STUDIES ON THE BIOSYNTHESIS AND ISOLATION OF THE BIOSYNTHETIC GENE CLUSTER FOR THE SALINOSPORAMIDES OF THE MARINE BACTERIUM *SALINISPORA TROPICA*

by

Laura Lynn Beer

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHARMACOLOGY AND TOXICOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Laura Lynn Beer
entitled Studies on the Biosynthesis and Isolation of the Biosynthetic Gene Cluster for the Salinosporamides of the Marine Bacterium *Salinispora tropica*
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____Date: June 27, 2006
Bradley S. Moore

_____Date: June 27, 2006
Laurence Hurley

_____Date: June 27, 2006
Elaine Jacobson

_____Date: June 27, 2006
Matthew Cordes

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: June 27, 2006
Dissertation Director: Bradley S. Moore

STATEMENT BY THE AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Laura Lynn Beer

FUNDING ACKNOWLEDGEMENTS

This research was supported in part by a Washington Sea Grant Industry fellowship (R/B-47) in partnership with Nereus Pharmaceuticals, La Jolla, CA.

ACKNOWLEDGEMENTS

I would like to acknowledge all of the employees of the College of Pharmacy with whom I have interacted over the past years and who have been very helpful and supportive. I would like to thank all members of my graduate committee for their critical review of this dissertation and guidance throughout my graduate studies and especially Dr. Moore for providing the opportunity to work on this fascinating project. A special thanks to Professor Bill Fenical and Dr. Paul Jensen at SIO, UCSD for providing the *Salinispora tropica* strains, protocols for growth and isolation of salinosporamide A and for standards of purified salinosporamide A. Thanks to Dr. John Kalaitzis for assistance with NMR analysis at the University of Arizona and Ray Lam, Barbara Potts and Kathy Reed at Nereus Pharmaceutical Company for assistance with the propionate feeding experiments and for providing protocols for purification of salinosporamides A and B and standards.

TABLE OF CONTENTS

TABLE OF CONTENTS - *continued*

TABLE OF CONTENTS - *continued*

TABLE OF CONTENTS - *continued*

LIST OF FIGURES

LIST OF FIGURES - *continued*

LIST OF FIGURES - *continued*

LIST OF FIGURES - *continued*

LIST OF FIGURES - *continued*

LIST OF TABLES

LIST OF ABBREVIATIONS

DMSO	dimethylsulfoxide
NMR	Nuclear magnetic resonance
HPLC	high pressure liquid chromatography
LCMS	liquid chromatography mass spectrometry
TLC	thin layer chromatography
UV	ultraviolet
HCl	hydrochloric acid
TRIS	trishydroxymethylaminomethane
ATCC	American Type Culture Collection
^{1}H	proton
^{13}C	carbon thirteen
PCR	polymerase chain reaction
PEP	phosphoenolpyruvate
E4P	erythrose-four-phoshate
P 450	cytochrome P450
CCR	crotonyl coenzyme A reductase
PDT	prephenate dehydrase
ChM	chorismate mutase
PKS	polyketide synthase
NRPS	non-ribosomal peptide synthetase
C18	carbon eighteen

LIST OF ABBREVIATIONS - *continued*

ORF	open reading frame
Bp	base pair
DAHP	deoxy-*arabino*-heptulosonate-7-phosphate
NCBI	National Center for Biotechnology Information

ABSTRACT

Salinosporamide A is a potent inhibitor of the 20S proteasome and has recently entered phase I human clinical trials for the treatment of multiple myeloma. This novel marine natural product is produced by the recently described marine bacterium *Salinispora tropica*. Salinosporamide A contains an unusual fused γ-lactam-β-lactone ring structure that is uniquely functionalized. Its biosynthesis has been examined through feeding experiments with ^{13}C-labeled precursors and a biosynthetic gene cluster has been isolated. Feeding experiments indicate that salinosporamide A is assembled from three distinct precursor molecules, namely acetate, a non-proteinogenic amino acid arising from a shunt in the shikimic acid pathway and a chlorinated tetrose molecule. Analysis of the biosynthetic gene cluster revealed that a mixed polyketide synthase (PKS) / nonribosomal peptide synthetase (NRPS) pathway is present. Three ORFs encode a large multifunctional type I PKS with a C-terminal NRPS C-domain (*salA*) an NRPS (*salB*) and a stand-alone type I KS domain (*salC*). Genes are also present for biosynthesis of the unusual amino acid including genes encoding a deoxy-*arabino*-heptulosonate-7-phosphate (DAHP) synthase (*salU*), a cytochrome P450 oxygenase (*salD*) and a potential prephenate dehydratase (*salX*). An unprecedented pathway for biosynthesis of chlorobutyrate is present that utilizes 5'chloro-5'-deoxyadenosine synthase (*salL*) in a pathway analogous to the fluorinase from *Streptomyces cattleya*. The cyclization and offloading may be facilitated by *salO* and *salF* that encode a cyclase and a thioesterase respectivy. A gene encoding a proteasome 20S β-subunit (*salJ*) may be involved in resistance and genes encoding transcriptional regulatory proteins in addition to genes

with no known function are also present in the cluster. Based on results from feeding experiments and the putative gene cluster, a biosynthetic scheme for salinosporamide A and some of the lesser produced salinosporamide analogs is proposed.

CHAPTER 1

INTRODUCTION

An Overview of the History of Natural Products as a Source of Medicine

Nature has long been the traditional source for medicines and the age-old convention of using plants as the basis of sophisticated medicine systems has been in existence for over 4000 years. Evidence of this exists in written records that date from approximately 2600 BC describing oils, herbs and juices from plants, many of which, to this day remain in use to treat a variety of ailments. The World Health Organization reports that up to 80% of the population in Africa uses traditional medicine for primary health care and that in China, traditional herbal preparations account for 30% - 50% of the total medicinal consumption (Cragg and Newman 2001). Some estimates suggest that approximately 80% of the world's inhabitants rely mainly on traditional medicines for their primary health care (WHO 2003). These plant-based systems are not lacking in the remaining 20% of the population, however. Of the prescriptions dispensed from pharmacies in the United States from 1959 to 1980, about 25% contained plant extracts or active principles derived from higher plants (Arvigo and Balick 1993). A more recent study using US-based prescription data from 1993 indicated that over 50% of the most-prescribed drugs in the United States were natural products or had a natural product predecessor in the design or synthesis of the agent (Grifo, Newman et al. 1997).

In 1958, the alkaloid compounds vinblastine and vincristine that were isolated

from *Vinca rosea* were found to have cytotoxic properties and have contributed significantly to the treatment of many cancers. These discoveries inspired the National Cancer Institute (NCI) to initiate a large scale screening program for antitumor agents from plant sources (Zubrod 1984) and over 35,000 plant samples were evaluated between 1960 and 1982 for potential anticancer activity. The most significant and well-known drug to emerge from the massive undertaking was paclitaxel (Taxol) from the bark of the Pacific yew tree *Taxis brevifolia*. Since then, the NCI-60, a panel of 60 human cancer cell lines has been used as the standard assay for screening compounds for anticancer activity. Much of the following information is based on a thorough review of the history and literature relating to natural products and their influence upon drug discovery by Newman, et al. (2000) and in the following paragraphs some of the significant highlights relating to the topic of natural products drug discovery and more specifically, the enormous potential of the marine environment as a source for biologically active molecules will be summarized.

Microbes as a Source for Therapeutics

With the discovery of the β-lactam antibiotic, penicillin G (Fig. 1.1) by Flemming in 1928 and the realization that microbes were a viable source for organic molecules with medicinal value, the search for novel biologically active molecules expanded, though it would be another 57 years before microorganisms (increasingly of marine origin) would be included in the NCI-60 screening. Soon thereafter, the aminoglycoside antibiotic streptomycin (Fig.1.1) was isolated from the actinomycete *Streptomyces griseus*.

Streptomycin was active against the then prolific tuberculosis vector *Mycobacterium tuberculosis* in addition to exhibiting a wide range of antibacterial activities. This discovery led to further work screening metabolites from soil microbes of the Phylum *Actinomycetales*, which led to the identification and isolation of numerous aminoglycoside antibiotics (Newman, Cragg et al. 2000).

While a fungus also makes penicillins, most natural antibiotics come from a group of soil dwelling actinomycetes, of which *Streptomyces* is the best-known genus. These organisms produce a remarkable array of so-called secondary metabolites that have evolved to give their producers a competitive advantage in the complex soil environment where they are exposed to stresses of all kinds (Challis and Hopwood 2003). The *Streptomyces* are spore forming and form filamentous, branched mycelia resembling that of the unrelated fungi. The *Streptomyces* are responsible for a large number of important metabolites including the antibiotics tetracycline and the 14-membered macrolides exemplified by erythromycin (Fig. 1.1) (Omura 1984). Further evidence of the rich chemical and biological diversity of the *Streptomyces*, (Omura, Ikeda *et al.* 2001; Bentley, Chater *et al.* 2002) include the immunosuppressant agents rapamycin (Fig. 1.1) (Cortes, Haydock et al. 1990) and FK506 (Kino, Hatanaka et al. 1987), antitumor agents mitomycin (Hata, Sano et al. 1956) and doxorubicin/adrianamycin (Arcamone, Cassinelli et al. 1969) and the veterinary agents thiostrepton and monensin (Beran and Zima 1993). Historically, microbes isolated from terrestrial sources have been the premier source for commercially successful drugs and the actinomycetes alone account for approximately 70% of the world's naturally occurring antibiotics (Newman, Cragg et al. 2000).

Compounds of microbial origins to recently enter clinical trials include the epothilones, which have shown anticancer activity as great as Taxol (Bollag, McQueney et al. 1995).

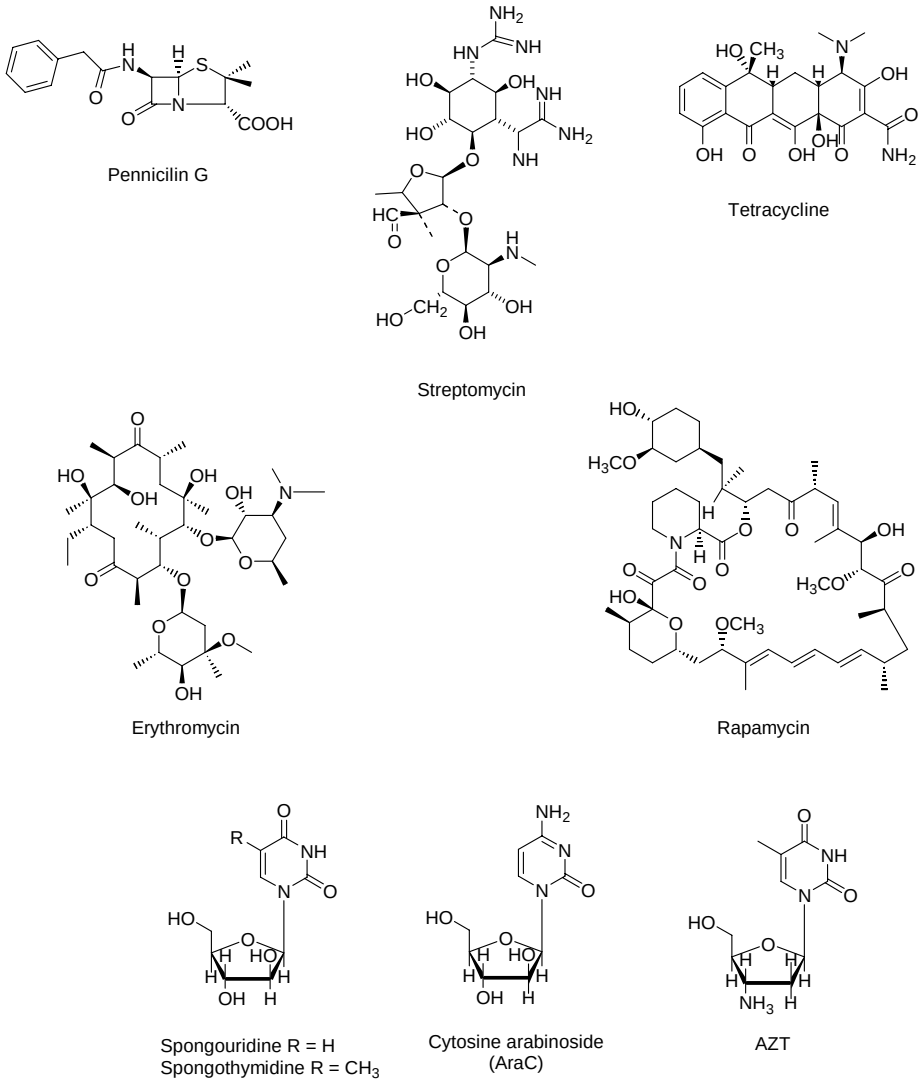


Figure 1.1. Examples of microbial natural products. The fungal and bacterial compound penicillin, and actinomycete products: streptomycin, tetracycline, erythromycin and rapamycin, the marine-derived compounds spongouridine and spongothymidine and the synthetic analogs AraC and AZT.

The Rise of Marine Natural Products as Potential Clinical Agents

The world's oceans are massively complex and the diverse assemblages of life forms that occur in environments of extreme variations in pressure, salinity and temperature require unique metabolic and physiological capabilities to ensure survival in such diverse habitats. They produce chemically and biologically novel metabolites that are not observed from terrestrial microorganisms which points to the ocean and its microbiological resources as an entirely new source for the discovery of new drug candidates (Fenical and Jensen 1993; Jensen and Fenical 1996; Burja, Banaigs et al. 2001). In the next millennium it is quite possible that marine microbial species may replace those from soil for the production of the next generation of antibiotics to treat infectious diseases.

The oceans cover over 70% of the earth's surface and attracted interest as a source for new chemical entities over 50 years ago with the first notable discovery of biologically active compounds from the sea. The compounds spongouridine and spongothymidine (Fig. 1.1) isolated from the Caribbean sponge *Crypotheca crypta* (Bergmann and Feeney 1950; 1951; Bergmann and Burke 1956) possessed antiviral activity and subsequent studies using synthetic analogs resulted in the development of cytosine arabinoside (AraC/Cytarabine, Fig. 1.1) as a useful chemotherapeutic agent. Even today AraC remains the single most effective agent for induction of remission in myelocytic leukemia (Garcia-Carbonero, Mayordomo et al. 2001). Spongouridine and spongothymidine demonstrated for the first time that naturally occurring bioactive nucleosides could be found containing sugars other than ribose or deoxy-ribose. They are

thought of as the prototypes of all synthetic nucleoside analogs and perhaps most notably, the Nobel prize winning anti human immunodeficiency virus (HIV) drug azidothymidine (AZT, Fig. 1.1), though no direct mention was ever made of the original arabinose-containing leads from natural sources (Suckling 1991).

An interesting case of "nature following chemistry" occurred when the analog arabinosyladenine (AraA, or Vidarabine) that had been chemically synthesized based on the spongothymidine structure in 1960 as a potential anti-tumor compound (Lee, Benitez et al. 1960) was later produced by fermentation of *Streptomyces griseus* and isolated along with spongouridine from a Mediterranean gorgonian in 1984 (Cimino, De Rosa et al. 1984). This example gave circumstantial evidence for the involvement of microbes in marine invertebrate-derived metabolites (Newman and Cragg 2006). In fact, recent work has shown that many bioactive molecules from invertebrate organisms are likely produced by bacterial symbionts. In the case of manzamine A (Rao, Kasanah et al. 2004), the base molecule has been isolated from a laboratory fermentation of a *Micromonospora* sp. cultured from the sponge from which manzamine A was isolated. Recent work by Piel *et al.* has shown that the pederin molecule (Fig. 1.2) is a product of a commensal pseudomad isolated from the *Paederus* beetle (Piel, Hofer et al. 2004; Piel, Hui et al. 2004; Piel, Butzke et al. 2005; Newman and Cragg 2006). Additional work with the marine derived pederin analogs; theopederins and onnamides (Fig. 1.2) from the sponge *T. swinhoei* has effectively demonstrated that this family of compounds are also most certainly derived from commensal microbes, sequestered and sometimes modified by the host invertebrate (Piel, Hui et al. 2004; Newman and Cragg 2006). The

biosynthetic gene cluster for onnamide A contains similar genes and architecture to that observed in the pederin gene cluster indicating that they encode the biosynthesis of closely related compounds. This work was the first isolation of natural product biosynthesis genes from a marine invertebrate host, confirming the long-suspected role of symbiotic bacteria in the production marine polyketides and peptides and raises further questions regarding the role of bacterial symbionts and the evolution of the biosynthesis of secondary metabolites (Piel, Hofer et al. 2004)**.**

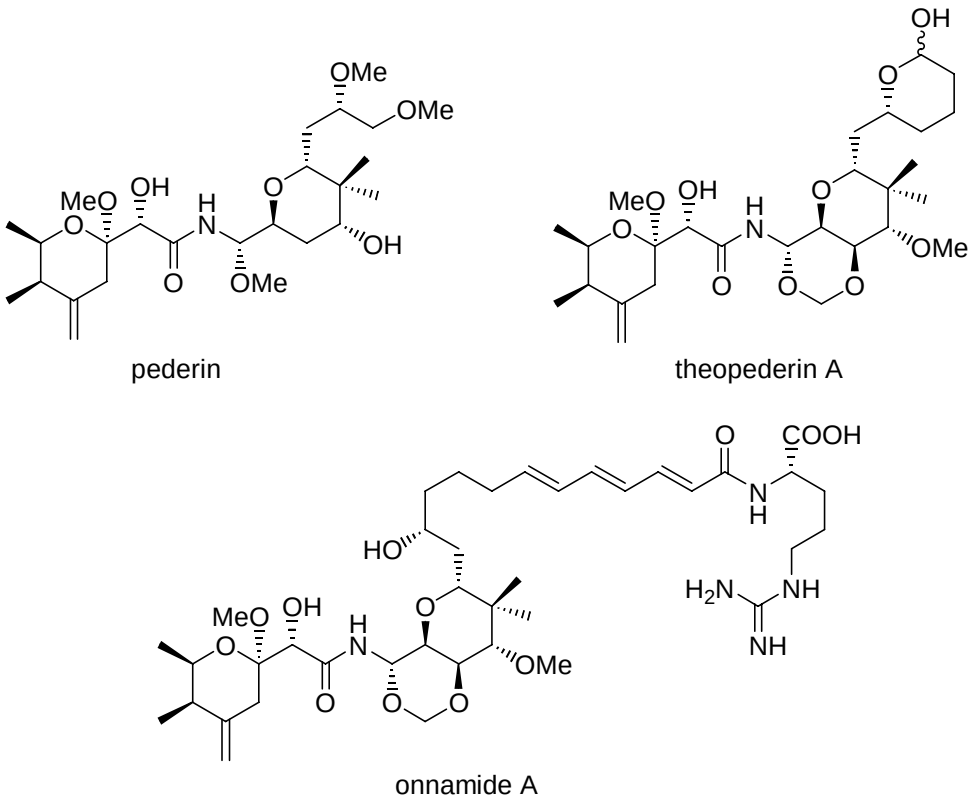


Fig 1.2. Examples of the pederin family of antitumor compounds derived from invertebrate symbionts. Pederin is derived from *Paederus* spp. Rove beetles. Theopederin A and onnamide A are derived from the marine sponge *Theonella swinhoei*.

Marine Natural Products as Anticancer Agents

The mid-1960s gave way to an explosion in the field of marine natural products research and since then over 10,000 new structures have been isolated from marine plants, phytoplankton, invertebrates and microorganisms, again emphasizing the potential of the marine environment as a source for bioactive compounds. Several marine-derived compounds, many isolated from marine microbes have since entered clinical trials as anticancer agents, though as yet, no marine natural product has been licensed for clinical use as an anticancer agent primarily due to issues with toxicity.

The list of targets for marine-derived antitumor compounds is expanding and marine natural products have shown antitumor activity as tubulin interactive agents, proteasome inhibitors, histone deacetylase inhibitors, DNA interactive agents, protein kinase and protein phosphatase inhibitors and protein synthesis/JNK modulators. A thorough review of marine natural products under clinical or pre-clinical investigation as anticancer agents has recently been published (Newman and Cragg 2006).

Recent data have shown that the discovery rate of new biologically active products has slowed in comparison to the increased rate of infectious diseases that are developing resistance toward traditional antibiotics and thus, it is imperative that the discovery rate of novel drug candidates increases (Newman, Cragg et al. 2000; Cragg and Newman 2001). The high throughput screening programs initiated by large pharmaceutical companies in the last decade have failed to yield the anticipated supply of new pharmaceutical candidates. In order to assure the constant flow of new compounds for drug discovery, academic and industrial researchers are constantly in search of new

natural product producing microorganisms from underdeveloped habitats such as the marine environment (Knight, Sanglier et al. 2003).

Marine Actinomycetes are the Premier Source for Novel Bioactive Molecules from the Sea

Early reports suggested that marine actinomycetes were derived from terrestrial sources and that they existed as metabolically inactive spores (Goodfellow and Haynes 1984). Consequently, the examination of marine actinomycetes was predicted to lead to a high isolation rate of known compounds. Therefore, streptomycetes isolated from the marine environment have been largely ignored until recently. Although known compounds are indeed encountered, novel compounds are being reported at a high frequency (Fenical and Jensen 1993; Davidson 1995) indicating that marine actinomycetes have the ability to produce novel metabolites which have the potential for development into useful commercial products (Fenical 1993; Moran, Rutherford et al. 1995; Jensen and Fenical 1996). Advances in molecular technology have led to the discovery of an increasing number of novel microbial species isolated from the marine environment with the potential to produce bioactive molecules (Courtois, Cappellano et al. 2003) and certain taxonomic groups have adapted to life in the sea (Helmke and Weyland 1984; Jensen, Dwight *et al.* 1991; Moran, Rutherford *et al.* 1995; Mincer, Jensen *et al.* 2002). The recent discovery of the major new marine actinomycete taxon *Salinispora tropica* in ocean sediments that is widespread, persistent and metabolically rich (Feling, Buchanan *et al.* 2003; Mincer, Fenical *et al.* 2005) underscores the

importance of marine bacteria as a vast untapped resource of novel biologically active metabolites, warranting further chemical as well as genetic evaluation.

The Marine Actinomycete Salinispora tropica

Salinispora tropica is a group of aerobic Gram-positive marine bacteria recovered from distinct tropical and subtropical ocean systems (Mincer, Jensen *et al.* 2002). The classification of the *Salinispora* as a new actinomycete taxon belonging to the family *Micromonosporaceae* is based on previously undescribed 16S rRNA gene sequences, an obligate requirement of salt, in particular sodium, for growth and morphological characteristics (Fig. 1.3). The *Salinispora* form extensively branched hyphae that carry smooth surfaced spores that can be single or clustered. They lack arial hyphae and pigmentation varies from bright orange to black after spores have formed (Maldonado, Fenical *et al.* 2005). Prior to the taxonomic characterization of the *Salinispora (S. tropica and S. aurenicola)*, only one marine genus, *Salinibacterium* (Han, Nedashkovskaya et al. 2003) and three marine species, *Dietzia maris* (Nesterenko et al. 1982 (Rainey, Klatte et al. 1995), *Rhodococcus marinonascens* (Helmke, Weyland et al. 1984) and *Williamsia maris* (Stach, Maldonado et al. 2004) had been described.

Preliminary screening of organic extracts of *Salinispora* strains demonstrated antibiotic and anticancer activities suggesting that they were an excellent resource for drug discovery warranting further chemical investigation (Feling, Buchanan *et al.* 2003).

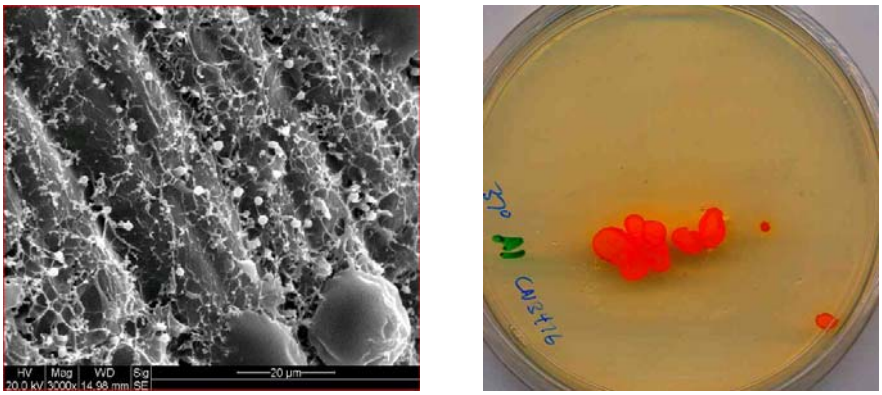


Figure 1.3. Electron micrograph showing *Salinispora tropica* spores and branched hyphae (left). SEM courtesy of T. Mincer. Agar plate with *Salinispora tropica* CNB476 grown on A1 seawater media showing orange pigmentation (right).

Structure and Activity of Salinosporamide A

Further studies with *S. tropica* strains led to the discovery of the highly bioactive small molecule salinosporamide A (**1**, Fig. 1.4), which was found to be a chemically unique molecule with a mass of 314 $[M+H]^+$ that shares its fused γ-lactam-β-lactone bicylclic ring structure with the streptomycete metabolite, clasto-lactacystin-β-lactone, commonly reffered to as omuralide (**2**, Fig. 1.4) (Fenteany, Standaert et al. 1994). Omuralide is derived from the *Streptomyces lactacystinaeus* fermentation product lactacystin (**3,** Fig. 1.4) which was originally reported for its ability to induce neuritogenesis in neuroblastoma cells (Omura, Matsuzaki et al. 1991). It was later shown to be the first natural product discovered that exhibited specific inhibition against the 20S proteasome (Fenteany, Standaert et al. 1995). The actual active molecule was found to be **2,** which forms spontaneously *in vitro* upon exposure of **3** to neutral aqueous media. Proteasome inactivation by **2** occurs by acylation of the catalytic N-terminal threonine residue on the β-subunit of the 20S proteasome (Corey and Li 1999) and both **2** and **3** are

routinely used as chemical reagents in laboratories worldwide to evaluate the function of the proteasome. Early studies with **1** displayed potent and highly selective activity in the NCI 60-cell-line panel with a mean GI_{50} (the concentration required to achieve 50% growth inhibition) value of less than 10 nM with its greatest potency observed against SK-MEL-28 melanoma and MDA-MB-435 breast cancer, all with LC_{50} values less than 10 nM (Feling, Buchanan *et al.* 2003).

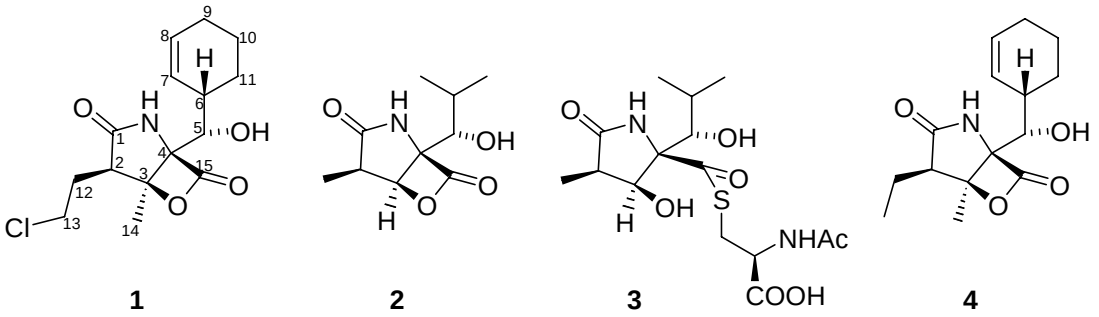


Figure 1.4. Structures of the proteasome inhibitors salinosporamide A (**1**), omuralide (**2**), lactacystin (**3**) and salinosporamide B **(4)**.

Although **1** and **2** share an identical bicyclic ring structure, **1** is uniquely functionalized and is 35 times more potent than **2** in proteasome inhibition assays (Fig. 1.5) (Feling, Buchanan *et al.* 2003). Structure activity studies in 1999 by Corey *et al*. demonstrated that the isopropyl group in **2** was essential for activity and that replacement by a phenyl ring abolished activity (Corey and Li 1999). Salinosporamide B (**4**), the des-chloro analog of **1,** also exhibits activity but is not as active against the proteasome as **1**. This suggested that **1** might interact with the 20S proteasome in a different manner than that of

2 or **4**. Salinosporamide A and other proteasome inhibitors including **2** (Tang and Leppla 1999) also show activity in mediating a toxic process in the cytosol of macrophages initiated by the lethal factor of anthrax lethal toxin, thus preventing cell lysis. Salinosporamide A is currently in development for the treatment of cancer by Nereus Pharmaceutical Company, La Jolla, CA under the name NPI-0052 (Chauhan, Catley et al. 2005) and recently entered phase I human clinical trials for multiple myeloma in May 2006.

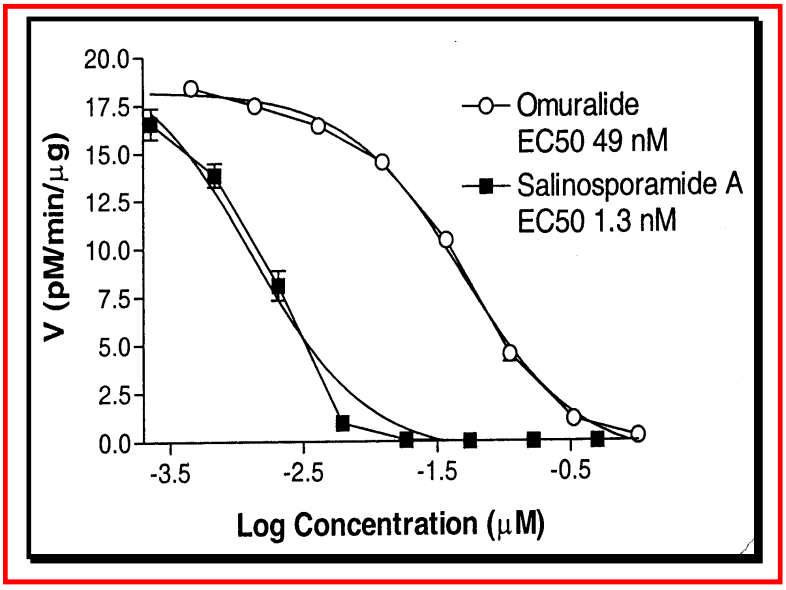


Figure 1.5. Proteasome inhibition assay comparing the ability **1** and **2** inhibit the chymotrypsin-like activity of the proteasome show that **1** is 35 times more potent than **2.** (Data courtesay of Nereus Pharmaceutical Company).

The 20S Proteasome: Structure and Function

Since the FDA approval of Velcade (Millenium Pharmaceuticals, Inc.) for the treatment of multiple myeloma in 2003, the 20S proteasome has represented a validated target for the treatment of cancer (Bross, Kane et al. 2004). The ubiquitin-proteasome pathway (UPP) is the primary mechanism by which potentially toxic misfolded proteins, proteins containing oxidized or abnormal amino acids and controlled proteolysis for regulatory proteins such as those involved in cell cycle progression and apoptosis occurs (Hochstrasser 1995; Coux, Tanaka et al. 1996). The sequential attachment of the ubiquitin peptide to targeted proteins generates a polyubiquitinated protein that serves as a substrate for destruction by the proteasome. The proteolytic activities of the mammalian proteasome are located within the 20S multisubunit structure consisting of four stacked rings arranged around an inner catalytic chamber (Fig. 1.6 A). Each of the outer rings consists of seven polypeptide α-subunits, whereas the two inner rings consist of seven β-subunits (Groll, Ditzel et al. 1997). The β-1, -2, and -5 subunits contain the postglutamyl peptidyl hydrolytic-like, tryptic-like and chymotryptic-like proteolytic activities of the proteasome, respectively (Groettrup, Kraft et al. 1996; Groll, Ditzel et al. 1997). These six catalytic subunits each possess an N-terminal threonine residue and are thus members of the N-terminal nucleophile hydrolase family of enzymes. During the immune response to challenges by foreign agents including tumors, the cytokine interferon-γ (IFN-γ) is released by activated T helper cells into the bloodstream, which mediates various immune responses including antiproliferative effects. When cells are exposed to interferon-γ, the β-1, -2, and -5 subunits are replaced by low-molecular-

weight protein (LMP) subunits LMP-2, -10 and –7, respectively, which preferentially cleave substrates after hydrophobic and basic residues (Fruh, Gossen et al. 1994; Groettrup, Kraft et al. 1996). The LMP-containing 20S proteasome has been dubbed the immunoproteasome owing to its ability to generate peptides that may be preferred substrates substrates for major histocompatibility complex (MHC) class I antigen presentation though its function in cancer is unknown (Teoh and Davies 2004).

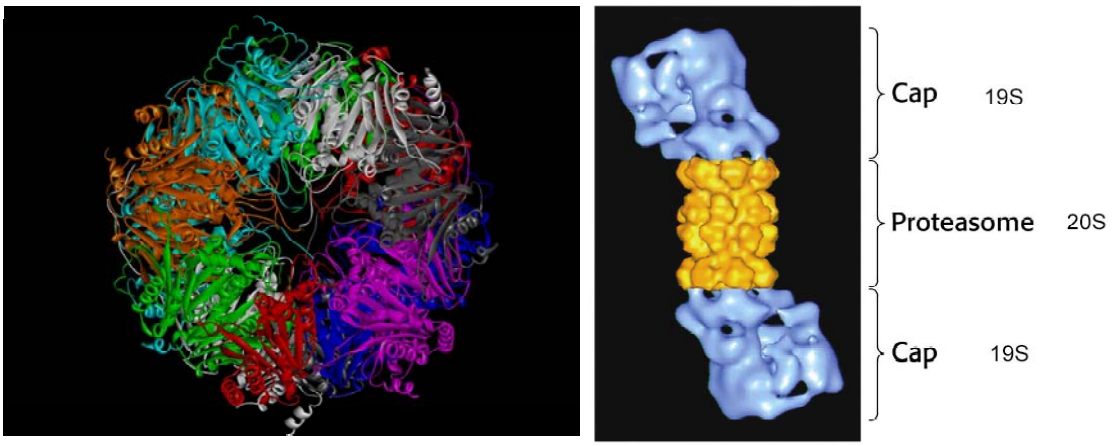


Figure 1.6. The 20S proteasome multisubunit structure showing the inner catalytic chamber. (left)**.** The 20S proteasome in complex with the 19S regulatory units forms the functional 26S proteasome (right). Pdb file: pdb1IRU, http://main.chem.ohiou.edu/courses/490/Lecture30-supplement.pddf

In order to process most proteins, the 20S proteasome requires the assembly of an activator or regulatory complex on the end of each outer ring of the 20S core particle. This 19S regulatory complex is the best-characterized proteasome activator and consists of a lid and base (Glickman, Rubin et al. 1998). The base directly associates with the α-rings and functions to bind and unfold the substrate, opens the channel of the 20S particle and catalyzes substrate translocation into the 20S proteasome. The lid contains isopeptidases that remove the polyubiquitin chain from the substrate allowing the release of free ubiquitin to be reused in further rounds of proteolysis. When capped by the 19S regulatory complex at each end, the 20S complex forms the core of the 26S proteasome (Fig. 1.6 B) (Glickman, Rubin et al. 1998).

The Development of Proteasome Inhibitors

Proteasome inhibitors were originally developed out of interest in the function of the proteasome during normal cellular processes and human disease and are represented by four general classes of molecules (Fig. 1.7) (Voorhees and Orlowski 2006). *N*-Acetyl-Leu-Leu-Norleucinal, also called calpain inhibitor I, was one of the first discovered proteasome inhibitors. It belongs to the class of proteasome inhibitors known as the peptide aldehydes (Wilk and Orlowski 1983), which are reversible serine and cysteine protease inhibitors that bind to the N-terminal active site threonine of proteolytically active subunits. The lack of specificity exhibited by the peptide aldehydes led to the replacement of the pharmacophore with a boronic acid functional group, generating a series of peptide boronic acid inhibitors (Adams, Behnke et al. 1998), the best known of

which is bortezomib/Velcade. Bortezomib is a dipeptidyl boronic acid that reversibly inhibits the chymotrypsin-like activity of the proteasome (Adams, Palombella et al. 1999).

As mentioned previously, the *Streptomyces* product **3** was the first identified natural proteasome inhibitor and did so by irreversible inhibition of the proteasome via a covalent modification of the amino-terminal threonine residue of the β-subunit. Another streptomycete natural product, epoxomicin, a peptide epoxyketone, (Hanada, Sugawara et al. 1992) has led to the development of other peptide α',β'-epoxyketones that inhibit the chymotrypsin-like activity of the proteasome and possess antiproliferative and anti-inflammatory properties (Elofsson, Splittgerber et al. 1999).

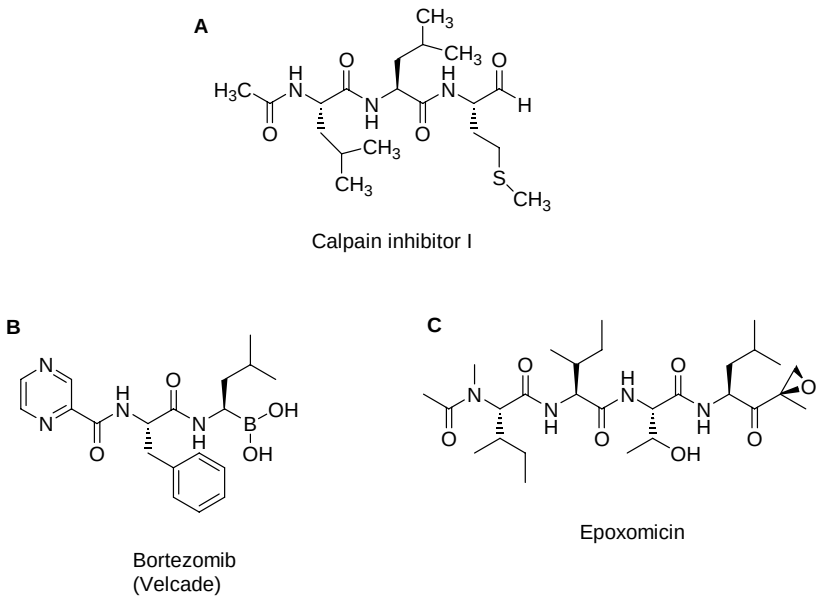


Figure 1.7. Structures representing the various classes of proteasome inhibitors. **A**. The peptide aldehyde calpain inhibitor I, **B.** The boronic-acid inhibitor Velcade and **C.** The actinomycete-derived epoxyketone, epoxomycin.

Most of the available proteasome inhibitors preferentially block the chymotrypsin-like activity of the proteasome, which is the rate-limiting proteolytic step (Kim and Crews 2004). Thus, the selective blockade of other proteolytic activities of the proteasome or simultaneous inhibition of several proteolytic activities could alter the activity or toxicity profile of proteasome inhibitors. Initial support for this possibility comes from studies with **1**, which induced apoptosis in myeloma cells that were resistant to bortezomib, and in combination, **1** and bortezomib led to synergistic induction of cell death (Chauhan, Catley et al. 2005).

Proteasome Inhibitors as Antitumor Agents

The first evidence that proteasome inhibitors possessed antitumor properties arose from studies with lactacystin and peptide aldehydes and showed that they potentiated tumor necrosis factor-mediated cell death (Vinitsky, Cardozo et al. 1994). Preclinical studies have repeatedly demonstrated a selective susceptibility of transformed cells to proteasome inhibition, though the proliferative status of a cell does not fully explain the increased sensitivity to proteasome inhibitors. Much of the activity comes from the ability of proteasome inhibitors to down regulate the nuclear factor-kappa B (NF-κB) pathway. The NF-κB transcription factor family plays an important role in tumerogenesis via the transactivation of genes involved in cell proliferation, apoptosis, tumor cell invasiveness, metastasis and angiogenesis (Voorhees and Orlowski 2006).

