGSLPASGPWAAALLSVYLERSVLALVPAPRLTVRIGAG----- 1742
PKSE|135E      DGRVVERWEGTLAVVRKRDGAGPWVPAMLSYLERGLERVLGGSRAVVVEAAG----- 1771
PKSE|145B      SGTVVERWEGTLAVVRKRGAGPWVPAMLSYLERGLERVLGGSRAVVVEPAPD---AA 1789
      *      * * * *      * *      : * :      * : * :

PKSE|MACR      GSQGSRRAGTAVALQRALGATAEVRYRDPGRPEIDGG-----LQVSAAHGLGVTLGVAA- 1818
PKSE|CALI      -----GPRDSGALLSRALGQPVVVRHRPDGRPEVDLP-----YTVSVAHSAPLDLAVAG- 1809
PKSE|009C      GSVAQRRGFTADAAARALGSPVAVRHRPDGRPELEPDRH---LSVSAAHGLGVTL SAVS- 1840
PKSE|NEOC      GSVAERRALTITIAASRTLGAAVTVRHRPDGRPEVDGG-----WHISASHGLELTVSAVA- 1856
PKSE|007A      ETTPERRAQTALAAGRAVDAPLEIRYRDPDGKPEADG-----VEVSASHSADLT LAIAG- 1829
PKSE|028D      GSVADRRRRRTAEAAASWALGRTTEVHHRPDGRPELADG-----RRISSSHAAGVTFTTVAD 1845
PKSE|054A      DHVARRRRARTRSAAGQVFGRPVEVRYRDPGRPEVPGG-----QNLVSAHAGGLTLCVSS- 1818
PKSE|132H      -----ERFGGS-----RSRHAGPADLSG-----RECR-SYQNGMVL SVSA- 1776
PKSE|135E      -AEADRRARTADAVARALGAPAE LRHRPDGRPELDG-----HTVSAHSDGLTLAVVG- 1823
PKSE|145B      TADQDRRSRTETAVGRALGRPVKLRHRPDGRPELDGGPGLEGRTVSASHDAGLTLAVVG- 1848
      :      : * * :      : :      : :

PKSE|MACR      -GRTVACDVEAVNV RTEADWTGLLG-EHAALAKLVAKETGEAPDTAATRVWSAAECLKKA 1876
PKSE|CALI      -DGTLACDAEPVAARPADVRDLVG-RHGAVAALLVAEVDGPPDVAATRVWCAEECLQKA 1867
PKSE|009C      -DTEVACDIEAVSMRSAHEWRGLLG-EHAVAELVAKETGEAPDTAATRVWGAVECLRKA 1898
PKSE|NEOC      -RAEVACDIEAVSMREPSEWQGLLG-EYA AVELVARETGEAPDTAATRVWSAVECLRKA 1914
PKSE|007A      AGR-IACDVETAIERTPEDWAGLLGEDLLAVGELLAADAREPLSVAHTRVWSALECVRKT 1888
PKSE|028D      AGRPLACDVEQVAERTAEQWAGLLGPDAERLAHLLAAERGEPLSTAATRVWGAVETLRKA 1905
PKSE|054A      -AETVGCDEVPVTGRSAQTWTGLLG-RHTDLARLLAAEAGEEIDVAATRVWTALECLQKA 1876
PKSE|132H      -AARVACDWEAVGRRTDDEWLLLVGSRFELIGQLRTMLTEPVTHTAARVWTAVECLSKI 1835
PKSE|135E      QGR-LACDAETVRPRRAEDWAALLGEAQLPVRDLLVAEAGDDPAVAATRVWCALECLRKS 1882
PKSE|145B      AGR-LACDVESVRERTAEDWDGLLGAGRLALRNLLATEAGEDRAVAGTRVWSALECLRKA 1907
      : * * *      * *      :      :      : : * * * * : *

PPTE <<
PKSE|MACR      GVMAGAPLT LAPRT--RDNWVFTAGALRIATFVTSLRGALDPAVFAFLTDGADDVPGVK 1934
PKSE|CALI      G-RPEGRLTLLPGA-LPDGWVLDAGDVRVATRVVAVAGAAAPAVVAVLSGAGR----- 1919
PKSE|009C      GIMAGAPLT VLP RR--KDAWVVFAGDLRIATFVTALRDALEPAVFAFLTHEPQSAQERG 1956
PKSE|NEOC      GAMAGTPLTVLPQK--KEAWVVFAGDLRIATFVTALRDALEPAVFAFLTRTPELLEGRS 1972
PKSE|007A      G-DMTQALT VHRVD--PDGWAVLSHGGA RIATWVTTVNDRTDPVVFVAVLQGEES----- 1939
PKSE|028D      G-HAVAALSLADGSGLP PGWVALRGGAHRIVSFVTALDGAADPVAFTVLTGGAR----- 1958
PKSE|054A      GRRSGDPLVLHPVT--RSGWSVLVSGDVRVATFATTVRDVTEPVVFAVLTEGR----- 1927
PKSE|132H      GYP PGVPLVLGGVY--DEGWVVLRTG SVTLVSTVVPISGADSPVAVAVLVAAPEGGDRG- 1892
PKSE|135E      A-ATGQALALDRVD--GSGWAVLSAGDAAIATWVTTVTDREDPVVFAFLAGKER----- 1933
PKSE|145B      G-ATTQALT LDRVH--PDGWAVLSAGDATVATWVTTVNGRTDPVVFVAVLAGKEN----- 1958
      .      * :      * : *      : : : :      * : : : *