Salinosporamide A Activity against the 20S Proteasome

Structure-activity studies using X-ray crystallography of the 20S proteasome core particle in complex with **1** and the des-chloro analaog **4** (Groll, Huber et al. 2006) show that the molecules form covalent adducts with the catalytic threonine residue as described for **2**. The crystal structures also reveal that the enhanced potency of **1** is in part due to the presence of the chlorine atom on C-13 that allows for the formation of a tetrahydrofuran ring between the C-3O and the C-2 side chain (Fig. 1.8). The resulting cyclic ether form of **1** is irreversibly bound to the proteasome and the C-5 cyclohexyl substituent in **1** may also contribute to its enhanced activity. The primary finding of these intensive studies was that the irreversible binding observed between **1** and Thr1O of the β-subunit was due to the formation of the tetrahydrofuran ring. The C-3OH on **2** and **4** allows for deacetylation or hydrolysis and regeneration of the β-lactone ring. In the case of **1,** the C-3O is tied up in the tetrahydrofuran ring and thus renders the proteasome irreversibly inhibited, which testifies to the importance of the contribution of the chlorine atom to the potency of **1.** In addition, the cyclohexene ring appears to play a role in binding with the 20S proteasome.

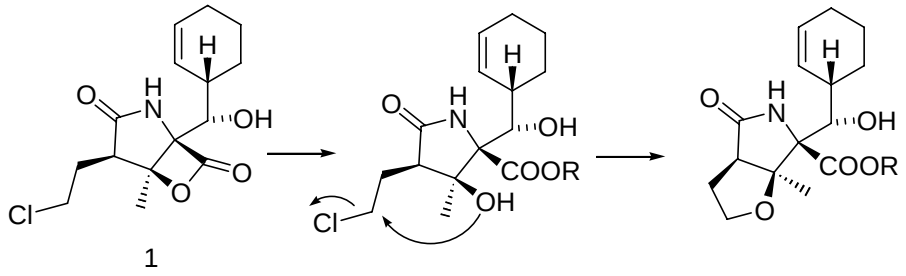


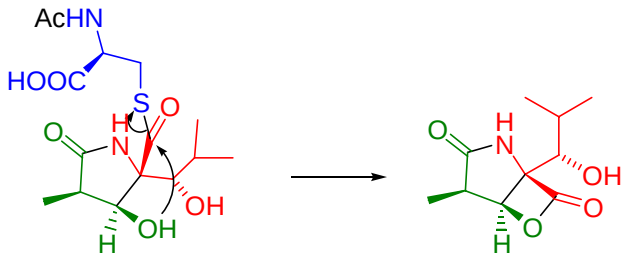
Figure 1.8. Mechanism of hydrolysis and further rearrangement of **1** in aqueous solution (R = H). An analogous reaction sequence occurs in the binding pocket of the proteasome core particle to form the tetrodydrofuran ring (R = Thre1$^{\gamma}$O^{6}).

It appears that nature has created a well-designed mechanism by which the proteasome can be regulated and the potent activity and novel structure of **1** have generated interest from the biosynthetic and synthetic communities and prompted its synthesis by two separate groups (Reddy, Saravanan et al. 2004; Endo and Danishefsky 2005; Reddy, Fournier et al. 2005; Reddy, Fournier et al. 2005). In fact, a recent paper from the Corey lab reporting the total synthesis of **1**, described salinosporamide A as one of the most fascinating biologically active natural products to have been recently isolated (Reddy, Fournier et al. 2005). The unique functionality of **1** suggests that novel biosynthetic pathways may be involved in the generation of the precursors and its assembly. Elucidation of the biosynthetic pathway is likely to reveal novel enzymatic reactions that may be applicable to the development of new biologically active small molecules using an *in vivo* approach. Isolating the biosynthetic gene cluster for **1** has several implications including potential regulation of the pathway in order to increase yields of **1** and minimize yields of fermentation analogs considered contaminants in pharmaceutical preparations. Access to the gene cluster potentiates the use of mutasynthesis, precursor directed biosynthesis and chemoenzymatic methods for the production of novel, bioactive and pharmaceutically relevant analogs of **1** (Kohli, Walsh et al. 2002; Xie, Uttamchandani et al. 2002; Moore, Kalaitzis et al. 2005). In order to generate structural analogs of **1** using a fermentation-based or chemoenzymatic approach, a basic understanding of the biosynthetic building blocks is necessary.

Chapter 2 of this work describes experiments in which a series of ^{13}C-labeled precursors were administered to *S. tropica* CNB476 cultures and the resultant enriched

salinosporamides were analyzed by nuclear magnetic resonance (NMR) to determine the biosynthetic building blocks incorporated into **1**. Chapter 3 describes the genomics approach that was undertaken in order to isolate the putative biosynthetic gene cluster for **1**. The feeding experiments and genomics approach were based on a biosynthetic scheme utilizing a mixed polyketide synthase (PKS) / non-ribosomal peptide synthetase (NRPS) biosynthetic mechanism and based on the reported biosynthetic pathways for structurally related molecules such as **2** and **3** (Takahashi, Uchida et al. 1995) and the widely studied β-lactams.

Feeding studies using ^{13}C-labeled precursors had shown that **3** is biosynthesized from the following three units, L-leucine, isobutyrate (from L-valine) as methylmalonic semialdehide and L-cysteine (Nakagawa, Takahashi et al. 1994). Formation of the β-lactone ring in **2** results from the attack of C-9O on C-4 followed by loss of the cysteine moiety (Fig. 1.9). However, the lack of available appropriate biosynthetic precursors and the absence of an analogous lactacystin-like precursor for **1** suggested that assembly might occur through a mixed NRPS/PKS-type mechanism.



2 **3**

Figure 1.9. The biosynthetic building blocks of **2** and **3**: green = methylmalonic semialdehide (from isobutyrate and/or L-valine), red = L-leucine and L-cysteine in blue.

NRPS and PKS - Derived Molecules and their Biosynthesis

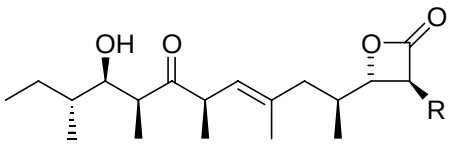
Non-ribosomal peptide synthetases and polyketide synthases are known to produce a wide variety of clinically relevant natural products (Finking and Marahiel 2004). Nonribosomal peptides refer to linear, cyclic or branched peptides, mostly consisting of less than 20 amino acid residues, which are often modified by acylation, glycosylation, epimerization, heterocyclization, or N-methylation of the amide nitrogen (Walsh, Chen et al. 2001). Many nonribosomal peptides are clinically important drugs such as cyclosporine A, penicillin, and vancomycin.

Polyketides make up one of the largest groups of natural products and are derived from sequential condensations of short carboxylic acids reminiscent of fatty acid biosynthesis (Hopwood 1997; Cane and Walsh 1999; Rawlings 2001; Moore and Hertweck 2002). Many polyketides are clinically valuable drugs including daunorubicin, erythromycin, lovastatin and rapamycin. NRPS and Type I PKS are multifunctional enzyme complexes containing sets of modules for the catalysis of each cycle of chain elongation (Cortes, Haydock et al. 1990; Donadio, Staver et al. 1991; Marahiel, Stachelhaus et al. 1997; Cane, Walsh et al. 1998; Schwarzer and Marahiel 2001; Mootz, Schwarzer et al. 2002) and recent work has focused on engineering these enzymes for combinatorial biosynthesis of "unnatural" natural products as either NRPS, PKS or hybrid NRPS/PKS products (Xiang, Kalaitzis et al. 2002; Du, Cheng et al. 2003; Kalaitzis, Izumikawa et al. 2003; Moore, Kalaitzis et al. 2005; Grunewald and Marahiel 2006). The genetic organization of PKS and NRPS genes makes them readily amenable to manipulation and there are several successful reports using the mixing and matching of

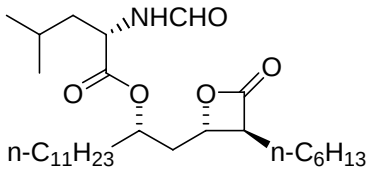
genes to create hybrid PKSs or NRPSs that result in the biosynthesis of novel molecules (Stachelhaus, Schneider et al. 1995; Long, Wilkinson et al. 2002). The PKS and NRPS genes responsible for the biosynthesis of a particular molecule are usually clustered on the genome arranged in modules and can be exchanged as cassettes, which contributes to the ease of working with these hybrid systems. A detailed description of NPRS and PKS biosynthesis is discussed in Chapter 3.

β-lactones as Therapeutics and their Biosynthesis

β-lactones occur naturally in a variety of organisms, and many possess potent biological activity including the streptomycete natural products ebelactone and lipstatin. (Fig. 1.10) (Umezawa, Aoyagi et al. 1980; Weibel, Hadvary et al. 1987). Orlistat (Roche), the tetrahydro analog of lipstatin, is a lipase inhibitor used for the treatment of obesity (Drent, Larsson et al. 1995). The mechanism of action involves the formation of covalent bond with the active serine residue site of gastric and pancreatic lipases, inactivating the enzymes and thus, rendering them unavailable to hydrolyze dietary fat in the form of triglycerides into absorbable free fatty acids and monoglycerides (Roche, 2005). Orlistat is also a novel and selective inhibitor of the thioesterase domain of fatty acid synthase (FAS) (Kridel, Axelrod et al. 2004), an enzyme strongly linked to tumor progression. FAS inhibition leads to halting tumor cell proliferation, inducing tumor cell apoptosis and inhibiting tumor growth in mice (Alo, Visca et al. 1996; Furuya, Akimoto et al. 1997; Alo, Visca et al. 1999; Bull, Ellison et al. 2001).



Ebelactone A (R = Me)
B (R = Et)



Orlistat

Figure 1.10. The naturally occurring streptomycete β-lactone ebelactone and Orlistat, the hydrogenation product of the natural product lipstatin.

Though Ebelactone and Orlistat are structurally similar, feeding studies using ^{13}C-labeled precursors showed that the biosynthesis of the carbon skeletons of ebelactone and orlistat differ significantly. Ebelactone is assembled from acetate and propionate precursors characteristic of a PKS biosynthetic pathway (Uotani, Naganawa et al. 1982), whereas Orlistat is assembled via a Claisen condensation of octanoyl-CoA with 3-hydroxy-δ-5,8-tetradecanoyl-CoA obtained by β-oxidation of linoleic acid (Schuhr, Eisenreich et al. 2002) and does not appear to be assembled by 2-carbon units that would indicate a PKS pathway. In both molecules formation of the lactone ring likely occurs when reduction of the carbonyl group generated by the Claisen condensation yields a hydroxy group and the β-lactone ring is formed with the CoA or an enzyme bound thio-

ester as a leaving group. The mechanism of action by which β- and γ-lactones inhibit enzyme activity relies on the enzymatic thiolate attack at the β- position or the carbonyl on the lactone and linear analogs show relatively very little activity (Lall, Karvellas et al. 1999) testifying to the importance of the ring. This mechanism is similar to that described for the binding of **1**, **2** and **4** with the 20S proteasome β-subunit.

β-lactam Biosynthesis

There are relatively few naturally occurring molecules that contain γ-lactone rings. However, a large number of naturally occurring microbial metabolites are β-lactams. Important classes of β-lactam antibiotics include the penicillins and the cephalosporins. All naturally occurring penicillins and cephalosporins produced by either eukaryotic or prokaryotic microorganisms share the basic structure of a thiazolidine ring connected to a β-lactam ring and are synthesized from the same three amino acids: L-α-aminoadipic acid, L-cysteine and L-valine (Fig. 1.11) The initial reaction of the penicillin and cephalosporin biosynthesis pathway is catalyzed by a single NRPS enzyme termed δ-(L-α-aminoadipyl)-L-cysteine-L-valine synthetase (ACVS) that condenses the amino acids to the tripeptide ACV. In the second step, oxidative ring closure of the linear tripeptide leads to formation of the bicyclic ring, i.e. the four-membered β-lactam ring fused to the five-membered thiazolindine, which is characteristic of all penicillins. The resulting isopenicillin N (IPN) is the first bioactive intermediate of both penicillin and cephalosporin pathways. This reaction is catalyzed by IPN synthase and IPN is the branch point of penicillin and cephalosporin biosyntheses

(Brakhage 1998; Brakhage and Caruso 2004). IPN synthase is a 2-oxoglutarate-dependent dioxygenase that requires 2-oxoglutarate and dioxygen as co-substrates. Based on the above-described pathways, many possible schemes for the biosynthesis of **1** were explored in the feeding experiments described in Chapter 2.

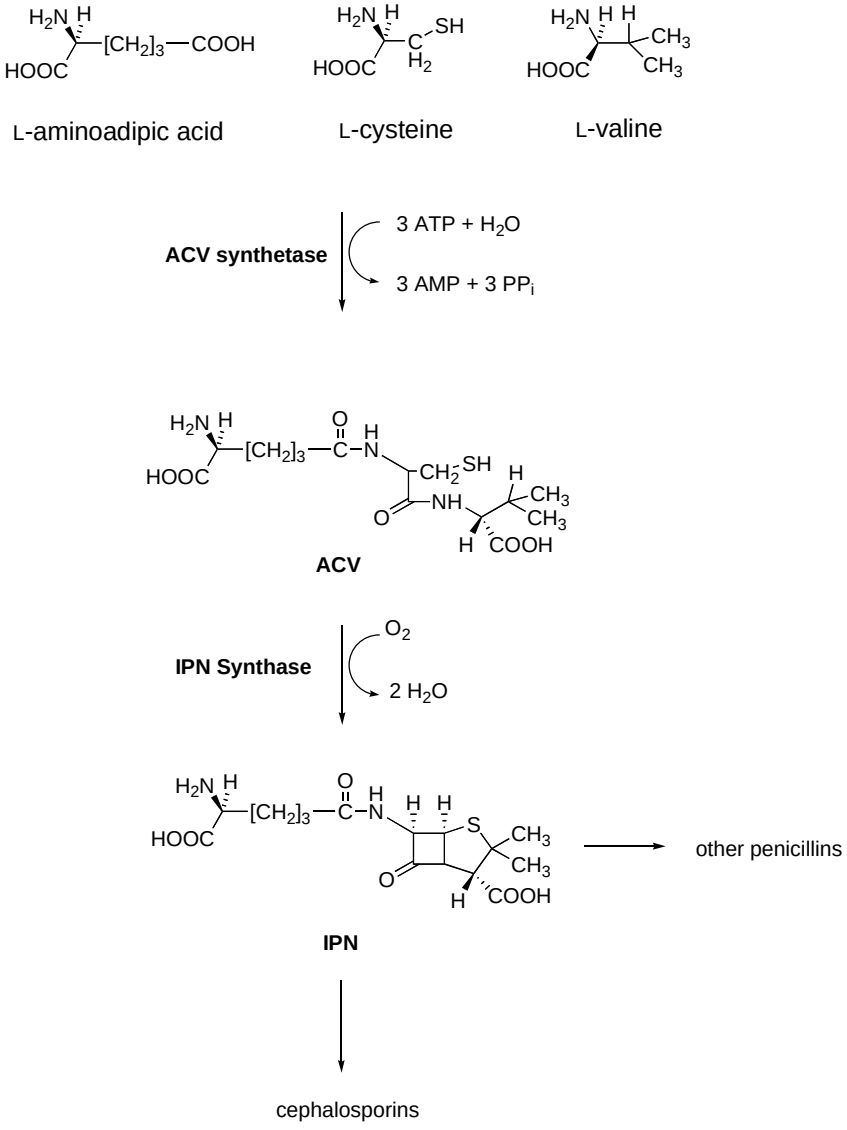


Figure 1.11. Biosynthesis of β-lactam antibiotics from the amino acid precursors L-α-aminoadipic acid, L-cysteine and L-valine.

CHAPTER 2

CHARACTERIZATION OF THE BIOSYNTHESIS OF SALINOSPORAMIDES USING ^{13}C-LABELED PRECURSORS

Introduction and Background

In order to generate novel salinosporamide analogs or optimize the production of **1**, an understanding of the basic biosynthetic building blocks and how they are assembled was necessary. Feeding experiments in shake flasks with the addition of stable isotope labeled precursors and analysis of the isotope-enriched product is a powerful tool to elucidate biosynthetic precursors and pathways involved in the assembly of both primary and secondary metabolites. The fate of individual atoms can be determined through nuclear magnetic resonance (NMR), which allows us to unequivocally identify specific atoms that show enrichment above the natural abundance levels of the respective isotope. To examine the biosynthesis of **1**, studies were performed in media to which specifically labeled ^{13}C-precursors were administered, and the resultant ^{13}C-enriched salinosporamides were was analyzed using 1D ^{13}C-NMR. Based on the results of feeding experiments, a proposed biosynthetic scheme is presented.

Proposed Biosynthesis of Salinosporamide A

The first biosynthetic feeding experiments were based on the structure and relatedness of **1** to **2** and **3**. Early studies with ^{13}C-labeled precursors had shown that **3** was derived from isobutyrate or valine, presumably as methylmalonic semialdehyde, L-leucine and L-cysteine (Nakagawa, Takahashi et al. 1994; Takahashi, Uchida et al. 1995)

and a hydrolytic cyclization reaction resulting in the loss of the acylated cysteine would yield **3** (Fig. 2.1 A). Thus, an analogous biosynthetic scheme for **1** was explored (Fig. 2.1 B). Based on known available biosynthetic precursors however, a scheme parallel to lactacystin biosynthesis was precluded by the presence of the methyl group (C-14) in **1**. This is due to the fact that there are no known naturally occurring methyl-extended molecules that would yield the bicycle with a methyl group on C-3 in **1.**

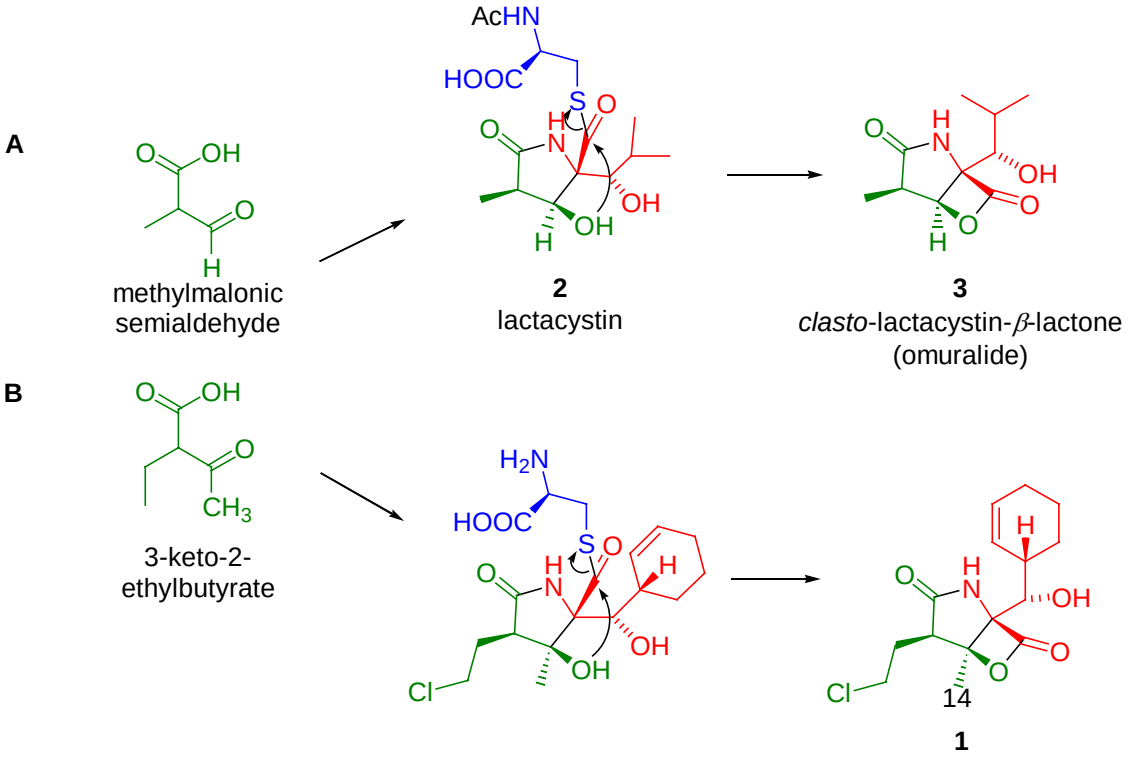


Figure 2.1. Proposed biosynthesis of salinosporamide A based on the lactacystin and omuralide biosynthetic pathway. (**A**) The biosynthetic building blocks of **2** and **3**: green = methylmalonic semialdehide: red = L-leucine and blue = L-cysteine. (**B**) Analogous biosynthetic scheme for **1**: green = 3-keto-2-ethylbutyrate; red = putative cyclohex-(2')-ene-3-hydroxyalanine (CHHA), blue = L-cysteine.

It was plausible that a scheme whereby the methyl ketone is made by a PKS, in which it is first coupled to the amino acid, not by an amide linkage, but rather through the α-carbon of the amino acid. However, the structure of **1** and the absence of an analogous lactacystin-like precursor from fermentations suggested that assembly might occur through a mixed PKS/NRPS biosynthetic pathway.

The γ-lactam-β-lactone ring system in **1** is unprecedented in a PKS/NRPS-derived product and its biosynthesis is likely to represent novel enzymatic reactions utilizing uncommon biosynthetic precursors (the gene cluster for lactacystin has not yet been reported). Based on hybrid PKS/NRPS biosynthetic pathways (Du, Sanchez et al. 2001; Silakowski, Kunze et al. 2001), a scenario was explored in which the alkyl portion of **1** (C-1, C-2, C-3, C-12, C-13, C-14) may arise from the condensation of acetyl-CoA (C-3, C-14) and ethylmalonyl-CoA (C-1, C-2, C-11, C-12) probably as the PKS-bound acyl-thioester, to give the diketide 3-keto-2-ethylbutyrate. The putative diketide would then couple through an amide bond to the putative non-proteinogenic amino acid (2*R*,3*S*,4*R*)-3-cyclohex-(2)-ene-3-hydroxyalanine (CHHA) by an NRPS to generate a linear enzyme-bound intermediate. Enzymatic off-loading with concomitant cyclization of the γ-lactam-β-lactone ring system via an unprecedented mechanism would yield the bicyclic core of **1** (Fig. 2.2 A).

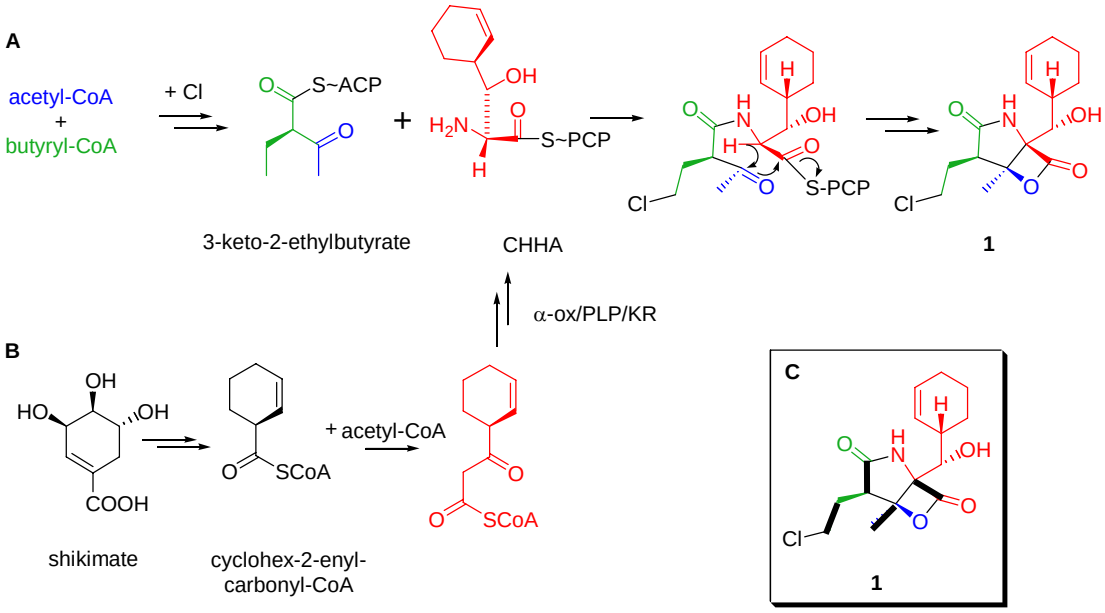


Figure 2.2. Proposed biosynthesis of **1**. **A**. Assembly of the diketide with CHHA to yield the linear enzyme-bound intermediate and cyclization to yield the bicyclic core of **1**. **B**. CHHA biosynthesis via a cyclohexenecarbonyl-CoA. Chlorination may occur prior to assembly of the diketide or elsewhere during biosynthesis. **C**. Predicted labeling pattern in **1** derived from [1,2-$^{13}C_2$]acetate based on the proposed biosynthetic pathway. Dark lines = acetate carbons. ACP: acyl carrier protein, PCP: peptidyl carrier protein, CHHA: cyclohex-2-enyl-3-hydroxyalanine. α-ox: alpha oxidation, PLP: pyridoxal phosphate aminotransferase, KR: ketoreduction.

Proposed Biosynthesis of the C-2 Alkyl Group

This biosynthetic scheme put forth in Figure 2.2 was based on several streptomycete PKSs, such as those involved in monensin, FK520 and tylosin (Fig. 2.3) production that utilize ethylmalonyl-CoA in polyketide chain assembly (Reynolds, O'Hagan et al. 1988; Gandecha, Large et al. 1997; Kakavas, Katz et al. 1997). In these systems ethylmalonyl-CoA is derived from butyryl-CoA and stable isotope incorporation

experiments have indicated the presence of multiple pathways for butyryl-CoA production in streptomycetes. One pathway involves isomerization of the valine catabolite isobutyryl-CoA to butyryl-CoA and is catalyzed by coenzyme B_{12}-dependent isobutyryl-CoA mutase (ICM) (Reynolds, O'Hagan *et al.* 1988; Zerbe-Burkhardt, Ratnatilleke *et al.* 1998). A second pathway involves the condensation of two acetyl-CoA molecules and the reduction of crotonyl-CoA to butyryl-CoA by crotonyl-CoA reductase (CCR) (Wallace, Bao et al. 1995) (Fig. 2.4). Functional copies of butyryl-CoA biosynthesis-associated genes, including CCR are present in some biosynthetic gene clusters and have been shown to provide the ethylmalonyl-CoA precursors required for the biosynthesis of molecules such as monensin and tylosin (Gandecha, Large et al. 1997; Liu and Reynolds 1999).

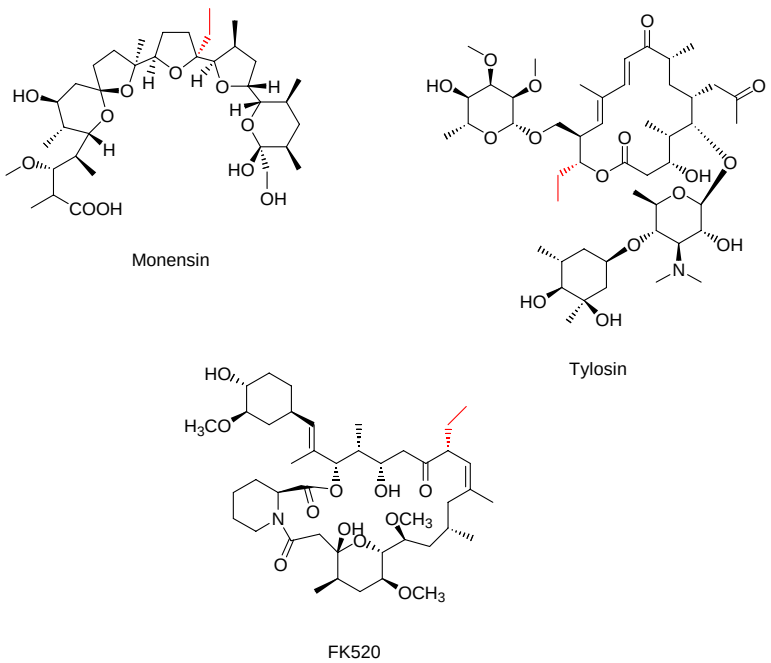


Figure 2.3. Streptomycete PKS derived molecules that utilize ethylmalonyl-CoA in polyketide chain assembly. The ethyl group that arises from butyryl-CoA is highlighted in red.

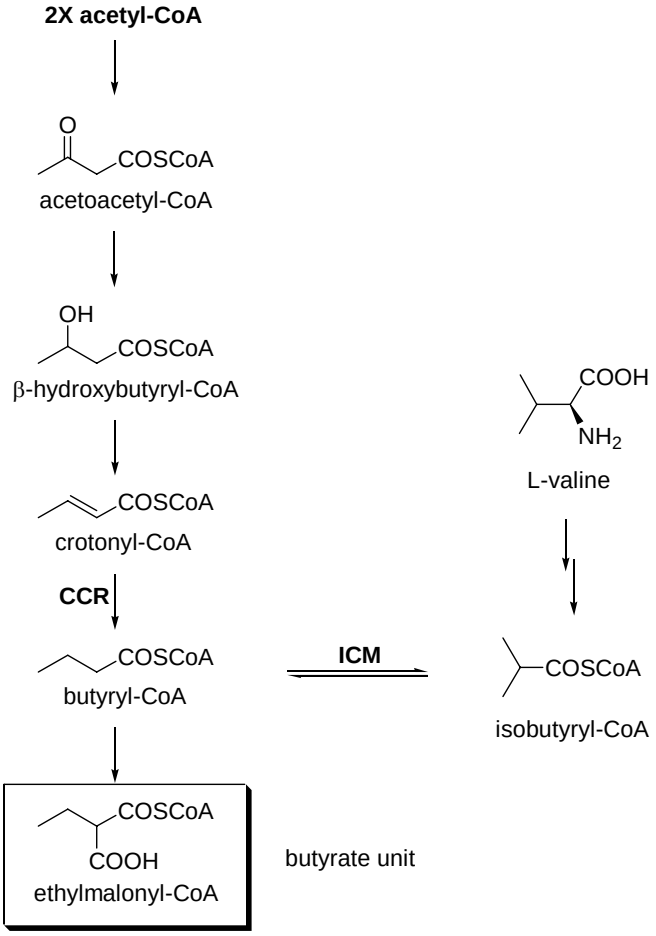


Figure 2.4. Pathways leading to ethylmalonyl-CoA in streptomycetes. CCR: crotonyl-CoA reductase; ICM: isobutyryl-CoA mutase.

Proposed Biosynthesis of the Amino Acid Moiety

The unusual amino acid moiety in **1** resembles the aromatic amino acids phenylalanine and tyrosine and the hydroxyl group at the β-position of the amino acid is a common modification found in many natural products including β-hydroxytyrosine

(vancomycin, novobiocin) (Sheldrick, Jones et al. 1978; Chen and Walsh 2001) and β-hydroxyphenylaline (katanosin B) (Kato, Hinoo et al. 1988). The oxidation is performed by a designated cytochrome P450, the genes of which are often associated with the NRPS in the gene cluster (Steffensky, Muhlenweg et al. 2000).

The shikimate pathway (Fig. 2.5) (Haslam 1993) is the essential metabolic route by which microorganisms and plants synthesize the aromatic amino acids and vitamins and the pathway is also the source of a vast number of shikimate-derived natural products (Floss 1997). Plants, and more commonly, microorganisms are known to produce natural products derived from shunt products of the shikimate pathway that diverge from the primary metabolic pathway at various branch-points giving rise to unusual and unique secondary metabolites such as 2,5-dihydrophenylalanine (Fig. 2.6). The biosynthesis of 2,5-dihydrophenylalanine involves an extra reduction step between the conversion of prephenate to phenylpyruvate (Fig. 2.5) (Shimada, Hook et al. 1978).

For the unusual amino acid moiety of **1** a pathway based on the biosynthesis of the cyclohexanecarboxylate starter unit in the streptomycete antibiotic ansatrienin A (Fig. 2.5) (Moore, Poralla et al. 1993) was explored. In the analogous pathway for the biosynthesis of the hypothetical amino acid CHHA, shikimate would undergo a similar series of dehydrations and double bond reductions to yield cyclohex-2-enylcarbonyl-CoA followed by condensation with acetate, α-oxidation, a PLP-like transamination and ketoreduction to yield CHHA (Fig. 2.1 B, C).

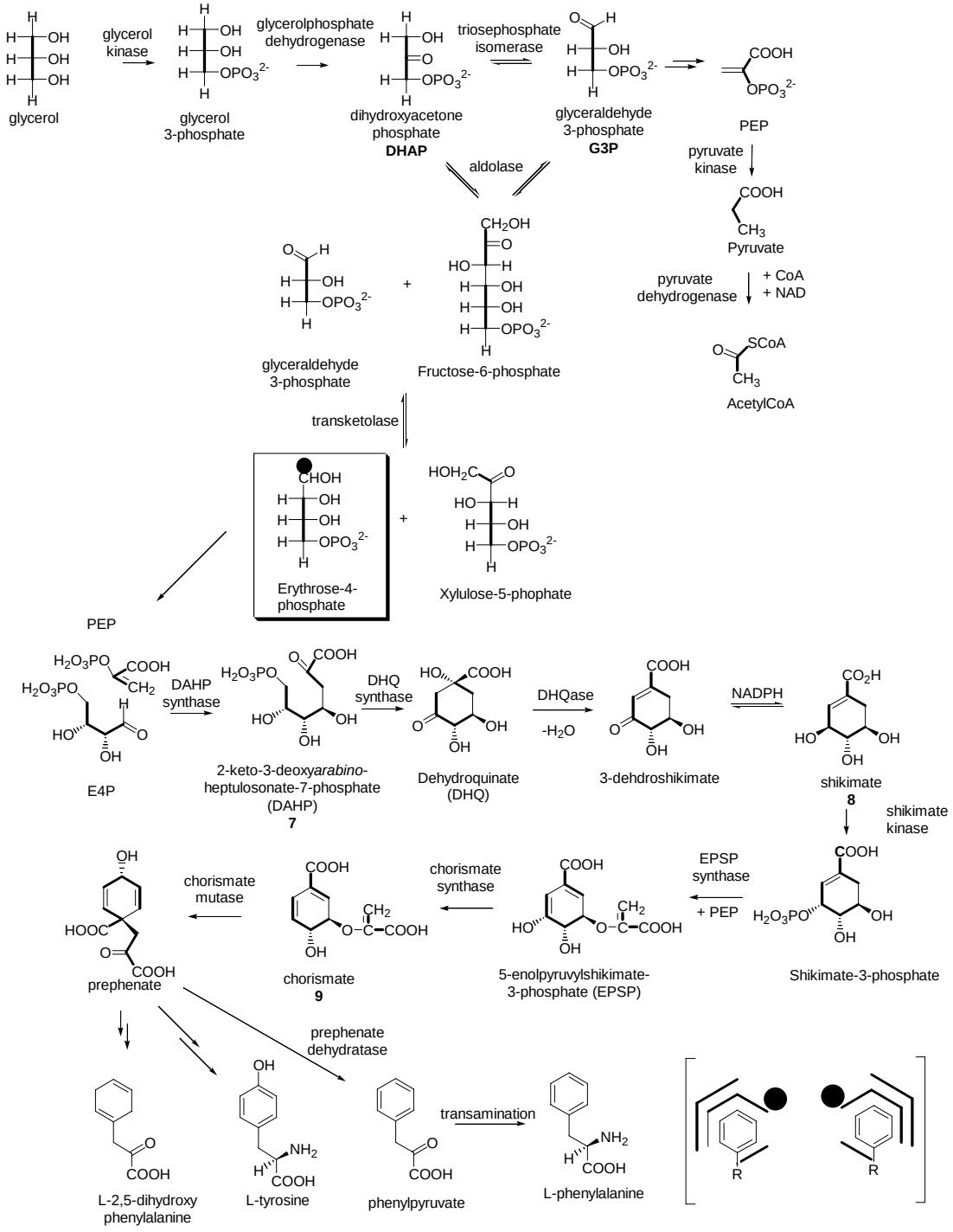
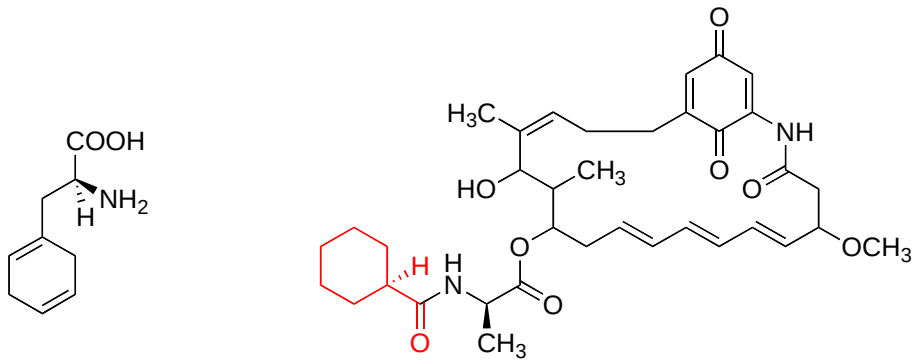


Figure 2.5. The shikimate pathway in aromatic amino acid biosynthesis. Glycerol is metabolized to acetate through glycoloysis. The pentose phosphate pathway leads to E4P with carbons 2, 3 and 4 arising from an intact 3-carbon glycerol via dihydroxyacetone phosphate (DHAP) or glyceraldehyde-3-phosphate (G3P). Feeding experiments with [U-$^{13}C_6$]glucose yield aromatic amino acids with a mixed-symmetrical labeling pattern on the aromatic ring. Intact carbons are indicated by dark line. Single carbon = dark circle.



L-2,5-Dihydrophenylalanine Ansatrienin A

Figure 2.6. Shunt products of the shikimate pathway. Ansatrienin A utilizes a cyclohexenecarboxyalte starter unit.

Materials and Methods

Chemicals

Reagents for fermentations were purchased from Difco, Sigma-Aldrich, Fisher and VWR. Amberlite® XAD7 resin was purchased from Sigma-Aldrich. Stable isotope compounds were purchased from Cambridge Isotope Laboratories, Inc. [1,7-^{13}C]shikimate was provided by Professor Heinz Floss at the University of Washington. [U-$^{13}C_{10}$]chorismate was produced *in vitro* using *Klebsiella pneumoniae* ATCC 2530

Strains

Salinospora tropica CNB476 and CNB440 were obtained from Dr. Paul Jensen and Professor Bill Fenical at the Scripps Institution of Oceanography, UCSD, La Jolla, CA (Maldonado, Fenical *et al.* 2005). All *S. tropica* fermentations were performed in sea

water medium A1BFeC (28g instant ocean, 4 g yeast extract, 10 g soluble starch, 1 g $CaCO_3$, 2 g peptone, 5 mL (8 g/L in H_2O) $Fe_2(SO_4)_3$ H_2O, 5 mL (20 g/L in ddH_2O) KBr, 10 mL Tris (pH 7), in dd H_2O to 1 L and autoclaved). Incubations were performed at 30° C with shaking at 250 rpm.

Chromatography

Analytical HPLC was performed on a Waters 600 HPLC system equipped with photodiode array detection. Fully automated screening was accomplished with the above instrument equipped with a Waters 717+ autosampler and its attached tray with monitoring by UV at 225 nm. Reversed-phase HPLC using a Phenomenex synergi hydro-RP C18 column (4.6 mm X 150) at a flow rate of 0.5 mL/ min was used to monitor the production of salinosporamide A and analogs. Semipreparative HPLC was performed using a YMC, C18 column (20 mm x 250 mm, 10 μm) reversed-phase HPLC at a flow rate of 9.5 mL/min with UV monitoring at 225 nm.

Nuclear Magnetic Resonance (NMR)

The ^{13}C-NMR spectra for unlabeled **1** and **4** and ^{13}C-enriched samples of **1** and [U-$^{13}C_{10}$]chorismate were recorded on a 300 MHz BRUKER-DRX NMR in the University of Arizona Cancer Center, field strength 75.47 MHz. Spectra for **4** enriched with [1,2-$^{13}C_2$]acetate and [U-^{13}C]glucose were recorded on a BRUKER-DRX 600 MHz NMR equipped with a microprobe in the Department of Chemistry at the University of Arizona, field strength 125.78 MHz. The spectra from [1-^{13}C]propionate feeding

experiments were recorded on a BRUKER-DRX 600 MHz NMR, field strength 125.78 MHz at Nereus. The ^{13}C-NMR spectra for **1** were acquired in pyridine. Spectra for **4** were acquired in DMSO. All spectra acquired at Nereus were acquired in DMSO. Spectra for [U-$^{13}C_{10}$]chorismate were acquired in acetone. Data are reported as follows: chemical shift, multiplicity, coupling constants (Hertz), assignment. Specific incorporations were calculated as the incorporation relative to the natural abundance of carbon 1 at 177.4 ppm after integration.