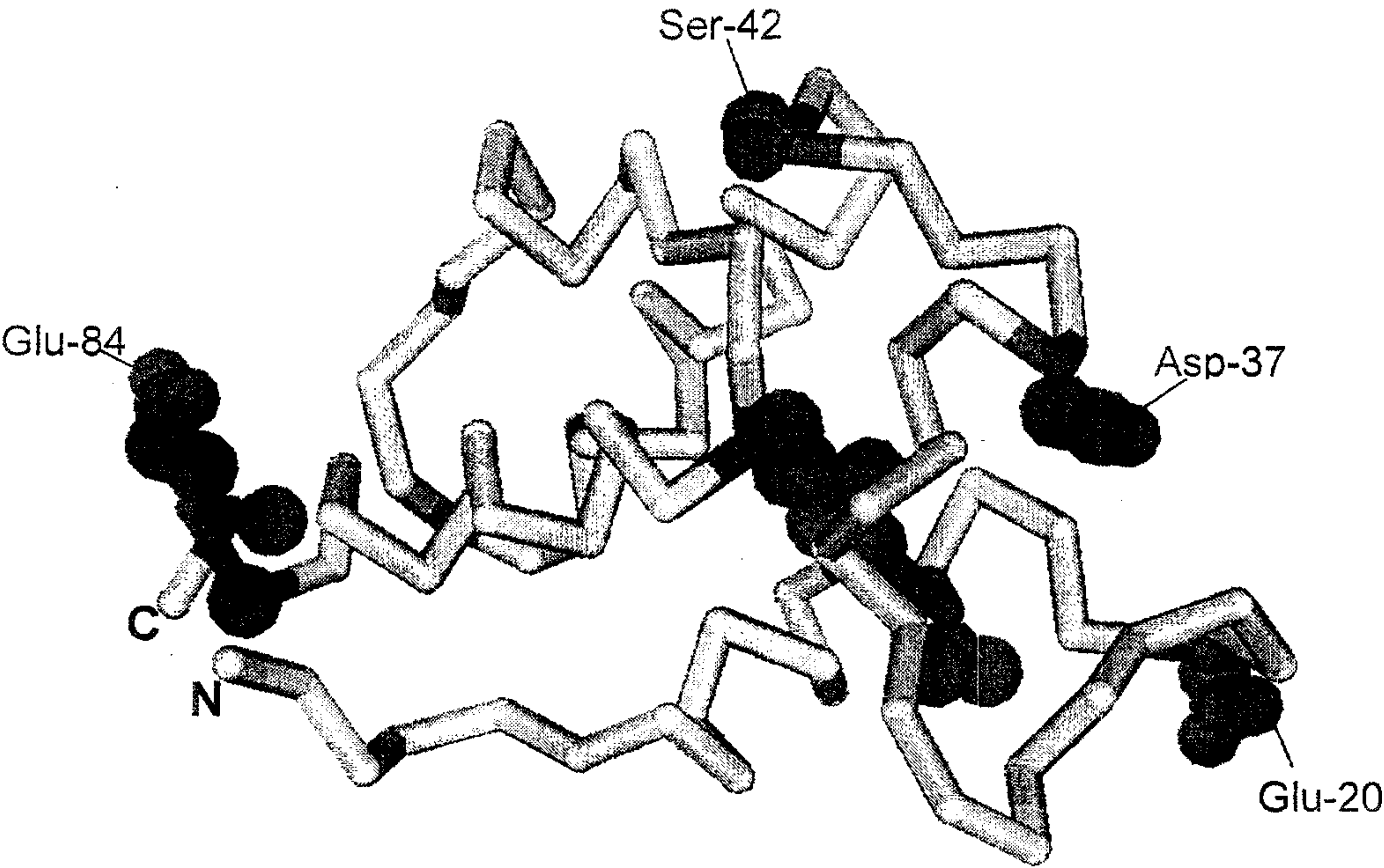
PKSE|MACR      GA--- 1936
PKSE|CALI      -----
PKSE|009C      -----
PKSE|NEOC      QDYVG 1977
PKSE|007A      -----
PKSE|028D      -----
PKSE|054A      -----
PKSE|132H      -----
PKSE|135E      -----
PKSE|145B      -----

```

Figure 9A

MACR	939	-----LEVLRA	VAERAELP-PELVADDSSL-LDDLHMSSITVGQIVNQAATRLGIG-	988
NEOC	975	-----LDVLSA	LVVERAELP-AHLVDPDSRL-LDDLHLSSITVGQIVNQAMAQLGIAP	1025
CALI	939	-----LELVRH	LVARAELP-VEVLRDDSRF-LDDLHMSSITVGQLVNEAARAMGLS-	988
1AF8	1	MATLLT	TDDLRRALVECAGETDGTDLSGDFDLRFEDIGYDSLALMETAARLESRYGVSI	60
		::	**	.*
			::	: ::*
				.**** : . .
				*::
MACR		-----AAHVPT	NFATATVAQLAEALE	1009
NEOC		-----AAQEP	TNFATATLAELAEALE	1046
CALI		-----AVAMP	TNFATATVREMAEALE	1009
1AF8		PDDVAGR	VDTPRELLDLINGALAEAA-	86
		.	* ::	:***

Figure 9B



PEVPLDPASDLYGGILFQGSRFQRLRRFHRMAARHVDADVTVRRPEGWFAFLPAEMLLADPGMRDALMHGNQVCVPDATLLPSGVERVHPLGNSGNVPDQILRYCAV
PAAPLVPATDLYGGVLFQGDRFQRLGTFHRAAARHVDADVAIGAPTGWFAFLPATLLMADPGMRDALMHGNQVCVPDATLLPSGIERLYPMAAGTDLPAKVRYCAT
PAVPLDPARDLYGDVLFQAGRFRRIKGYRQVVAARVAEAEVVTSDAASWFS AFLPGRVLGDPGARDAFMHG IQVCVPDATLLPEGIDRWSAGPKLSATEAVTMTAR

ERSRDGDTYYVDIAVRDAEGTVVERWEGTLTHAVRKTNGSGPWVAPLLGPYLERFLEEVLGAIHIAVTVEPHGDNPAQSV AERRALTIIAASRTJIG ---
 ERYRDGDTYYVDIAVRTEGTVVERWEGTLTHAVRKTGSGPWVEPLLGSYLERFLEEVLGSHIAVAVEPDADPGSQGRRAGTAVALQRALG ---
 EREQHGTAYYVDVVVRDATGAVIEHMMGLRLRAVRPHAPRGSWPPALLPLLQRRLAEVFPGDIAVAAAAPGGGPRD-S-----GALLSRALG ---
 -----MKIYGIYMDRPLSQEENERFMFTFISPEKREKRRF-----YHKEDAHRTLLGDVLVRS-VISRQYQLDK

AAVTVRHRPPDGRPEV-DGGW-HISASHGLELTVSAVARAEVACDIEAVS-----MREPSE-WQGLLGEYAAVAELVARETGEAPDTAAATRVWSAVECLRKA
ATAEVRYPDGRPEI-DGGL-QVSAHGLGVTGLGVAAGRTVACDMEAVN-----VRTEAD-WTGLLGEHAALAKLVAKETGEAPDTAAATRVWSAVECLRKA-
QPVVVRHRPPDGRPEV-DLPY-TVSVASHAPLDLAVAGDGTACDAEPVA-----AR-PADVRRDLVGRHGAVAAALLVAEVDGPPDVAAATRVWCAEECLQKA-
SDIRFSTQVYQKPCIPDLPDHAHNISHSGRWVIGAFDSQPIGDIIEKTKPISLEIAKRFFSKTEIYS-----DLLAKDKDEQTD-----YFYHIWSMKTESFIKQE

73 76 89/90 107/109 127 147 151 155

β_2 β_3 β_4 α_4 α_5

4
 -----GAMAGTP--LTVLP-QKKEAWVFTAGDLRIATFVTALRDALEPAVFAFLTRTPELLEGRSQDYVG
 -----GVMAGAP--ITLAP-RTRDNWVFTAGALRIATFVTSLRGALDPAVFAFLTDGADDVPGVKGA---
 -----GRPEGR--LTLPGALPDGWVLDAGDVRVATRVVAVAGAAAPAVVAVLS-G---AGR-----
 -----GKGLSLPLDSFVRLHQDGQVSIELPDSHSPCIKTYE-VDPGYKMAVCAAHPDFPEDITMVS-----YEELL-RAAAA
 [B5] → [B6] → [B7] → [B8] → [α6]

Figure 10B

Figure 11

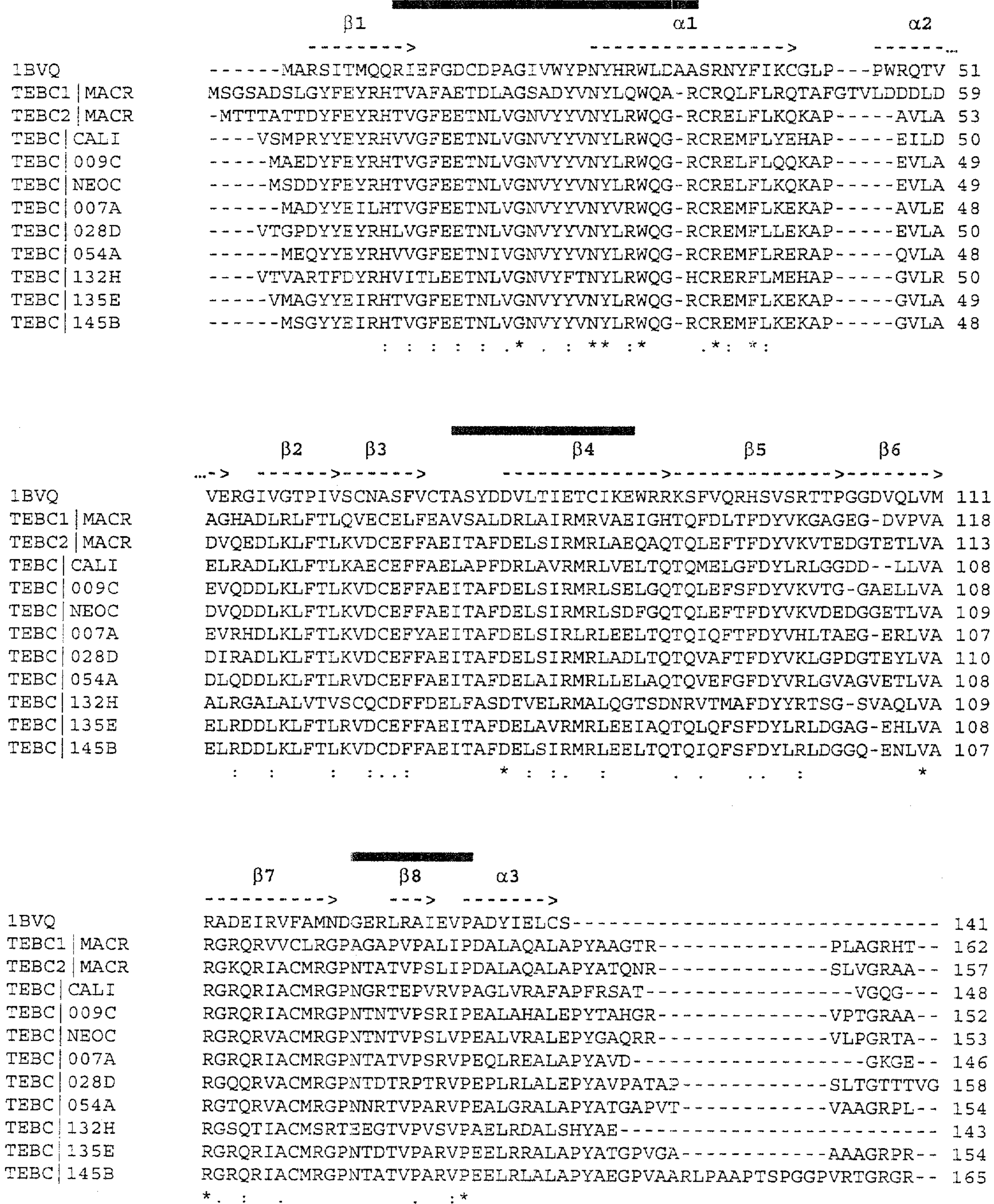


Figure 12A

```

UNBL|MACR      -VSMLRALRRRVLTTPNVRETHLDRRGFHVKNPEAKNQLETVGATFLOQYAYAVEARSAAE 59
UNBL|CALI      MPANWRTIRQYALTPGMAQTTFATRGFRARDEPTRERLESVGAHFLTGYGHAVGARGPDE 60
UNBL|009C      VSSVFGALRRRLLTPPVSETTMEVRGFHVKNNAEAKKRLEYIGEVLFRGYAYAVEAGSPA 60
UNBL|NEOC      VSTAFGALRRRLLTPPISETTMKVVRGFHVKNNTESKKRLEQVGETFLOQYAFAVEAGSSAE 60
UNBL|007A      LGNGWRTIRRRMITPDVSETSLDKRGFHKKSPAAQELLETVGEKFLLYAHAVEARSVEQ 60
UNBL|028D      MSGGWRAVRRRLLTPSHNETKLSTRGFHLKDDAARTNLETVGGTFLDGYAIAVEARDQDE 60
UNBL|054A      -MKAWRTLRRRVLTTPDVSQTLMSVRGFHVKDQAGRERLENVGRYFLTGYAAAAEARTAAE 59
UNBL|132H      -----VAALGQELDQVDFGRRRFRRLRTGPARETLERAGRSFLEGFNAAVAYPADDR 51
UNBL|135E      VPSPLRMLRRRILTNPVAETQLARRGFHVKTPDARELLETVGSRFLEGYAYAMEAGTPAE 60
UNBL|145B      VPSAWRKLRRRILTTPSTSETLLEKRGFHRKTPAAQQLLETVGERFLEGYGYAMEARDTGS 60
               :      :      :      * * :      :      * * * * * :      *

UNBL|MACR      AEEWLETVPRAYRGFAYEGAGMGAVMLDSLTSK - -RLAGFLEGEGARHDYMIWVGVGWA 117
UNBL|CALI      AVGALETVPDLRGFAYEGAAMGLAVLDGLTGGR - -RIARFLAGPAARHVYMHVVGVGWA 118
UNBL|009C      AQERLETVPRDVRGFAYEGAGMGAVVHDALPGHGG - RLQGLLAGQGRHHDYMIYVGIGWA 119
UNBL|NEOC      AEELLETVPRDFRGFAYEGAAMGAVIHDALPGHGG - RLAGLLSGRGYHDYMIHVGIGWA 119
UNBL|007A      AEEWLEQIPVKYRGFAYEGAGMGYGLDGLPGGRRHVERFLDGPGEKHDIYIYVGVGWA 120
UNBL|028D      AHERLERIPVRYRGFAYEGAAMGLAMLDGLPLPGNDRVARFLAGHGAPHDYMVHVVGVGWA 120
UNBL|054A      AEPVLEAVAAPYRGFAYEGAAMASAVRDALPGGGR - HVRDLLAGRGDRHVYMAVVGVGWA 118
UNBL|132H      LASEIERIDVPLRGFAYEGAGMACAVLDILTSLGGRNTRALLRGPASDYPHLVHVVGVGWA 111
UNBL|135E      AETRLETVPARFRGFAYEGAGMGFAMLDGLPLPGRGRVGEFLAGRGARHNYMVYVGIGWA 120
UNBL|145B      AENLLEGVPIRFRGFAYEGAGMGFAMLDGLPLSGSGSVARFLAGRGADHVYMVYIGVGWA 120
               : * :      * * * * * . * :      : * * . : : : * : * *

UNBL|MACR      MARLP-KFLWPDVTEIDPVLRLWLDGYGFHQAYFKTDAYVRKPHLTHPFSWKGGDD-TY 175
UNBL|CALI      MARLP-RWRRHAIQPADRLLGWLALDGYGFHQAYFHTRRYVWSHRRDEVLPWPGDPIGRW 177
UNBL|009C      MARLP-KPLWPDISATDPLLRLWLDGYGFHQAYFRTDAYVRNPHLQHPFRWRGCHN-HY 177
UNBL|NEOC      MARLP-RPLWPDIRTTDPLLRLWLDGYGFHQAYFKTSAYVRDPSPAPFRWNGCHN-HY 177
UNBL|007A      MARLP-RFRWPSAEDFDPLLRLWLDGYGFHQAYFKTAKYVDGQFQDPDFSWPPGND-GY 178
UNBL|028D      MARLP-RFRWASIAPPDPLLRLWLDGYGFHQAYFRTARYVHQHRESDFPWPBGDTGRY 179
UNBL|054A      MARLP-RPLWGRLLHAPDPLLRLWLDGYGFHQAYFKTSRYVDGQYRHDRFPWPVEGPADY 177
UNBL|132H      FARLRLRPGWGRAVVRDPLLRLWLDGYGFHQGFHTDRVIGGKVVEHGLTE---D--- 164
UNBL|135E      MARLP-RFRWPDITGLDPLLRLWLDGYGFHQAYFRTEQYVHGRHRERSFPWPADDS PGY 179
UNBL|145B      MARLP-RFRWPDIDALDPLLRLWLDGYGFHQAYFRTARYVHEQYRDPAPFPWPAHDS PSY 179
               : * * :      * : * * * * : * * : : : :

UNBL|MACR      SQRVLDQIGIGRALWFGGTDPDVVGGLIAAYPEHRHGDLYAGSGLACTYAGSADETELKR 235
UNBL|CALI      TGRVVDQIGIGRALWFGGADTDRIADTVDGFPDRHEDLYSGVALAATYAGGAPPEDLRR 237
UNBL|009C      TANAIDQIGIGRALWFGGTDPDVVSGLIRAYPEHRHGDLYAGAGLACAYAGGAGEDELAH 237
UNBL|NEOC      TANAIDQIGIGRALWFGGTDPDTVAGLIRSYPEHRHGDLYAGAGLACAYAGGATEQELAL 237
UNBL|007A      ALRAIDQIGIGRALWFGGTDVDLVADTVARFPERRHGDLYAGIGLASTYACGVTGDELLK 238
UNBL|028D      AGHAIDQGVGRALWFGGTDPAVVADLVLDGYQADRHADLWAGVGLAACYAAGATDAELRL 239
UNBL|054A      AARVVDQGVGRATWFGGTDARRVVQTFDRFAADRRPDLYAGAGLAATYAGGAGAQLTW 237
UNBL|132H      QRAIRDQGVGRSLWFGGECADPEAVALRIDDFPRNRRPDLSGVGLAATYAGGVRADELES 224
UNBL|135E      VSRAIDQGVGRALWFGGTDPDVATLIEKFERRHSDLYGGAGLAAAYAGGVDEAELRA 239
UNBL|145B      AGRAIDQIGIGRALWFGGTDADLVATMIEKFERRSDLYSGAGLAATYAGGVDEAELRA 239
               * * : * :      :      :      : * : * : * . * * . * :

UNBL|MACR      FAEHAGKHLPNLAQGSFAFASAEARVKAGTIIDHTKMASRVLCAGRTPPEASRVCLETRPTG 295
UNBL|CALI      LRERGGAYAPAMAQGSFAFAAEARERAGLTTAHTAVATDVFCG-APPAAEAAVTQAALADL 296
UNBL|009C      FAEGAGEHRWALAQGAFAAEARERAGLTKAGTVIDHTHLAARVVCG-TTAEKASQVCRDTRPQA 296
UNBL|NEOC      FARQAGEHRWALAQGAFAAEARVRAGLVSEHTHLASRVVCG-VSVEQASRMCNGLRPSV 296
UNBL|007A      LAEFAGEHRGNLAQGSFAFAAEARVRAGLLIPETEVAATRAICG-LPAERAAAITQEVVPAT 297
UNBL|028D      LLDRAGPHRPQLAQGAFAAATARIEAGLLTEHAEAAATAVLCG-LTPQQAAGVCTRARPRP 298
UNBL|054A      LRDAAGPYAADLAQGAFAAGARVRAGLVVPHNEVATRI LCG-LPTQAAAATVDEARIDL 296
UNBL|132H      LALLAGEYRADLAQGCSEFACEARRVSGVPEHTRLAAPILAG-VTADVAGSWANRAQHAL 283
UNBL|135E      FQDRAGIHRAIVAQGAFAFACEARIRAGLLGPHTELAAQVLCG-TDAAAAAKVTQDLRPTG 298
UNBL|145B      FWERAGSHRAMVAQGSFAFAAEARERAGLPVVHTELATRVFCG-MTPAQAAARVTQEVVPAG 298
               :      . * :      : * * : * * * * :      :      :      *

```

Figure 12B

UNBL	MACR	CDGG-AIPAFETWRQEIARQIVVPAYSQKGASA	327
UNBL	CALI	DRDG-PEPAYLVWRQRIAKQFVTLG----	RC-- 322
UNBL	009C	PDVHGAFPAYEKWRRDIAAQLASSSLVSKGADQ	329
UNBL	NEOC	ASRN-ALPAYENWRRDISARLASESTLRKGADQ	328
UNBL	007A	VVDG-ELPAFETWRQRIAEAILSG-----	GAGK 324
UNBL	028D	AVDG-PVPAYEVWRQEIADRITDLA---AG	AHR 327
UNBL	054A	AAVG-ATPVYEVWRQRIKSTFASTR----	RA-- 322
UNBL	132H	GPADGTSAQYQQWRAGIRNLWADNME---	GQPS 313
UNBL	135E	RSTG-PVPSYETWRQAIADTFTAAR----	S--- 323
UNBL	145B	PVPG-ALPAYEVWRQAIADRFTNVG----	GC-- 324
		. : ** *	

Figure 13A