Salinispora tropica Fermentation and Isolation of the Salinosporamides

S. tropica strains were stored in 40% glycerol at -80 °C and maintained on 4% agar plates. Seed cultures were inoculated with a loop full of glycerol stock into 5 mL A1BFeC meduim in a test tube containing glass beads and incubated for 48-72 hours. Incubation with glass beads prevented mycelium from forming and resulted in more viable cultures. A 25 mL starter culture was inoculated with 1 mL of the seed culture that was maintained at 30° C with shaking at 250 rpms and used to reseed a new starter culture every 5-6 days. In general 1-2 mL of starter culture was used to inoculate 100-200 mL of medium in 250 or 500 mL Erlenmeyer flasks respectively. Yields decreased with cultures larger than 200 mL, possibly due to aeration issues. Amberlite® XAD-7 non-ionic polymeric resin (20 mg/L washed with 2:1 methanol:H_2O x 3, rinsed several times with dd H_2O, autoclaved and dried) was added after 24 hours of growth. Cultures were harvested and the resin was extracted 72-96 hours later when cultures reached a dark orange/brown color.

Preparation of Crude Extracts

Crude extracts were prepared by applying the entire culture to a glass column stopped with a small amount of glass wool. Following the removal of all culture medium from the column by vacuum, the resin was extracted three times with ~100 mL acetone and the elutant was filtered through Whatman Grade 4 filter paper into a 500 mL round bottom flask and dried under reduced pressure. The dried crude extract was then dissolved in ~1 mL acetone 3x and transferred to a small glass tube, redried and weighed. The production of salinosporamides A and B was monitored by analytical HPLC or by thin layer chromatography (TLC).

Monitoring the Production of Salinosporamides by Analytical HPLC

The production of **1** in crude extracts was monitored by reversed-phase analytical HPLC using the following conditions: 30% acetonitirile (MeCN) in water increasing to 70% MeCN over 20 min, increased to 100% MeCN from 20-21 minutes and held at 100% MeCN from 21-31 min. The column was then brought back to 30% MeCN over 1 min and equilibrated for 5 min prior to injection of the next sample. Compound **1** eluted at ~12 min. Midway through the project, conditions were altered to also monitor the production of **4** using reversed-phase HPLC with isocratic elution at 30% MeCN, 70% H_2O for 25 minutes at a flow rate of 1 mL/min with similar UV monitoring. Compounds **1** and **4** eluted at 16.9 and 6.8 min respectively.

Purification of Salinosporamides

The dried crude extracts were dissolved in a small amount of acetone and applied to an acetate equilibrated Waters C18 mini cartridge, eluted three times with 10 mL acetone and dried under reduced pressure. The resultant semi-pure sample was dissolved in 0.5-1 mL acetonitirile (MeCN) and injected onto the column in 250 mL aliquots. Semipreparative reversed-phase HPLC conditions used a gradient elution from 30% MeCN in H_2O to 70% MeCN over 20 min at a constant flow rate of 9.5 mL/min and UV monitoring at 225 nm. Purified salinosporamide A eluted at ~ 5 min at which time the column was washed with 100% acetonitrile and equilibrated for 5 min prior to injection of the next sample.

Separation and purification of **1** and **4** was also performed using modified normal phase isocratic flash column chromatography methods provided by Nereus. This method was also sufficient to reisolate samples of **1** from NMR solvents. Dried crude extracts were dissolved in a small amount of acetone and applied to a column containing ~ 200 mL silica gel equilibrated with 40% EtOAc in hexanes. The same solvent mixture was used to elute the compounds. Typically 3 x 25 mL fractions were collected then fractions of 5-10 mL were collected. The fractions were analyzed against standards by TLC in the same solvent system. Compounds **1** and **4** typically eluted in fractions 8-17 and fractions 20-27, respectively. Yields from either method were generally 60-100 mg/L for **1** and 10-20 mg/L of **4** from flash chromatography. The same solvent system was used to purify **1** from the NMR solvents pyridine and DMSO. The NMR sample was applied directly to the column and 1-2 mL fractions were collected. For these experiments **1**

typically eluted in fractions 70-100. Fractions containing the compound of interest were dried under reduced pressure. Both **1** and **4** appeared as white powder.

For the propionate feeding experiments the crude extract from a 4 L culture (915 mg) was separated first by normal phase isocratic HPLC for the isolation of compound **1** with a flow rate of 25 mL/min. Mobile phase was EtOAc (24%) in hexanes (76%) and performed on a Gilson HPLC equipped with a 170 Diode Array detector and a 215FC fraction collector with ELSD monitoring. Fractions of 12.5 mL were collected. A normal phase silica column (Phenomenex Luna Si 10mm, 250 x 21.2 mm id) was used. After compound **1** eluted from the column after 27 minutes, the solvent system was ramped to 100% EtOAc over 1 minute and held at 100% EtOAc for 4 minutes for the elution of compound **4**, **5** and **6** from the column. LCMS was used to verify compound identity. These methods were sufficient to purify compounds **1** (70 mg) and **4** (7.4 mg). Compounds **5** and **6** eluted together under these conditions and were separated using normal phase HPLC with a gradient (same instrumentation and flow rate) that started with 40% EtOAc in 60% hexanes increased to 100% EtOAc over 20 min, which yielded 0.8 mg **6** and 3.1 mg of compound **5**. All compounds were verified by UV, LCMS and analyzed by 1D ^{1}H and ^{13}C NMR.

Feeding Experiments with ^{13}C-labeled Precursors

Experiments were generally performed in volumes of 200 mL. Single doses of labeled precursor were administered as sterile solutions in water to the fermentations along with the Amberlite resin after 24-36 hours growth. Fermentations were grown and

harvested and extracted as described. The concentrations at which labeled compounds were added is as follows: [1,2-$^{13}C_2$]acetate: 50 mg/L; [U-$^{13}C_6$]glucose: 50mg/L; [1,7-$^{13}C_2$]shikimate: 125 mg/L; [7-$^{13}C_1$]phenylalanine: 125 mg/L, [1-$^{13}C_1$]butyrate: 250 mg/L; [1,3-$^{13}C_2$]glycerol: 100 mg/L; [U-$^{13}C_3$]glycerol: 100 mg/L; [1-$^{13}C_1$]2-methylbutyrate: 100 mg/L [1,2-$^{13}C_2$]alanine: 125 mg/L. Feeding experiments with [U-$^{13}C_{10}$]chorismate used 100 mg/L administered between two additions at 24 hours and 48 hours.

Preparation of [U-$^{13}C_{10}$]Chorismic Acid

Klebsiella pneumoniae ATCC 2530 (formerly *Aerobacter aerogenes* strain 62-1) was grown as described by Gibson (Gibson and Gibson 1962; 1964) and modified by Rieger et al. (Rieger and Turnbull 1996). Growth from an overnight culture in LB medium was added (1 mL) to 250 mL of growth medium (2 g yeast extract, 2 g casamino acids, 41 mg tryptophan, 2 g citric acid, 0.2 g anhydrous magnesium sulfate, 12.5 g Na_2HPO_4, 1 g KH_2PO_4 and 2 g NH_4Cl in 1 L distilled water and autoclaved) in a 1 L Erlenmeyer flask and incubated at 30° C with shaking at 250 rpm for ~ 6 hours until the culture showed an OD_{625} > 1.8 nm. The cells were harvested by centrifugation and resuspended in 250 mL unsterilized accumulation medium (19.2 g Na_2HPO_4, 20.5 g KH_2PO_4, 2.7 g NH_4Cl, 21 mg $MgCl_2 \cdot 7H_2O$, and 3 mg tryptophan in 1 L distilled water). Rather than the 10 g glucose prescribed in the literature, 4 g [U-$^{13}C_6$]glucose in 5 mL water was added at the time of inoculation. The cells were incubated for ~ 16 hours at 30° C and 250 rpm until a 1:20 microfuged aliquot (in H_2O) showed an OD_{275} > 0.8 nm and were harvested by centrifugation. The cells were discarded and the spent medium

was adjusted to pH 7.4 at 4° C with 10N NaOH. The solution was passed through a Dowex 1-Cl column (25 g washed with 1N HCl, rinsed with H_2O, washed with 1 N NaOH and rinsed with H_2O) and washed with 50 mL distilled water. Chorismate was eluted with 1 M NH_4Cl (pH 8.5) at a flow rate of ~ 5 mL/min. Fractions were collected in 10 mL aliquots and those with an absorbance at 274 nm of greater than 0.3 (1:100 dilution in H_2O) were pooled, acidified with 1 N HCl to pH 1.5 at 4° C and extracted 3 times with an equal volume of ether. The ether layer was dried with anhydrous $MgSO_4$ for 1 hour on ice and concentrated under reduced pressure yielding yellow oil. In order to crystallize chorismate a small volume (2 mL) of dichloromethane was added and the solution was chilled on ice. Three volumes of hexanes were slowly added and chorismate was precipitated on ice for 30 minutes. Two more volumes of hexanes were added and the reaction mixture was incubated on ice for an additional hour to assure complete crystallization. The sample was dried slowly under reduced pressure in an ice bath. Yield from a 500 mL culture was 39 mg of [U-$^{13}C_{10}$]chorismate. ^{13}C-NMR and LCMS were used to confirm chorismate production and 100% ^{13}C-enrichment at all carbons (Appendix A, Fig. 9, 10, 11).

Optimization of Salinosporamide A Yields and Purification Techniques

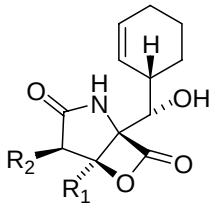
Given that the structure of **1** had previously been determined by NMR and x-ray crystallography, the next step was to perform feeding experiments to elucidate its biosynthetic precursors. Prior to undertaking any biosynthetic feeding experiments it was paramount that the fermentation conditions, isolation and purification methods be

optimized. Typical yields of **1** at the onset of this project were ~1 mg / L, which for our purposes was not practical. Since analysis on a 300 MHz NMR generally requires ~ 5 mg of pure compound in order to obtain high quality ^{13}C-NMR spectra to unambiguously show ^{13}C-enrichment and the satellite peaks that result from ^{13}C-^{13}C coupling, it was pertinent that yields of **1** be increased at least 10-fold. Data from time course experiments showed that the production of **1** peaked at day 3 during fermentation and was essentially degraded by day 4. Also and it appeared that feedback inhibition may have been limiting the production of **1**.

The modified seawater medium was used to optimize growth and it was found that the addition of Amberlite XAD-7 non-ionic resin to adsorb the salinosporamide compounds followed by isolation and extraction of **1** from the resin, resulted in slightly increased yields of **1** (P. Jensen, personal communication). The modified fermentation and isolation protocol consistently yielded 80-100 mg / L of **1** and 10-20 mg / L of **4**, an overall improvement of >800 %. The procedure was also extremely efficient in that the time from resin collection to obtaining a purified sample sufficient for NMR analysis was approximately 4 hours.

During the course of the biosynthetic studies, data suggested that **1** and **4** were biosynthesized from different precursors and that it was unlikely that **4** was a direct precursor of **1**. This implied that there might be promiscuity in the biosynthetic enzyme complex that allows alternate substrates to be loaded onto the putative PKS. Several minor salinosporamide analogs have been isolated from crude extracts including the

stereoisomers of **1** and **4**. Though some analogs are products of hydrolysis or degradation during preparation and purification, the compounds NPI-2063 (**5**) and NPI-2079 (**6**) (Fig. 2.7) are consistently produced by fermentation. Compound **5** has a methyl group at C-2, and compound **6** has an ethyl group at C-3, which suggested they were derived from propionate. Biosynthetic studies were undertaken to determine if labeled propionate was a precursor for these 3-carbon units, though even with the improved fermentation and isolation techniques described above, **5** and **6** were produced in amounts too low for our available detection methods. Thus, in order to obtain enough labeled sample of **5** and **6** for NMR analysis, fermentations and processing of the crude extracts were performed at Nereus as described previously. The yields from these experiments are shown in Table 2.1.



5 R_1 = Met R_2 = Met
6 R_1 = Et R_2 = ClEt

Figure 2.7. Salinosporamide analogs containing propionate units. NPI2063 (**5**); NPI2059 (**6**).

salinosporamide analog	mg purified	% of product **1** = 100%
1 (NPI-0052)	70	
4 (NPI-0047)	7.4	106
5 (NPI-2063)	3.1	4.4
6 (NPI-2065)	0.8	1.1

Table 2.1. Yields of salinosporamides from [1-$^{13}C_1$]propionate feeding experiments using HPLC fractionation.

Feeding Experiments and NMR Analysis

Once optimal growth and purification conditions were established, feeding experiments could be conveniently undertaken on a relatively small scale. ^{13}C-NMR experiments were generally performed using ~10 mg of **1** and the observed ^{13}C-enrichment ranged from 2-5%. The 1D ^{13}C-NMR analysis for the acetate and glucose enriched **4** were performed using 2-4 mg of purified compound on a 600 MHz NMR equipped with a microprobe.

Results

In order to unambiguously determine how **1** is assembled, feeding studies using ^{13}C-labeled precursors were designed to determine the biosynthetic precursors utilized during the assembly **1**. Results of the ^{13}C-labeled precursor feeding experiments for **1** are summarized in Table 2.2.

	Incorporation (%)				J_{CC} (Hz)		
δ_{cc} ppm[a]	[1,7-$^{13}C_2$] shikimate [b,c]	[1-$^{13}C_1$] butyrate[b,c]	[1,2-$^{13}C_1$] alanine [b,c]	[1,3-$^{13}C_2$] glycerol [b,d]	[U-$^{13}C_3$] glycerol [a]	[1,2-$^{13}C_2$] acetate [a]	[U-$^{13}C_6$] glucose [a]
177.4					-	-	s, 47, 2*
46.7				1.8	36	-	36, 42*
86.8		2.2			43	43	43
80.8					36	-	44, 47
71.5				1.4	36	-	44, 3
39.8	1.2				41	-	42
129.6				1.5	43	-	42
129.2				1.3	0	-	s, 39*
25.9				1.5	31	-	34, 36*
22.2					-	-	33
27.0				1.5	31	-	33
29.5					-	-	36
43.8				1.8	36	-	36
20.5			2.3	1.6	43	43	43
169.9					-	-	47, 4

Table 2.2. Summary of results from 1D ^{13}C-NMR analysis of ^{13}C-enriched **1** from labeled precursor feeding studies. [a] referenced to d_5-pyridine. [b] % incorporation = (A-B)/B*A = intensity of enriched carbon and B = intensity of natural abundance carbons. [c] Incorporation relative to the natural abundance carbon at 177.4 ppm (C1). [d] Incorporation relative to natural abundance carbon at 22.2 ppm (C10). * two coupling species, s = enriched singlet.

Acetate Feeding Experiment

If the proposed biosynthetic scenario outlined in Scheme 2.2 were correct, a total of four intact acetate units would be incorporated into **1** including three in the "diketide" or alkyl portion of the molecule and one in the amino acid (Fig. 2.2 C). To test this hypothesis, feeding experiments with [1,2-$^{13}C_2$]acetate were undertaken. Analysis of the resultant **1** by 1D ^{13}C-NMR showed C-3 and C-14 as enhanced and coupled doublets (J = 43 Hz) verifying that these carbons are directly bonded to one another and were incorporated as an intact 2-carbon unit (Fig. 2.8), presumably as acetyl-CoA or an enzyme bound acyl-thioester.

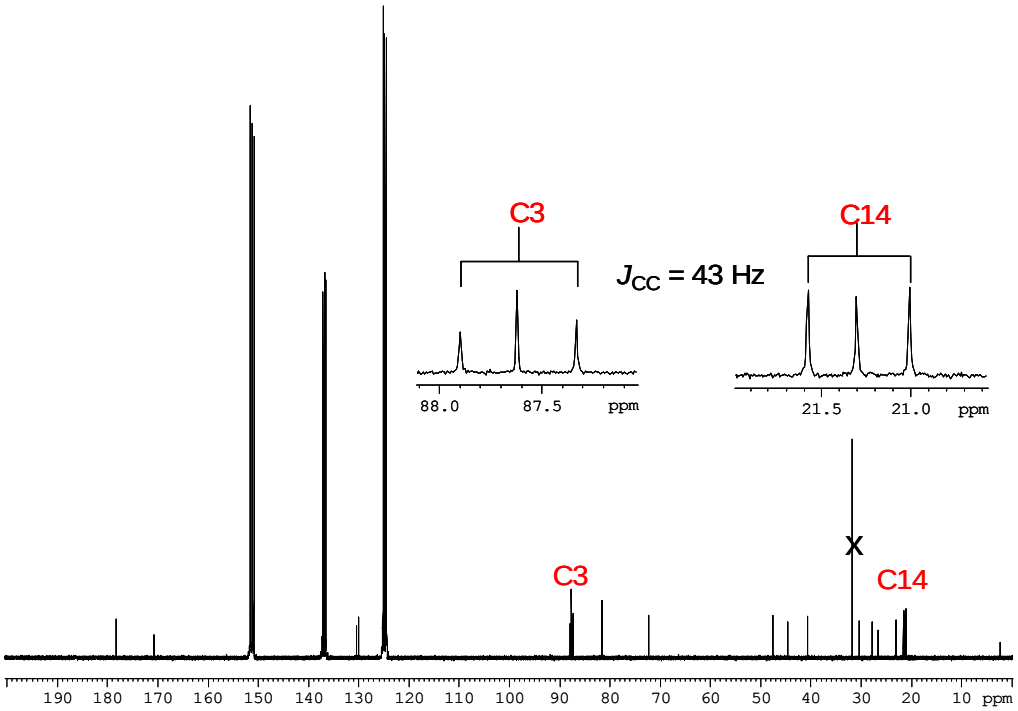


Figure 2.8. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1,2-$^{13}C_2$]acetate-enriched **1**. Individual resonances for the acetate-derived carbons are shown.

No detectable ^{13}C-enrichment was observed elsewhere in the molecule and since all known pathways for ethylmalonyl-CoA biosynthesis in actinomycetes proceed through butyryl-CoA and utilize acetate, the data suggested that the 4-carbon putative ethylmalonyl precursor in **1** must arise from a pathway other than butyryl-CoA via ICM or CCR. The lack of incorporation observed in the amino acid portion of **1** also eliminated the cyclohexenylcarboxylate pathway for CHHA biosynthesis outlined in Figure 2.2 B. However, the alternate pathway for the biosynthesis of cyclohexyl

molecules is via the shikimic acid pathway. Thus, based on its similarity to phenylalanine, and the lack of acetate incorporation, experiments were designed to test whether the CHHA moiety arose via a shunt in the shikimic acid pathway.

Feeding Experiment with the General Precursor Glucose

Since glucose is a general precursor that when metabolized produces metabolites via glycolysis and the Krebs cycle, it was a powerful tool to explore the origins of the remaining carbons. Feeding experiments using [U-$^{13}C_6$]glucose were undertaken and analysis of ^{13}C-enriched **1** by 1D ^{13}C NMR revealed that all carbons in **1** were highly enriched by [U-$^{13}C_6$]glucose and each carbon coupled to at least one of its neighboring carbons as observed by the satellite peaks in the 1D ^{13}C-NMR spectrum (Fig. 2.9). Expansions of individual carbon resonances: Appendix A, Figure 3.

Glucose Feeding Exeperiment Results: the Alkyl Carbons

The ^{13}C-NMR signals at C-3 and C-14 appear as enhanced and coupled doublets (J = 43 Hz) confirming the incorporation of the intact 2-carbon unit observed in the [1,2-^{13}C]acetate feeding experiment. The ^{13}C-NMR signal at C-1 is a doublet of doublets with a major doublet (J = 47 Hz) split by long-range coupling (J = 2 Hz). The enhanced natural abundance singlet for C-1 is much more intense than the surrounding doublet which represents a second labeled species with single enrichment. The ^{13}C-NMR signal at C-2 is a strong doublet (J = 36 Hz) with a weaker doublet of doublets (J = 42 Hz, J = 36 Hz) in a ratio of 3:1, also indicative of two coupling species. The signal at C-12 is a

triplet (J = 36 Hz) indicating it has two coupling partners and is part of an overall 3-carbon unit. A minor doublet is also observed at C-12 (J = 38 Hz), which may be an artifact of the high overall incorporation of glucose on the neighboring carbon. The ^{13}C-NMR signal at C-13 is a doublet (J = 36 Hz).

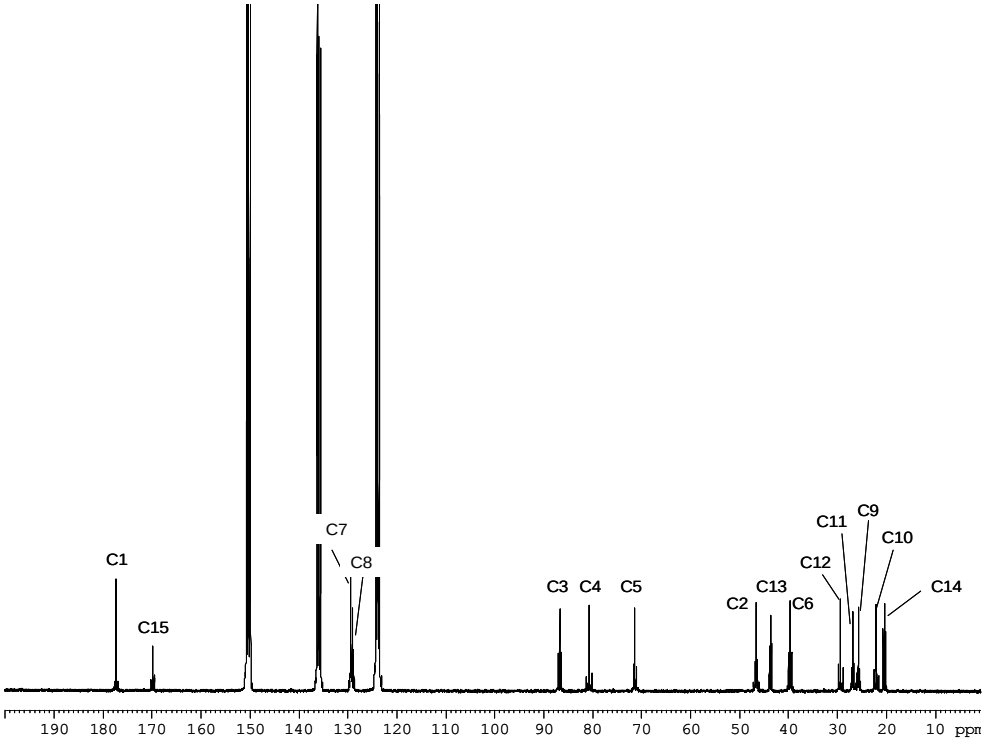


Figure 2.9. 1D ^{13}C-NMR in d_5-pyridine of [U-$^{13}C_6$]glucose-enriched **1**. Expansions of individual resonances for all carbons can be found in Appendix A. Figure 1.

Results from the 1D ^{13}C-NMR data show that carbons 1, 2, 12, and 13 are part of an overall four-carbon unit and that a second species exhibiting a "3 + 1" labeling pattern

consisting of a three-carbon unit: C-2, C-12 and C-13, and strong single enrichment at C-1 is also present. The observed ^{13}C-^{13}C coupling pattern from the glucose experiment confirmed that if an ethylmalonyl-CoA precursor was involved in the biosynthesis of **1,** it must arise from a presumably novel pathway for butyryl-CoA biosynthesis or an alternate pathway that excludes butyryl-CoA all together for the formation of ethylmalonyl-CoA.

Glucose Feeding Experiment Results: the CHHA Moiety

For the CHHA moiety, the ^{13}C-NMR signal at C-4 is a triplet (J = 43 Hz). The signals at C-5 and C-15 are identical doublets of doublets consisting of a major doublet (J = 43 Hz) that is split into a second smaller doublet (J = 3 Hz) indicating long-range coupling to each other. Carbons 5, 4 and 15 are part of a three-carbon AMX spin system with C-4 at the center. The ^{13}C-NMR signal at C-6 is a doublet (J = 42 Hz) and the signal at C-7, which overlaps with the adjacent methylene C-8, is also a doublet (J = 42 Hz) indicating that C-6 and C-7 couple to one another and are part of a two-carbon unit.

The 1D ^{13}C-NMR signals for the remaining ring carbons show C-11 as a doublet (J = 33 Hz) and carbon 10 is a triplet (J = 33 Hz) indicating that it has two coupling partners with the same J-values, one of which is C-11. Two coupling species are also observed at C-9 as a doublet of doublets: the first coupling (J = 34 Hz) to C-10 and a second coupling (J = 36 Hz) to C-8. At C-8 there is enhanced single enrichment in addition to a second coupling species, a minor doublet (J = 39 Hz) that likely couples with C-9 though its coupling constant is slightly larger than the 36 Hz value observed in the C-9 triplet. These data indicate that carbons 8-11 are part of an overall four-carbon

unit in addition to a second 3+1 species where C-9, C-10 and C-11 are part of an overall three-carbon unit and C-8 shows single enrichment in a ratio of 3 to 1 in favor of the singly enriched 3+1 species over the 4-carbon unit.

The observed labeling pattern is consistent with the shikimic acid pathway (figure 2.5). In shikimic acid biosynthesis, condensation of erythrose-4-phosphate (E4P) and phosphoenolpyruvate (PEP) via DAHP synthase yields 2-keto-3-deoxy-D-*arabino*-2-heptulosonate-7-phosphate (DAHP, **7**), which undergoes a cyclization reaction to give shikimate (**8**) and attachment of a second PEP molecule leads to chorismate (**9**). At **9**, the pathway separates into three branches, each leading specifically to one of the aromatic amino acids. The 1D ^{13}C-NMR data from [U-$^{13}C_6$]glucose enriched **1** show that carbons 6-7 and carbons 8-11 originate from PEP and E4P respectively. The second species that shows single enrichment at C-8 and C-9, C-10 and C-11 as part of a three-carbon unit occurs when E4P is produced via products of the pentose-phosphate pathway. In the pentose phosphate pathway glycerol metabolism yields fructose-6-phosphate (F6P) formed from two distinct three-carbon units arising from dihydroxyacetone phosphate (DHAP) or glyceraldehyde-3-phosphate (G3P). A transaldolase reaction yields xylulose-5-phosphate and E4P with the 3+1 labeling pattern (Fig. 2.5). Finally, the ^{13}C-^{13}C coupling observed between C-4, C-5, and C-15 is indicative of an intact 3-carbon unit and is consistent with the second PEP addition in the biosynthesis of **9** that becomes the amino acid alkyl group during aromatic amino acid biosynthesis. The observed glucose incorporation into **1** is summarized in Figure 2.10.

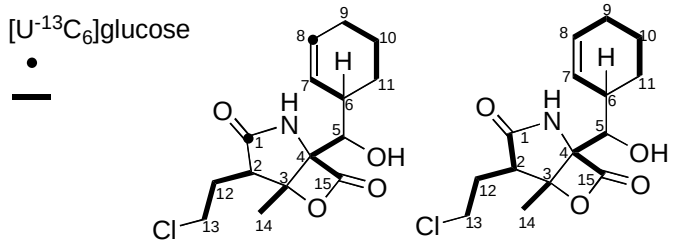


Figure 2.10. Summary of [U-$^{13}C_6$]glucose incorporation in **1** showing two observed labeling patterns. Solid line: intact incorporation. Circle: single enrichment.

The CHHA Moiety: Glycerol Feeding Experiments

Results from the [U-$^{13}C_6$]glucose feeding experiment were highly suggestive that CHHA arises from a shunt in the shikimic acid biosynthetic pathway prior to aromatization of the ring. If the CHHA moiety proceeded through an aromatic intermediate, a symmetrical labeling pattern would have been observed on the cyclohexene ring, which occurs when stereospecific control on the ring is lost after aromatization. This results in scrambling of the ^{13}C label and can be observed in the aromatic amino acids phenylalanine and tyrosine and other molecules that proceed through the aromatic intermediate phenylacetate (Fig. 2.7). The glucose experiment showed only one labeling pattern and thus, stereospecific control was retained.

Closer examination of the satellite peaks from C-6 and C-11 showed a second minor splitting of the ^{13}C signal with a coupling constant of 30 Hz. The most likely explanation was that the splitting resulted from the high overall incorporation of glucose on the adjacent carbons rather than that there was an additional coupling species. In

order to conclusively eliminate the existence of a third coupling species, [1,3-^{13}C]glycerol was administered to *S. tropica*. As mentioned above, glycerol is metabolized in the cell to PEP and acetate via glyceraldehyde-3-phosphate (G3P). The fate of the enriched carbon atoms can be traced through the pentose phosphate pathway and through reentry into glycolysis as shown in Figure 2.5.

An experiment in which [1,3-$^{13}C_2$]glycerol was administered showed enrichment C-14 as expected. The remaining carbons showed the predicted ^{13}C-enrichment in the erythrose portion of the molecule where incorporation was observed on C-2 and C-13, consistent with the incorporation of E4P. Labeling on the CHHA moiety was as predicted for a shikimic acid pathway precursor though no enrichment was observed on the carbonyl carbons C-1 or C-15. This is likely a result of dilution of the label in vitro or a result of nuclear Overhauser effects that occured in the NMR experiment.

The results from the [1,3-$^{13}C_2$]glycerol feeding experiment (Appendix A, Fig. 4) were consistent with the glucose data. However, the incorporation observed on the cyclohexyl ring did not convincingly explain the ^{13}C-^{13}C coupling between C-6 and C-11. To resolve the issue, an experiment using uniformly labeled glycerol was performed. The expected ^{13}C incorporation profile from the [U-$^{13}C_3$]glycerol experiment (Appendix A, Fig. 5, 6) should be similar to what is observed with uniformly labeled glucose with the exception that the intact 4-carbon E4P unit would not be observed. The 1D ^{13}C-NMR data from [U-$^{13}C_3$]glycerol enriched **1** was as predicted though a similar effect was observed at the C-1 and C-15 carbonyl carbons as in the [1,3-$^{13}C_2$]glycerol experiment, The data conclusively showed that the CHHA moiety arose from a non-aromatic product

of the shikimate pathway and reiterate that the 4-carbon alkyl moiety exhibited an erythrose-like labeling pattern (Fig. 2.11).

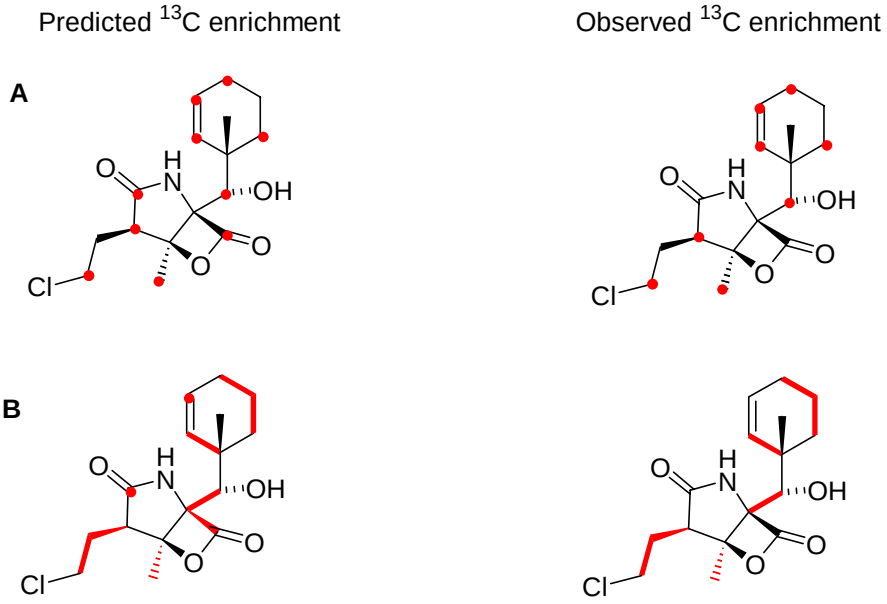


Figure 2.11. Predicted and observed ^{13}C enrichment for **1** with **(A)** [1,3-$^{13}C_2$]glycerol and **(B)** [U-$^{13}C_3$]glycerol. The lack of incorporation observed on the carbonyl carbons (C-1 and C-15) may be due to dilution or nuclear Overhauser effects.

Phenylalanine and Shikimate Feeding Experiments

In order to further probe the biosynthesis of the amino acid moiety of **1** [1-$^{13}C_1$]phenylalanine and [1,7-$^{13}C_2$]shikimic acid were fed. As expected, no enrichment was observed in the phenylalanine experiment (Appendix A, Fig. 8). An intense signal was observed on C-6 of **1** in the shikimate experiment indicating that shikimic acid was incorporated (Fig. 2.12). The single ^{13}C-enrichment observed on C-6 was consistent with

the loss of C-7 as CO_2 during the conversion of chorismate to prephenate. The incorporation of shikimate in combination with the glucose data confirmed the involvement of a non-aromatic amino acid intermediate and suggested that CHHA is biosynthesized through a previously unreported shunt in the shikimic acid pathway that branches from either chorismate or prephenate (Fig. 2.5).

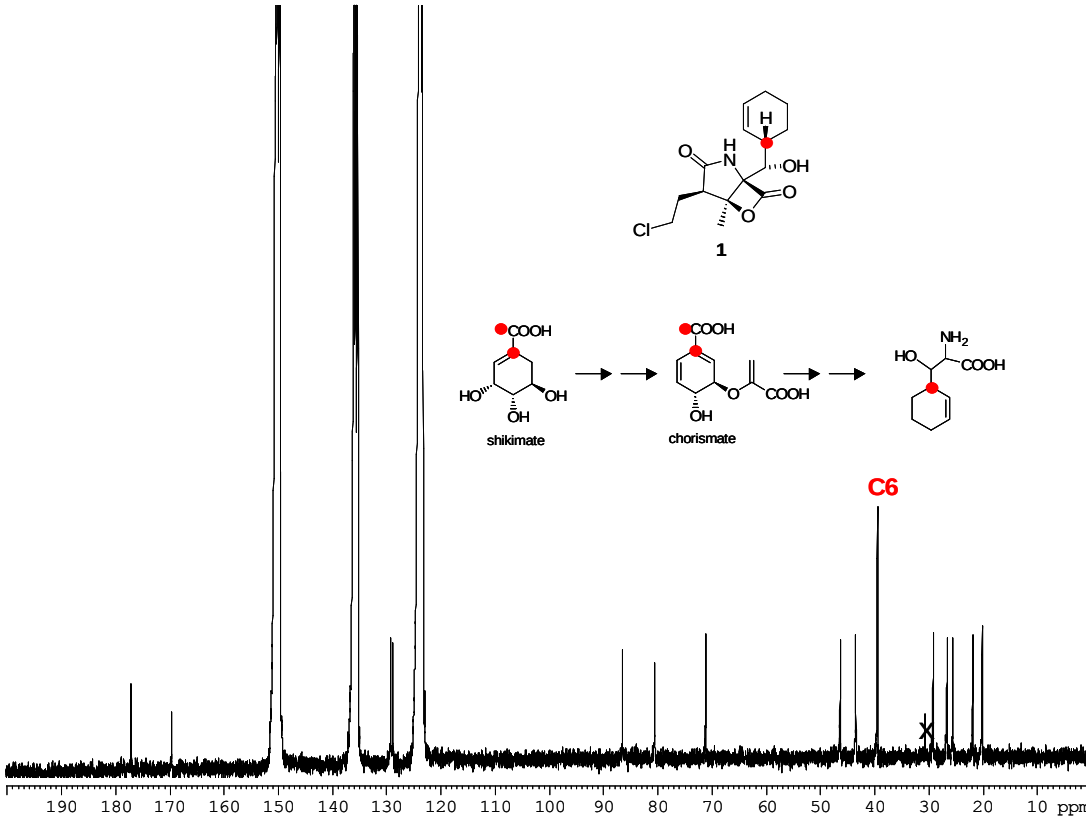


Figure 2.12. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1,2-$^{13}C_2$]shikimate enriched **1**. Enrichment is observed on C-6. Inset, conversion of shikimate to chorismate and the hypothetical amino acd showing the loss of C-1 from shikimate.

Chorismate Feeding Experiment

Based on the glucose and shikimate data it appeared that the pathway must proceed through chorismate and possibly prephenate on the route to CHHA biosynthesis. To explore possible the possible shunt pathway, the next obvious step was to feed ^{13}C-labeled chorismate and prephenate. However, chorismate and prephenate are not available commercially as ^{13}C-labeled compounds. Thus, in order to test the shunt-pathway hypothesis, chorismic acid was prepared *in vivo* from D-[U-$^{13}C_6$]-glucose with *Klebsiella pnuemoniae* (62-1) which is deficient in the shikimic acid pathway enzyme chorismate mutase (Gibson and Gibson 1964; Rajagopalan, Chen *et al.* 1992; Rieger and Turnbull 1996) which results in the over-production of chorismate. A 500 mL culture to which 2 g of [U-$^{13}C_6$]glucose was added as the sole carbon source produced 39 mg of uniformly labeled chorismate (100% enrichment) which was confirmed by ^{13}C-NMR (Appendix A, Fig. 9, 10) and LCMS (Appendix A, Fig. 11). The labeled chorismate was then fed to *S. tropica* CNB476 cultures. Several attempts at purifying **1** using flash column chromatography were unsuccessful and an unknown impurity consistently eluted with **1**. Analysis of the impure sample by NMR showed enrichment and coupling on the acetate carbons 3 and 14 that resulted from the *in vitro* catabolism of chorismate.

Origin of the 4-Carbon Alkyl Moiety

The lack of ^{13}C-enrichment observed in C-1 C-2, C-12 and C-13 of **1** in the [1,2-$^{13}C_2$]acetate feeding experiment precluded the possibility of the putative ethylmalonyl-CoA precursor arising from the known streptomycete butyryl-CoA pathways, or via a

PKS/fatty acid synthase-like mechanism. An intriguing observation from the [U-$^{13}C_6$]glucose feeding experiment was that the labeling pattern observed in C-1, C-2, C-12 and C-13 was identical to the labeling pattern observed in the 4-carbon E4P unit (C-8-11) in the cyclohexene ring where two coupling species were observed: a four-carbon unit and a second species with the 3+1 labeling pattern. This put forth the possibility that a previously unidentified pathway utilizing an erythrose-like precursor may function to provide the putative ethylmalonyl-CoA moiety. To test whether this unusual tetrose-like precursor arose from a novel pathway via a butyryl molecule, feeding experiments using [1-$^{13}C_1$]butyrate were performed (Appendix A, Fig. 12). No enrichment was observed on the carbons of interest, though enrichment was observed on C-3 (5.1%), which resulted from β-oxidation of [1-$^{13}C_1$]butyrate to [1-$^{13}C_1$]acetyl-CoA and incorporation as previously demonstrated with [1,2-$^{13}C_2$]acetate. This observation established that C-3 corresponds to the carboxyl carbon of acetate and that the β-lactone ring oxygen most likely arises from acetate.

To further study the origins of the unusual 4-carbon unit, a series of ^{13}C-labeled compounds including [3-$^{13}C_1$]alanine, [1,2-$^{13}C_2$]glycerol and 2-methyl[2-$^{13}C_1$]butyrate were fed. To test whether metabolites of branched-chain amino acids were precursors, singly labeled [3-$^{13}C_1$]alanine, which is metabolized to pyruvate and acetate, was administered and enrichment was observed in **1** at C-3 (Appendix A, Fig. 13), substantiating the previously observed acetate incorporation. Given that pyruvate is the precursor of valine, the data eliminated any variation of butyryl-CoA biosynthesis through the ICM pathway. The possibility that a threonine metabolite such as α-

ketobutyrate that eventually leads to isoleucine was also eliminated given that pyruvate is a precursor of threonine. As described previously, glycerol can enter glycolysis via G3P or F6P through the pentose phosphate pathway. The observed labeling pattern from [1,2-$^{13}C_2$]glycerol experiments was consistent with the conversion of glycerol to pyruvate and E4P through glycolysis and the pentose phosphate pathway as illustrated in Figure 2.7 and consistent with the [1,3-$^{13}C_2$]glycerol and [U-$^{13}C_3$]glycerol experiments. Variations on the branched chain amino acid biosynthetic pathways were therefore eliminated as the origin of the 4-carbon unit.