```

<Putative N-terminal signal sequence>
UNBV MACR MTQNPVSWLRRQAAGIVALVVMVSAFYAVKPDSSAAEKRELAJSFQFEPMSIAMPAGFK 60
UNBV CALI -----MFRRLAGLVALVLLTGMVVLVRQPEANADERRAMAEPYRFTPMSLPMPPGLP 53
UNBV 009C -VTVAKNWIRRNNSPGIVALALMVGTFFYGVRLPESSAAEIDKLAKDFSFEPMISIALPAGFE 59
UNBV NEOC -VTMAKNWLRNNSPGIVALTLMASVFYVVRLEPSAADVRESAADFAFEPMTIAMPGGFP 59
UNBV 007A -MTATLGWLRKQLAGIVALVLMAGLFVVAQLPTVSTAEADTMASKYAFEPLTIALPEAAK 59
UNBV 028D --MSPRRRRFRLVPGLVTVLVAASMFFVIRSSVAVAGG-DDAAAAAYRFKELPIAMPPGYE 57
UNBV 054A LPSQPVSPRLRLMAPILVLVLAATMFLVARLPSASADTRAELIAGRFAFTELPIELPPGLP 60
UNBV 132H --VLSAPGFRFRFVPKIAAALCCVLAWVLAQPPSVAEDRARLATRFGFDHHAISPARDG 58
UNBV 135E -MPTVLGRIRLQLPGIIAVLAIVTGYFLVLPPTTSAAEQDRMASRFHFTALPIELPPASK 59
UNBV 145B --VAYRERFRRLIPGLVVIVVATSLFFAVRTSVAVAGG-EQVAKQYGFKEMPIAMPPGYD 57
      : * . . . . : . * : * . : .

UNBV MACR QQ---TVRKVNKAYKHIEAWISSVGAGVAMNDIDNDGLPNDLCITDPRIDQAVITPAPVP 117
UNBV CALI QQ---SIRRVNGAYQHIAAWISSVGAGAAMNDLDGDGLANDLCVTDPRVDRVVTPAPT- 109
UNBV 009C RQ---EVRKVNKAYQHIEAWISSVGAGVAVNDLDGDGLANDLCVNDVRIDQAVVTPAPT- 115
UNBV NEOC TQ---KIRQVNKAYEHIDAWISSVGAGIALNDMDGDGLSNDLCITDPRIDQAVVTPAPS- 115
UNBV 007A SQ---SIRTVNKEYEHIRAWISSVGAAIAVNDLDGDGLANDLCFVDPKSDQVITPTPG- 115
UNBV 028D SRPMNTIRKVNPAVEKIRAWISSVGASIAINDVTGHGLADGMCIVDTRTDSVVVTWTPTA 117
UNBV 054A ER---SIRTVNPKYEHIRSWISSVGASIAVNDLDGQGVANDMCLVDTRSDAVIVTPVPTD 117
UNBV 132H DR---RMRAVAPVYERIRNWSSVGAGAGLFAVDGGVSHDCLVDPRTDVTVTEPAPT- 114
UNBV 135E HQ---TIRTVNQDYEHIRAWISSVGAAITMTDLRGTGKSGDLCLVDTRTDQVVVTPVPG- 115
UNBV 145B RQPMNTVTRTVNPAVYQKIRSWISSVGASIAINDLTGHGVADGMCIVDTRTNSVIVTYTPTA 117
      : . * * * : : . : . : . : . : .

UNBV MACR GRKSATYEPFALDPAPLPKSKYSAPIGCVPGDFNEDGATDLLVYYWGRTPVIFQKKMAW 177
UNBV CALI -AGADRYQPFVLDPAPLPMNPYVAPMGCLPGDLNADGRTDLLVYYWGRTPVVFLAR---- 164
UNBV 009C --RKDAYEPFALDPAPLGTSKTMAPMGCMGPDYNEEDGRMDLLVYYWGRTPVLFMNE---G 170
UNBV NEOC --RGKAYEPFALDPAPLGISDTMAPMGCVPGDFNEDGAIDL VYYWGRTPVIFQNE---G 170
UNBV 007A -KGGDRYAPFALDPAPLPMGKYIAPMGCVPADYNEEDGRVDLLAYYWGRTPIFLSK---- 170
UNBV 028D P-EADRFTPFVLDGRPLPMDTMAPTGCTPGDFNGDGRNDFLVTYWGRTPVLFMAR---- 172
UNBV 054A ---GATYQPFVVDPAPLPMGSTIAPMGCTPGDFNLDGRMDLLVIYWGRTPVVFLQK---- 170
UNBV 132H --TGERYAAFTLKPATLPYAGYVAPMGCLPADLNEDGWQDVVYYWGRSPVLFRLT---- 168
UNBV 135E -SDGTRYAPFALDPAPLPMNEYIAPMGCVAGDFNEDGRTDLLVYYWGRTPVLFGLR---- 170
UNBV 145B R-PADRFTPFVLDPAPLPMDTMAPTGCTPGDFNGDGRMDLLVTYWGRTPIELMAK---- 172
      : . * . . . * * * * . * * * * . : : : : :

UNBV MACR DKKMPPKMECFEPIELVPG--AGGGIYTGPLWNSNAAAVADFDDGDGHKDLIYIGNYFPESP 235
UNBV CALI ADATGLSRAAYHPVELVPGAATGGSRYDGPKNWNTNAATLADFDDGDGHLDVYIGNYFPDSA 224
UNBV 009C EKGKPLTADSFTPTTELLPG--KPGRTYTGPLWNSNAAAVADFDDGDGHDDIYIGNYFPDSP 228
UNBV NEOC GRGEPLTASSFTPTTELLPG--KPGPRYTGPLWNSNTAAVADFDDGDGHDDIYIGNYFPDSP 228
UNBV 007A PGATKLEPKAYEPVELVPGNNKNGEYSGPLWNTNAASVGDFDDGDGHQDIFIGNYFPDSA 230
UNBV 028D ADAAAPAAESYVPRELVASE-SLDGRYHGPRWNTDAAYVGDLGSGHPSIIIGNYFPDSD 231
UNBV 054A EGATALSNDTYQPVLPVQRTDDGVYRGPLWNTNAVTVADFDDGDGRPDIALENYFPDSQ 230
UNBV 132H -PGSAPAAAGFSERELVS-----PQVWNTNAATTADLDGDGHLDLFFGNYPFDGA 218
UNBV 135E PDATTLDAHAFQPVLPVGPNETDGKYTGAQWNTNTATVADFDDGDGHQDIFIGNYFPNGP 230
UNBV 145B SDATTPSASSYVPRELVPSQ-SLDGKYHGPRWNTDADYVADLDGSGHPSIVIGNYFPDSD 231
      : . * : . * : : . * : : : : :