Finally, the possibility that the alkyl moiety in **1** proceeded through a mechanism analogous to that seen in lactacystin and omuralide biosynthesis was explored (Fig. 2.2). In lactacystin biosynthesis, methylmalonyl-semialdehyde is the proposed biosynthetic precursor for the alkyl moiety (Takahashi, Uchida et al. 1995). In isoluecine metabolism 2-methylbutyrate is converted to 2-ethylmalonyl semialdehyde that can then be metabolized to butyryl-CoA or ethylmalonic acid and eventually ethylmalonyl-CoA. In addition, some methanogenic bacteria can synthesis isoluecine directly from 2-methylbutyrate by reductive carboxylation (Robinson and Allison 1969; Ikiel, Smith et al. 1984). Therefore, if the pathway proceeded via a product of isoleucine catabolism, and if this pathway is present in *S. tropica*, 2-methylbutyrate could be incorporated into **1**. The feeding experiments with 2-methyl[2-$^{13}C_1$]butyrate however, showed no enrichment in **1** (Appendix A, Fig. 14). Ultimately, pathways utilizing butyryl-CoA via CCR, ICM or a novel pathway involving branched chain amino acid metabolism in addition to all known possible pathways from which an ethylmalonyl-like molecule could

occur were systematically eliminated. The E4P-like labeling pattern left questions as to whether it arose from an erythrose precursor or an alternate tetrose or pentose molecule. However, there were no additional labeled precursors available that could provide further biosynthetic information.

Biosynthesis of Salinosporamide Analogs

An alternative option to further explore the biosynthetic origins of the alkyl portion of **1** was to examine some of the minor fermentation analogs including **4**, the des-chloro analog of **1**, and analogs **5** and **6**. Originally, it was thought that **4** might be a direct precursor to **1** or that **4** was the dechlorinated product of **1**. However, in a small feeding experiment with [U-$^{13}C_4$]butyrate, analysis of **4** by LCMS revealed a major molecular ion *m/z* 284.0 ($[M+H+4]^+$), which suggested that butyrate is incorporated as a 4-carbon unit into the alkyl moiety in **4**. A minor molecular ion *m/z* 286.0 ($[M+H+6]^+$) was also observed which suggested that butyrate is incorporated as a 4-carbon unit into the alkyl moiety and as acetate in **4** (Tsueng and Lam, personal communication). The feeding experiments with labeled butyrate showing no incorporation into **1**, together with the LCMS data suggested that the 4-carbon units of **1** and **4** arose from alternate sources. For analogs **5** and **6** where $R_1 = CH_3$ or $R_2 = CH_2\text{-}CH_3$, respectively, it seemed likely that a propionate unit (as propionyl-CoA) would be incorporated. Feeding experiments with [1-$^{13}C_1$]propionate and analysis of compounds **1**, **4**, **5** and **6** by LCMS and 1D ^{13}C-NMR showed ^{13}C-enrichment at the C-1 position in **5** and at C-3 in **6** (Fig. 2.13) and no propionate enrichment in compounds **1** or **4**.

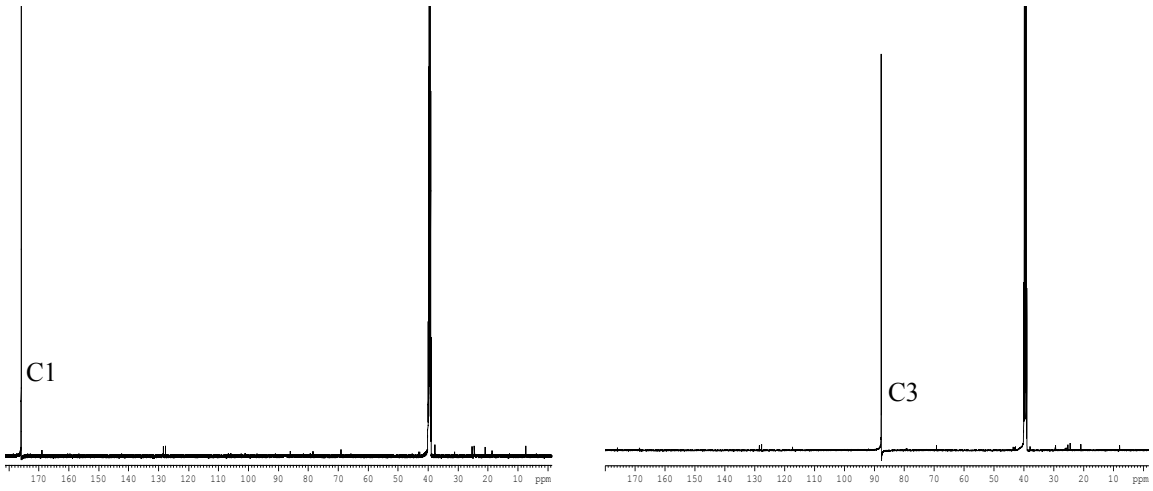


Figure 2.13. 1D ^{13}C NMR of [1-^{13}C]propionate enriched **5** (left) and **6** (right). ^{13}C-enrichment was observed at C-1 and C-3 for **5** and **6** respectively.

These feeding experiments confirmed that salinosporamide analogs with a 3-carbon unit utilized a propionate starter unit and the MS data suggesting that a 4-carbon butyryl moiety was the precursor for **4** (R. Lam, personal communication) required NMR analysis of 13C-enriched **4** to verify that this was truly the case. Feeding experiments with [1,2-$^{13}C_2$]acetate were repeated and the resultant ^{13}C-enriched **4** was analyzed by 1D ^{13}C-NMR. Enrichment and ^{13}C-^{13}C coupling were observed between C-3 and C-14 in concurrence with the acetyl-CoA incorporation reported for **1**. Coupling was also observed between C-1 and C-2 (J = 46) and C-12 and C-13 (J = 35). The data unequivocally showed that the 4-carbon moiety in **4** indeed arises from two acetate units and is consistent with the known biosynthetic pathways for ethylmalonyl-CoA

biosynthesis outlined in Figure 2.4. A [U-$^{13}C_6$]glucose feeding verified the data from the [1,2-$^{13}C_2$]acetate experiment though the overall incorporation percentage was too low to observe coupling for the remaining carbons (Table 2.3, Appendix A, Fig. 18, 19). These data verified that compounds **1** and **4** incorporate distinct biosynthetic precursors for the alkyl moiety and that **4** utilizes butyrate, which arises from the established biosynthetic pathways that utilize acetate.

Carbon I.D.	δ_c (ppm)	[1,2-$^{13}C_2$] acetate [a]
1	175.8	46
2	49.1	46
3	86.1	42
4	78.6	
5	69.1	
6	37.7	
7	128.5	
8	127.7	
9	24.6	
10	21.0	
11	25.3	
12	18.0	35
13	12.4	35
14	20.2	42
15	168.9	

a = referenced to DMSO. b % incorporation = (A-B)/B. A = intensity of enriched carbon and B = intensity of natural abundance carbons.

Table 2.3. 1D ^{13}C NMR data for [1,2-$^{13}C_2$]acetate-enriched **4**. Overall incorporation % is not reported in the experiement due to the small amount of material available for analysis.

Discussion of Results and Conclusions

PKS derived molecules often show variation in structure resulting from promiscuity in the biosynthetic enzymes. The feeding experiments have shown that two- three- and four-carbon precursors can be incorporated into the various salinosporamide analogs from acetate, propionate and butyrate respectively and confirmed that **4** is neither a precursor nor a degradation product of **1**. The extensive feeding studies suggest that biosynthesis of the putative 4-carbon chlorobutyryl-moiety in **1** must through a novel mechanism toward chloroethylmalonyl-CoA and the erythrose-like labeling patterns suggest that the precursor has a tetrose origin. The low levels at which the minor analogs are produced with respect to **1** is highly suggestive that the salinosporamide biosynthetic enzymes (the PKSs) show selectivity for the chlorinated tetrose-like precursor and acetate as chloroethylmalonyl-CoA and acetyl-CoA respectively. The amino acid moiety arises from a shunt in the shikimic acid pathway though whether the branch-point is a chorismate or prephenate intermediate remains to be determined.

Based on the results of feeding experiments and available precursors a proposed biosynthetic scheme for **1, 4, 5** and **6** is depicted in Figure 2.14. In this scenario, the diketide arises from the condensation of two alkyl groups represented by R_1 and R_2. The incorporation of acetate into the 4-carbon unit of **4** and the butyrate incorporation observed by MS analysis it is likely that the PKS utilizes an ethylmalonyl-CoA precursor that arises via the CCR pathway. For **1** it is plausible that a chlorinated tetrose molecule is converted to 4-chloro-ethylmalonyl-CoA which is then loaded onto the PKS. The NRPS bound amino acid, CHHA would then extend the diketide to yield the linear

salinosporamide precursor. An unusual cyclization reaction and off-loading would yield the bicyclic salinosporamide molecule.

The biosynthetic pathway for **1** likely represents novel enzymatic reactions for precursor biosynthesis, in particular, the putataive 4-chlorotetrose unit, and the CHHA amino acid moieties are unprecedented in the literature and the unusual cyclization reaction to form the γ-lactam-β-lactone ring is also unique. By elucidating the biosynthetic building blocks for **1**, these experiments may facilitate the development of new approaches for precursor directed biosynthesis for the production of new salinosporamide analogs with improved biological activity. Knowledge of the multiple pathways by which salinosporamide compounds are assembled can be taken into consideration when optimizing fermentation conditions in order to reduce or eliminate the production of the minor salinosporamide analogs presently considered contaminants in the large-scale industrial production of **1.** This is especially true given that the activity of **1** is so finely tuned to its functional groups. Including the interactions of the cyclohexene ring with the binding pocket, the β-lactone ring which is necessary for binding to the proteasome and the opening of the ring that occurs to form the tetrahydorfuran ring, in addition to the necessity of the chlorine atom for the formation of the stable tetrahydrofuran ring that provides the basis for the irreversible binding of **1** to the proteasome β-subunit.

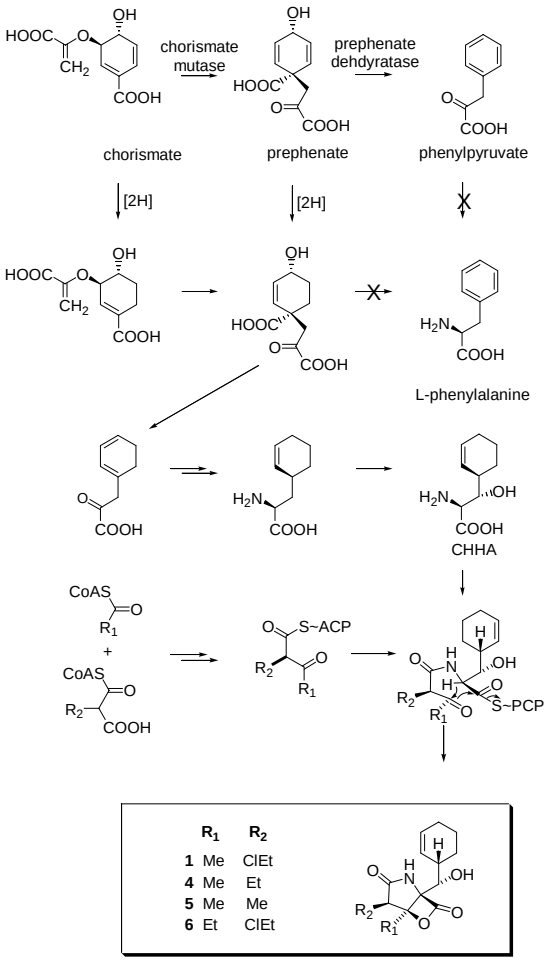


Figure 2.14. Proposed biosynthetic scheme for **1**, **4**, **5** and **6** based on feeding experiments with ^{13}C-labeled precursors.

CHAPTER 3

CLONING AND SEQUENCING THE SALINOSPORAMIDE BIOSYNTHETIC GENE CLUSTER

Introduction and Background

In order to carry out genetic experiments and generate salinosporamide A analogs, the DNA sequence of the genes that encode salinosporamide A biosynthesis and regulation are required. Based on the available biosynthetic precursors it appeared that **1** was biosynthesized via an alternate mechanism than that observed for the structurally similar lactacystin and omuralide. On the basis of general biosynthetic principles, and results of feeding experiments with ^{13}C-labeled substrates, the proposed biosynthesis of salinosporamide would be accomplished via a hybrid polyketide synthase (PKS) nonribosomal peptide synthetase (NRPS) enzyme complex. This chapter describes the isolation and sequencing of the putative salinosporamide biosynthetic gene cluster.

The Genetic Organization of PKSs and NRPSs

Type I PKSs and NRPSs genes share a modular organization where each module is a functional building-block responsible for the incorporation and modification of one carboxylic or amino acid unit respectively. The enzymatic units that reside within a module are called domains and these domains catalyze the steps of substrate activation, covalent binding and carbon-carbon or peptide bond formation (Cane 1997). Domains of equal function share a number of highly conserved sequence motifs and these "core-

motifs" allow the identification of individual domains on the protein level (Schwarzer, Mootz et al. 2002).

A typical type I polyketide module consists of an acyltransferase (AT) domain responsible for selection and activation of starter or extender units, an acyl carrier protein (ACP) for extender unit loading and a β-ketoacyl synthase (KS) domain that catalyzes the decarboxylative condensation between the aligned acyl thioesters to elongate the growing polyketide chain (Fig. 3.1 A) (Cane 1997; Cane, Walsh et al. 1998; Staunton and Wilkinson 1998; Shen 2000). PKSs utilize a wide assortment of starter units, such as malonyl-CoA, short chain (branched) fatty acids (methyl-, methoxy-, ethyl- or propyl malonyl-CoA) (Katz 1997), alicyclic and aromatic acids and amino acids in the assembly of their products (Moore and Hertweck 2002).

Non-ribosomal peptide synthetases are arranged in a similar fashion to modular PKS enzymes. A typical NRPS module consists of an adenylation (A) domain which is responsible for selection and activation of amino- or carboxy-acid substrates as an amino acyl adenylate, while ATP is consumed (Dieckmann, Lee et al. 1995; Stachelhaus and Marahiel 1995; Mootz and Marahiel 1997; May, Kessler et al. 2002). The activated amino acyl adenylate is then loaded onto the peptidyl carrier protein (PCP), also known as the thiolation (T) domain (Stachelhaus, Huser et al. 1996; Ehmann, Shaw-Reid et al. 2000). A condensation (C) domain catalyzes peptide bond formation between the aligned amino acyl thioesters of adjacent PCP molecules to elongate the growing peptide chain (Fig. 3.1 B) (Cane 1997; Cane, Walsh et al. 1998; Stachelhaus, Mootz et al. 1998; Cane and Walsh 1999; Bergendahl, Linne et al. 2002).

The activation of both ACPs and PCPs occurs via a post-translational modification of the *apo*-carrier protein to the *holo*-carrier protein by the attachment of a 4'-PP prosthetic group which is catalyzed by a family of 4'-phosphopantetheinyl transferases (PPTases) (Fig, 3.1 C) (Lambalot, Gehring et al. 1996; Walsh, Gehring et al. 1997). During the elongation process the growing molecule remains tethered to the carrier protein in a thioester linkage via the sulfhydryl group of the 4'phosphopantetheine group. Termination of polyketide or non-ribosomal peptide synthesis is normally catalyzed by a thioesterase (TE) domain, which is usually located C-terminal of the last module. Product release is achieved through hydrolytic cleavage of the thioester bond on the carrier protein to release a linear product or a macrocycle, which appears to be the favored mechanism (Fig. 3.1 D). An additional discrete TE (TEII) has also been found to associate with several NRPS and PKS gene clusters and TEII domains have been shown to function in an editing capacity whereby they liberate misprimed carrier proteins.

Hybrid peptide-polyketide metabolites are biosynthetically derived from amino acids and short carboxylic acids. Based on the biosynthetic mechanisms by which the amino acid, or peptide, and carboxylic acid, or polyketide, moieties are incorporated into these products, hybrid natural products are divided into two classes: those whose hybrid peptide-polyketide backbone is assembled by a NRPS-PKS system that mediates elongation of a NRPS bound peptidyl intermediate by a PKS module or vice versa (Fig. 3.1 E, F) and those whose hybrid peptide-polyketide backbone is assembled via mechanisms that do not require direct functional hybridization between the NRPS and PKS proteins (Du, Sanchez et al. 2001).

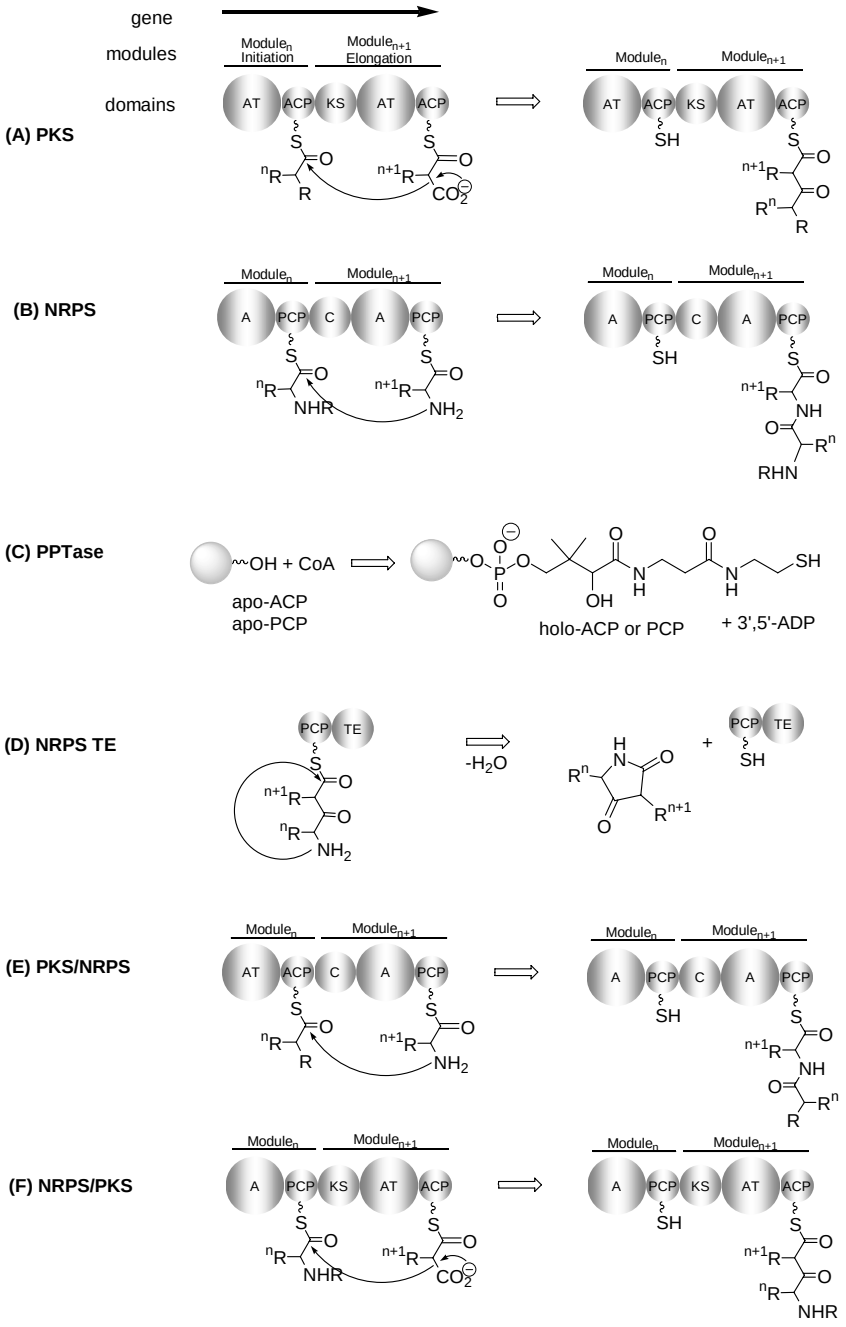


Figure 3.1. The modular organization of PKS and NRPS and comparison to hybrid NRPS/PKS and PKS/NRPS. (A) C-C bond formation for polyketide biosynthesis catalyzed by two hypothetical PKS modules. (B) C-N bond formation for peptide biosynthesis catalyzed by two hypothetical NRPS modules. (C) Posttrasnlational modification of apo-carrier protein into holo-carrier protein by a PPTase. (D) Example of TE off-loading and cyclization mechanism for hypothetical peptide. (E) C-N bond formation for hybrid polyketide-peptide biosynthesis catalyzed by a hypothetical PKS/NRPS hybrid. (F) C-C bond formation for hybrid peptide-polyketide biosynthesis catalyzed by a hypothetical NRPS/PKS hybrid.

Tailoring Enzymes in NPRS and PKS Biosynthesis

In addition to the PKS and NRPS architecture, domains have also been identified for the modification of the acyl substrates or the amino acyl and or peptidyl substrates during the assembly process. Examples of these "tailoring enzymes" include epimerization (E) domains for the conversion of an amino acid from the L- to the D-configuration (Marahiel, Stachelhaus et al. 1997), methyltranferases (MT) for N- or C-methylation of amino acid residues (Gehring, DeMoll et al. 1998; Reimmann, Patel et al. 2001), cyclization (Cy) domains that replace C-domains for the formation of heterocyclic rings (Konz, Klens et al. 1997) reductases (R) for reductive release of an aldehyde product (Ehmann, Gehring et al. 1999) and oxidation domains (Ox) (Du, Chen et al. 2000; Du, Sanchez et al. 2000; Molnar, Schupp et al. 2000; Du, Sanchez et al. 2003).

PKS tailoring enzymes that determine the reductive state of the growing polyketide molecule are analogous to those used by fatty acid synthases (FASs) and include β-keto reductases (KR) (Cane 1997), dehydratases (DH) and enoyl-reductases (ER) (Cane 1997), which determine whether synthesis stops at the stage of the hydroxyl group, double bond formation or complete reduction. Methyl transferase domains introduce methyl groups on the β-hydroxyl or enol group (O-MT) (Silakowski, Schairer et al. 1999) or introduce a methyl branch into the α-position (MT). An acyl CoA ligase (AL) domain may be present for priming of the loading module with an unusual starter unit (Duitman, Hamoen et al. 1999) and amino-transferase (AMT) domains that covert an activated fatty acid into an amino acid have also been noted and are uniquely located between a PKS and NRPS module in hybrid PKS/NRPS systems (Duitman, Hamoen et

al. 1999; Tillett, Dittmann et al. 2000). Genes encoding enzymes for further modification through acylation, glycosylation and halogenation may also be present in or near a biosynthetic gene cluster (Du, Sanchez et al. 2003).

The Colinearity Rule

Early work with NRPS and type I PKS led to the dogma of the colinearity rule: that the hierarchical arrangement of the modules correlated to the position of the building blocks in the final product (Marahiel, Stachelhaus et al. 1997; Keating and Walsh 1999). There are several classes of NRPSs and PKSs however, that involve iterative use and non-linear arrangements of the modules, contradicting the colinearity rule (Mootz, Schwarzer et al. 2002). Although the general set of genes required for assembly of the carbon skeleton from PKS and NRPS derived molecules follows the general scheme outlined above, consideration for the unexpected must be made when predicting the arrangement of modules and domains for NRPS and PKS products.

Proposed Biosynthetic Gene Cluster for Salinosporamide A

Based on other hybrid NRPS/PKS biosynthetic gene clusters (Silakowski, Kunze et al. 2001; Silakowski, Nordsiek et al. 2001), the salinosporamide megasynthetase would putatively chain extend acetyl-CoA with a 4-carbon precursor such as chloroethylmalonyl-CoA via a PKS mechanism. The putative diketide would then couple through an amide bond to the putative nonproteinogenic amino acid, (2R,3S,4R)-3-cyclohexene-3-hydroxyalanine by an NRPS to generate the linear enzyme-bound

salinosporamide A precursor. Enzymatic off-loading with concomitant cyclization to form the γ-lactam-β-lactone ring system would yield the salinosporamide skeleton. Thus, the minimum set of genes required for the proposed biosynthesis of carbon skeleton of **1** would include two PKS modules for assembly of the alkyl moiety, an NRPS module for loading and extension of the amino acid moiety and a TE domain for off-loading and/or cyclization of the linear peptide intermediate (Fig. 3.2).

The first PKS, or loading module, would activate acetyl-CoA and consist of an AT and an ACP domain. A second PKS extender module would conatin an AT and an ACP domain for activation of the putative ethylmalonyl unit and a KS that would catalyze the decarboxylative condensation reaction between the ethylmalonyl- and acetyl-thioesters. The NRPS module would consist of A and PCP domains for activation and loading the amino acid and a C-domain to catalyze the formation of the amide bond with the polyketide. A TE domain would be required for off-loading and possibly cyclization to form the salinosporamide bicyclic core.

Genes encoding tailoring enzymes including a chlorinase for attachement of a chlorine atom on the ethyl moiety and a cytochrome P450 β-hydroxylase for hydroxylation of the amino acid at the β-position were also predicted to be present in or near the cluster (Fig. 3.2). Other genes that were expected in the gene cluster included those for biosynthesis or modification of the putative ethylmalonyl-CoA and amino acid starter units. As described in Chapter 2, a number of streptomycete PKSs, such as those involved in monensin, FK520, tyolosin and niddamycin production use ethyl malonyl-CoA during polyketide extension (Reynolds, O'Hagan et al. 1988; Byrne, Shaffiee et al.

1993; Gandecha, Large et al. 1997; Kakavas, Katz et al. 1997; Liu and Reynolds 1999). In addition to the CCR gene located with primary metabolic genes involved in acetate assimilation, *ccr* homologs (Wallace, Bao et al. 1995) have been observed within the biosynthetic gene clusters of tylosin, niddamycin, coronafacic acid and monensin A, which all contain an ethylmalonyl-CoA precursor (Gandecha, Large *et al.* 1997; Rangaswamy, Mitche *et al.* 1998; Stassi, Kakavas *et al.* 1998; Liu and Reynolds 1999; 2001). Other natural products that utilize shunt products of the shikimic acid pathway (Blanc, Gil et al. 1997; Floss 1997; McDonald, Mavrodi et al. 2001; Arakawa, Muller et al. 2002) are known to possess copies of genes associated with the shikimate pathway in their gene clusters including deoxy-*arabino*-heptulosonate-7-phosphate (DAHP synthase), chorismate mutase (ChM) and prephenate dehydrogenase (PDH) (He, Magarvey et al. 2001). Therefore, the possible involvement of a *ccr* homolog or shikimate pathway genes was explored. In addition to the biosynthetic genes, genes encoding resistance, transport and transcriptional regulation were expected within or surrounding the gene cluster.

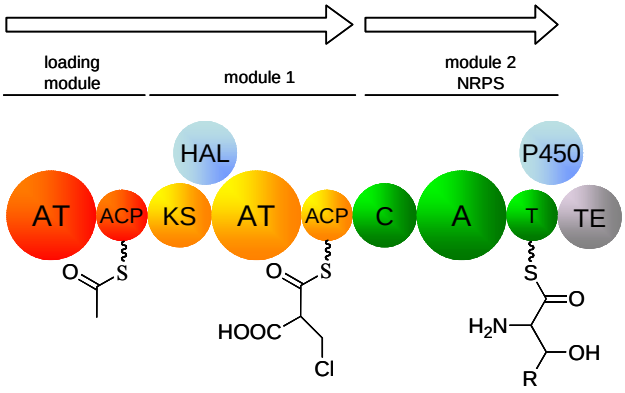
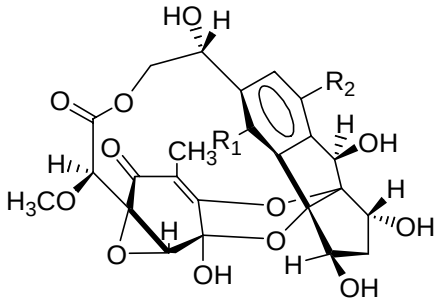


Figure 3.2. The proposed minimal biosynthetic gene cluster required for the assembly of **1**. The PKS loading module (red) contains AT and ACP domains. Module 1: The second PKS (orange) contains AT, ACP and KS domains. Module 2: The NRPS (green) contains A and PCP domains. Tailoring genes are in blue: Cytochrome P450 β-hydroxylase and halogenase and the TE domain (gray).

Isolating a biosynthetic gene cluster usually entails creating a genomic or gDNA library in a cloning vector that can uptake large DNA fragments such a cosmid or fosmid that can handle DNA inserts of 40 ~ 60 kilobases (kb) respectively. The library is then screened by Southern hybridization with PCR generated probes for the genes of interest and the clones that cross-hybridize with the probes are selected as candidates for containing the putative gene cluster. Multiple probes can be used to generate specificity and increase the likelihood of isolating a cosmid clone that contains the gene cluster of interest. Once a putative gene cluster is isolated and verified by DNA sequencing, its function must be verified via heterologous expression in addition to functional characterization of the individual genes at the protein level.

The use of degenerate PCR primers to amplify NRPS and PKS genes based on higly conserved core sequences has proved to be a useful method to locate biosynthetic gene clusters (Turgay and Marahiel 1994; Sosio, Bossi et al. 2000). The NRPS A-domain is particularly valuable for this approach because it is primarily responsible for amino acid selectivity, which can be predicted based on the nonribosomal code (Stachelhaus, Mootz et al. 1999; Challis, Ravel et al. 2000). Considering that the salinosporamide A producing *S. tropica* CNB476 strain also produces the sporolide macrolide compounds (Fig. 3.3) (Buchanan, Williams et al. 2005), that are likely to be biosynthesized via a PKS pathway, several copies of PKS genes were expected to be present in the genome. Given that salinosporamide A and its analogs were the only other molecule isolated from the strain thus far that appeared to be biosynthesized by a mixed

PKS/NRPS pathway, the initial search for the gene cluster for **1** was focused on the NRPS.



A R_1 = Cl R_2 = H
B R_1 = H R_2 = Cl

Figure 3.3. Structure of sporolide A and B isolated from *S. tropica*.

Materials and Methods

Reagents and strains

Restriction enzymes and polymerases were purchased from Invitrogen or New England Biolabs (NEB). *Escherichia coli* strain XLIBlue and DH5α were routinely used as a host for DNA cloning and sequencing unless otherwise noted. *E. coli* was grown overnight in LB (Luria-Bertani) medium with ampicillin at a final concentration of 50 μg/mL at 37° C with shaking at 250 rpm. *Salinispora tropica* was grown according to procedures described in Chapter 2.

E. coli strain transformation, and other DNA manipulations were performed according to standard methods (Sambrook, Fritsch et al. 1989). High molecular weight

genomic DNA (gDNA) was isolated from *S. tropica* CNB476 using a modified procedure for streptomycete DNA (Kieser, Bibb et al. 2000). *S. tropica* cultures were grown in A1BFeC medium until turning a bright orange color. Cells were pelleted by centrifugation at 5000 xg for 5 min at 4° C, resuspended in 50 mL lysis buffer (25 mM Tris pH 8.0, 25 mM EDTA, 15% sucrose) and 50 mg lysozyme and incubated at 37° C for 1 hour. One mg proteinase K and 1 mL 10% SDS were added and the cells were incubated 30 minutes at 50° C and chilled on ice at 4° C. DNA was isolated via phenol chloroform extraction followed by chloroform extraction, and isopropanol precipitation. The air-dried pellet was redissolved in 1.5 mL TE (pH 8.0) and RNAse A (50 μg/mL). Plasmid preparations were performed with the Qiagen Miniprep Kit according to manufacturer's instructions with elution buffer heated to 60° C to elute cosmid DNA

High molecular weight gDNA from *S. tropica was* partially digested with MboI (200 μL gDNA, 272.5 μL ddH_2O, 52.5 μL, 10X NEB 3 buffer, and .2.5 U MboI). The reaction was pipetted slowly and constantly through a wide bore 1 mL pipette tip to prevent shearing and promote digestion while held at 37° C. 100 μL of the digestion reaction were removed and placed in tubes containing 10 μL of 10 mM EDTA on ice to terminate enzyme activity at 3, 6, 9, 12, and 15 min. 10 μL of each aliquot was viewed on a 0.3% agarose gel with λ and λ HindIII digested markers at 60 volts for 4 hours. The 12 and 15 min digestions were combined and dephosphorylated (100 μL digested DNA, 90 μL H_2O, 10 μL 10X NEB3 buffer and 10U CIP) and incubated at 37° C for 30 min at which time 10U CIP were added and incubation continued for an additional 30 min.

Dephosphorylated MboI digested gDNA was purified via a phenol/chloroform and chloroform, precipitated with ethanol and resuspended in 15 μL H_2O.

Approximately 20 ug pOJ446 was digested with HpaI in a total volume of 200 μL for 2 hours followed by dephosphorylation with 10U CIP and incubated at 37°C for 30 minutes 10U CIP were added and the reaction mixture was incubated an additional 30 minutes. The vector was purified via phenol chloroform extraction and ethanol precipitation and digested with BamHI, phe/chl purified and resuspended in 15 μL H_2O.

Ligations were performed with ~ 0.5 μL gDNA, ~1 μg HpaI/BamHI digested pOJ446, 1X T4 DNA ligase buffer and 10U T4 DNA ligase in a total volume of 15 μL and incubated at 16°C overnight. A gel was run to confirm ligation fragments of appropriate size. The reaction was packaged using Gigapack III gold packaging extract according to Stratagene® protocol and transformed by electroporation into XLI Blue MRF' cells using titers of 1:10 and 1:50 dilutions. The 1:10 packaging reaction dilution resulted in more efficient titers so the reaction was repeated and plated on LB-apramycin (50 μg / mL) plates and grown ~ 10 hours overnight at 37° C.

Construction and Screening the S. tropica Cosmid Library

Hybond-N+ nylon transfer membranes (Amersham Pharmacia) were placed on the master plates containing cosmid clones. The membranes were marked with a needle and placed colony side up on fresh LB apr50 plates and incubated at 37° C ~ 10 hours until colony growth was optimal. Master plates were also reincubated at 37° C for 4 hours after colony lifts and placed in at 4° C. The membranes were removed from plates

and placed colony side up on 3mm Whatman filter paper prewetted with 0.5 M NaOH for 30 seconds. Filters where then placed on a sheet of Whatman paper (grade 4) prewetted with 1 M Tris-HCl (pH 7.6) for 30 seconds then transferred to another sheet of Whatman paper prewetted with 1 M Tris HCl (pH 7.6), 1.5 M NaCl for 30 seconds. The membranes were then baked at 80°C for 2 hours to cross-link DNA to membranes. To remove *E. coli* debris membranes were boiled in 20X SSC in a sealed bag for 10 minutes then rinsed several times at room temperature in fresh bags containing 2X SSC.

Biotinylated probes were prepared per the ECL kit (Amersham Pharmacia) and membranes were incubated, washed, and illuminated per ECL protocol. Primary wash buffer consisted of 0.3X SSC, 0.4% SDS and the secondary wash buffer was 2X SSC. X-ray film exposure times varied per experiment but average the time for optimum detection was ~ 20 minutes.

PCR amplification and cloning of DNA probes

Degenerate oligonucleotide primers were purchased from SIGMA/Genosis and Qiagen. *Taq* polymerase and dNTP's were purchased from Invitrogen (Carlsbad, NM). All reactions were performed with 0.5 ug *S. tropica* gDNA. Subcloning and sequencing of PCR products was performed using the TOPO TA cloning kit (Qiagen) according to manufacturer's protocol. Primers are listed in Appendix B, Table 1. Ketosynthase (KS) domains were amplified using the primer set (KSF1/KSR1) and the crotonyl-CoA reductase was amplified with the primer set (CCRF2/CCRR2). Conditions were as follows: 5 minute denaturation at 95°C; 1 min at 95°C; 1 min at 60°C for KS and 70°C

for CCR; 1 min at 72°C; 30 cycles; 5 min at 72°C. The primer pair GN1_111_1F/ A8-2R successfully amplified the NRPS probe (5 min denaturation at 95°, 28 cycles of 95° 1 min, 50° 1 min, 72° 1 min, 10 minute extension at 72°). The prephenate dehydratase fragment was amplified using the primer set sal101F1/sal101R2 (94° denaturation for 2 min, 30 cycles of 94°, 1 min, 58° 1 min, 72° 45 s, final extension at 72° for 4 min). The DAHP synthase fragment was amplified using the DAHPF1/DAHPR1 primer set (94°denaturation for 4 min, 30 cycles of 94° 1 min, 50° 30 sec, 72° 45 sec, 72° extension for 2 minutes).

DNA Sequencing and Analysis

The cosmid DNA was prepared for subcloning by several methods including nebulization, sonication and restriction endonuclease digestion. Nebulization was performed with a reaction mix containing 500 μL cos23, 200 μL 10X TM buffer, 300 μL H_2O, 1mL 50% glycerol) that was placed in a modified nebulizer on ice. The nebulizer was modified with a rubber seal to block the top opening and rubber tubing was attached to the small opening and a nitrogen tank. The solution was nebulized at 15 psi for 5 minutes. Samples were divided into 5 tubes, precipitated with ethanol, dried and resuspended in 27 μL 1X TM buffer. End-repair was performed with 5 uL kinase buffer, 5 μL 10 mM ATP, 7 μL 0.25 mM dNTPs, 1 uL (3U / μL) T4 polynucleotide kinase, 2 μL (5U / μL) Klenow DNA polymerase and 3 μL Klenow buffer. The samples were incubated at 37° C for 30 minutes and separated on a 1% low melting point (LMP) gel for

1 hour at 100 mAmps. The bands corresponding to ~1.5 kb were cut out, gel purified with the Qiagen gelQuick kit and eluted with 20 μL elution buffer. Ligations containing 5 μL HincII digested CIP dephosphorylated pGem3Zf- vector (Stratagene) 10 μL cos23 DNA, 20 μL 10X ligase buffer and 2 μL T4 DNA ligase in a total volume of 200 μL were incubated overnight at 15°. The ligations were transformed into DH5α *E. coli* and incubated on LB/ampicillin plates overnight.

Shearing of cosmid DNA was performed by sonication using a sonicator on power level 2. Genomic DNA 100 μg was diluted in 10 mL Tris (pH 7) and pulsed for 5 seconds. The sample was precipitated with ethanol, dried and resuspended in 25 μL TE buffer and run on a 0.7% agarose gel. The fragments ranging from 1.5 - 2 kb were gel purified using the Qiaquick gel purification kit, end filled and ligated into CIP dephosphorylated pBluescript II SK(-) (Stratagene), transformed into *E. coli* XLIBlue. Colonies were grown overnight and Plasmids were isolated and sequenced with the T7 and KS primers.

Sequencing was performed using the BigDye Terminator Ready Mix and an ABI PRISM® 3730 DNA Sequencer (Applied Biosystems) in the Genomic Analysis and Technology Core at the University of Arizona. The sequences were edited and assembled using the Sequencher program. Translation of DNA sequence used the Bioedit and FramePlot (Ishikawa and Hotta 1999) programs. Open reading frames were determined using FramePlot and Vector NTI. Alignments were performed using ClustalW and Bioedit. Sequence homologies were determined using NCBI BLAST and the CDD database.

Transposon insertions were used to obtain sequence from some of the larger fragments. Transposon insertion reactions were performed using the Epicentre EZ::TN <T7KAN-2> transposon insertion kit according to the manufacturer's instructions. Plasmids were restriction digested with BamHI to roughly determine transposon insertion position and reaction efficiency. The transposon insert contained one BamHI site whereas the plasmids alone contained at least two. Therefore, digests yielding at least three fragments were sequenced with the Kan-2-FP-1 and Kan-2-RP-1 primers provided by Epicentre. Translations of open reading frames can be found in Appendix B, Fig. 1.1 and the complete the *sal* gene cluster sequence is in Appendix B, Fig. 1.2.