UNBV MACR -VLDDTKDG-GVTMNDMSHAQNGGGGHFFLKTPG--GYKWIP---ADKVVP-EGREKG 286
UNBV CALI -VLDDTVHG-GVAMNRSMSNGLNGGEDHVFRWTGGTAGATPSASFAEVPDVFD-TKVSRG 281
UNBV 009C -VLDPKDG-DVTMNASLSHAQNGGGGHFFRWTE--GYEKV-----DDAIP-QSLNKG 277
UNBV NEOC -VLDPKNG-DVTMNDLSHAQNGGGGHFFRWTES--GFEKT-----DDAIP-QGLNKG 277
UNBV 007A -VLDDRVS-GVEMNKSMHADNAGGKYILRFTGATQGAKPSATFALDDKAIP-ADSQGG 287
UNBV 028D -VLDPQGI-RVQMNDSLSAKNAGGDHVLRFHSSTAGAAPDARFVEEKDAIA-FDASTG 288
UNBV 054A -VLDPGLP-NVQMNHSMHGHNAAGAHVLRWSGATSGDDPTVTY-EEQVALD-PRYATG 286
UNBV 132H RVLDPTAQQPELVMTDSLSGDHNGGTHRYFRFASATGGSTPDVRYAEAVDPVEGDSRTTG 278
UNBV 135E -VLNDQVSG-GVVMNHSMSHAQNSGGKYILRRTGGNVGDKLSAGFECSDDAFP-DEAKHG 287
UNBV 145B -VLDPHGLN-NVVMNNSLSARNAGGDHVLRWYRSTSGPEPTVSYVEEKDAIP-YSASTG 288
      ** : . * : * . * * : . *

```

Figure 13B

UNBV	MACR	WTLAASATDVGDLLPELYLGLDFGSTLLHNRSTPGELKFVEVKAS-HNGIMPKSKEIG	345
UNBV	CALI	WTLAVAANDLDGDQLPELYVANDFGPDRLHNRSERGRIAFAPVESPGLPGLTPKSKRLG	341
UNBV	009C	WTLGASAADLDGDGLPEMFLAHDFGTSALLHNTSTPGRLRFSEVKSE-HTATIPKSKELG	336
UNBV	NEOC	WSLGASAADLDGDRLPEIFLAHDFGTSALLHNTSRPGRIEFREVKAV-HSGTVPKSKEIG	336
UNBV	007A	WSLAASATDVGDNLPELYIGNDFGHDRLLYNKSRPGHVEFAEVKGI-RGPNEPKSKVIG	346
UNBV	028D	WTLAIAGADLTGDALPEVYIANDFGHAHLLHNVSTPGRIRFEEATGE-RTPPTPKSFVLG	347
UNBV	054A	WTLGAASADLDGDLLPELYLANDFGQDRFFHNVSTPGRIRFNLAEGV-RTPLTPKSLVLG	345
UNBV	132H	WTLAAAAQDQDGLPELYVANDFSPDQLLVNVSTPGQIRFREAHGE-RHALTPKSKVVG	337
UNBV	135E	WSLASSAIDLDGDQLPELYVANDFGNDRMLHNVSSPGHPKFVTVTGP-RDATTPKSKILG	346
UNBV	145B	WTLAISGADLTGEGLPDLYIANDFGHAHLLYNRSTPGHISFTEAKGE-RTPPTPKSFVLG	347
		:. :. *: : **::::. **. :. * * *. * . . . *** :*	

Figure 13C

UNBV	MACR	RSSFKMGIDWADLNGDGILDAFLSNITTSFGIQESNYTFISKAKNLDDLHRKMSDKKAP	405
UNBV	CALI	HDSFKMGVDFGDDIDGDMFDLYVGNITTSFGIQESNFAFVNTAADTAALRAALWAGEAP	401
UNBV	009C	RSSFKMGVDFGDDLHDGLYDMFVSNITTSFGIQESNFAFISDAGSRSELQGRFAEGEAP	396
UNBV	NEOC	RSSFKMGVDFGDDLHDGLYDMFVSNITTSFGIQESNFAFINKAGDKADLRSRFENGAP	396
UNBV	007A	NDSFKMGVDFADLDHDGLYDLVSNITTSWGIEESHFQFMNTAKDTADLRGRLOQGEAP	406
UNBV	028D	KGSFKMGVDFGDDVGNDSFDMVSNITVAWGLEESNFWINQAKDPAEMKRKLTDRIAP	407
UNBV	054A	HDSFKMGVDFADLHSGRFRDMFVSNITESWGLEESNFWHNTAASPEAAREQLSRGIAP	405
UNBV	132H	NDSFKMGASFAIDLNDGMPDILVSNITEPYALQESNFAFISTGD-----RDALRRGVAP	392
UNBV	135E	NDSFKMGVDFGDDLDRKGLYDLFVSNITASFGIEESNFWQFMNTARDKADLRAQMDDGTAP	406
UNBV	145B	NGSFKMGVDFGDLGHDGRFDMVSNITVPWGLEESNFWINQAKDNADMRRKLSSGVAP	407
		..***** :.: * * :.*** :.:***: : . . : : **	
UNBV	MACR	WKDESASLNTAWSGWDAKMAFDNDGRPEITQATGFVKGRSRWAQLQELATANDLLV	465
UNBV	CALI	WHDRSAELGLAWSGWSWDVKFGDFTNRGDPQIVQTSFVKGEVNRWAQLQEAATANDDLL	461
UNBV	009C	YKDKSTGLGLAWSGWWDVKMGDFDNGDLEIVQALGFVKGNRWPQLQELATANDALV	456
UNBV	NEOC	YRDESTDLGLAWSGWWDVKMGDFDNGDLEITQALGFVKGNRWPQLQELATSNDALV	456
UNBV	007A	WVDRSAQAGTAWSGWWDVKIADYDNGGESVITQATGFVKGDVNRWPQLQELATSNDLL	466
UNBV	028D	FTQEAADHGVAVTGWGWDKMGDFRNSGQQDILQADGFVKGNIDRWPWLQEMAMTNDLL	467
UNBV	054A	FENRAARRNLAWVGWWDVKMAFDNNGGLEVVQAAGFIKGDINRFNWLQELAMANDLML	465
UNBV	132H	FDDRSEELGLSRSGWSWDVKAADFNDGAAEVMHATGFIKGTNRWPQMGEAAMSNDLIL	452
UNBV	135E	FEDRSAAAGTAWSGWWDVKMGDFDNGQLAIVQATGFVKGGTNRWPQLQELATANDALL	466
UNBV	145B	FTQEAQQYGMVAVTGWGWDKMGDFLNSGDLVQLQADGFVKGNIDRWPWLQEMAMTNDLL	467
		:.: : : **.*. * :. * * :.: **:* .*: : ** * : ** :	
UNBV	MACR	RHPGAWPKVEAGDDLQYLRFWTRN--GGKYDNISPELGLDVPVPTRGITATGDADGDG	523
UNBV	CALI	ANPRWWPKVEQDDIAGGQHLAFHVRG-ADGRYEDLSHELGLADRVPSRGITATADADGDG	520
UNBV	009C	PNPQWWPNVEHGGDLQAGSQRMRFFAKDPDSGRYHNLVALGIGEPITRGITATGDVGDG	516
UNBV	NEOC	ANPTWWPNVRQDDLAGSQRMRFFAKDQDTGRYINLSTALGLGDPVPTRGITATGDVGDG	516
UNBV	007A	KHPYFWPNMVAGDDVGGDHTLHFVAKSS-DGRYTDLAPRLGLAVPVPTRGITATGDADGDG	525
UNBV	028D	SNPKLWPNVPGDDLAGDETMAFYARTD-SGKFANISKQLGLDVPITRGITATDTTGTG	526
UNBV	054A	REPAMWPNAPGDDISGNPNVAFWVRED-NGRYVNLSPELGLDEDTPSRGISVADPDGDG	524
UNBV	132H	GNPALWPRFTEEDGLSGHNRNTFFTRDG-AGRFADVADLVGVGTDAVSRAFAVGVDGDG	511
UNBV	135E	DNPWWPNVQAGDDLQAGDQTLRFFVKGT-DGHYSNLGRLGLAVPVPTRGITATGDYNGG	525
UNBV	145B	SNPAMWPNVQPGDDIAGDEAIAFYAKTP-DGRYVNISQLGIKVPTRGITATDTTGTG	526
		. * ** . * :. * :. * :. * :. * :. * :. * :. * :. * :. * :	
UNBV	MACR	RLDFVVARQWDAPVFYCNLSPAKNDFLGLKLVDK-----	558
UNBV	CALI	RLDLVVARQWDAPVFYRNDSPDTGSFLTLLRLHEQA-----PAAG-----P-L	562
UNBV	009C	RLDLVVARQWGEVFLHNVSKSPGAYLGLKLTHPD-----	551
UNBV	NEOC	RLDIAVARQWDEPVFYRNTAPEPGSWLELVFTHPD-----	551
UNBV	007A	KLDFAVARQWEQPIFYRNVSPGTGSYLNLKLVDKA-----SADG-----P-L	567
UNBV	028D	ALDFAIARQWGPPAFYANQSANLGHDLTLRLYRPATDTATTGTTGAGATAATDATAGPGL	586
UNBV	054A	AQDIAVARQWGAPAYFRNTRGSDNHLSLRLSRPA-----LAADG-----RTP--	567
UNBV	132H	RLDFVANQWAQSTLYRNTSQSSAEFVGLRLRQPADVG-----TCAG-----N	554
UNBV	135E	RLDFAVARQWDAPVFYRNDSPDPGAYLGLRLTYDTP-----DAAG-----P-L	567
UNBV	145B	ALDFAVARQWGPPAFYANTSPNLGDYLNLRLYRPA-----GAGG-----AGQGL	570
		.:..*** . * :. * :	
UNBV	MACR	---G-SPVVGAQVRVNFDPDGKVLGRVDGGSGHSGKRSTDVHIGLG-KVSGPVQASISWR	613
UNBV	CALI	AGAG-SPVVGAQVRVTPDGRVLIDRVDGGSGHSGRRSNEVSLGLD-DVTGPVSVHLTWR	620
UNBV	009C	---G-SPVVGAQVRVTPDGRVLIDRVDGGSGHSGKRSTDVHIGLGKDVDPVPATITWR	607
UNBV	NEOC	---G-APVVGAQVRVTPDGRVLRVDGGSGHSGKRSTDVHIGLGEEAQGEVSGTVTWR	607
UNBV	007A	KAAG-TAAIGAQVTVPDGGKYMVRVDGGSGHSGKRSTHEIQIGLGK-VTGPVKVCLQWR	625
UNBV	028D	ATTG-TPAYGATVCTTPDGRKQIQQLDGGGGHGGFRSFDVRFGLGTQSG-PVTVDLAWR	644
UNBV	054A	STTGTSPAYGAHVTTTADGRTOVGQLDGGGGHSGRRSFDVFFGLGDAADRPVSVQLCWR	627
UNBV	132H	SEGADRPAGATAVVTTPDGTKHSQQVYPANGHNGVNPDLVFGLDVDRDGPPLVLSWR	614
UNBV	135E	PAPG-SPVIGAQIEVTTTPDGRKLIDRVDGGSGHSGKRSHYVHIGLGQGVSGPLVVKLQWR	626
UNBV	145B	VNTG-SPAYGATVRITTPAG-TQISQLDGGGGHGGFRSFDVRFGLGTYHG-PVTADLTWR	627
		. . . ** : . * :. . **.* :. : ** . : : **	

Figure 13D

UNBV	MACR	DTTGQVRKQALTLSPGWHSVQLGTEAKEK----	642
UNBV	CALI	DRSGAPHEQELTLAPGRHTLTLGSQAREVSR--	651
UNBV	009C	DRAGEVHEQKLKLTPGWHSIQLGTEAKEK----	636
UNBV	NEOC	DREGDVHEQEVRLAPGRHSFELGSQVKEK----	636
UNBV	007A	DLTGQIRTQEVQLTPGDHTFQLGAQAKEK----	654
UNBV	028D	DNAGGLHTETRQLSAGSHTLMLTDDIQEVAAR-	676
UNBV	054A	DLNGQTHRQTIDLTAGRHDLLLTDRAEELNRR-	659
UNBV	132H	DGCGRRHTATVNVAPGWHRIILLHADGTTMVEDK	647
UNBV	135E	DRTGQVHTQTVTLSPGWHSIQLGSQAKEK----	655
UNBV	145B	DAGGTLHQTTKQLSPGTHSLMLTSDVQEVPSR-	659
		* * : : : * * . *	

Figure 14A