Results

PCR Amplification of Biosynthetic Genes from S. tropica CNB476

Others in our lab designed degenerate primers based on the conserved NPRS A domain motif A3 (SGTTGxPKG) core sequence (Gocht and Marahiel 1994) and the invariant serine residue (LGGxS) (Schlumbohm, Stein et al. 1991) of the PCP that is the site of thioester formation and of 4'PP binding. The primers were used to PCR amplify a 549 bp region of an NRPS adenylation domain. Results of a BLAST search showed the sequence was most closely related to saframycin Mx1 synthetase (42% / 58%) (Identity/similarity) (Pospiech, Cluzel et al. 1995) and contained the AMP binding motif characteristic of NRPS domains (CDD:40636). An alignment showed that the amino acid signature of the active site residues was similar to A-domains with a phenylalanine signature (Conti, Stachelhaus et al. 1997).

Given that 16S RNA sequence data had shown that *S. tropica* was closely related to the *Micromonosporacae* (Mincer, Jensen *et al.* 2002), the codon bias from the NRPS fragment and *Micromonospora chalcea* (Nakamura, Gojobori et al. 2000) was used to design primers against the well conserved ketosynthase domains LAMDPQQ and VEAHGTGT positioned up and downstream of the type I KS active-site cysteine respectively (Beyer, Kunze et al. 1999). Primers were also designed for a CCR based on alignments of the conserved domains (RHDLPYHVIGSD and DNRYLWMKL) from streptomycete CCR proteins associated with secondary metabolite biosynthetic gene clusters (Wallace, Bao et al. 1995; Gandecha, Large et al. 1997).

At least two distinct PKS KS fragments were amplified (Fig. 3.4) although several more were expected to be present in the genome based on the production of the sporolide compound. The ks1 and ks2 PCR products only varied by 1 amino acid residue which may indicate that they are the same fragment and an error occurred during PCR or sequencing. It is likely that many more PKS genes are present and that this particular primer set lacked the degeneracy to amplify them.

```
KS1 235 aa        REAKAMDPQQ  RLLLETSWEA  LERAGIDPTS  LRGSPTGVFT  GVMYHDYASR  LPEVPEELAG
KS2 230 aa        -----MDPQQ  RLLLETSWEA  SERAGIDPTS  LRGSPTGVFT  GVMYHDYASR  LPEVPEELAG
KS3 235 aa        -----MDPQQ  RLMLELAWEA  LEDAGIAPGN  LAGTDTGVFV  GTSWCDYSAL  AHEAAAQFEI
Clustal consensus      *****  **:** :***   * *** * .  * *: ****.  *. : **::     *.. ::

KS1 235 aa        --YLGNGSMA  SVASGRVSYT  LGLEGPAVSV  DTACSSSLVA  IHLAVQALRS  GECSLALAGG
KS2 230 aa        --YLGNGSMA  SVASGRVSYT  LGLEGPAVSV  DTACSSSLVA  IHLAVQALRS  GECSLALAGG
KS3 235 aa        GPHAATGMHD  SIIANRVSYA  LALQGPSMAI  DTACSSSLVA  VHLACQSLWS  GESELALAGG
Clustal consensus   : ..*     *: :.****:  *.*:**:::: **********  :*** *:* *  **..******

KS1 235 aa        VTVMSTPDTF  VEFSLQRGLA  ADNRCKSLAA  GADGTAMSEG  AGMLLLERLS  DAERHGHQVL
KS2 230 aa        VRVMSTPDTF  VEFSLQRGLA  ADNRCKSFAA  GADGTAMSEG  AGMLLLERLS  DAERHGHQVL
KS3 235 aa        VTLNIFGPHY  LAMNETGALS  PDGRCKTFDA  RADGTVRGEG  AAVVVVTPLR  VALERGLPVY
Clustal consensus **:      :  : :.   .*:  .*.***:: *   ****. .**  *.::::  *    * .:*  *

KS1 235 aa        AVVRGSAINQ  DGASNGLTAP  NGPSQQRVIR  QALAAANVRP  DQVDLVEAHG  TGTTPPG--
KS2 230 aa        AVVRGSAINQ  DGASNGLTAP  NGPSQQRVIR  QALAAANVRP  DQVDLVEAHG  TGTTTPG--
KS3 235 aa        CLIRGSAVNN  DGLSNGLTAP  NPRAQETLLR  TAYRRAGLAP  WLVDYVECHG  TGTPNSGDP
Clustal consensus .::****:*:  ** *******  *  :*: ::*   *   *.: *    ** **.**  ***. .*
```

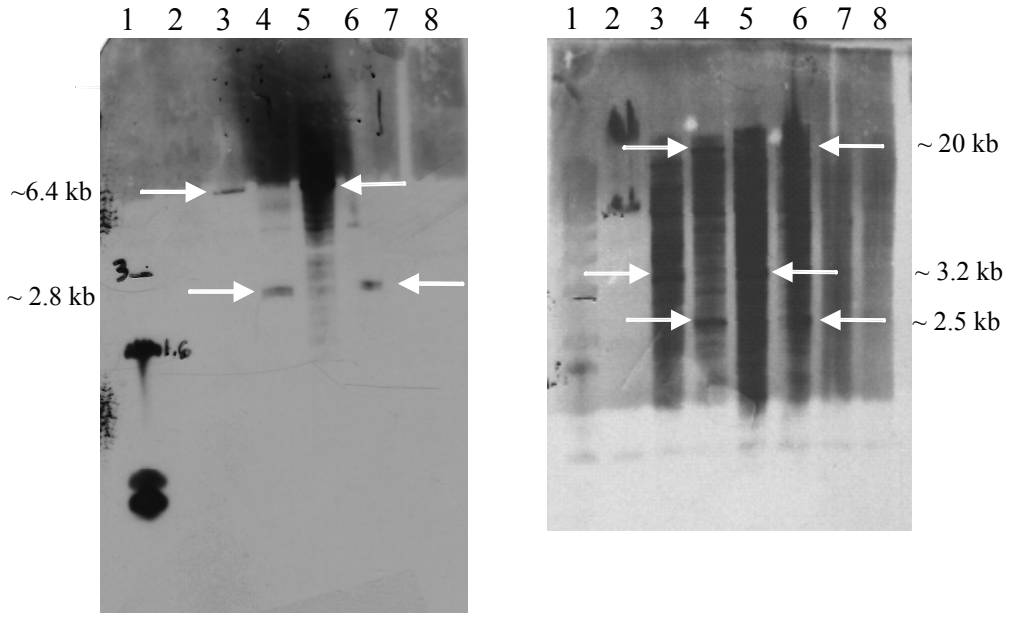
Figure 3.4. ClustalX alignment of PCR amplified PKS-KS fragments from *S. tropica* CNB476.

A 756 bp *ccr* gene fragment was amplified that was 68% identical and 81% similar to the CCR of *Streptomyces hydroscopicus* (AAR32675, pfam 00107) at the protein level. Two copies of the *ccr* gene were expected given that there would be one gene designated for primary metabolism and the additional gene associated with the natural product cluster. However, since the sequences for CCRs associated with primary and secondary metabolism (at least for the *Streptomyces*) are very similar in nucleotide and amino acid sequence, the amplified CCR was considered a viable probe.

Sequences from cytochrome P450 proteins that are known to perform hydroxylations at the β- positions on amino acids, specifically on aromatic amino acids, were used to design degenerate primers that amplified a 273 bp (90 aa) gene-fragment that aligned most closely with the cytochrome P450 monooxygenase (BAD59492) from the actinomycete *Nocardia farcinica*. The putative protein was not closely related to the

heme-based NADH-dependent β-hydroxylating P450s and since many other copies of P450s were expected to be present on the genome; this fragment was not used as a probe.

The NRPS and CCR PCR products were used in positive control experiments as biotinylated probes against restriction digested *S. tropica* CNB476 and CNB392 gDNA, another salinosporamide A producing strain. *Streptomyces coelicolor* and CNH099 gDNA were used as negative controls. Although the Southern hybridization conditions were not yet optimized, two distinct bands corresponding to ~ 2.8 and ~ 6.4 kb were visible with the NRPS probe against the BamHI and SacI digested *S. tropica* DNA respectively. For the CCR probe there was a high level of non-specific annealing though a distinct band could be seen that corresponded to a DNA fragment size of ~ 3.2 kb in the BamHI digests and two bands at ~ 2.5 and ~ 20 kb in the SacI digests. No annealing was observed in either of the negative controls for the NRPS or the CCR probe (Fig. 3.5).



Lane	
1	1 kb DNA ladder
2	Hind III digested λ DNA ladder
3	*S. tropica* CNB476 Bam HI digested gDNA
4	*S. tropica* CNB476 SacI digested gDNA
5	*S. tropica* CNB392 Bam HI digested gDNA
6	*S. tropica* CNB392 SacI digested gDNA
7	CNH099 BamHI digested gDNA
8	*S. coelicolor* BamHI digested gDNA

Figure 3.5. Agarose gel of Southern hybridization experiment with NRPS probe (left) and CCR probe (right) amplified from *S. tropica* CNB476. Two strains of *S. tropica* were used as positive controls. Strain CNH099 and *S. coelicolor* were used as negative controls. Arrows show significant bands resulting from annealing of probe.

Construction of the S. tropica CNB476 Cosmid Library

Based on the proposed biosynthesis for **1** and published data on the sizes of PKS, NRPS, and tailoring genes (Finking and Marahiel 2004), the gene cluster was predicted

to be ~30 kb in length. A cosmid library, which can contain insert fragments of ~ 40 kb was constructed using the cosmid shuttle vector pOJ446 (Bierman, Logan et al. 1992) which has been engineered to possess the origins of replication for both *E. coli* and *Streptomyces* and can therefore be utilized for library construction in *E. coli* and for later genetic experiments including expression in a heterologous host such as *Streptomyces lividans*. Initial attempts to partially digest the gDNA using the restriction enzyme Sau3AI resulted in very little gDNA cleavage. *Micromonospora* gDNA is characteristically difficult to digest with Sau3AI, which may be due to DNA methylation. MboI digests methylated DNA and is often used successfully against Micromonospora DNA (Tohru Dairi, personal communication) and was used to partially digest the *S. trolpica* gDNA to fragments of ~ 40 kb. The DNA was then cloned into pOJ446, packaged and used to transfect XLIBlue MRF' *E.coli* cells.

Data from complete genome sequences of Actinobacteria including the *Mycobacteria* and *Streptomyces* indicate that genome sizes can range from 4.3 megabases (Mb) for *Mycobacterium bovus* to 9.12 Mb for *Streptomyces avermitilis* MA-4680 (National Center for Biotechnology Information (NCBI)). Based on its relatedness to the *Streptomyces*, the genome size of *S. tropica* was estimated to be approximately 8 Mb in length. Thus, in order to obtain 10X coverage of the genome, a diverse cosmid library consisting of 2000 cosmid clones was required. Once the appropriate number of cosmid clones was achieved, the diversity of the library was verified by randomly selecting 20 cosmids and digesting them with BamHI restriction endonuclease. The resultant

fragments were analyzed by agarose gel electrophoresis, which confirmed that the library was comprised of cosmid clones containing distinct DNA inserts (Fig. 3.6).

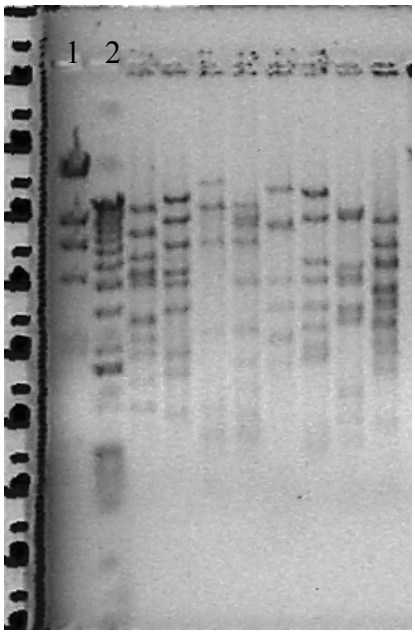


Figure 3.6. Agarose gel of BamHI digests of 8 of the 20 randomly selected cosmids showing library diversity. Lanes 1 and 2 contain λ-Hind and 1 kb ladder respectively.

Screening the S. tropica Cosmid Library by Southern Hybridization

The PCR-amplified products were then used as biotinylated probes to isolate several potential salinosporamide A biosynthetic gene cluster-containing cosmids. Colony lifts from agar plates were first probed with the NRPS fragment, which cross-hybridized to 14 cosmid clones, 2 of which were also positive for the CCR probe. The 14 NRPS-positive cosmid clones were then digested with BamHI restriction endonuclease

and reprobed with the NRPS oligonucleotide, which positively identified three cosmids (Fig. 3.7), which also cross-hybridized with the CCR and PKS probes. Based on these results, the three cosmid clones were selected as potential carriers of the salinosporamide A gene cluster and further analysis. The restriction digests and results of Southern hybridization experiments showed that all three cosmids were nearly identical except the clone designated pLBcos23 contained a slightly larger DNA insert, which led to its selection for further sequence analysis (Fig. 3.8).

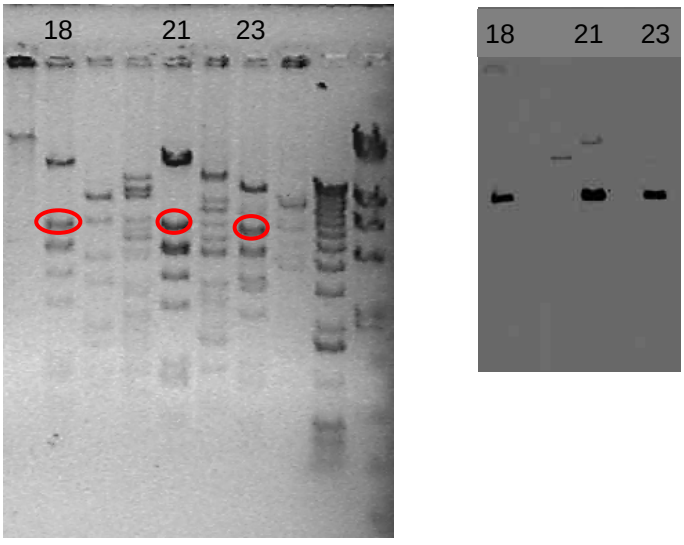
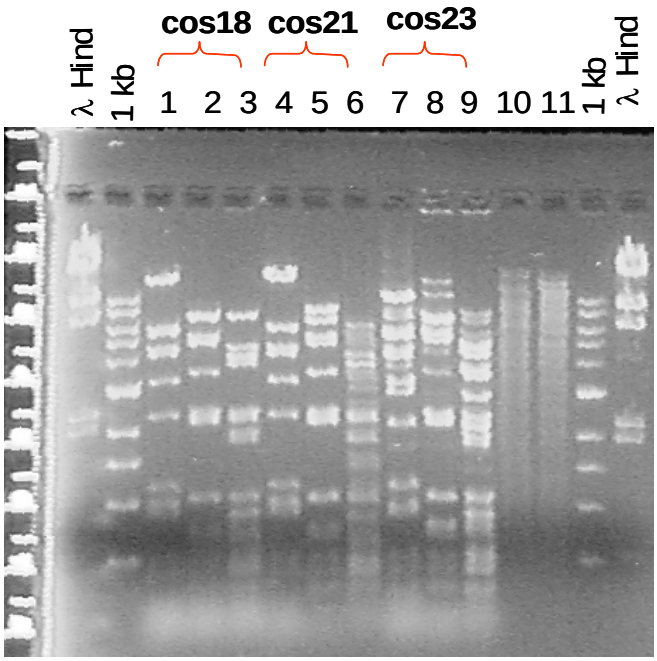


Figure 3.7. Agarose gel of BamHI digested NRPS positive clones from colony lifts that hybridized with the NRPS probe (left). Scanned image of Southern hybridzation results of NRPS probe binding to cosmids 18, 21 and 23 (right). Red circles indicate corresponding fragment from membrane.



Lane #	Sample
1	cos18 Bam HI
2	cos18 Sac I
3	cos18 Bam/Sac
4	cos21 BamH
5	cos21 Sac I
6	cos21 Bam/Sac
7	cos23 Bam HI
8	cos23 Sac I
9	cos23 Bam/Sac
10	gDNA Bam HI
11	gDNA Sac I

Figure 3.8. Agarose gel of the 3 NRPS, PKS and CCR positive cosmids digested with BamHI,, SacI and doubly digested. Based on the number and sizes of fragments, cosmid 23 appeared to have a slightly larger insert.

`

Sequencing the Putative Salinosporamide Gene Cluster

Cosmid pLB23 was subjected to BamHI restriction digests and the resulting DNA fragments were subcloned and end-sequenced with the following results:

700 bp	cytochrome P450
943 bp	PKS ACP domain followed by a KS domain
1260 bp	PKS AT domain followed by a KS domain
1080 bp	NRPS C-domain
1100 bp	NRPS adenylation domain

Additional sequence analysis yielded two-component response regulating genes, an acyl dehydratase related to nodulation proteins of the monoamine oxidase (MAO)

family, a flavoprotein involved in the oxidative deamination of primary, secondary and tertiary amines, and sequence with homology to DNA binding proteins and slime mold topoisomerase were identified. Considering that either by Southern hybridization or by direct sequencing, cosmid 23 had thus far been consistent with the biosynthetic proposal in that it contained PKS, NRPS, CCR, and P450 genes, it was selected for complete sequence analysis.

Initially several attempts were made at shotgun sequencing using different methods to shear the cosmid DNA including sonication, aspiration (nebulization), and restriction endonucleases digests. Sub-cloning and sequencing of the resultant DNA fragments consistently yielded a limited number of inserts even though DNA concentrations were adequate. It appeared that the high overall G-C content which is > 76% for some regions in the cosmid resulted in biased shearing of the DNA which carried over into the cloning and sequencing. Ultimately, sequencing of the entire cosmid was accomplished through primer walking which also proved problematic due to the high G-C content. Difficulties arose while designing primers for sequencing through some of the extremely G-C rich regions (> 80% G-C), so the transposon insertion method was used to obtain sequence through some of these regions.

The completely sequenced cosmid 23 insert spans 4 base pairs and contains 17 open reading frames (ORFs) and 1 partial ORF at the N-terminus. The recently completed genome sequence of the related strain, *S. tropica* CNB440, also contained the putative salinosporamide A gene cluster (Copeland, Lucas et al. 2006). This enabled the cosmid sequence to be extended from the partial *salU* gene. The entire gene cluster

encompasses 26 ORFs designated *salA* - *salZ* spanning 40,370 bp and has an overall G-C content of 66.6%. The sequence from the cosmid and the cluster from the genome are 99% identical at the nucleotide level and ~ 99% identical at the protein level. The results of sequence analysis are graphically presented in Figure 3.9 and summarized in Table 3.1.

Promotor regions were analyzed by using the Shine-Dalgarno sequences associated with streptomycete transcriptional operons that are typically located 5-12 nucleotides upstream of the initiation codon (Shine and Dalgarno 1975). The conserved Shine-Dalgarno sequence for many streptomycete genes is often (a/g)GGAGG although there are many variations and some do not appear to have organized RBSs at all (Strohl 1992). Putative promotor regions for the *sal* genes that appear to be transcribed as single operons are listed by their putative ribosomal binding site (RBS) in Table 3.2 along with the distance of the RBS upstream of the start codon. The start codons for *salB* - *salH* overlap with the stop codons of the preceding ORF suggesting that they are under the control of a single transcriptional operator. The RBS for *salA*, (GAGGG) is located at position -11 from the ATG start codon. The RBS for *salM*, *salN*, *salR* and *salZ* are identical to the conserved (a/g)GGAGG sequence. Characterization of the -35 regions was less clear for all ORFs which is often the case for streptomycete promotor regions (Strohl 1992).

As expected from preliminary sequencing data, the gene cluster contained modular type I PKS and NRPS genes and a bacterial halogenase. A P450 β-hydroxylase, a TE and a CCR were found to flank the PKS and NRPS genes and the remaining ORFS

appeared to be involved in the generation and/or modification of the biosynthetic precursors that are loaded onto the PKS and NRPS. Genes for transcriptional regulation, and resistance along with possible transport and housekeeping genes in addition to genes with an unknown function are also present in the gene cluster. Fragments of transposon genes at the C-terminus define the gene cluster, which is consistent with many natural product biosynthetic gene clusters that have boundaries flanked by transposable elements suggesting the origin of horizontal gene transfer (Hochhut, Lotfi et al. 2001).

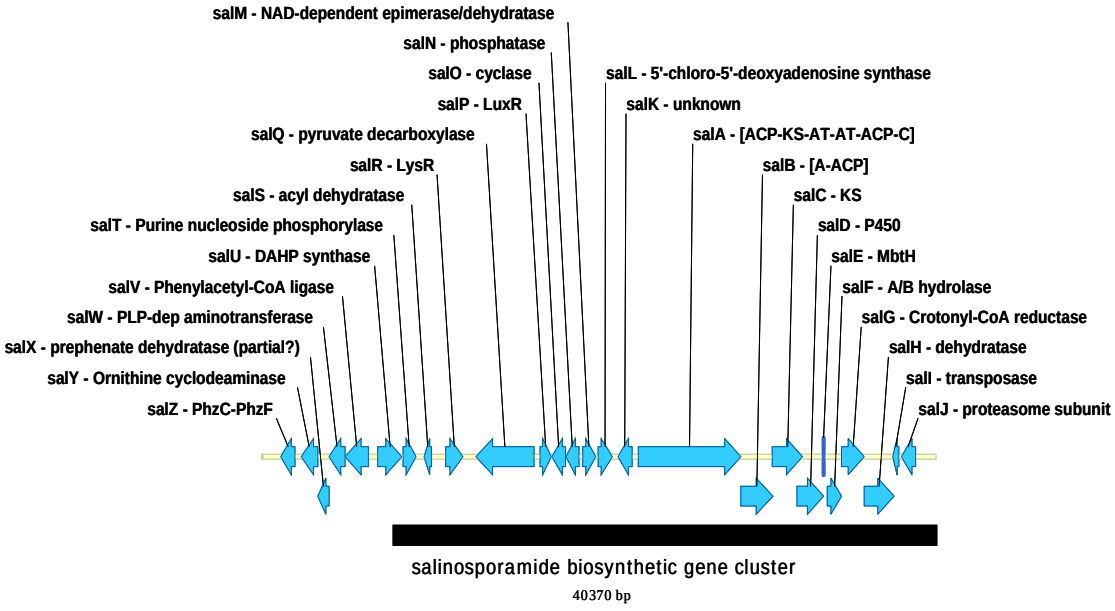


Figure 3.9. Graphical representation of the open reading frames from the putative salinosporamide biosynthetic gene cluster from *S. tropica* CNB440. Predicted functions are based on results of BLAST search. Sequence from *S. tropica* CNB476 cosmid is denoted by black bar.

Name	residues	size kDa	proposed function	sequence similarity (protein, origin)	Identity/similarity (%)	conserved domain	protein accession number or Locus
salA	2053	220.24	PKS/NRPS ACP, KS, AT, AT, ACP C	StiA protein, *Stigmatella aurantica*	ca. 40/60 domain homology	40636, 00833, 40779 40636, 31223	CAD29805
SalB	632	67.02	NRPS A, PCP	Mx1, *Myxococcos xanthus*	45/59	40588, 40636	AAF15891
salC	597	64.81	B-ketoacyl synthase	FabB, 3-oxoacyl-ACP	39/58	30652	ZP_01186252
salD	528	58.44	Cyp450	Quinaldate-3-hydroxylase, *Micromonospora* sp. ML1	51/65	32307	CAJ34365
salE	71	8.24	MbtH	CdaorfX, *Streptomyces coelicolor* A32	41/61	43545	CAB38589
salF	286	30.75	Abhydrolase	Putataive, *Streptomyces coelicolor* A32	51/67	30941	SCO3233
salG	460	56.03	crotonylCoA reductase	CCR, *Streptomyces hygroscopicus*	68/83	30949	AAR32675
salH	572	60.87	dihydroxyacid dehydratase	IlvD, *Mycobacterium avium*	51/65	40995	NP_962565
salI	128	13.77	transposase	Putative transposase, *Streptomyces avermitilis*	68/72	34992	NP_821430
salJ	282	29.4	proteaseome B-subunit	Proteasome B-type subunit, *Rhodococcus erythropolis*	59/75	48439	AAC45736
salK	218	30.2	hyp proten	Conserved hypothetical protein, *Streptomyces coelicolor*	47/60	n/a	CAB40695
salL	283	30.15	5'chloro-5'-deoxyadenosine synthase	5'fluoro-5'-deoxyadenosine synthase, *Streptomyces cattleya*	36/54	41907	2C2WC
salM	255	26.58	NAD-dependent dehydrogenase	Short-chain dehydrogenase/reductase SDR, *Mycobaterium* sp. MCS	44/60	40206	ZP_00765853
salN	212	27.06	phosphatase	Hypothetical protein, *Enterococcus faecalis*	34/52	31878	AAO82843
salO	227	28.88	cyclase/enoyl-coAdehydratase/	Putative cyclase, *Jannaschia* sp.	31/46	32062	XP_509279
salP	225	24.29	LuxR, transcriptional regulation	LuxR: regulatory protein, *Rubrobacter spheroides*	37/58	32379	NP_823547
salQ	1149	123.64	pyruvate decarboxylase	Pyruvate ferredoxin oxidoreductase, *Nocardoides* sp JS614	53/66	33956	ZP_00656820
salR	320	36.31	LysR, transcription regulation	LysR, *Pseudomonas putida*	29/50	30928	NP_744719
salS	119	15.67	Acyl dehydratase	Dehydratase, *Corynebacterium glutamicin*	52/66	48045	NP_59954`
salT	267	28.57	5'-methylthioadenosine phosphorylase	5'-methylthioadenosine phosphorylase, *Streptomyces coelicolor*	60/71	40973	NP_627402
SalU	485	52.88	DAHP synthase	AroG, *Mycobacterium avium*	56/72	41520	NP_960850
salV	471	51.14	Phenylacetyl-CoA ligase	Phenylacetyl-CoA ligase, *Photorhabdus luminescens*	39/59	31730	NP_930273
salW	320	34.16	PLP-dependent aminotransferase	Branched-chain amino acid transferase, *Rubrobacter xylanophilus* DSM 9941	44/61	29568	ZP_00599544
salX	217	22.3	prephenate dehydratase (fragment)	Putative prephenate dehydratase, BacA, *Bacillis subtilis*	37/50	40878	AAM90568
salY	329	34.67	Ornithine cyclodeaminase	Ornithine cyclodeaminase, *Silicibacter pomeroyi*	40/56	32552	AAV97046
salZ	278	29.45	PhzC/PhzF	PhzC/PhzFprotein. *Brevibacterium linens*	57/72	30733	ZP_00378489

Table 3.1. Annotation of the *sal* gene cluster from *S. tropica* CNB476 and CNB440.

gene	putative RBS	position relative to start codon
salA	TGAGGG	-12
salJ	TGGAAC	-10
salK	AGGAGC	-11
salL	GGAGG	-13
salM	AGGAGG	-12
salN	GGGAGG	-15
salP	TCGAG	-12
salQ	AGGAG	-15
salR	GGGAGG	-13
salU	TGGCCG	-12
salV	TGGGGGC	-12
salY	GGGAGC	-12
salZ	GGGAGG	-11

Table 3.2. Putative ribosomal binding sites for *sal* genes from *S. tropica*.

The Salinosporamide Megasynthetase: PKS and NRPS Core Genes

Three ORFs encode a large multifunctional type I PKS with a C-terminal NRPS C-domain (*salA*), an NRPS (*salB*) and a stand-alone Type I KS domain (*salC*). The *salA* gene encodes a bimodular PKS (6159 bp) with the atypical domain arrangement: ACP-AT-AT-KS-ACP-C suggesting that its 220,240 Da product performs the function of activation and loading of the starter unit and extension and condensation of the growing polyketide chain with the putative NRPS-bound amino acid. Though the order of the PKS domains deviates from the predicted co-linear arrangement (AT-ACP-KS-AT-ACP), its composition is consistent with the PKS predicted for the biosynthesis of **1**. A similar domain organization has been observed in three hybrid PKS/NRPS biosynthetic gene clusters from Myxobacteria, namely the loading modules from the type I PKS

multienzymes SorA, MxaF, and MtaB, which catalyze chain initiation and the first step of elongation in the biosynthesis of soraphen (*Sorangium cellulosum)*, myxalamid (*Stigmatella aurantiaca*), and myxothiazol (*Myxobacterium fulvus*) respectively. (Schupp, Toupet et al. 1995; Silakowski, Nordsiek et al. 2001).

A unique feature of the *salA* PKS gene is the presence of the NRPS C-domain located at the C-terminus. Amino acid sequence analysis confirmed that the conserved core motifs associated with C-domains were present including the invariant active site histidine His147, although the C3 motif is degenerate compared to the conserved C3 motif (HHVAADA for SalA vs HHxxxDG) (Konz and Marahiel 1999). The additional important structural residues (R62 and D151) that are involved in amino acid condensation were present confirming that this was a *bona fide* NRPS C-domain (Fig. 3.10) (Bergendahl, Linne et al. 2002).

APVAHGQAAIWFMQQLSLDGAAYNLMFGARVPDKIDESALRRAAAAVVERHPAL**R**TVFVEA

GGHPYQVIKADPGYEFDTVDGTGLDDAALTDLLAEHGHRPFDLDQGPLLRLVLVNRGAADN

* C3

CLLLVI**H**VAA**D**AASVDIVVRDLREFYGEAQRGTLLTQAPETPYTEFVEWEREWLGSPAAE

AALHWWSQQLAQPPAHLDLPRLSAPDAAADKRPAKERDGVVSYAGEDASFRWDAADTRLLK

DFAVRAGVSVSTLVMAGFVATLGRATGARDIVLGTAVAQRSDAGRESAVGYYLNTIPVRAR

PRPDIT

Figure 3.10. Partial amino acid sequence of the condensation domain located at the C-terminus of SalA. Residues in red denote conserved core motifs. The residues involved in amino-acid condensation are in bold. An asterisk denotes the conserved catalytic histidine residue.

A search using the PKS manipulation site (Ansari, Yadav et al. 2004) predicted that the loading AT_1 is most closely related to AT domains specific for malonyl-CoA and that the extender AT_2 is similar to AT-domains that recognize methylmalonyl-CoA. However, Clustal analysis and results of a neighbor-joining tree (Fig. 3.11) showed that these predictions were not concrete.

Just downstream of *salA,* the *salB* gene encodes a 632 amino acid NPRS protein with two distinct domains. The N-terminal domain of *salB* is highly homologous to NRPS A-domains and includes the conserved core motifs common to A-domains. The C-terminus of *salB* encodes a PCP, which includes the conserved 4'PP binding core motif LGGxS. On the basis of the nonribosomal consensus code (Stachelhaus, Mootz et al. 1999; Challis, Ravel et al. 2000) the substrate specificity of the *salB* NRPS A-domain did not predict any known substrates though it was most closely related to phenylalanine and tyrosine-like activating A-domains. A homology model of the A-domain with the phenylanine activating GrsA NRPS (Fig. 3.12) revealed that the binding pocket of the cos23 A-domain was slightly larger than the active sites of A-domains known to activate aromatic amino acids. Given that the space requirement for a cyclohexene ring occupying the active site of the NRPS would be greater than that required for the phenyl or hydroxy-phenyl group of phenylalanine or tyrosine respectively, and its context in the gene cluster adjacent to the bimodular PKS, the proposed function of SalB is activation of the CHHA amino acid moiety.

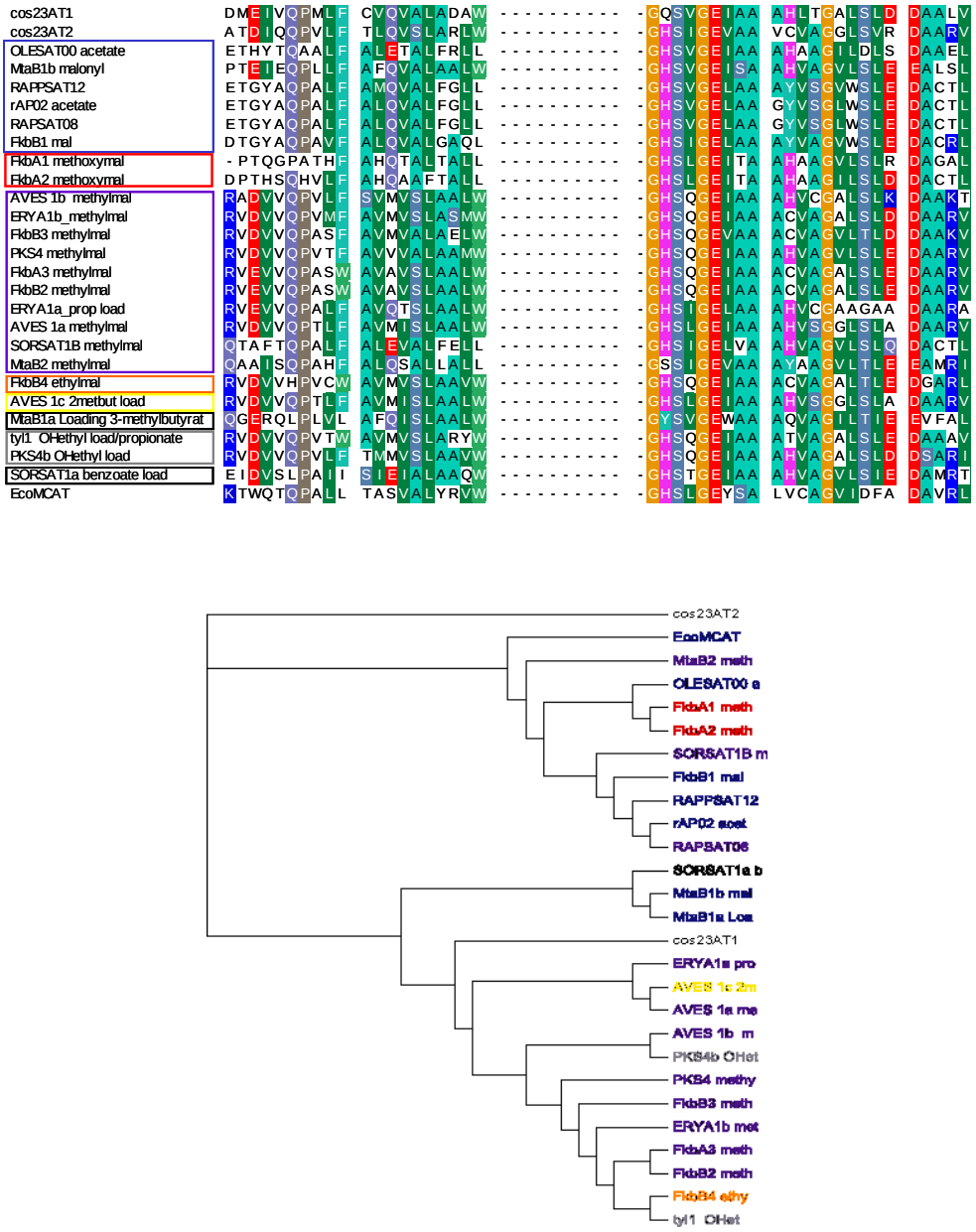
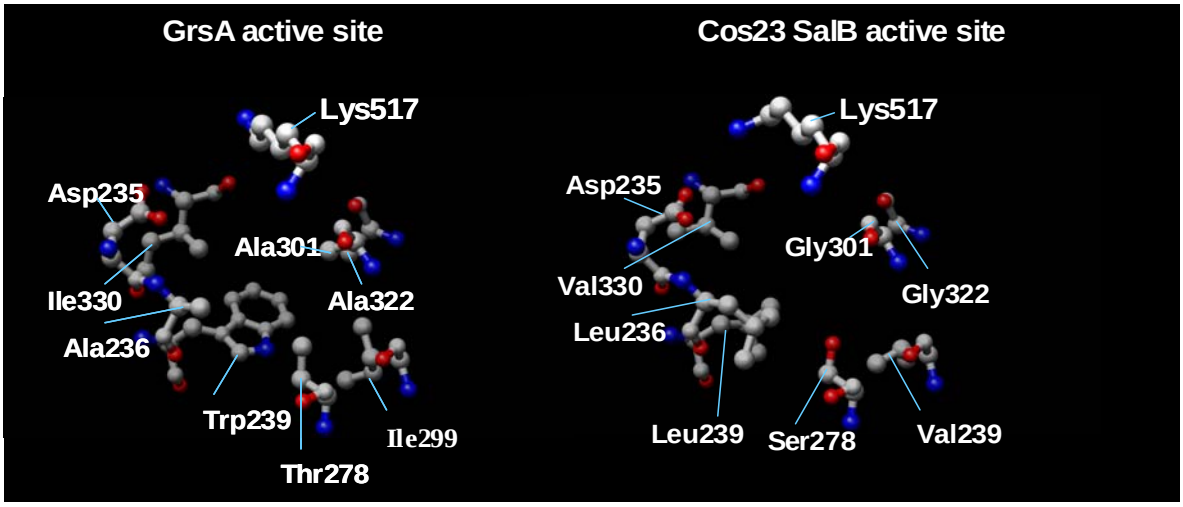


Figure 3.11. ClustalX alignment of the SalA AT domain conserved regions (cos23AT1 and cos23AT2) with ATs of known loading specificity (Top). Neighbor-joining tree of the same sequences (bottom), CoA starter units are indicated as follows: Blue: malonyl; Purple: methylmalonyl; Red: methoxymalonyl; Orange: ethylmalonyl; Yellow: 2-methylbutyrate; Gray: hydroxyethylbutyrate; Black: benzoate/other.



Protein	Module	activated AA	Residue position according to GrsA numbering							
			235	236	239	278	299	301	322	330
GrsA	M1	D/L-phe	D	A	W	T	I	A	A	I
Cos23		?	D	L	L	S	V	G	G	V

Figure 3.12. The GrsA active site (left) and the modeled *S. tropica* CNB476 A-domain active site (right). The CNB476 A-domain is proposed to activate the nonproteinogenic amino acid (2R,3S,3R)-3-cyclohexene-3-hydroxyalanine. Table comparing the active site residues of GrsA and the CNB476 A-domain.

Downstream of *salB* is the 1791 bp *salC* gene that encodes a type I KS domain. The occurance of a stand-alone KS is unique for a type I PKS or hybrid PKS/NRPS system. Database searches and protein alignments confirm that *salC* is most similar to the KS domains of modular type I PKSs (CDD:30652). Phylogentic analysis of Type I polyketide KS domain protein sequences has suggested that they cluster phylogenetically into two functional groups. One group representing KS domains which use acyl-CoA's as their starter or extender unit and the distinct group of type I KS domains comprises

sequences isolated from hybrid or mixed PKS/NRPS systems which condense amino acids onto a polyketide extender unit (Du, Sanchez et al. 2001; Moffitt and Neilan 2003). The starter units for type I PKS chain elongation can be derived from simple aliphatic acids such as acetate or propionate (Wiesmann, Cortes et al. 1995) and more commonly, the loading modules of PKSs consist of a KSQ, an AT and an ACP domain, where the AT recruits a dicarboxylic acid (malonate) onto the ACP and the KSQ domain accomplishes its *in situ* decarboxylation (Bisang, Long et al. 1999). KSQ refers to the active site residue of the KS domain and amino acid sequence alignments show that in KS domains known to possess decarboxylative activity, the conserved active site cysteine (C) has been replaced by a glutamine (Q). Furthermore the activation of more complex alicyclic and aromatic starter units are selected by CoA ligase-like domains which activate the carboxylic acid to the corresponding acyl-adenylate utilizing ATP before covalent attachment to the AT domain via the ACP domain.