```

UNBU|MACR      -----MSSHTSAVPPTAAAPTAAAGSPAKSAKPAAPPRHDP-----KVIKALQRF 45
UNBU|CALI      -----MTQKPAK-----DPR-----ITALRRF 17
UNBU|009C      -----VTMSAPH-----TAARHNP-----KVTALRRF 23
UNBU|NEOC      -----VTMSAAQPLS-----QTAGPRHNP-----KVVITALRRF 28
UNBU|007A      -----MTVMAEQTLTAPKSN-----KTITALRRF 24
UNBU|028D      -----MTATDISGVATR-----AVRPHAACKQP-----NRDPRYLALRNF 35
UNBU|054A      MTDVKPVEDARGVLTASPDTAGRNG-GNPVPRQQAGGDRPGVPPRVDSRDPYMALRNF 59
UNBU|132H      -----VKTLPRKTKDQ-----GDSAGHRKP-----KDPRSAALRRF 31
UNBU|135E      -----VTDMAKAPTTKQTAHQAPPEAAARPATASRHEP-----KVVIALRRF 44
UNBU|145B      -----MSTTAQPRPKGV-----AAAATAGPAKP-----KDPRYLALRNF 34
               *                                     **:.

```



```

               1A                               1B
...-----
UNBU|MACR      AVSISVLNVLGYTVLGFEGWLWPFIAVATGYTTEIVLEWISAKGEGRAPRYAGGGRKGL 105
UNBU|CALI      AISITILNIAGYTVLGFEGAWWPLIALATGYAVELGLEIIIGARAERRPPRFRGNGLRGL 77
UNBU|009C      AISISVLNVFGYTVLGFEGPWLWPFVALATGYTVEIALELLGARAEGRAPRFTGGGFKGL 83
UNBU|NEOC      AISISVLNIVGYTVLGFEGPWLWPFIALATAYTVELTLEAVGARAERATFRGGGFRGL 88
UNBU|007A      AISITIFNIIGYTVLGFEGPYTYPFIALATAYATEILLEIIIGARAQGRGVFRGNFKGL 84
UNBU|028D      ALSMSVFNIFGYTLLGFEGPWLWPLICAPFGYVVEIVLELISAWAQORTPRFRGNNGRGL 95
UNBU|054A      AMSMTVFNILGYTVLGFEGPWTWPFALAIQYATEILVETVAARATGRSAAYSGNGMWGM 119
UNBU|132H      GLSISVLTIVGHTLLGFEGAYLTPVVAVLVALGTEVLLESVEAASLGRRPRYLQAG-AV 90
UNBU|135E      AISITVFNILGYTVLGFEGPWTWPLIALATAYTTETALELVGAKVEGRRPRFLGNLPGM 104
UNBU|145B      AISISVFNIFGYTLLGFEGPWLWPIIAVLTAACEICFELISAWAQRRAPRFRGNVGRGV 94
               .:*.:.:.: *.:.***** : *... . * . * : * * : * . .:

```



```

               2                               3
-----
UNBU|MACR      MEFLLPAHITSLAVNMLTYTNDRWVMMFGVIVAVGTHVLRAPVKGRMRHYMNPSNFGI 165
UNBU|CALI      VEFLLPAHITALAVNMLLYVNDRVPMFVAVVAISGKWLFRVPVNGRLRHFMNPSNFGI 137
UNBU|009C      VEFLLPAHITALAVNMLSYVNDKMWVMVFGVIVAVGTHVLRAPVKGRMRHFMNPSNFGI 143
UNBU|NEOC      MEFLYPAHITALAVNMLTYVNDQVWVMVFGVVAVGAKWVLQAPVRGRMRHFMNPSNLGI 148
UNBU|007A      VEFLLPAHITGLALNMLTYVNDQVLVMMFGVVAVGAKWVLQAPVRGRMRHFMNPSNFGI 144
UNBU|028D      YEFLLPSHITALAANMLLYANDRLLPILLAVFIGVAAKHVLQAPVYGRMRHFMNPSNFGI 155
UNBU|054A      YTFLLPTHITALAANMLLYANDNFWPIGFAVVVAVGQKALLQAPIRGRMRHFMNPSNFGI 179
UNBU|132H      VDFLLPSYIGGLACAMMLLYANDRLMPTVLAVVIAVASKYLIRVKVNGRLRHVLNPSNTGI 150
UNBU|135E      LEFLYPHITGLALNMLTYVNDRLPVMMFGVVFAVSAKWVLRVVGRGRMRHFMNPSNFGI 164
UNBU|145B      YEFLLPAHITALAVNMLTYGNNQLLPVFFGVVGVGGKHALQAPIAGMRHFMNPSNFGI 154
               ** *:.* .** ** * *:. .:.* ... * :. : **:** :***** **

```