An alignment of the KS domains from *salA* and *salC* with KS domains of known functions revealed that the KS catalytic active site residues from both proteins are cysteines. Sequence analysis of the active sites show that the *salA* KS domain is most similar to the KSs from the loading modules of SorA, MxaF and MtaB that possess a similar ACP-KS-AT-AT-ACP domain arrangement (Fig. 3.13). The function of the stand-alone *salC* is unclear though it may play a role in modification of the amino acid by interacting with the β-hydroxylase encoded by the adjacent *salD* gene or in the cyclization reaction. When the PCR-amplified KS fragments were compared to the SalA and SalC sequences, the SalA KS and the PCR-amplified ks3 fragment were identical and

the ks1/2 PCR product was not similar to either of the KSs present in the gene cluster. All of the PCR amplified products contained the conserved active site cysteine residue.

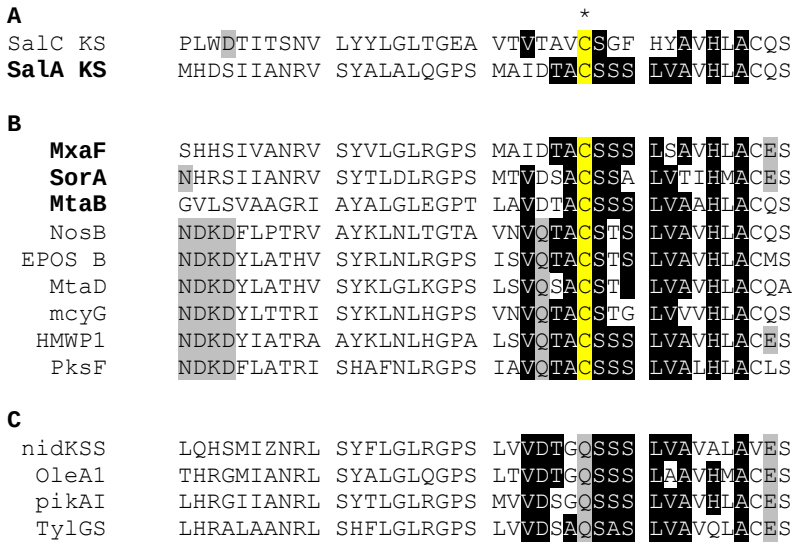


Figure 3.13. Alignment of the active sites of the ketosynthase (KS) domains from cosmid 23 containing the putative salinosporamide A gene cluster and type I KS domains with distinct functions. Conserved residues are highlighted in *black*. Residues conserved within each functional group are highlighted in *gray*. **A**. The stand-alone KS domain (SalC) from cosmid 23 and the PKS KS domain from SalA. **B**. Representatives of KS domains specific for amino acid starter units in mixed or hybrid systems. **C**. Representatives of KSQ domains with decarboxylative activity.

Genes for the Biosynthesis of Chloroethylmalonyl-CoA

The chlorobutyryl-moiety of salinosporamide A is unusual and feeding studies with labeled precursors indicated that this 4-carbon unit arose from a tetrose-like molecule. The ^{13}C-enrichment was identical to that observed on the erythrose-4-phosphate (E4P) portion on the cyclohexene ring of the amino acid. This led to speculation that a novel pathway was involved in the biosynthesis of the 4-chlorobutyrate moietiy that is incorporated into **1**. Several genes are present in the *sal* cluster that

support a pathway toward chlorobutyryl-CoA via a novel pathway involving S-adenosylmethionine (SAM). The *salL* gene (849 bp), located upstream of *salA* encodes a protein that is 36% identical and 54% similar to a bacterial fluorinase (FlA) from *Streptomyces cattleya* (O'Hagan, Schaffrath et al. 2002; Deng, Cobb et al. 2006). The fluorinse pathway of *S. cattleya* leads to the biosynthesis of fluoroacetate and 4-fluorothreonine. The FlA protein, 5'-fluorodeoxyadenosine synthase, mediates the reaction between inorganic fluoride and S-adenosyl-L-methionine (SAM) to generate 5'-fluoro-5'-deoxyadenosine (5'-FDA) (Fig 3.14). The next intermediate in the fluorinase pathway, 5-fluoro-5-deoxy-*D*-ribose-1-phosphate (5-FDRP), is generated by a nucleoside phosphorylase designated FlB. However, the steps from 5-FDRP to fluoroacetaldehyde and 4-fluorothreonine remain to be elucidated (Schaffrath, Cobb et al. 2002; Huang, Haydock et al. 2006). This pathway was the first reported example whereby a carbon-fluoride bond is formed in bacteria.

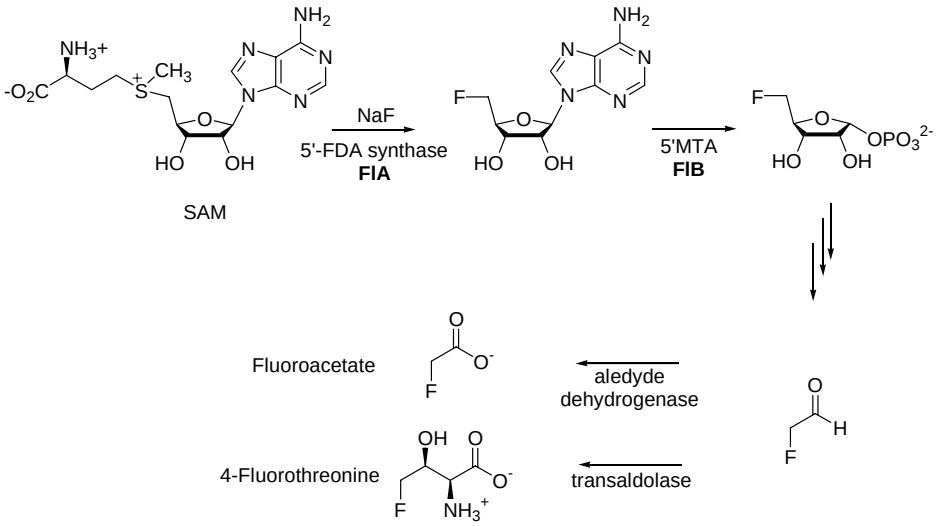


Figure 3.14. Proposed Biosynthesis of Fluoroacetate and 4-Fluorothreonine in *S. cattleya* catalyzed by 5'-fluorodeoxyadenosine synthase (FlA) and 5'-methylthioadenosine phosphorylase (FlB). Scheme adapted from Huang et al. 2006.

A sequence alignment of SalL with F1A indicated that substitutions occur in some of the active site residues in SalL (Fig. 3.15). Table 3.1 lists the residues reported for the F1A active site residues that participate in hydrogen bonding with 5'FDA compared to SalL (Dong, Deng et al. 2003; Dong, Huang et al. 2004). The consequence of such substitutions in SalL would be a slightly larger pocket in the active site that may accommodate the larger chloride ion.

```
salL              1   MFLSRFRRRSLRFVVCTTTNPEVGFRPIKRRSLTMQHNLIAFLSDVGSADEAHALCKGVM
S. Cattleya       1   ------------MAANSTR-----RPI-----------IAFMSDLGTTDDSVAQCKGLM
Clustal Consensus 1               :...:*.     ***           ***:**:*::*:: * ***:*

salL              61  YGVAPAATIVDITHDVAPFDVREGALFLADVPHSFPAHTVICAYVYPETGTATHTIAVRN
S. Cattleya       32  YSICPDVTVVDVCHSMTPWDVEEGARYIVDLPRFFPEGTVFATTTYPATGTTTRSVAVRI
Clustal Consensus 26  *.:.* .*:**: *.::*:**.*** ::.*:*: **  **:.: .** ***:*:::***

salL              121 EKGQ----------------------LLVGPNNGLLSFALDASPAVECHEVLSPDVMNQP
S. Cattleya       92  KQAAKGGARGQWAGSGAGFERAEGSYIYIAPNNGLLTTVLEEHGYLEAYEVTSPKVIPEQ
Clustal Consensus 69  ::.                       : :.******: .*:    :*.:** **.*: :

salL              159 VTPTWYGKDIVAACAAHLAAGTDLAAVGPRIDPKQIVRLPYASASEVEGGIRGEVVRIDR
S. Cattleya       152 PEPTFYSREMVAIPSAHLAAGFPLSEVGRPLEDHEIVRFNRPAVEQDGEALVGVVSAIDH
Clustal Consensus 91    **:*.:::**  :****** *: **  :: ::***:  .:..:   .: * *  **:

salL              219 AFGNVWTNIP-THLIGSMLQDGERLEVKIEALSDTVLELPFCKTFGEVDE-GQPLLYLNS
S. Cattleya       212 PFGNVWTNIHRTDLEKAGIGYGARLRLTLDGVLP--FEAPLTPTFADAGEIGNIAIYLNS
Clustal Consensus 127 .******** *.*  : :  * **.:.::.:    :* *:  **.:..* *:  :****

salL              277 RGRLALGLNQSNFIEKWPVVPGDSITVSPRVPDSNLGPVLG
S. Cattleya       270 RGYLSIARNAASLAYPYHLKEGMSARVEAR-----------
Clustal Consensus 162 ** *::. * ::   : :  * *  *..*
```

Figure 3.15. Amino acid sequence alignment of SalL with the fluorinase, FlA from *S. catleya*. Active site residues involved in hydrogen bonding are highlighted in gray. Alignment performed with clustalX.

Amino acid Residue (#, aa)	
F1A	**SalL**
Asp 21	Asp 50
Ser 23	Ala 52
Phe 156	Trp 102
Ser 158	Gly 104
Asp 16	Asp 45
Tyr 77	Tyr 106
Trp 50	Phe 79
Ala 279	Gln 286
Phe 254	Phe 262
Asn 215	Asn 222
Asp 210	Asp 217
Phe 213	Phe 220
Ser 269	Ser 276
Arg 270	Arg 277

Table 3.3. Comparison of the active site residues from the chlorinase SalL vs. the *S. cattleya* fluorinase F1A. Residues that differ are highlighted in gray.

The next step would be catalyzed by the *salT* (801 bp) gene product that shows high homology to 5'-methylthioadenosine phosphorylases (5'-MTAP) and is 71% similar to the homolog from *Streptomyces coelicolor* (SCO3188). 5'-MTAP catalyzes the reversible phosphorolysis of 5'-deoxy-5'–methylthioadenosine (MTA) to adenine and 5-methylthio-D-ribose-1-phosphate (Deng, Cobb et al. 2006). The reactions catalyzed by SalL and SalT, are analogous to what has been observed in the fluorination pathway (Fig. 3.14) (O'Hagan, Schaffrath et al. 2002; Deng, Cobb et al. 2006). In the chlorinase mechanism, the thiomethyl group of SAM is replaced by a chlorine atom to form 5'-chloro-5'-deoxyadenosine (5'ClDA) via a nucleophilic substitution process. In the analogous reaction, SalT would yield cyclic 5'-chloro-5'-dexoyribose-1-phosphate (Fig. 3.16).

The remaining genes proposed to have a role in the biosynthesis of the chloroethyl moiety are based on their protein sequence homologies to enzymes present in the methionine salvage pathway (Sekowska, Denervaud et al. 2004) and branched-chain amino acid metabolic pathways. The ORF designated *salN* (636 bp) is similar to uncharacterized proteins conserved in bacteria that may be related to metal chelating phosphoesterases or phosphatases (34% / 48%, identity/similarity). The most conserved regions are centered on the metal chelating residues (CDD:31878). This domain family includes the serine/threonine protein phosphatases that are one mechanism by which signaling processes can be terminated by removal of phosphate groups. The methionine salvage pathway has two genes, *mtnC and mtnX* encoding phosphatases involved in reactions that ultimately yield the 3- and 4-carbon units that reenter the pathway

(Sekowska, Denervaud et al. 2004). In the fluoroaceate biosynthetic gene cluster a small 216 aa protein, FlD, shows similarity to phosphatases and haloacid-dehalogenases (Huang, Haydock et al. 2006). The authors propose that the protein may act as a resistance mechanism by defluorinating fluoroacetate and/or 4-fluorothreonine or that it plays a role in further metabolism of 5'FDRP in the pathway leading to fluoroacetate. SalN is proposed to dephosphorylate 5-chlororibose-1-P to yield 5-chloro-5-deoxyribose. The next step would be catalyzed by the product of *salM* (765 bp), which is immediately upstream of *salL*. Its deduced amino acid sequence shows homology to short chain NAD dependent dehydrogenases (CDD:40206) and other closely related members of this family of proteins are the FabG-like 3-oxoacyl-acyl-carrier-protein reductase (COG1028) and FabI enoyl-acyl carrier protein oxidoreductases (COG0623). A homolog of SalM from the liver of *Sus scrofa* (EC: 1.1.1.115) has been shown to catalyze the NAD+ or NADP+ dependent reaction of D-ribose to D-ribonate and NADH H+ or NADPH+ (Schiwara, Domschk et al. 1968). SalM would catalyze the dehydrogenation of 5-chloro-5-deoxyribose to 5-chloro-D-ribonate.

The s*alH* gene (572 bp) encodes a protein that is similar to the IlvD single domain family of dihydroxyacid dehydratases (53% / 65%) (CDD:40995). Dihydroxyacid dehydratase catalyzes third step in the common pathway leading to biosynthesis of branched-chain amino acids, specifically, the dehydration of 2,3-dihydroxy-3-methylbutyryCoA to 3-methyl-2-oxobutyrylCoA and H_2O in valine and isoluecine biosynthesis (Velasco, Cansado et al. 1993). The 6-phosphogluconate dehydratases are also members of this family and all members of the family catalyze a similar reaction

whereby a 2,3-dihydroxy moiety is converted to a 2-dehydro-3-oxy moiety. A possible role for SalH is the dehydration of 5-chloro-D-ribonate to 2-oxo-3-deoxy-5-chlororibonate.

The large open reading frame (3447 bp) designated *salQ,* has a deduced amino acid sequence highly homologous to pyruvate oxidoreductase (POR) and indolepyruvate ferredoxin oxidoreductase (IOR). SalQ is most similar to a pyruvate ferredoxin oxidoreductase from *Nocardoides* sp JS614 (53% / 66%) (CDD: 33956). POR catalyses the final step in the fermentation of carbohydrates in anaerobic microorganisms (Kletzin and Adams 1996). This reaction involves the oxidative decarboxylation of pyruvate with the participation of thiamine followed by the transfer of an acetyl moiety to coenzyme A for the synthesis of acetyl-CoA. Bacteria utilize IOR in energy production and for the conversion of pyruvate to acetate and peptide fermentation (Mai and Adams 1994). POR and IOR are thiamine dependent oxidoreductases that catalyzes the ferredoxin-dependent oxidative decarboxylation of arylpyruvates such as indolepyruvate or phenylpyruvate, which are generated by the transamination of aromatic amino acids, to the corresponding aryl acetyl-CoA. The proposed function of SalQ is the oxidative decarboxylation of 2-oxo-3-deoxy-5-chlororibonate and transfer of the acetyl moiety to CoA to give 3-hydroxy-4-chlorobutyryl-CoA.

The next step may involve *salS* (357 bp) that encodes a putative acyl dehydratase that is similar to MaoC-like dehydratases (CDD:48045) that share similarity with a wide variety of enzymes including estradiol-17-β-dehydrogenase-4, peroxisomal hydratase-dehydrogenase-epimerase and FAS β-subunit. When the context of *salS* homologs were

examined it showed that these genes are located near genes involved in fatty acid metabolism and/or branched chain amino acid metabolism (e.g. *Mycobacterium avium subsp. paratuberculosis* K-1, MAP3538). In fatty acid metabolism β-hydroxyacyl-ACP dehydratase (FAS β-subunit) primarily catalyzes the dehydration of β-hydroxyacyl-ACP to trans-2-acyl-ACP, the third step in the elongation phase of the bacterial type II, fatty-acid biosynthesis (Chirala and Wakil 2004). The enzymes are also bifunctional and can catalyze the isomerization reaction of trans-2-acyl-ACP to cis-3-acyl-ACP. The analogous reaction performed by SalS would be the dehydration of 3-hydroxy-4-chlorobutyryl-CoA to 4-chlorocrotonyl-CoA.

The deduced amino acid sequence for the *salG* gene (1380 bp) is highly homologous to crotonyl-CoA reductases (CCR) from streptomycetes and most closely related to the CCR from *Streptomyces hydroscopicus* (68% / 85%) (Wu, Chung et al. 2000). Members of this family of enzymes are Zn-dependent alcohol dehydrogenases/oxidoreductases (CDD:30949) and include the NADPH quinone oxidoreductases that carry out 2-electron reductions on quinones in the cell. The reduction reaction catalyzed by SalG would yield a fully reduced 4-chlorobutyryl-CoA molecule. A carboxylase from outside the cluster or from primary metabolism would yield 4-chloroethylmalonyl-CoA that would then be loaded onto the PKS and extended with acetate to yield the diketide.

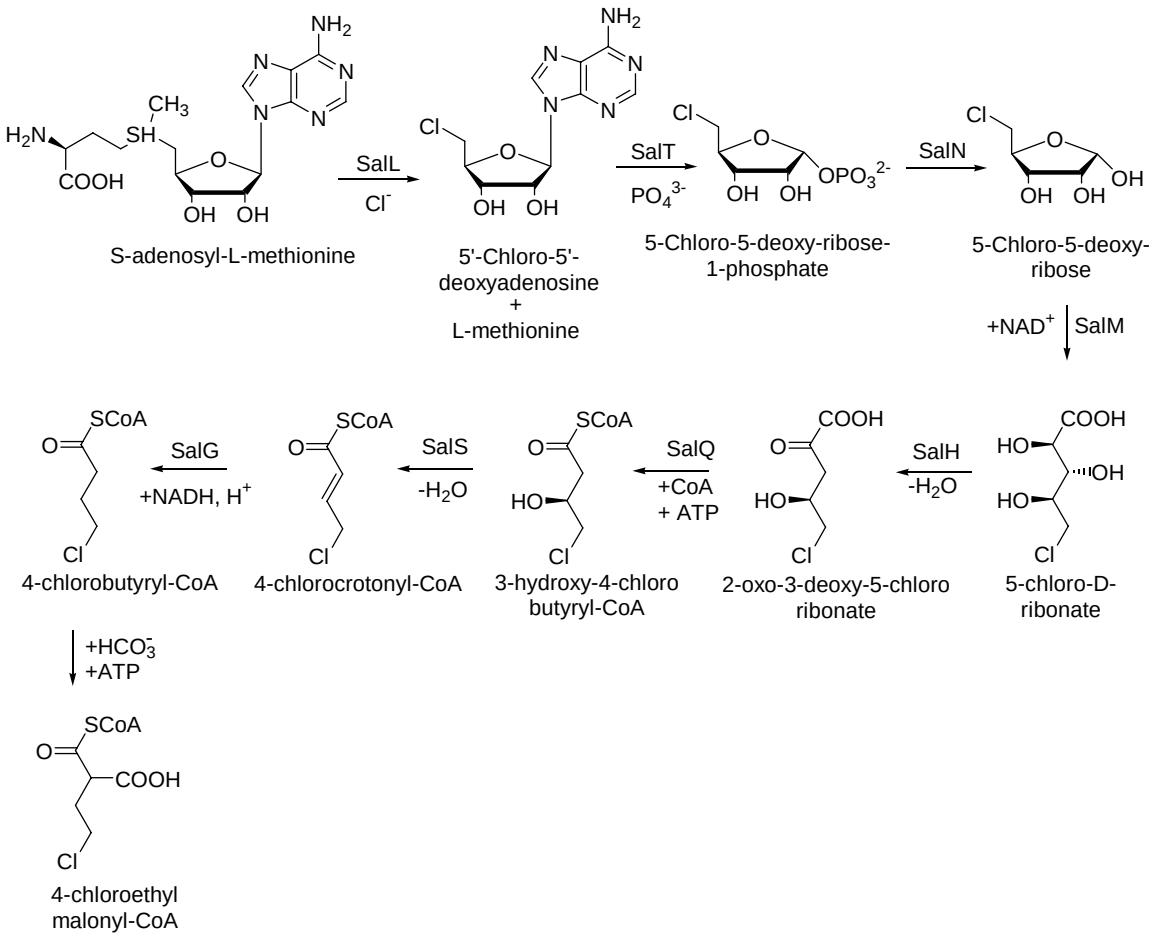


Figure 3.16 Proposed biosynthesis of 4-chloroethylmalonyl-CoA for salinosporamide A biosynthesis based on the *sal* genes and information from feeding studies.

Genes Probably Involved in Biosynthesis of the Amino Acid Moiety

Genes that are probably involved in the biosynthesis of the amino acid have also been found in the cluster. The *salU* gene (478 bp) encodes a 3-deoxy-7-phosphoheptulonate (DAHP) synthase that is similar to AroG (56 % / 72%) from *Mycobacterium avium* (Li, Bannantine et al. 2005). DAHP synthases are aldolases (CDD:41520) that catalyze the committed step in shikimate synthesis whereby phosphoenolpyruvate (PEP) and erythrose-4-phosphate (E4P) are condensed to form DAHP (Bentley 1990). Subsequent reactions yield shikimate, chorismate and eventually the aromatic amino acids. The putative DAHP synthase would provide a dedicated pool of DAHP for biosynthesis of the putative CHHA amino acid that would then proceed via the primary metabolic pathway to prephenate. An as yet unknown series of reductions would yield the α-keto acid precursor of the cyclohexenyl-alanine.

The next gene proposed to play a role in biosynthesis of the amino acid is *salX* (369 bp) that encodes a 122 amino acid protein. In a BLASTP search, the SalX protein aligns with 3 putative prephenate dehydratases (PDT) (CDD:40878). As described previously, PDTs are involved in phenylalanine biosynthesis and catalyze the decarboxylation of prephenate to phenylpyruvate (He, Magarvey et al. 2001). Full-length PDTs are 200+ amino acids in length whereas *salX* is truncated at the N-terminus eliminating the first 90 amino acids. A pairwise alignment with one of the three putative PDTs, ywfB (CAB15801.1) from *Bacillis subtilis* with the PDT from *Streptomyces lividans*, indicates that the two proteins are not closely related. In fact, they are only 4% identical and 13% similar (BLOSUM 62 similarity matrix). Given that there is no

experimental evidence for the function of the putative PDTs that SalX aligns with in the BLAST search, and that they do not possess any of the conserved motifs from characterized PDTs it is easy to speculate that this gene is not functional or encodes only a partial protein. A functional PDT would support the proposed biosynthetic scheme and thus, the proposal that *salX* encodes a functional protein is at present, loosely based.

The 320 amino acid gene product of *salW* (960 bp) is a PLP-dependent amino transferase (CDD:29568) that is most similar to a protein from *Rubrobacter xylanophilus* (44% / 61%). Members of this family of enzymes catalyze the reversible reaction whereby L-branched chain amino acids are converted to the corresponding α-ketoacid (Sugio, Petsko et al. 1995). SalW would perform the transamination of the α-keto acid to the amino acid.

SalD (508 aa) aligns well with cytochrome P450 oxidases (CDD:32307) including designated amino acid β-hydroxylases and is most similar to quinaldate 3-hydroxylase from a *Micromonospora* species (51% / 65%) (Lombo, Velasco et al. 2005). When compared to other P450s, SalD has an extra ~100 residues at the N-terminus which does not share homology with any known linker regions (Fig. 3.17.) The cytochrome P450-mediated oxygenations catalyzed by this class of enzymes occur via a heme-based NADPH dependent mechanism utilizing molecular oxygen and they are common tailoring enzymes associated with secondary metabolite biosynthesis and the related gene clusters (Walsh, Chen et al. 2001). In SalD the conserved cysteine residue presumably serving as the heme attachment site is present at position 402 and the putative threonine

oxygen-binding site is position 295. SalD is proposed to perform the β-hydroxylation of the amino acid.

Just downstream, the *salE* gene (213 bp) encodes a 71 amino acid protein that is 41% identical to CdaorfX and related to other MtbH proteins with conserved domains of unknown function (CDD:43545). These domains contain 3 fully conserved tryptophan residues and many of the members of this family are found in known antibiotic synthesis gene clusters and associated with the NRPS or mixed PKS/NRPS biosynthetic genes though their function is unknown (Stegmann, Rausch et al. 2006). Based on the biosynthesis of aromatic amino acids via the shikimate pathway, the results of feeding experiments and the predicted functions of the *sal* genes, a proposed biosynthetic scheme for the amino acid moiety is shown in Fig. 3.18.

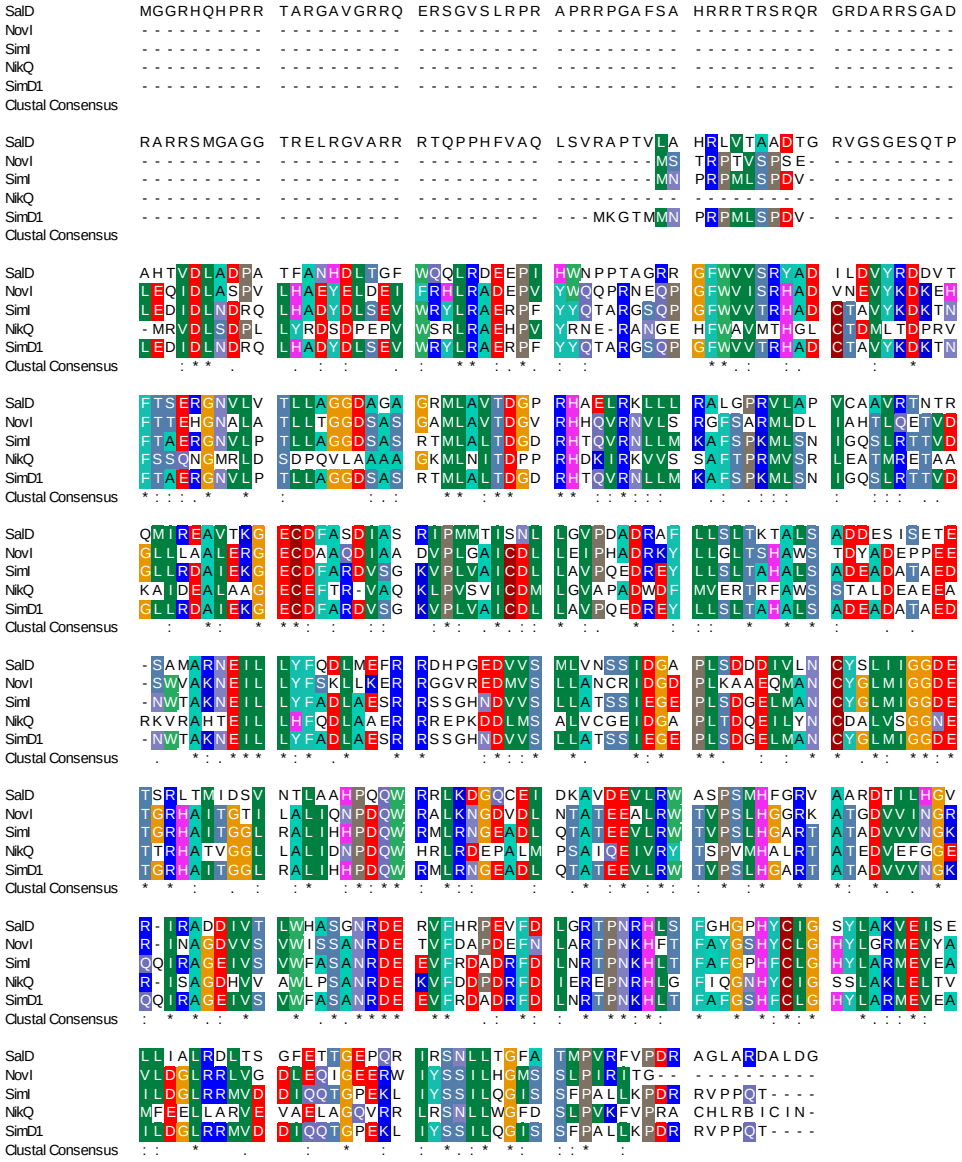


Figure 3.17 Alignment of the SalD cyP450 with P450's known to catalyze the hydroxylation of amino acids at the β-position. SalD has an extra ~100 amino acids at the N-terminus.

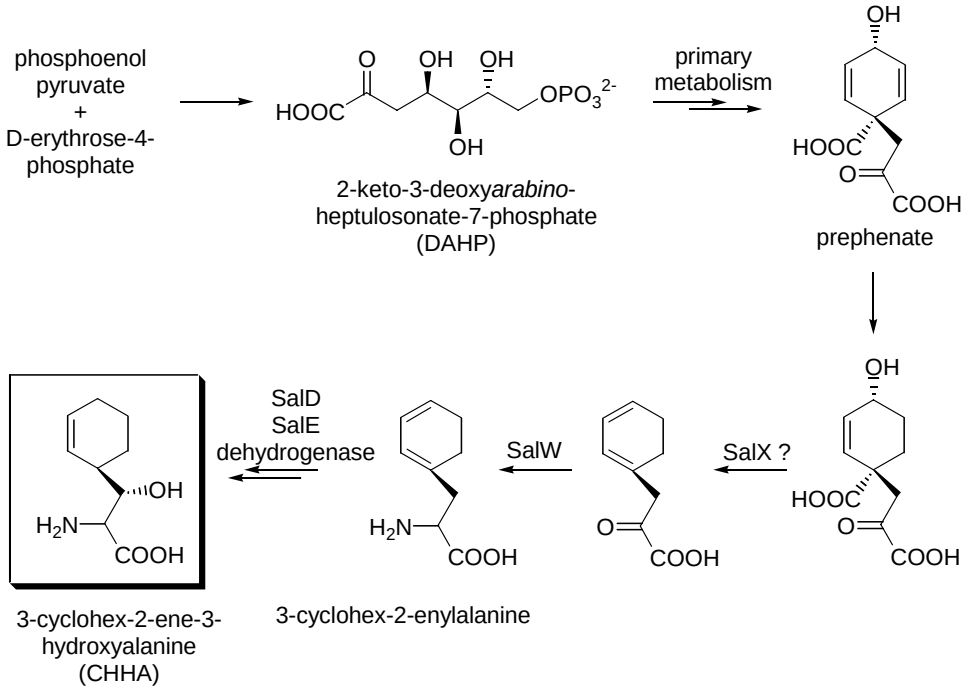


Figure 3.18. Proposed biosynthesis of the amino acid moiety in the salinosporamides based on putative functions of the *sal* genes.

Genes Probably Involved in Cyclization and Offloading Reactions

One of the fascinating features of salinosporamide biosynthesis concerns formation of the γ-lactam-β-lactone bicycle. A putative cyclase and a thioesterase, SalO and SalF respectively may be involved in these steps. The *salO* gene (681 bp) has a putative gene product that shows homology to a large number of poorly characterized putative metal dependent hydrolases that are putative cyclases or enoyl-CoA dehydratases involved in antibiotic synthesis (CDD:32062) (Lomovskaya, Doi-Katayama

et al. 1998). The conserved motif HxGTHxDxPxH that is considered part of the active site on these proteins is also present on SalO. SalO aligns most closely with a putative cyclase gene (Jan1317) from *Jannaschia* sp. CCS1 and when the context of the homolog was examined the genes surrounding Jan1317 were similar to those surrounding *salO*. In fact, homologs of *salM*, *salP*, *salQ*, *salS*, and *salT* were found nearby. Similarly, in *Nocardia farcinica* (YP_120682) and *Chlorobium tepidum* TLS (CC2269), genes that appear to be involved in branched chain amino acid metabolism such as succinate dehydrogenase, oxidoreductases, thioesterases and genes for transcriptional regulation are present near *salO* homologs. Based on these findings, a possible role for SalO is that it may be involved in biosynthesis of 4-chlorobutyrylCoA. Alternatively, a more attractive role for SalO, would pertain to the cyclization reaction. Once the linear salinosporamide precursor is assembled on the NRPS, SalO might facilitate the cyclization and offloading reaction by deprotonation of the α-carbon on the amino acid (C-5 in **1**). Nucleophilic attack by C-5 on the C-3 carbonyl would form the lactam ring. Subsequent deprotonation of the hydroxyl group and nucleophilic attack by the C-5 oxygen on the amino acid carbonyl (C-15) possibly by a separate enzyme, followed by hydrolytic release from the enzyme thioester linkage could yield the bicyclic core. The latter reaction could be catalyzed by SalF.

The *salF* gene (861bp) encodes a putative protein with similarity to a/b hydrolases that are a large family of enzymes with conserved domains also found in thioesterases (TE) and haloperoxidases (CDD:30941). Amino acid sequence comparison shows that SalF is most similar to TEs associated with modular type I PKS or mixed PKS/NRPSs

from streptomycetes. *SalF* is most closely related to the gene SCO3233 from *Streptomyces coelicolo*r A32 (51% / 67%), a putative thioesterase who's predicted function is based on sequence comparisons and its location C-terminal to an NRPS biosynthetic gene cluster (Redenbach, Kieser et al. 1996). The position of *salF* relative to the core PKS, NRPS and the cyP450 is contextually consistent with the off-loading type I TEs typical for NRPS or hybrid gene clusters. A second group of TEs, the Type II TE domains are usually located within a gene cluster and perform an editing role in the mispriming of substrates (Kim, Cropp et al. 2002). The conserved domain GxSxG that is found ~100 residues from the N-terminus and the GxG motif located downstream are well conserved in all TE-like enzymes rendering predictions of function based on sequence alignments difficult. More commonly, the function of TE domains is predicted based on their position in relation to the other biosynthetic genes. SalF is unusual in that it is does not directly follow the PKS or NRPS genes and its stand-alone architecture is more similar to a TEII, though overall sequence homologies suggest it is more closely related to TEI proteins that are positioned at the end of gene clusters and that play a role in off-loading rather than an editing function. Based on sequence alone, it is not clear whether the cyclization occurs in discrete steps catalyzed by two or three enzymes potentially including SalO and SalF and SalC, or in a concomitant fashion initiated by a single enzyme.

Genes Probably Involved in Regulation of Salinosporamide Biosynthesis and Resistance

The *salP gene* (675 bp) and the *salR* gene (960 bp) encode proteins similar to the LuxR and the LysR family of transcription regulators respectively. LuxR and LysR are transcriptional regulators that contain a substrate binding domain (CDD:43392) that are usually found N-terminal to a DNA binding effector domain and a regulatory helix turn-helix (HTH) DNA binding motif (CDD:40266). The LuxR family of transcriptional regulators activate transcription in response to the quorum-sensing acyl homoserine lactone molecules (AHLs) (Eberhard, Burlingame et al. 1981). The actinomycetes use a similar mechanism which utilizes γ-butyrolactones rather than AHLs to regulate the production of antibiotics and other secondary metabolites (Chater and Horinouchi 2003). This type of lactone-dependent transcriptional activation would allow the cells to coordinate expression of the *sal* genes with cell density. It is conceivable that **1** or **2** or their degradation products may in fact act in a similar fashion to the somewhat structurally similar AHLs and γ-butyrolactones and play a role in regulating transcription.

An unanticipated finding was the *salJ* gene (837 bp) at the C-terminus of the gene cluster that encodes a 20S proteasome β-subunit. The deduced amino acid sequence is most similar to the proteasome β-subunit from *Rhodococcus erythropolis* (62% / 77%) (Tamura, Nagy et al. 1995). The N-terminal sequence for SalJ (TTIVAVTF) is identical to the deduced N-terminal amino acid sequence of other eubacterial β-subunits after removal of a characteristic propeptide (Chen and Hochstrasser 1996). The propeptides of β-subunits are thought to facilitate the initial folding of the β-subunits and promote the maturation of the holoproteasomes when the N-terminal active site threonine residue

becomes available after autocatalytic release of the propetide (Chen and Hochstrasser 1996). Sequence comparison reveals that the conserved HG residues that precede the processing site are present in addition to a central consensus sequence: SSFX(D/E)(F/Y/L)LX$_4$PEXLP, that is thought to be essential to the function of eubacterial propeptides are present on SalJ. The length of the propeptide usually ranges from 53-65 residues and in SalJ the propeptide is 48 amino acid residues. The discovery of a proteasome β-subunit associated with the putative salinosporamide A biosynthetic gene cluster is of particular interest considering that salinosporamide A exerts its activity through irreversible covalent binding to the N-terminal active site threonine residue of the proteasome β-subunit (Groll, Huber et al. 2006). The function of SalJ may be related to resistance and preliminary modeling studies with SalJ have indicated that the site of salinosporamide binding contains a mutation that may inhibit the binding interaction and thus render the *salJ* product resistant to inhibition by salinosporamide A (B.S. Moore, personal communication).

Genes of Unknown Function

SalV (1413 bp) encodes a putative protein that is homologous to phenylacetate-CoA ligases (CDD: 31730) and is most similar to PaaK from *Photorhabdus luminescens* (39% / 59%) (Duchaud, Rusniok et al. 2003). This family also includes the EntE domain (CDD:31224) found in proteins classified as peptide arylation enzymes involved in secondary metabolite biosynthesis, transport, and catabolism. At this time no function for SalV is proposed.

The *salY* gene (987 bp) encodes a putative protein related to ornithine cyclodeaminases (OCD) (CDD:32552) and is most closely related to the putative protein from *Silicibacter pomeroyi* (40% / 56%) (Moran, Buchan et al. 2004). This family of proteins are deaminating cyclases and catalyze the conversion of L-ornithine to L-proline by an NAD+ dependent hydride transfer reaction that culminates in ammonia elimination (Costilow and Laycock 1971). Closer inspection of sequence alignments of SalY with putative ornithine cyclodeaminases indicated that a serine residue in SalY has replaced the conserved Glu56. SalW also contains an alanine residue rather than a glycine residue at position 6 of the NAD binding site (GAGVQA vs GxGxxG) of the ornithine decarboxylase. These two modifications suggest that *salW* may be more closely related to the alanine dehydrogenase family of deaminases (CDD:31030) (Smith, Mayhew et al. 2003). Alanine dehydrogenases catalyze the interconversion of pyruvate and alanine and also contain the lysine-2-oxoglutarate reductases, which is part of the lysine degradation pathway. The mechanism however, remains similar in that an NAD dependent reaction an imino group is formed between the α-amino and the α-carbon of an amino acid. The alanine dehydrogenase reaction then proceeds via a water-mediated hydrolysis of the imino intermediate, whereas in the OCD mechanism, the imino intermediate is subject to direct attack at the C2 position by the α-amino group followed by deprotonation and elimination of ammonia and hydride transfer from NADH yielding the L-pro product. A possible function for SalY may be in the cyclization reaction or biosynthesis of the amino acid.

The *salZ* gene (834 bp) encodes a putative protein that is 72% similar to PhzC/PhzF epimerases (CDD:30733). PhzC/PhzF is involved in the decarboxylative dimerization of two 2,3-dihydro-3-oxo-anthranilic acid molecules to create phenazine-1-carboxylate by *Psuedomonas fluorescens* and is thought to play a structural role in stabilizing intermediates in the pathway (Mavrodi, Ksenzenko et al. 1998). At this time it is not clear what the functions of SalY and salZ are in the biosynthesis of **1** and they may be housekeeping genes associated with other metabolic pathways.

A small ORF, s*alI* (384 bp) encodes a transposase fragment that is most closely related to a putative transposes of *S. avermitilis* (Omura, Ikeda et al. 2001). Transposase proteins are necessary for efficient DNA transposition and often located on the periphery of biosynthetic gene clusters (Hochhut, Lotfi et al. 2001). Transposases contain domains that are members of the DDE superfamily (CDD:34992), which contain three carboxylate residues that are believed to be responsible for coordinating metal ions needed for catalysis. The catalytic activity of this enzyme involves DNA cleavage at a specific site followed by a strand transfer reaction. Based on its small size compared to complete transposon sequences in the database (405 for *Streptomyces avermitilis*) and that it does not contain the conserved residues required to be a part of the DDE superfamily, *salI* appears to be a gene-fragment rather than a fully functional ORF.

Upstream of the *salA* gene and transcribed in the opposite direction as *salA-H* is *salK* (654 bp) that encodes a conserved hypothetical protein found in Actinobacteria. A BLAST search revealed only 12 homologs with no reported conserved domains and an unknown function. Closer examination of the context of the *salK* homologs in

Streptomyces coelicolor (SCO6745) (Redenbach, Kieser et al. 1996) and *Mycobacterium avium* (MAP0138c) (Li, Bannantine et al. 2005) showed that the genes are associated with other genes involved in transcriptional regulation and with fatty acid or isoprenoid processing but as yet, not reported to be associated with NRPS or PKS genes. Based on biosynthetic feeding studies and analysis of the *sal* gene cluster, a summary of the proposed biosynthesis of the salinosporamides is outlined in Fig. 3.19.

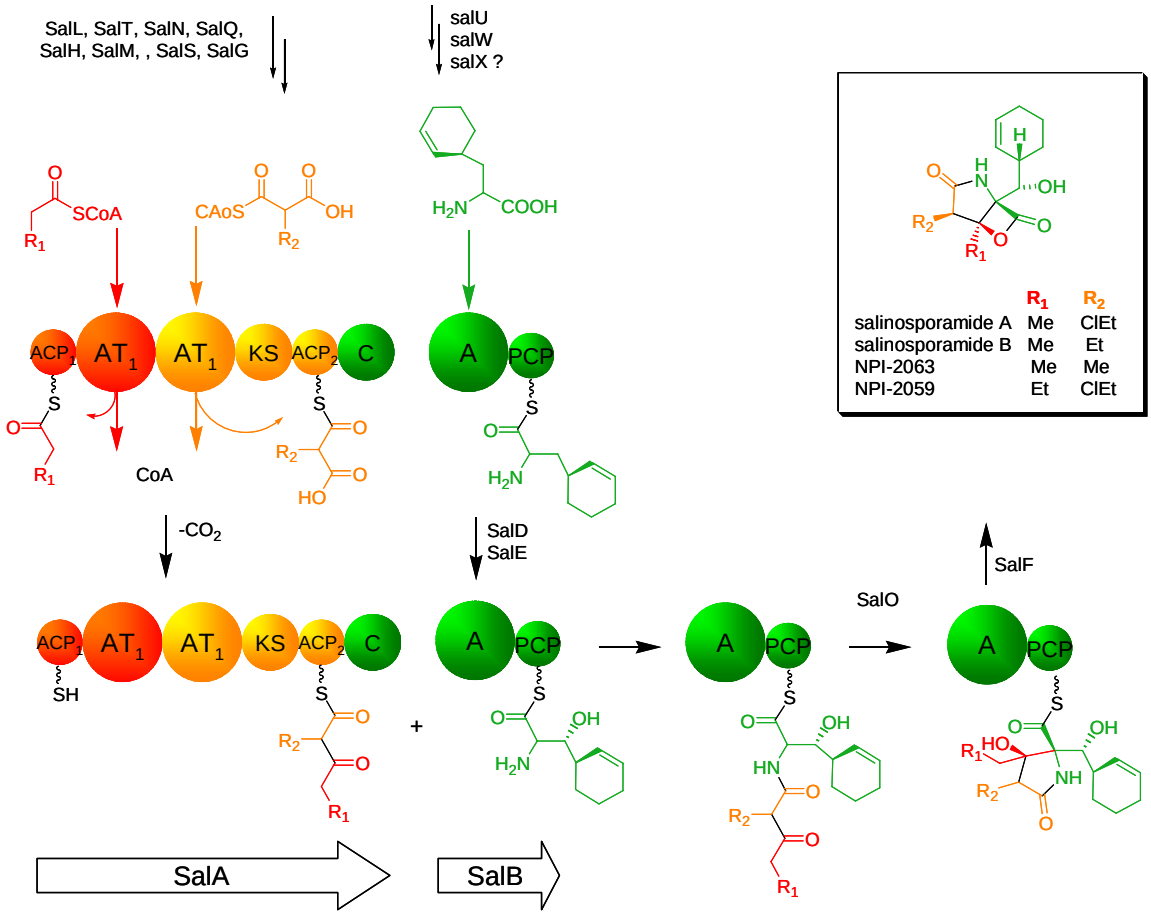


Figure 3.19. summary of the proposed biosynthesis of the salinosporamides based on predicted functions of the *sal* genes.

CHAPTER 4

DISUSSION AND SIGNIFICANCE OF RESULTS

The salinosporamide biosynthetic gene cluster is much larger and complex than originally hypothesized based on the minimum number of genes that would have been necessary for biosynthesis. The organization of the PKS and NRPS and tailoring genes exhibited by *salA-F* suggested that salinosporamide biosynthesis might be a straightforward hybrid PKS/NRPS system. However, feeding studies suggested otherwise and analysis of the complete *sal* gene cluster confirmed that novel mechanisms are utilized for the biosynthesis of salinosporamide and its precursors. The pathway for synthesis of the 4-carbon chlorobutyryl moiety appears to be initiated by an unprecedented chlorination reaction when compared to the biosynthesis of other chlorinated marine natural products. The pathway by which the amino acid moiety is synthesized and the exact mechanism of cyclization and offloading remain elusive, though based on the results of the feeding studies described in Chapter 2 and analysis of the putative salinosporamide A gene cluster, a general biosynthetic scheme for **1**, **4**, **5**, and **6** based on the predicted functions and arrangements of the *sal* genes has been proposed in Chapter 3 (Fig. 3.18).

The putative chlorination reaction catalyzed by SalL is central to the biosynthesis of the chlorobutyrate moiety and explains the erythrose-like labeling pattern observed from the feeding experiments. The published major classes of halogenases that have been identified that incorporate chlorine atoms into natural products include the heme-

and vanadium-dependent halogenases that utilize H_2O_2 and an active site iron or vanadium ion respectively to form a hypohalous species from the chloride anion to introduce the chlorine atom at a point of unsaturation (Groves 2003). Metal free halogenases require H_2O_2 to catalyze the formation of a peroxy-acid which subsequently forms the hypohalous species to react with an unsaturated substrate (Hecht, Sobek et al. 1994). Another common mechanism involves the $FADH_2$ dependent halogenases that require molecular oxygen to form an epoxide intermediate which is opened via nucleophilic attack with the chloride anion (van Pee 2001). The marine natural product barbamide, contains a trichloroleucine which has been proposed to proceed via radical mechanism for the polychlorination of the methyl group on leucine (Chang, Flatt et al. 2002) and an enzymatic chlorine radical process has recently been reported in pseudomonas syringae (Vaillancourt, Yeh et al. 2005).

The fluorinase from *S. cattleya* was unique because of the rarity of fluorinated compounds in nature. The enzymatic nucleophilic substitution mechanism overcomes the low solubility of naturally occurring fluoride salts, the low nucleophilicity due to tight solvation and the extremely high oxidation potential of fluoride. *In vitro* studies with the *S. cattleya* fluorinase have shown that the enzyme is also a chlorinase (Deng, Cobb et al. 2006) although the reaction is only observed in coupled enzyme assays that prevent the reverse reaction from occurring. The slightly larger space in the binding pocket of SalL resulting from the Ser158Gly substitution may more easily accommodate and possibly confer preference for chloride, which is less electronegative (χ = 3.2 vs 4.0) and has a larger atomic radius (1.75 vs 1.47) than fluoride. Preliminary *in vitro* studies with the

heterologously expressed SalL have shown that it indeed functions as a chlorinase and a brominase (B.S. Moore, personal communication). Furthermore, studies undertaken at Nereus showed that by increasing bromide concentrations in fermentations, the chloride atom on **1** could be replaced by bromide though production was at very low concentrations relative to **1.** High concentrations of fluoride were toxic to the cultures and no incorporation was observed in studies with iodide (R. Lam, personal communication). Given that the total salinity of seawater (3.5%) is 55% chloride, 0.19% bromide and < 0.01% fluoride, the relative abundance of reported chlorinated marine natural products is not surprising. Further analysis of the kinetic parameters of the reactions with SalL will illustrate which substrates are preferred. Since the high concentrations of sodium as NaCl in seawater is a growth requirement for *S. tropica* it is not yet known whether limiting the chloride concentrations in fermentations would further alter the equilibrium of the reaction to favor bromide or another halide ion and/or decrease the relative concentions of **1** vs **4**, **5** or **6**. While an analogous reaction using a combination of nucleophilic chloride ion and SAM is responsible for the production of chloromethane has been reported in fungi (Saxena, Aouad et al. 1998; Ni and Hager 1999), SalL appears to be the first *bona fide* chlorinase discovered in bacteria that utilizes the nucleophilic substitution/SAM mechanism.

Feeding studies also showed that the 4-carbon moiety of salinosporamide B (**4**) is biosynthesized via the known biosynthetic pathways reported for streptomycetes that utilize acetyl-CoA and proceed through butyryl-CoA and ethylmalonyl-CoA as described previously. The high ratio at which **1** is produced compared to **4**, **5** and **6** suggests that

the PKS AT_2, shows preference for the chlorinated precursor molecule over butyryl-, or propionyl- molecules but also shows that there is promiscuity with respect to the specificity of the PKS. In addition, the results confirmed that **4** is neither a degradation product nor a direct precursor of **1** as originally hypothesized. The observation that **1** and **4** are not precursors or degradation products of each other is also contrary to the hypothesis that the SalN phosphatase might be involved in dechlorination reactions, which is proposed to be the function of the phosphatase-like protein in the fluoroaceate gene cluster from *S. cattleya*. If SalN were a dechlorinase, we would expect to observe compound **4** with the same labeling pattern observed in **1**. The labeling pattern in **4** also poses the possibility that the CCR, *salF* may be used strictly for the biosynthesis of the ethylmalonyl-CoA precursor as seen for molecules such as monensin and tylolosin. Genetic manipulations whereby the activity of *salF* is abolished will determine the extent of its involvement in biosynthesis of the individual salinosporamide analogs.

Halogenated compounds are of particular interest to the pharmaceutical industry and the mechanism observed by which salinosporamide A binds the proteasome irreversibly by tetrahydrofuran ring formation can be used as a model system for other drugs. Coupled with the fact that the PKS AT2 domain appears to show preference for the chlorobutyryl precursor, the halogenase and the PKS may be genetically engineered to accept alternate substrates, potentially leading to novel compounds. The pesticide industry relies heavily on halogenated compounds and current methods to supply these molecules, especially those that are fluorinated requires reactions that utilize and produce

environmentally hazardous chemical by-products. An enzymatic approach could be used to reduce the amount of chemical waste associated with the production of pesticides.

The unusual CHHA amino acid arises from a shunt in the shikimate pathway though the exact branch point could not be determined by the experiments described herein. Early feeding experiments ruled out an aromatic precursor, which would have required highly unusual reductive chemistry, and confirmed that shikimate was a precursor of **1**. Unfortunately the chorismate experiments were inconclusive due to the inability of obtaining a totally pure sample of [U-$^{13}C_{10}$]chorismate-enriched **1**. This was puzzling since the purification method using flash column chromatography had previously been consistent and reliable. Even by collecting fractions of < 1 mL with TLC analysis showing pure samples of **1,** a clean spectrum by NMR nor MS was unattainable. To confirm that chorismate or prephenate were not co-migrating with **1** a TLC experiment with chorismate, prephenate, **1** and **4** confirmed that the labeled chorismate was not eluting with **1**. The experiments were repeated a second time with similar results. Further studies using reversed-phase semipreparative HPLC to purify the compound may be required to obtain a pure sample of **1** but was not available at the time the studies were being performed. The discovery of the partial prephenate dehydratase gene (*salX*) in the gene cluster was encouraging to the shikimate shunt pathway hypothesis. However, whether this is a functional gene encoding a functional protein remains to be evaluated. It is conceivable that one of the reductases purported to be involved in biosynthesis of the alkyl moiety, such as *salM* could be involved in reductive chemistry on the amino acid.

Perhaps the most surprising and intriguing feature of the *sal* gene cluster was the presence of the *salJ* gene located at the C-terminus that encodes a 20S proteasome β-subunit. Close inspection of the *S. tropica* CNB440 genome revealed that a full set of 26S proteasome genes are located elsewhere on the chromosome suggesting that *salJ* is a second copy and not required for assembly of the primary proteolytic machinery. By binding and sequestering **1** or **4** SalJ may play a role in resistance or act as a transcriptional regulator along with SalP and SalR. In fact, preliminary modeling studies with SalJ have shown that a mutation occurs in the binding pocket that would inhibit the binding interaction with **1** (B.S. Moore, personal communication). This lends to the possibility that SalJ has evolved as a resistant copy of the 20S proteasome and may indeed be the resistance mechanism against salinosporamide toxicity in *S. tropica*. A knockout experiment or an experiment whereby the mutant status of *salJ* is restored to a "normal" or non-mutant β-subunit could elucidate its function.

Regarding the fermentation of *S. tropica* and salinosporamide production, cultures grown without resin produced **1** at very low levels (1-2mg/L). In addition, all of compound **1** was degraded by day 4. The cells sporulated shortly thereafter and cultures died suggesting that **1** may be toxic to the cells and/or feedback inhibition of **1** production was occurring. When the XAD7 resin was added to the fermentations after 24 hours when primary growth was shifting to secondary metabolism, the cultures remained healthy for 7-9 days and production of **1** increased to ~100 mg/L. This corroborates the hypothesis that **1** may be toxic to the cells, that feedback inhibition is occurring and/or that when present a low concentrations, the salinosporamides may serve as quorum

sensing molecules. Further studies to test the effects of **1** on growth and the production of other natural products in a non-producing strain of *S. tropica*, in addition to studies on the function of *salJ* are necessary to test these hypotheses.

A key requirement for PKS and NRPS activation is the PPTase and since the cluster lacks its own copy of a PPTase gene, there must be a copy elsewhere on the chromosome that is utilized during secondary metabolite biosynthesis. Examination of the S. tropica CNB440 genome showed that there is a possible candidate (SCO23) though it is only 25 % identical and 42% similar to the *sfp* PPTase from *B. subtilus*.

Ultimately, knowledge of the biosynthetic building blocks and access to the genes responsible for salinosporamide A biosynthesis imparts the opportunity to increase yields through regulation of the pathway or to decrease yields of compounds considered contaminants in fermentations. This could potentially be accomplished through increasing or reducing the amounts of individual substrates that are intermediates in the pathway or through genetic engineering of transcriptional regulators and/or resistance mechanisms. Furthermore, the possibility of generating novel analogs through genetic engineering and combinatorial biosynthesis with improved or novel therapeutic properties is possible. The pathways described for biosynthesis of the 4-chlorobutyrate moiety, the unusual amino acid, and the cyclization reaction all present opportunities to use biotechnology in order to create novel chemistry. The potential use of the chlorinase to generate halogenated precursors not available by the more common halogenase reactions is enormous and has potential for pharmacological, industrial and agricultural application. The γ-lactam-β-lactone ring structure is highly unusual and also

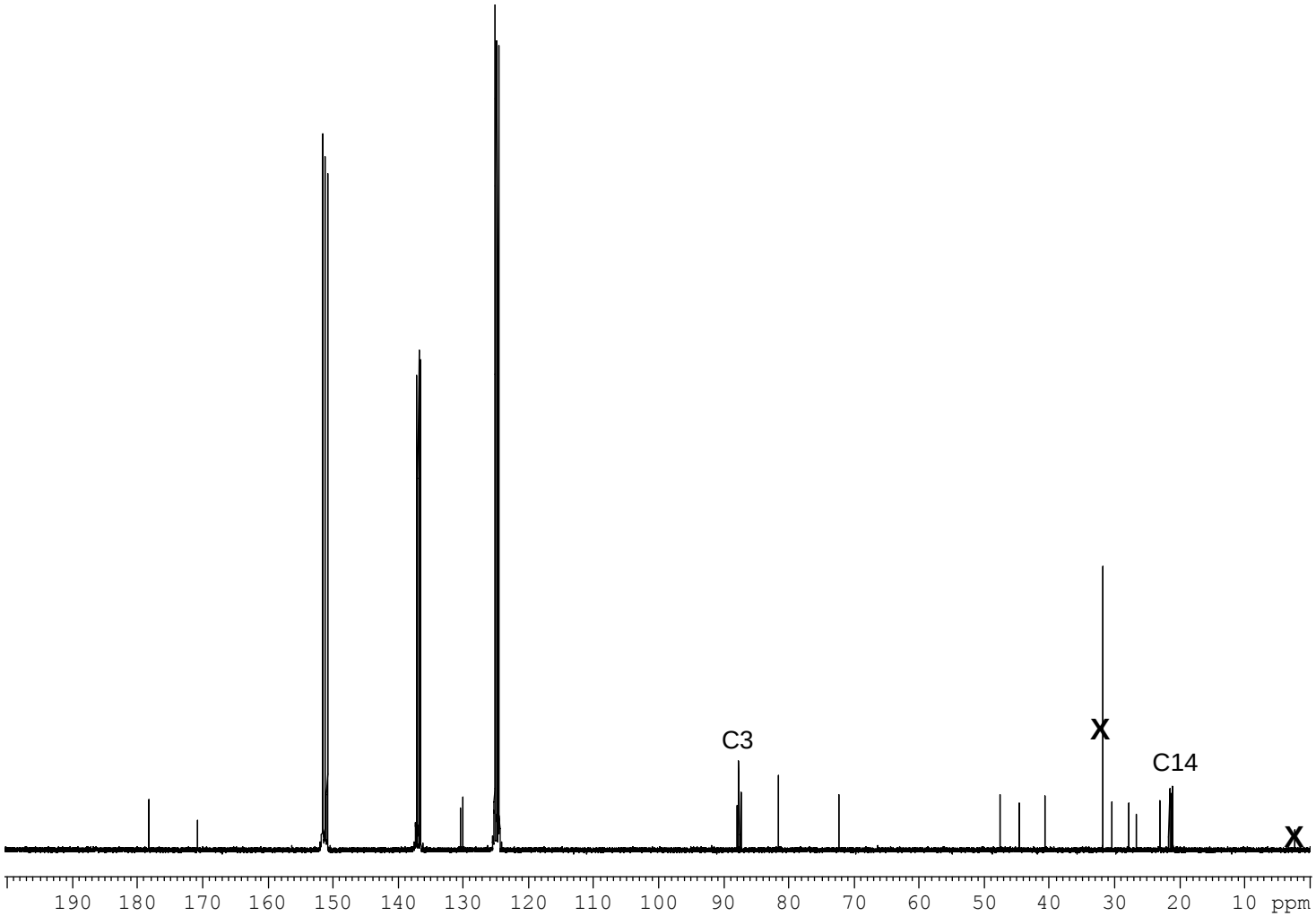
unprecedented in a mixed PKS/NRPS pathway. Consequently, studies centered on the functionality of the bicyclic core offers tremendous potential as specific positioning of functional groups could regulate the stability or instability of the bicycle as a potential pharmacophore or pro-drug.

Further functional analysis of the *sal* genes will undoubtedly yield novel biochemistry that can be harnessed through combinatorial methods to provide a framework of intermediates for additional synthetic chemical modification to yield novel compounds. Although salinosporamide A appears to be an exquisitely designed natural product as an inhibitor of the proteolytic activity of the 20S proteasome, it is reasonable to speculate that derivatives may possess entirely new biological targets and activities.

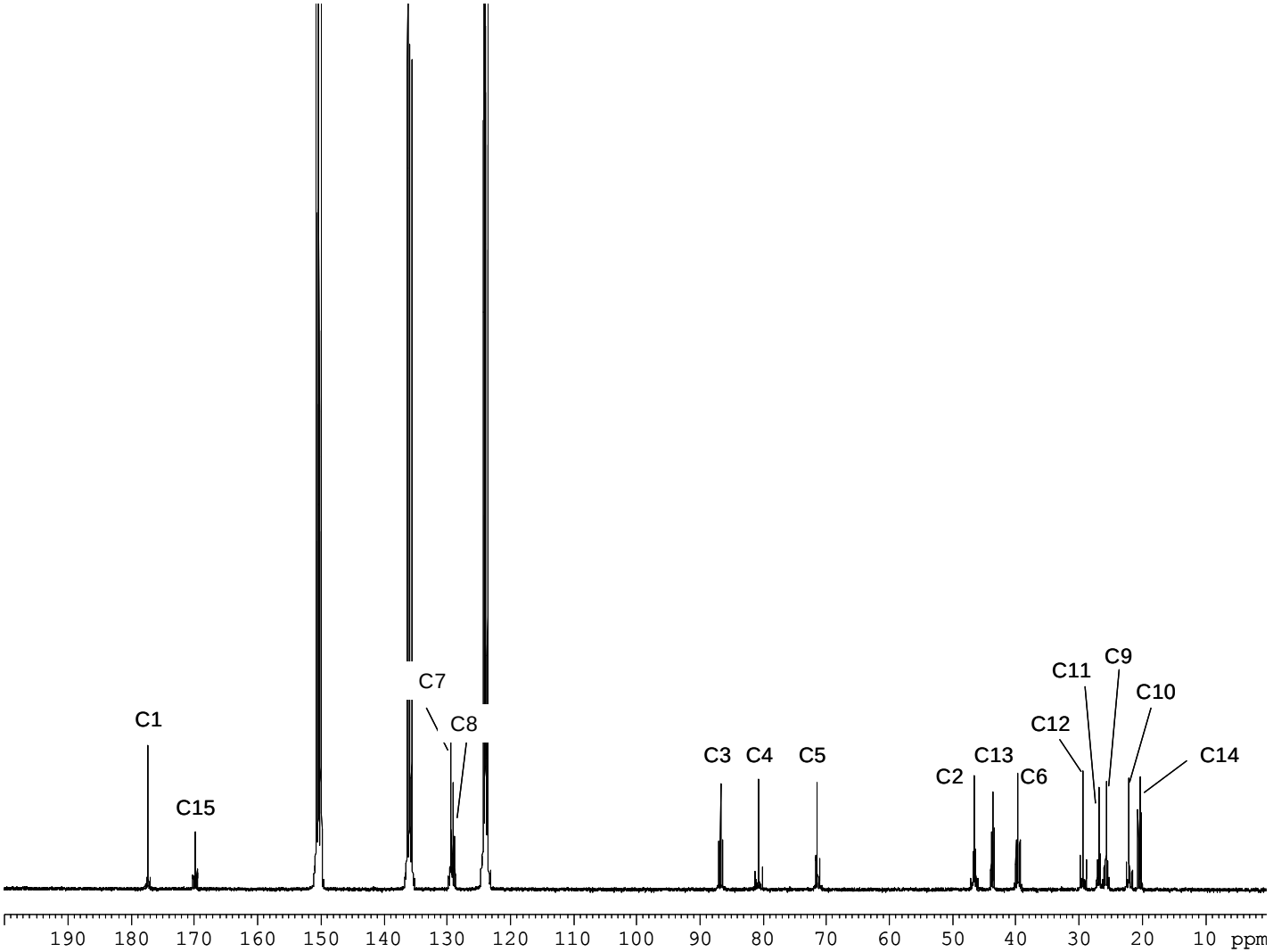
The vast amount of information gained from the examination of just a single biosynthetic gene cluster from *S. tropica* is proof of principle that the potential for discovering new compounds and even entirely new bacterial genera from the sea is enormous and should not be overlooked. In fact, analysis of the *S. tropica* CNB440 genome indicates that there at least 15 putative gene clusters for biosynthesis of secondary metabolites, in addition to several "orphan" genes not associated with any particular cluster and may have interesting activities. Analysis and evaluation of each of these genes is necessary to fully understand the potential of this organism. The salinosporamides are only one example that the under explored marine environment offers great promise as a resource for novel organisms that produce novel compounds possessing potent and pharmacologically relevant activity. Furthermore, marine natural products that once showed promising activity but were abandoned due to low production

levels or toxicity issues should be systematically revisited using modern fermentation and genetic techniques in order to fully harness the pharmaceutical potential of the marine environment.

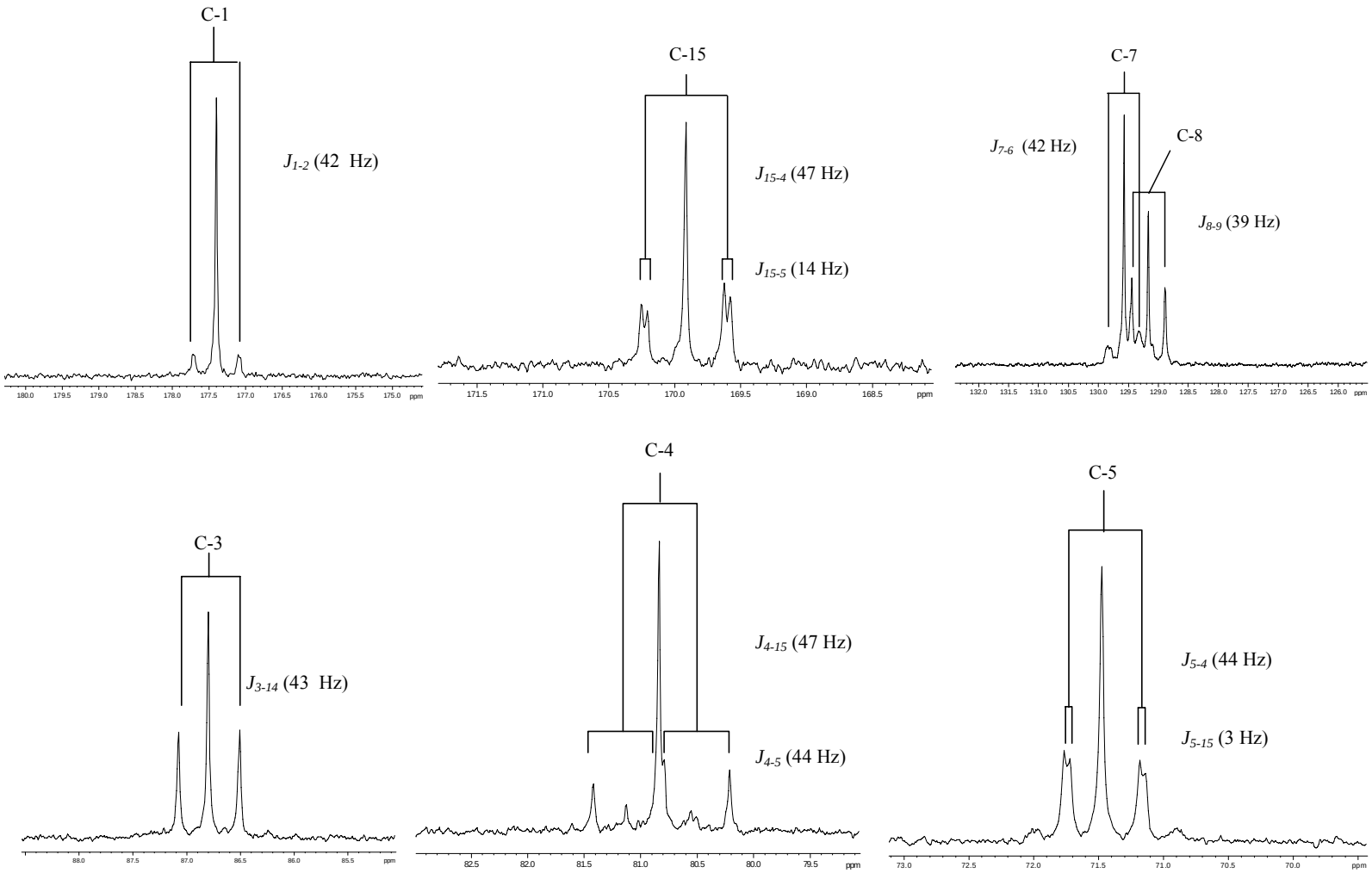
APPENDIX A



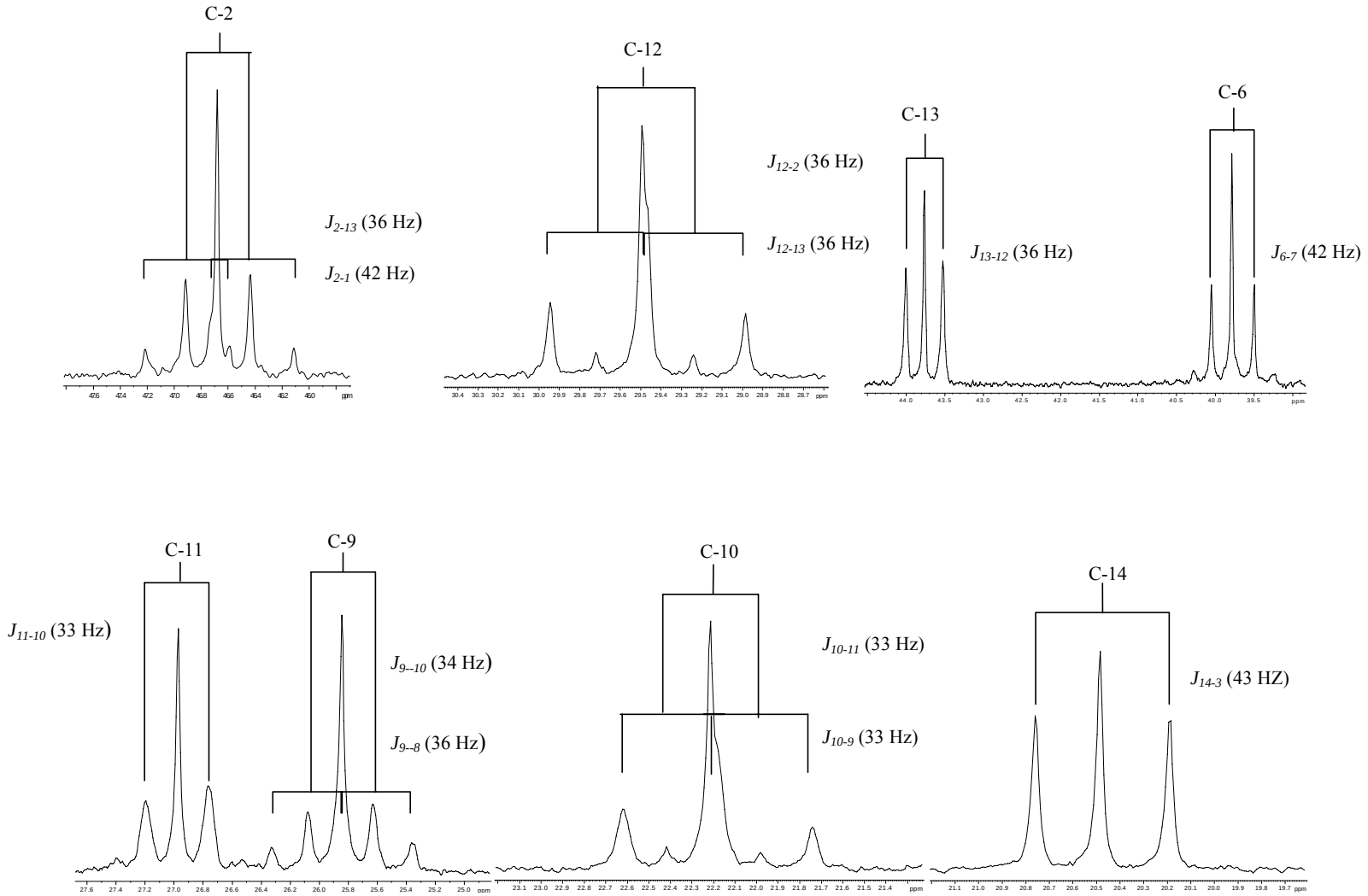
Appendix A, Figure 1. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1,2-$^{13}C_2$]acetate-enriched **1**. X = solvent peak from work-up.



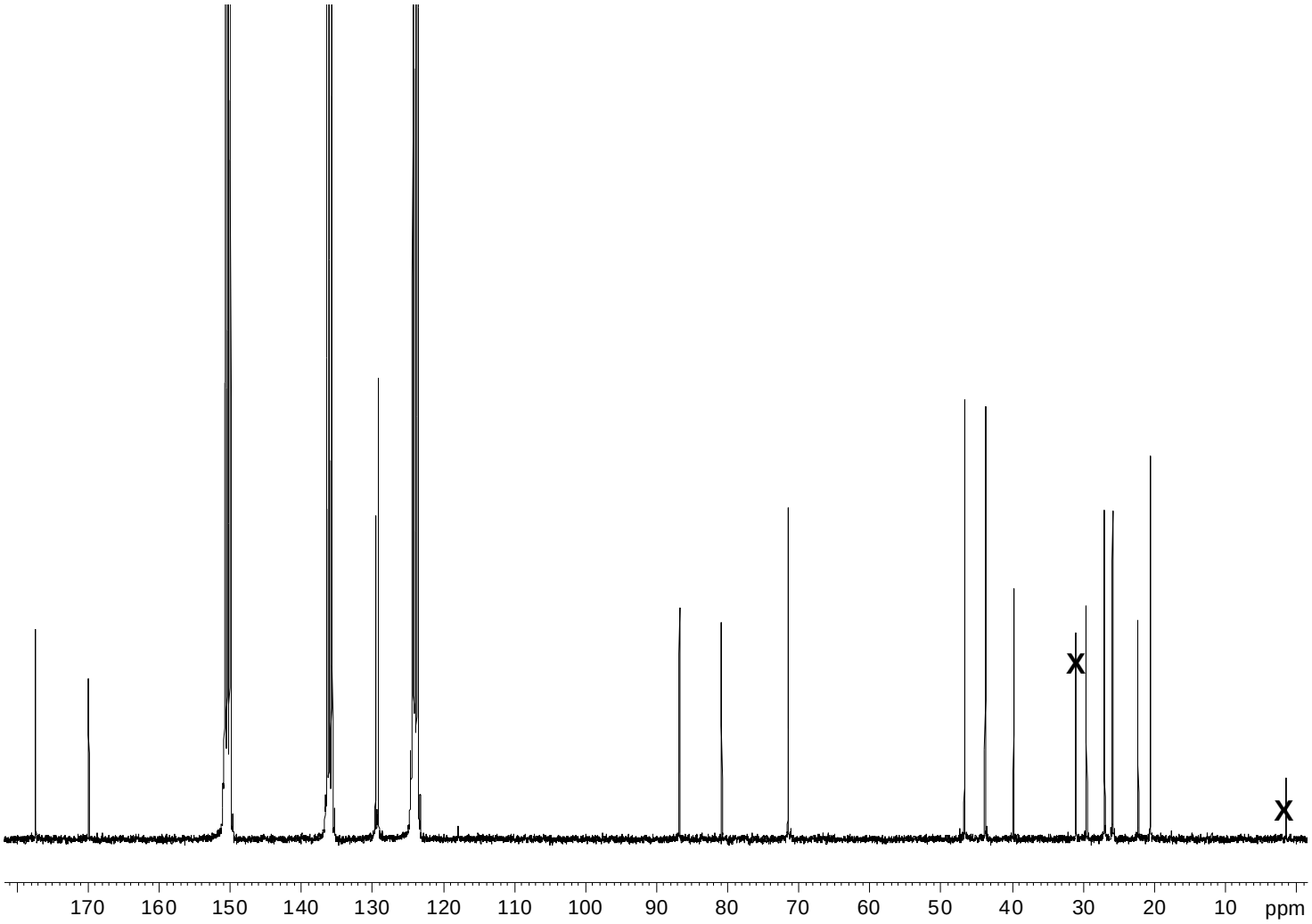
Appendix A, Figure 2. 1D ^{13}C-NMR spectrum in d_5-pyridine of [U-$^{13}C_6$]glucose-enriched **1**.



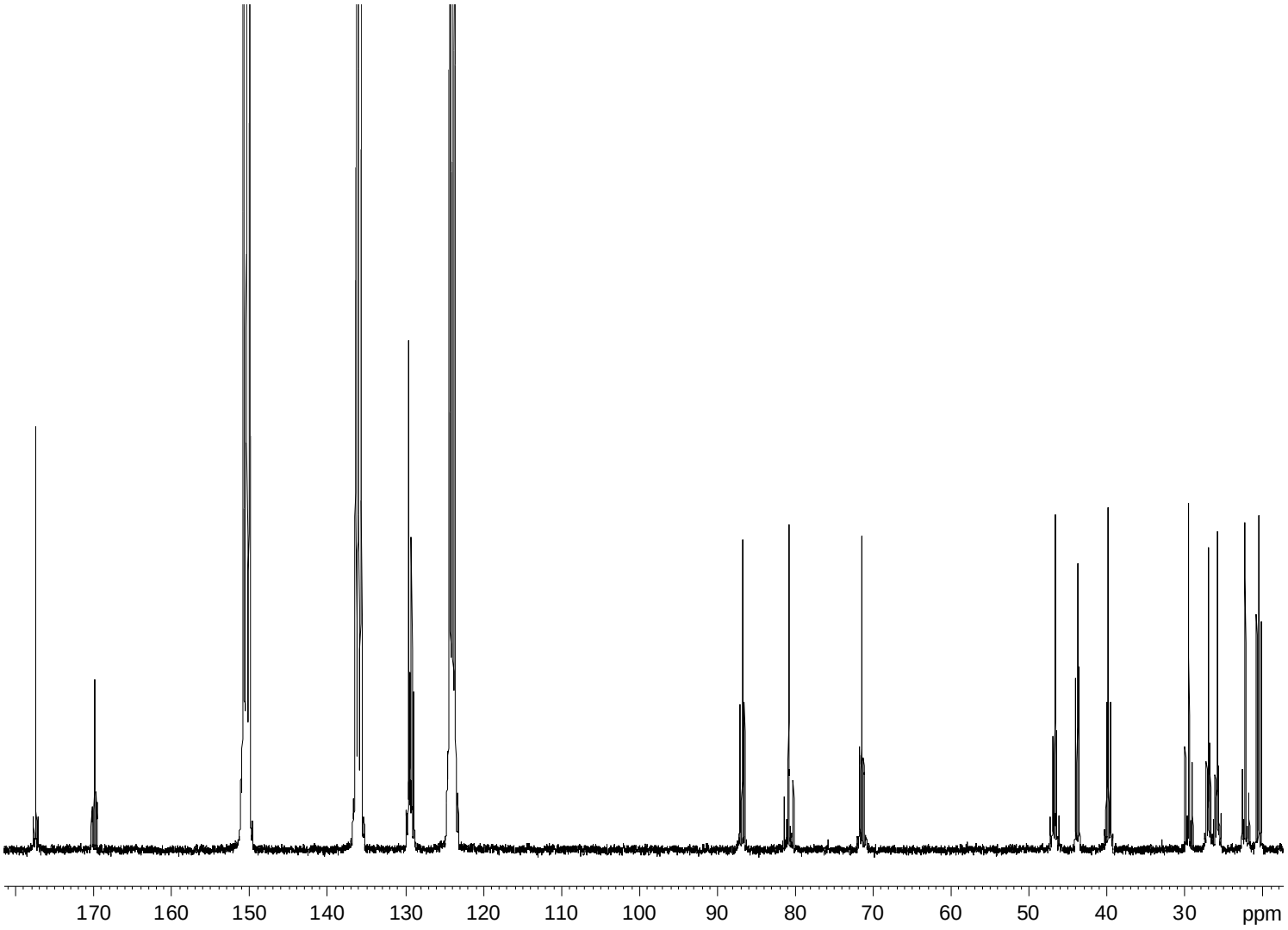
Appendix A, Figure 3. Expanded spectra of peaks from [U-$^{13}C_6$]glucose-enriched **1**.