```

               3                               4                               5
-----
UNBU|MACR      MIILVLF-PWASIAPPYHFTEYLTGTTGPGDWILPAVITLTGTMLNAKLTNRMPLIMGWL 224
UNBU|CALI      AVVLLLF-PWISIAPPYQFTEYLEG---PADWIVPAVIVVFGTMLNAKLTRGMWLIAGWL 193
UNBU|009C      AVILVLF-PWASIAPPYHFTEYLDG---GFDWLVPAILITLTGTMLNAKLTERMWLIAGWL 199
UNBU|NEOC      AVILLLF-PWASIAPPYHFTEYVDG---GIDWLVPAILITLTGTMLNAKLTERMWLIVAWV 204
UNBU|007A      TIILLVF-PWASIAPPYHFTEQVDS---WVGWLIVGIIIVSGTVLNAMLTQRMWLIGAWL 200
UNBU|028D      TMLVLFGSWISIAPPYEFTENANT---FFRIGIPLIITTAGTVINAMLTKRVPPLIVGWL 212
UNBU|054A      TATLLVF-SWVNVAAPPYHFTEVND---MISILIPVIVTAGTVLNAMLTKKVPLIVGWL 235
UNBU|132H      VVLLLVF-PWVSIAPPYQFTEWTS---VVDALIPVLLLAAGTMLNAKLTKKIPLILGWV 206
UNBU|135E      AAILVLF-PWASIAPPYHFTEHVNT---PFGWLIVLGILMSGTILNAKLTRGMWLIAGWL 220
UNBU|145B      TMSLLCFGSWFISIAPPYEFTENANT---YFRVMIPLIITAGTVINALLTKRTPLIVGWL 211
               *: * . * .:*****.*** : : **:** ** : ** .:

```

Figure 14B

	5	6	
	-----	-----	
UNBU MACR	VG	FALQAIIRGWMFDT	SIPAALGMMTGTAFLFTNYMVTDPGTSPSKKSSQIAFGAGVAA 284
UNBU CALI	GV	FVLQSVVRGLVLD	TAILPALATMTGVAFVLFTNYMITDPGTTPSRPLSQFAFGGGVAL 253
UNBU 009C	GG	FALQAIVRGLLF	DTSIPAALAMMTGVAFVLFTNYMITDPGTTPSSKWGQIAFGGGVAA 259
UNBU NEOC	GG	FAAQAVVRGLLF	GTSIPAALAMMTGVAFVLFTNYMITDPGTTPSSKWGQIAFGGGVAA 264
UNBU 007A	IT	FALQAIIRGLV	FDTAIPGALGMMTGVAFLYTNMVTDPGTTPSKPASQILFGSGVAL 260
UNBU 028D	GG	FVIQALLRHWI	WDVAIFSALGPMMSGVAFVLYTNMVTDPGTTPSKGRNQFMFGSSVAM 272
UNBU 054A	GA	FIIQALVRHWI	WDVSLWGALVPITGVAFVLFTNYMITDPGTTPTAGWMQFMFGASVGM 295
UNBU 132H	GG	FVLQAVLRS	AFTDLSIVSAVLPIGTAFILFTNYMITDPSTSPSKPRNQVLFGLATAA 266
UNBU 135E	GT	FALQAVLRGLV	FGTSIPAALGMMTGVAFLFTNYMITDPGTTPSKPANQVAFGAGVAL 280
UNBU 145B	GA	FAIQAFIRHWI	WHVALFSALGVMTGVAFLFTNYMISDPGTTPMKGRAQFVFGSSVAF 271
	*	*::*	. :: *: :*:**:*:*****:***.*:* *..** ...

Figure 14C

7

```

-----
UNBU|MACR      VYGLLMAVNITYGIFFFATAIVCGIRGLFLWFLHLQAKQQAAAKTAPVRPVEALVAVPSPA 344
UNBU|CALI      VYGVLTGASIVYGLFFATAIVCLIRGGFLWSLHAVRVA-----ARDGKGTPPAA- 302
UNBU|009C      LYGVLTAMSIAYGLFFATALACAIRGAFLWTADIVAKK-----RAEEALELAAVTR 310
UNBU|NEOC      AYGLLTALSVAYGIFFFATALVCGVRGGFLWISDIVSR-----RTEEALAVAAVG- 314
UNBU|007A      AYGFFMVVHVAYGLFLATALVCLIRGMFLWGLHFSKKA-----TEKWEAEQAK- 308
UNBU|028D      VYGVLMFLFNVVYTLFFATTIVCGLRGAGWWVAHFRNRR-----KQGGGPVEVP- 320
UNBU|054A      VYGVLMVFNIVYTTFFAVTIVCLLRGLFWWGKWLLEK-----HRDVSLDPAP- 343
UNBU|132H      AYAVLVQLHVVFGLFFALVAVCALRGVGLAILSWRQSVEVP-----AQVAQKHDAVVR 319
UNBU|135E      LYALFMIAHIAYGIFFFATALTCLIRGLYLWYVHFRDRG-----HDERQATVLR- 328
UNBU|145B      VYGILMVFNVVYTLFFATTIVCGLRGLGWAAHLLIKRA-----RQADATGAES- 319
      *...:   :...:  *:* . . * :**

UNBU|MACR      PAP-AEAPTVPLAAATEDACAAGTCGHEKCAAARVAAPAAAVPPAAVSPAVPAAASAAAE 403
UNBU|CALI      -----P-----GTVDGTVPPAVAR----- 316
UNBU|009C      SVSRAEAPSG-----AQPDAAAPAQAEAPAPCACPADACSCPAPANTEAAP 355
UNBU|NEOC      -LT-SEKPS-----VERTPIEHAETLPCVN----- 337
UNBU|007A      -----SAEITSLPKPAEK----- 321
UNBU|028D      -----SGQSGAPERPMGN----- 333
UNBU|054A      -----PPAAVALPTAR----- 354
UNBU|132H      -----DRIEEVGVPVFGAN----- 333
UNBU|135E      -----AAAASAPAAEAS----- 340
UNBU|145B      -----SGGTAAQPQQSQV----- 332

UNBU|MACR      DACAAGSCQHGXCAAMRAEAAKEKKVAVSL 433
UNBU|CALI      -----EMVRA-- 321
UNBU|009C      CACPADACSCP---APANTEADPAKTPVAA 382
UNBU|NEOC      -GCAEGACSCSG--NPKDSEADDRIVVSA 364
UNBU|007A      -----PET--GAVAA-- 329
UNBU|028D      -----EAVAA-- 338
UNBU|054A      -----
UNBU|132H      -----GRV---- 336
UNBU|135E      -----GKVAAV- 346
UNBU|145B      -----GAVAA-- 337

```


Figure 16

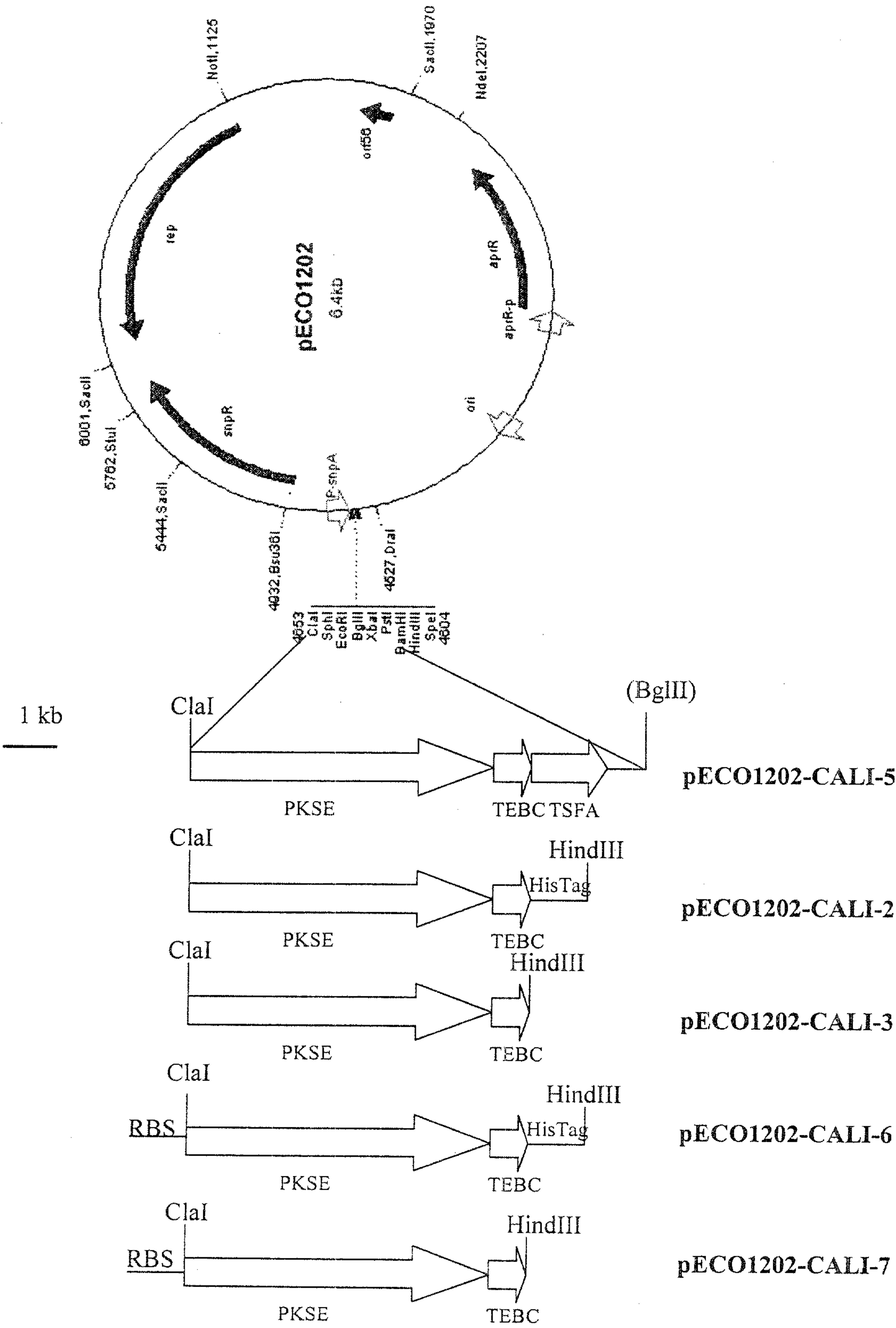


Figure 17

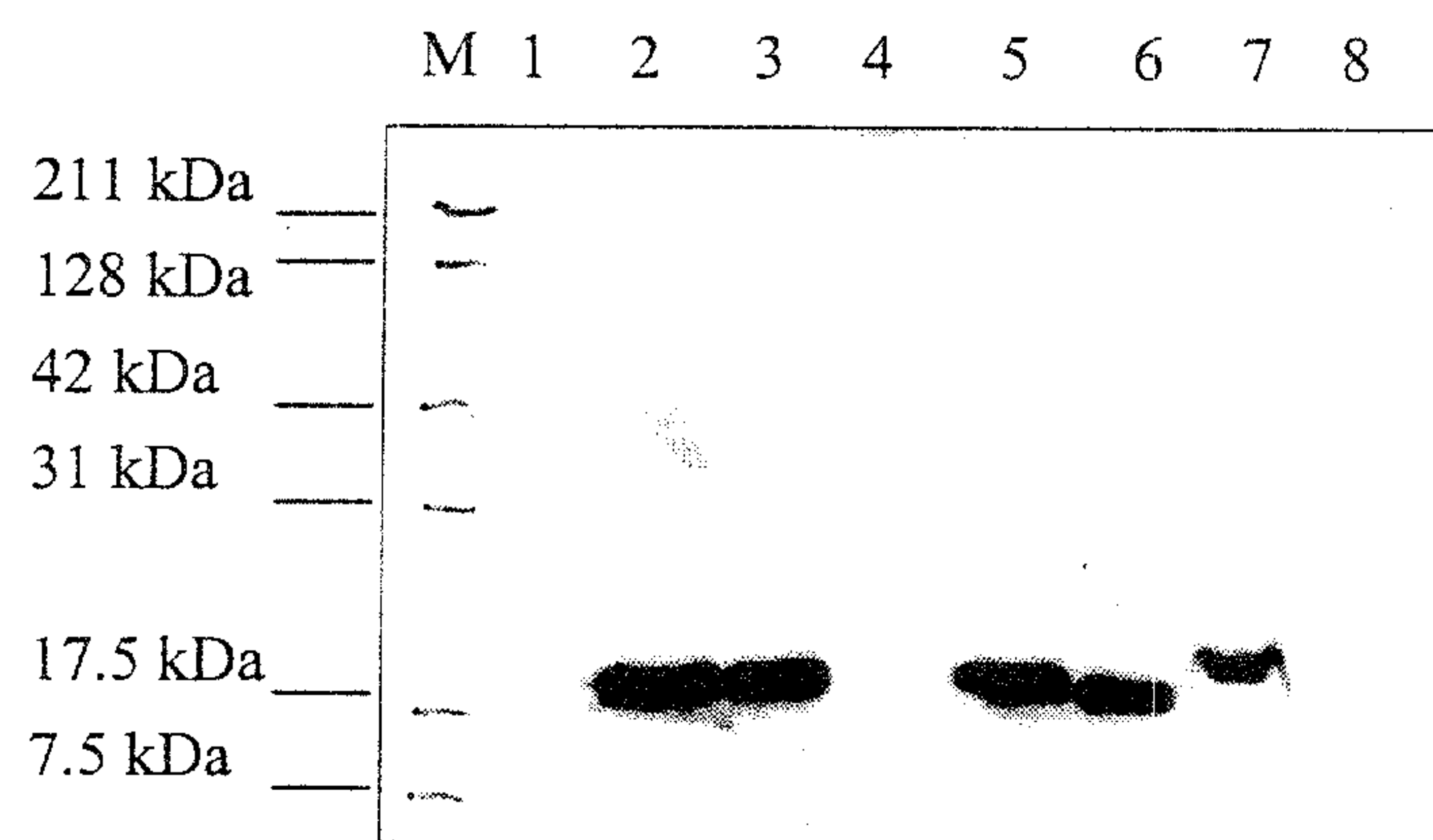
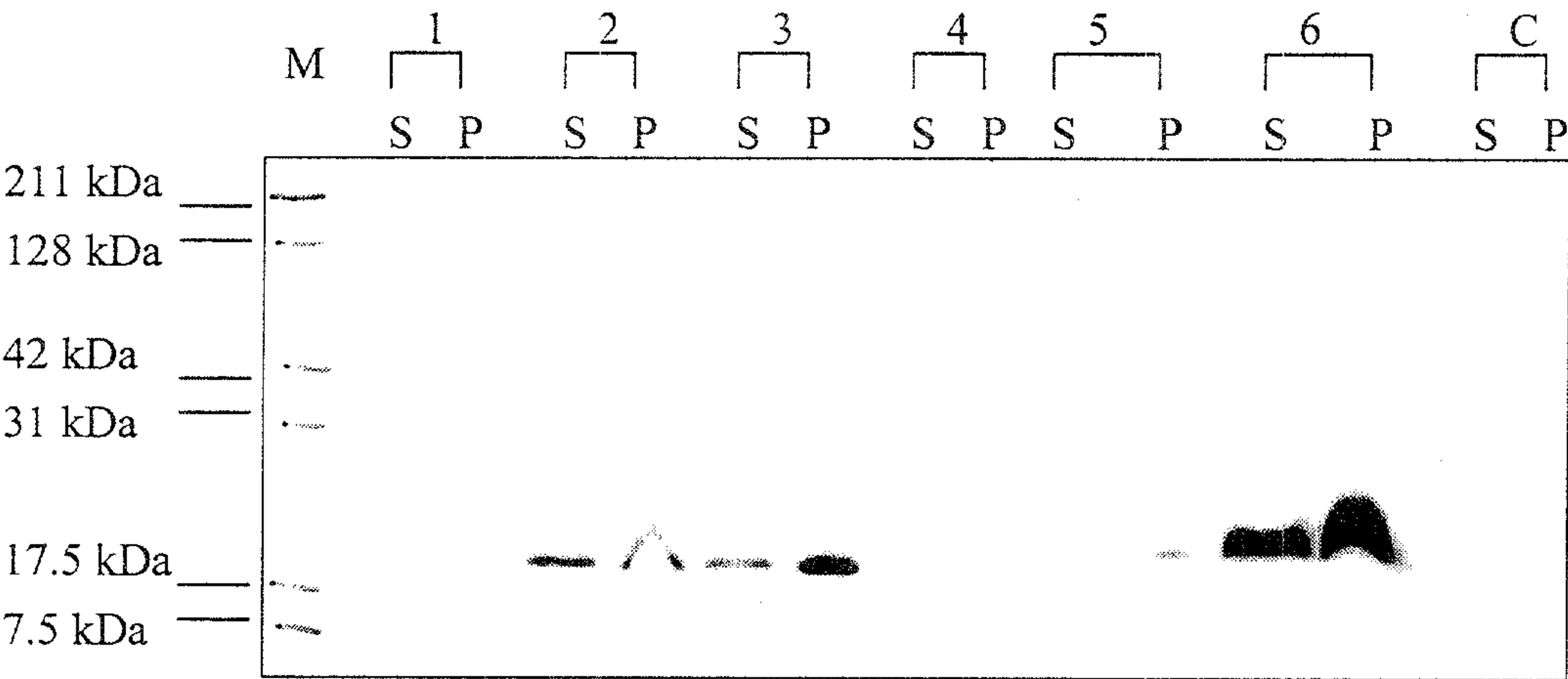


Figure 18



Macromomycin
Streptomyces
macromyceticus



Calicheamicin
Micromonospora
echinatus



009C
Unknown enediyne
S. Ghanaensis



Neocarzinostatin
S. carzinostaticus



007A
Unknown enediyne
Amycolatopsis
orientalis



028D
Unknown enediyne
Kitasatospora sp.



054A
Unknown enediyne
M. megalomicea



132H
Unknown enediyne
Saccharothrix
aerocolonigenes



135E
Unknown enediyne
S. kaniharaensis



145B
Unknown enediyne
S. citricolor



PKSE



TEBC



UNBL



UNBV



UNBU