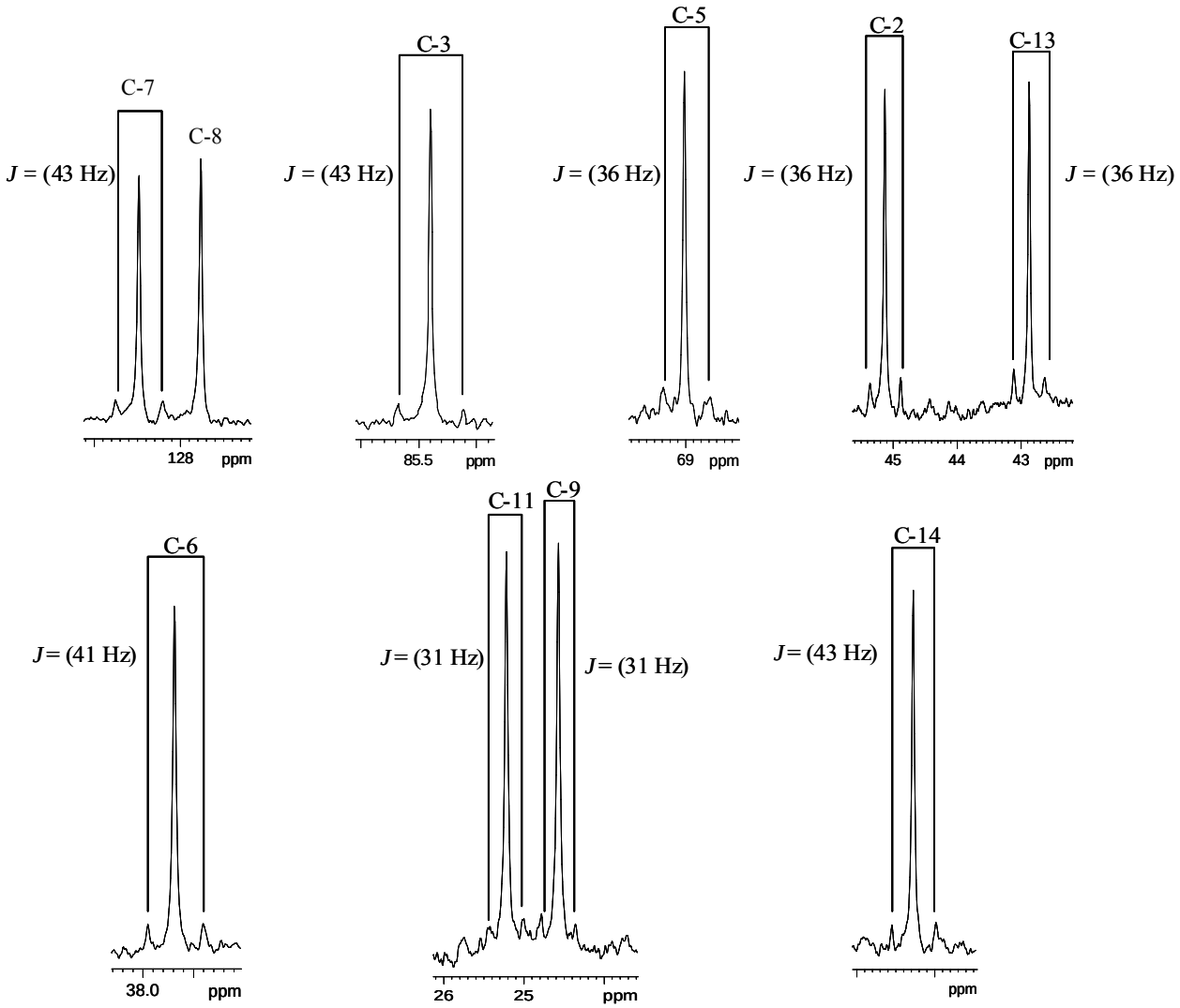
Appendix A, Figure 3 - continued. Expanded spectra of peaks from [U-$^{13}C_6$]glucose-enriched **1**.



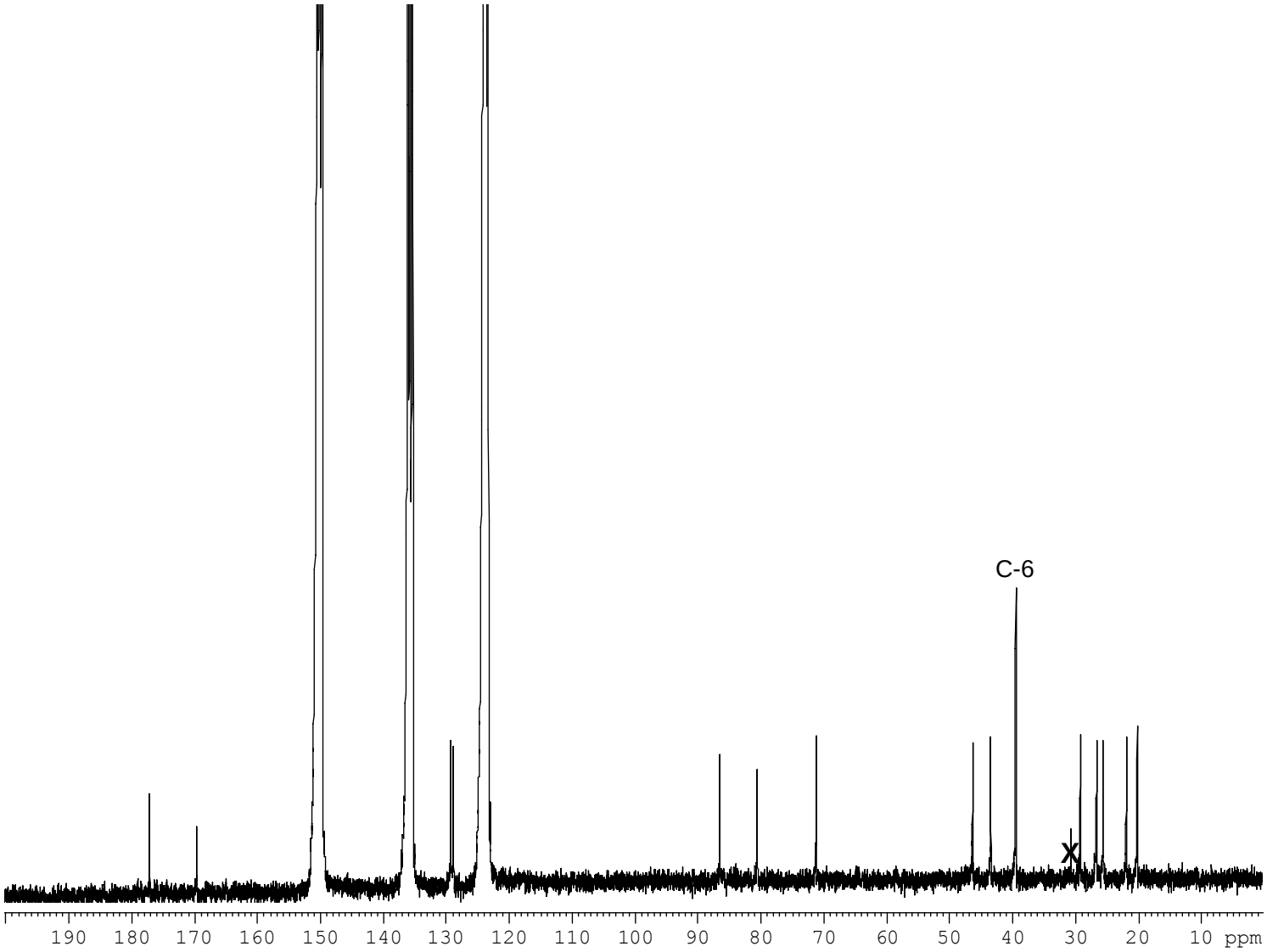
Apendix A, Figure 4. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1,3-$^{13}C_2$]glycerol-enriched **1**. X = solvent peaks from work-up.



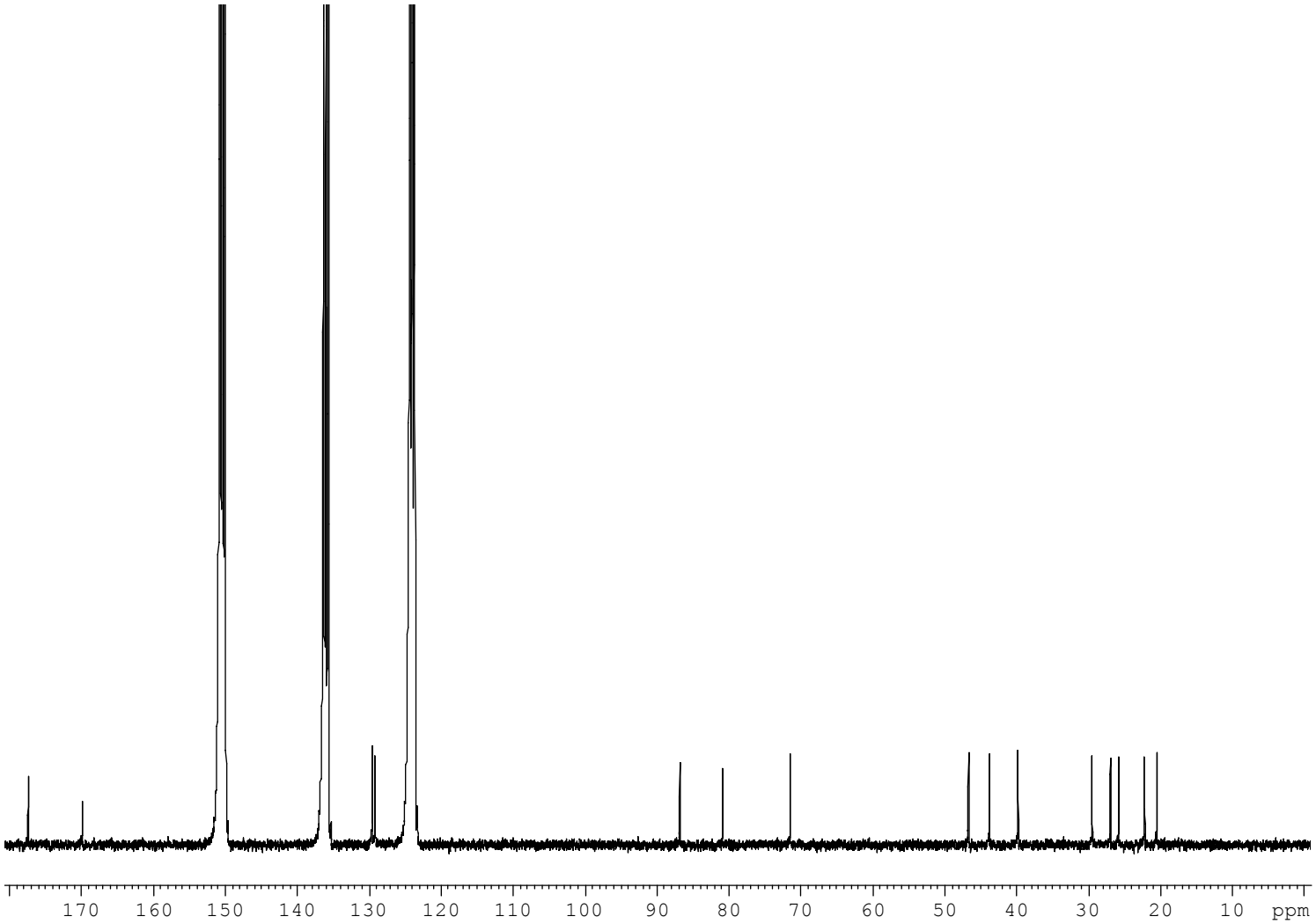
Apendix A, Figure 5. 1D ^{13}C-NMR spectrum in d_5-pyridine of [U-$^{13}C_3$]glycerol-enriched **1**.



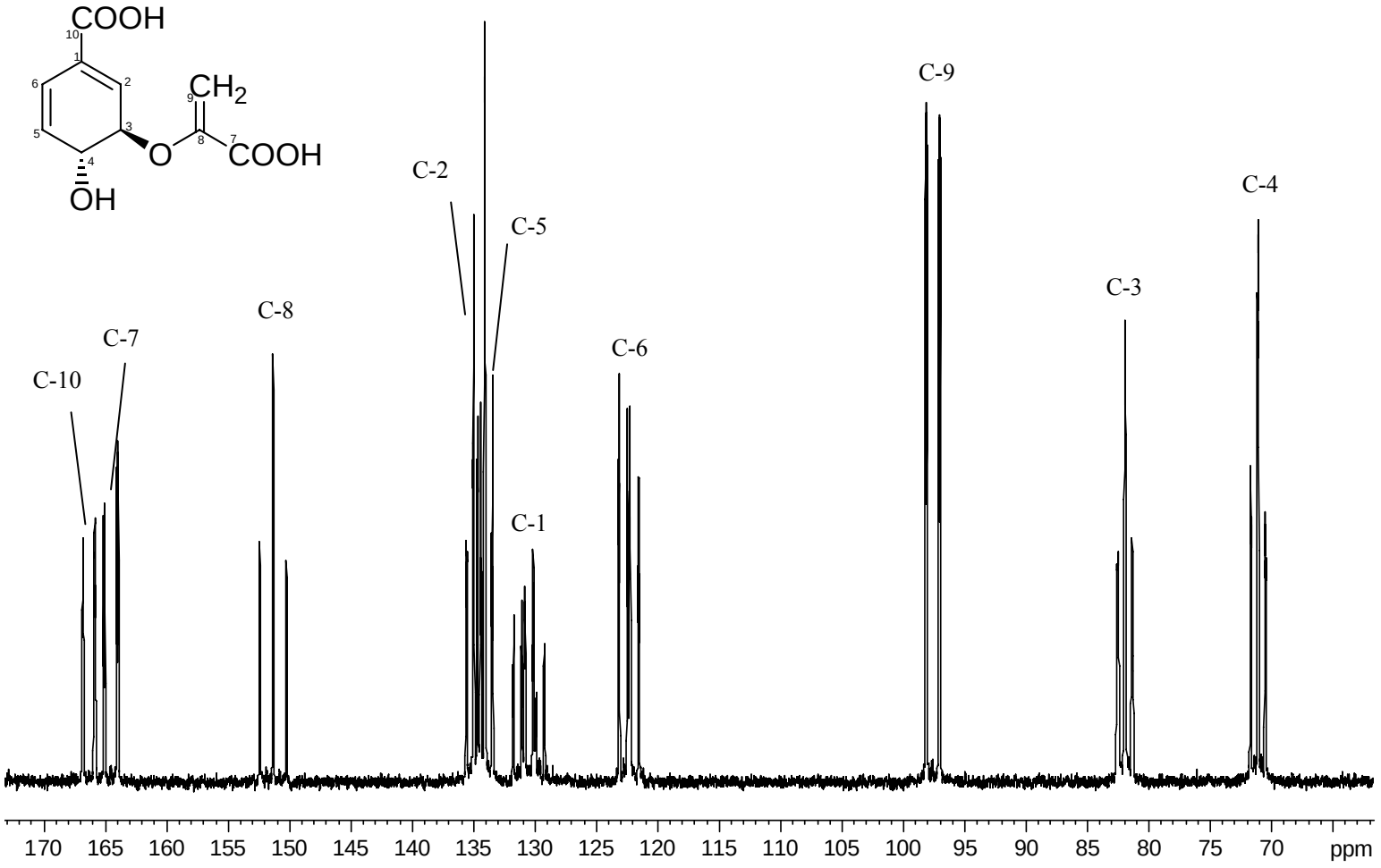
Apendix A, Figure 6. Expansions of 1D ^{13}C-NMR spectra in d_5-pyridine of [U-$^{13}C_3$]glycerol-enriched **1.**



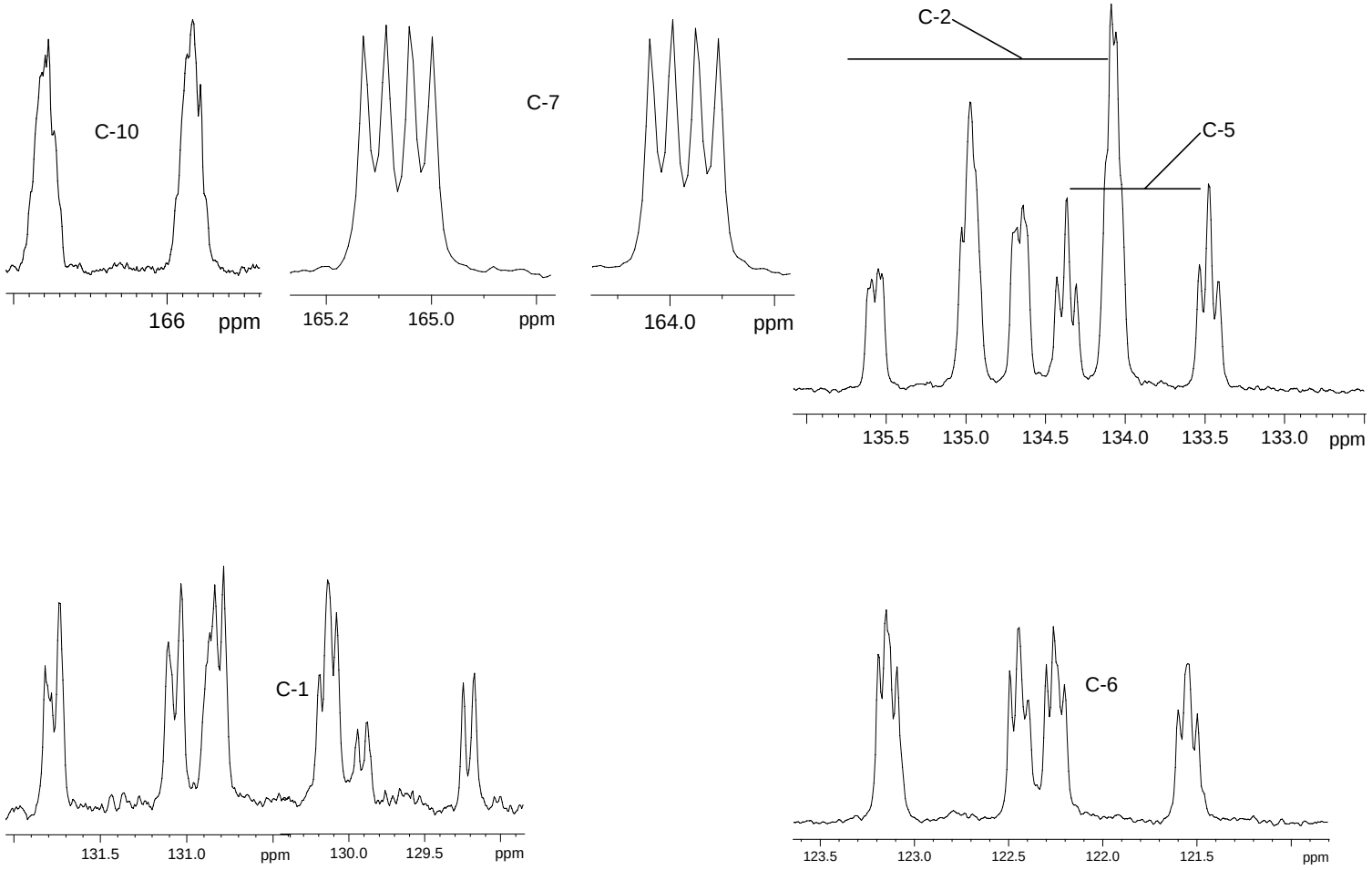
Apendix A, Figure 7. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1,7-$^{13}C_2$]shikimate enriched **1**.



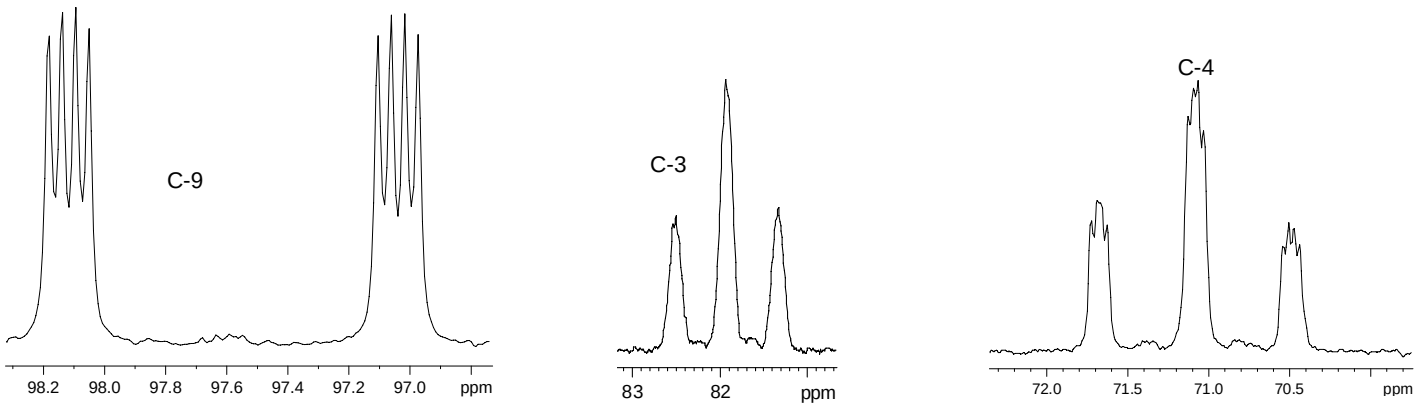
Apendix A, Figure 8. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1-$^{13}C_1$]phenylalanine-enriched **1**.



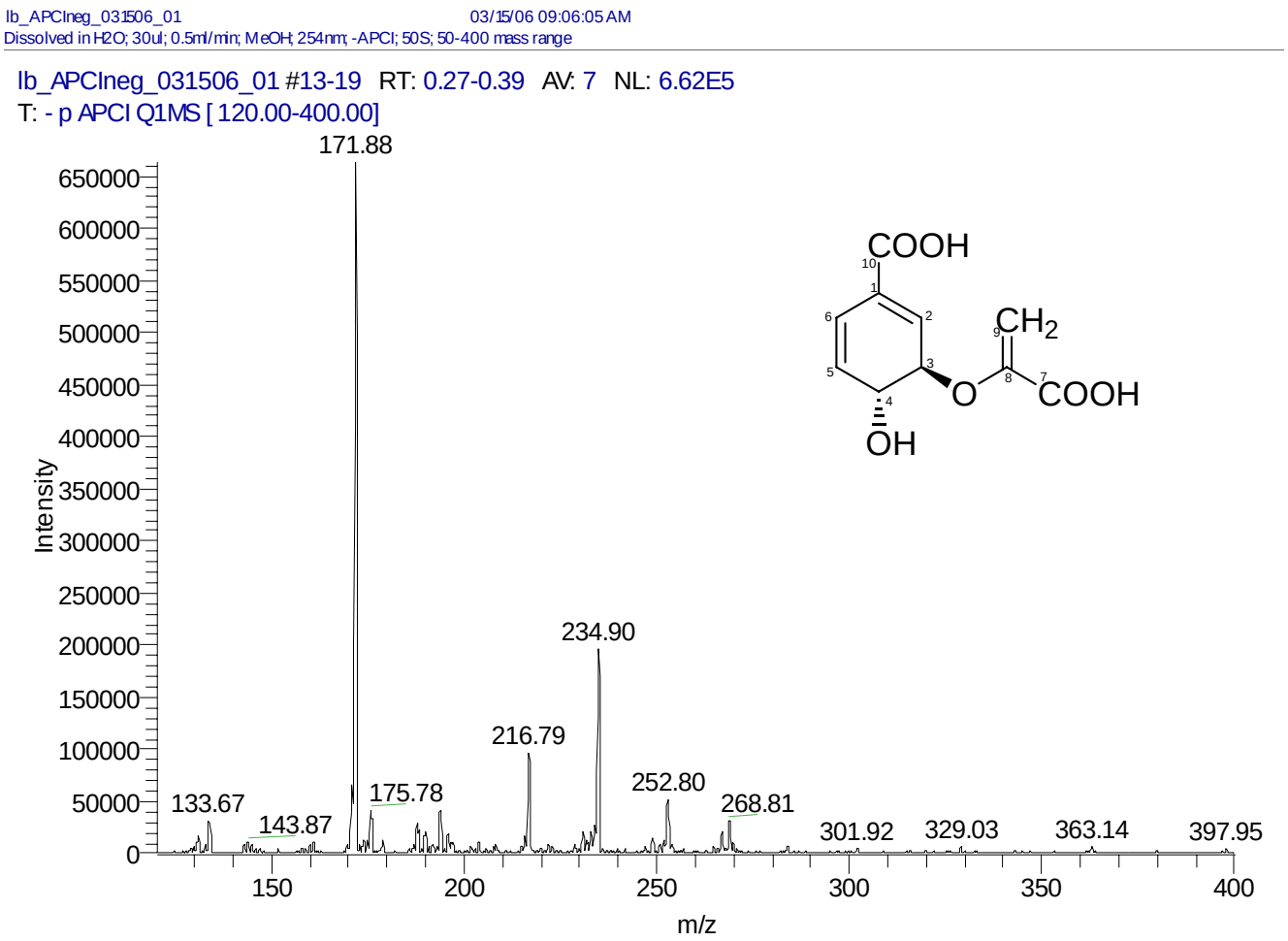
Appendix A, Figure 9. 1D ^{13}C-NMR spectrum in d_6-acetone of [U-$^{13}C_{10}$]chorismate.



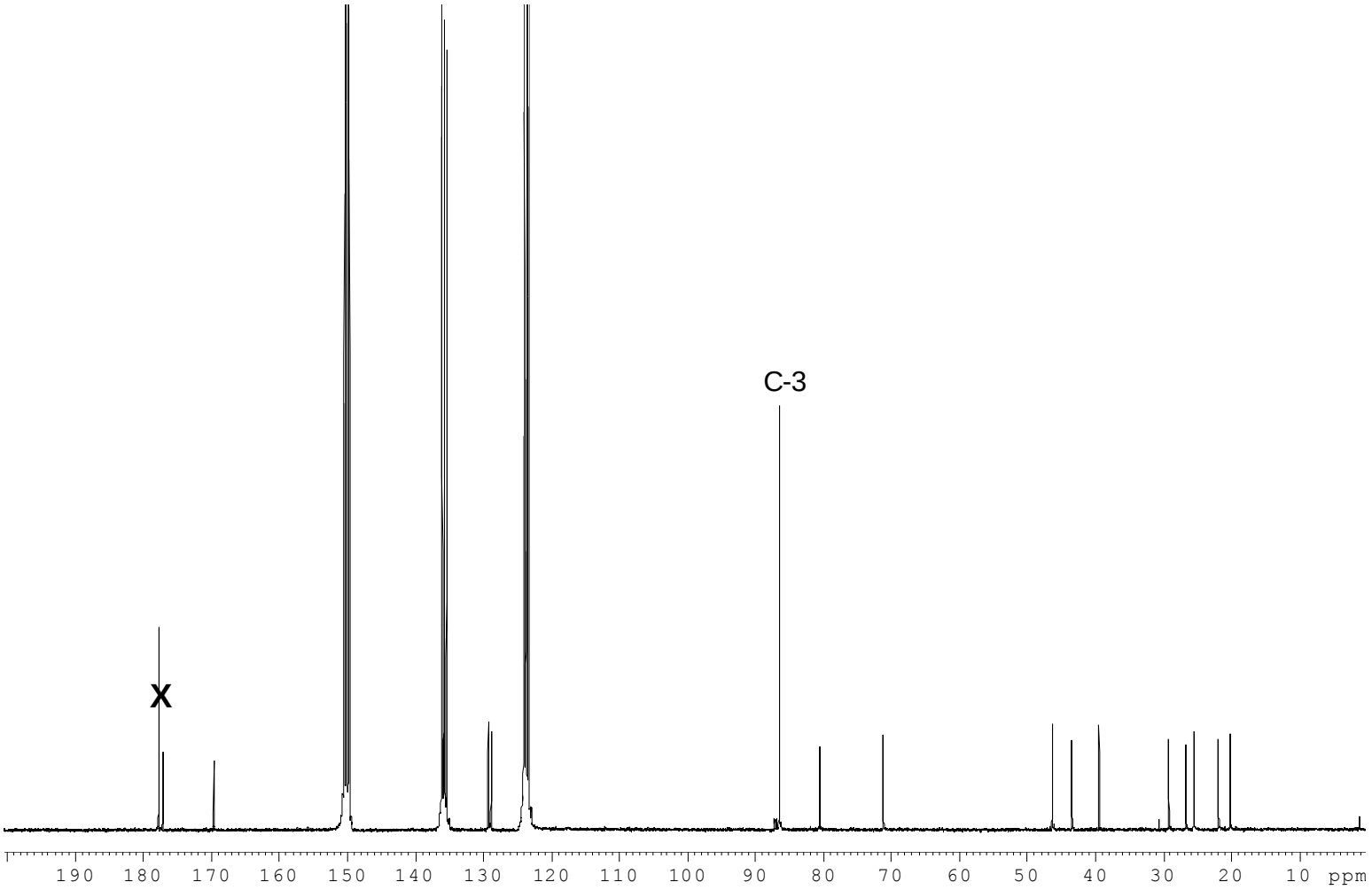
Appendix A, Figure 10. Expansions of indiviual carbon data from 1D ^{13}C-NMR spectrum of [U-$^{13}C_{10}$]chorismate in d_6-acetone.



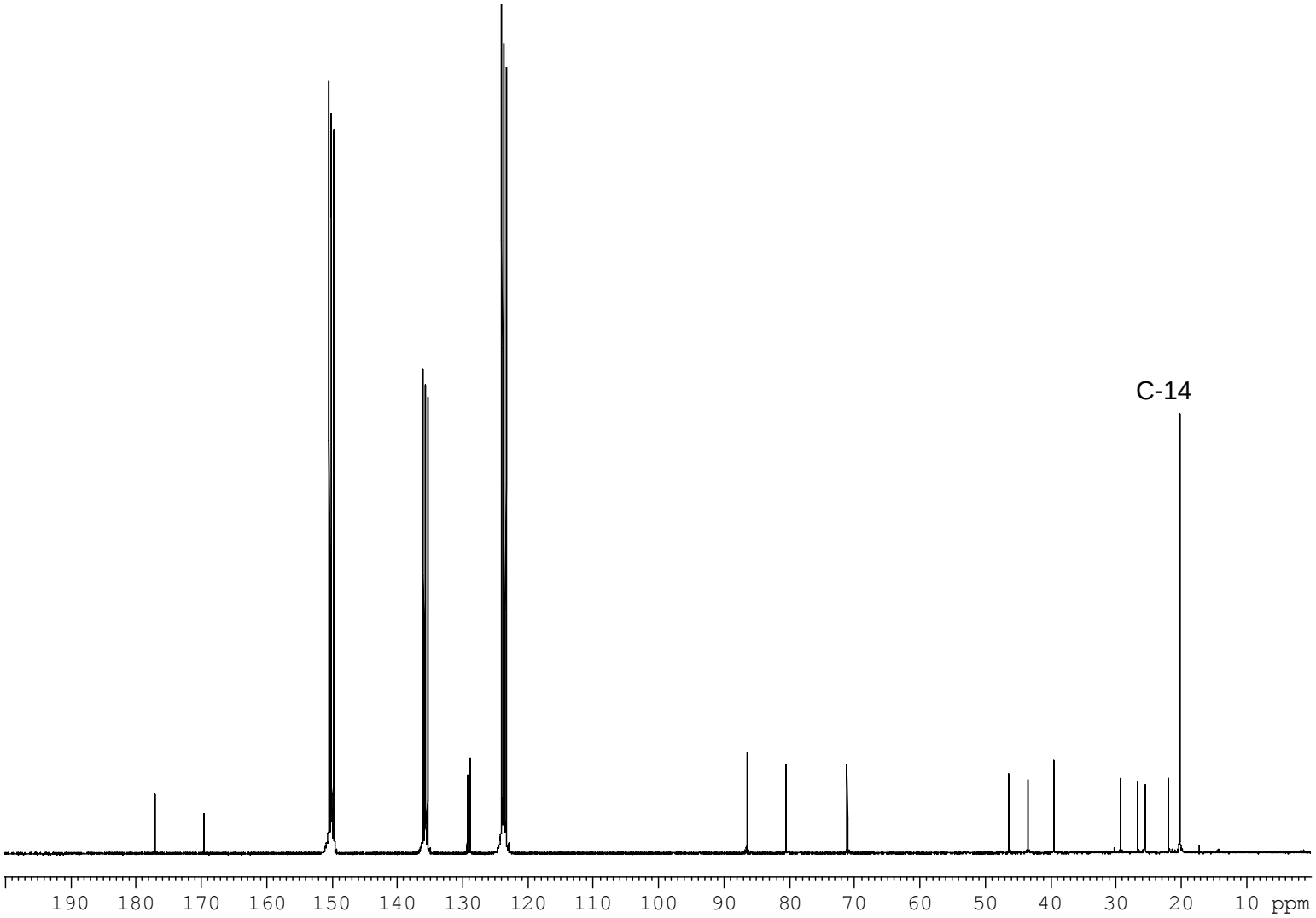
Appendix A, Figure 10 - continued. Expansions of indiviual carbon data from 1D ^{13}C-NMR spectrum of [U-$^{13}C_{10}$]chorismate in d_6-acetone.



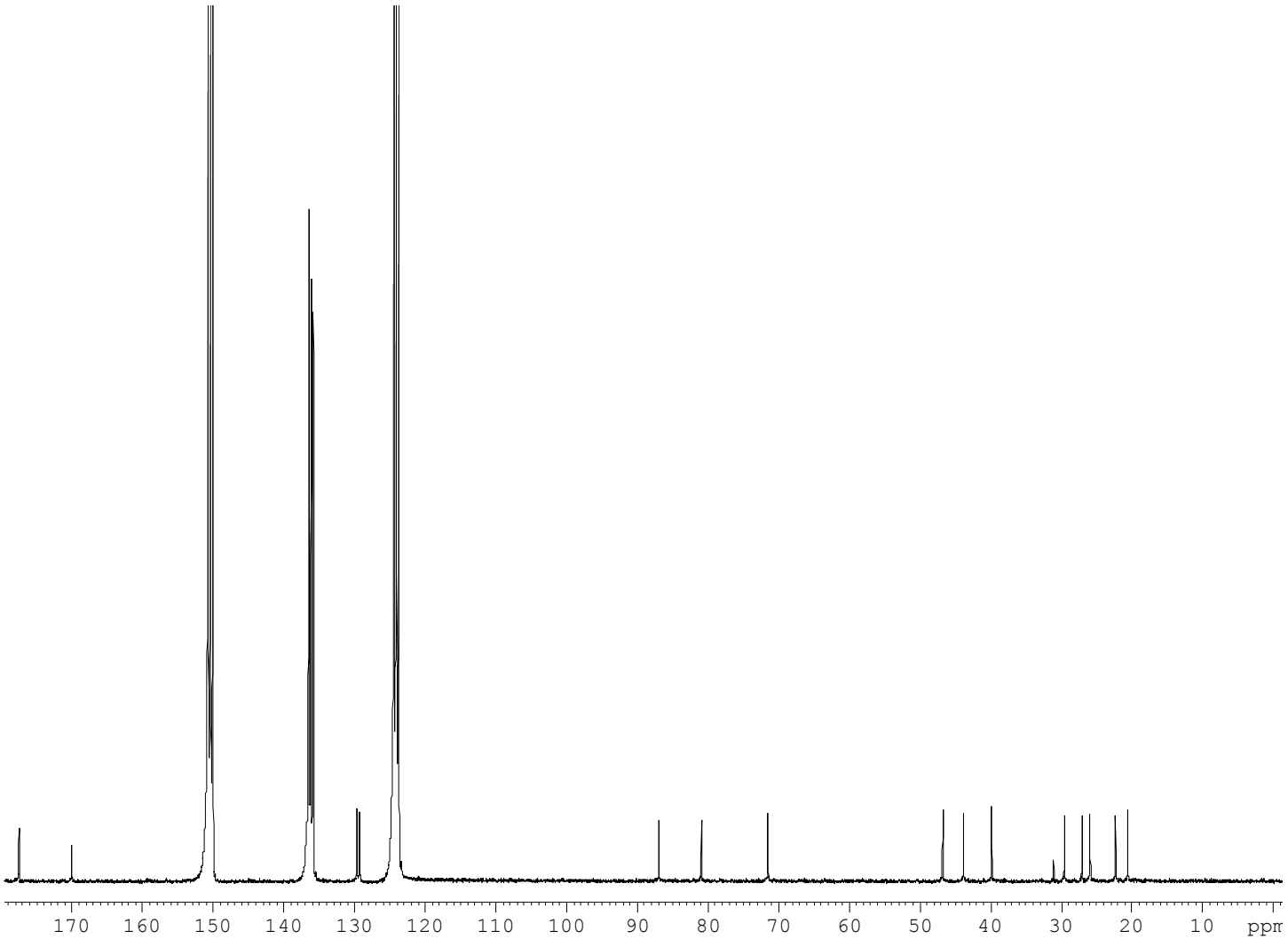
Apendix A, Figure 11. LC/APCI/MS chromatogram of [U,-$^{13}C_{10}$]chorismate. Acquired in negative mode with monitoring at 254 nm. The calculated molecular weight of unlabeled chorismate is 225.18. The peak at 234.90 correlates to uniformly labeled chorismate $[M-H+10]^-$. The mass of 171.88 results from the loss of $CH_2CHCOOH$ including carbons 7, 8 and 9.



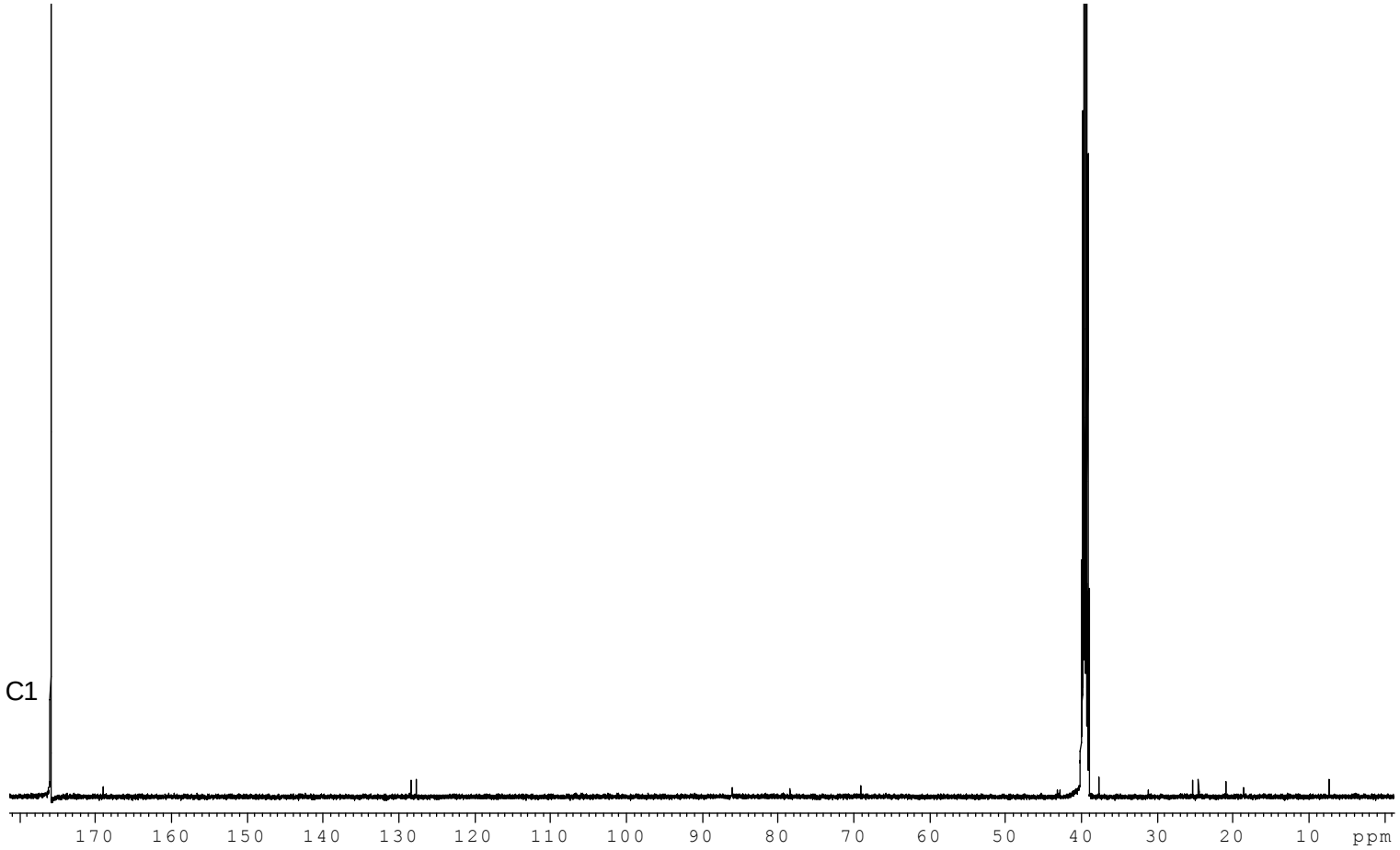
Apendix A, Figure 12. 1D ^{13}C-NMR spectrum in d_5-pyridine of [1-$^{13}C_1$]butyrate-enriched **1**. X = solvent peak from work-up.



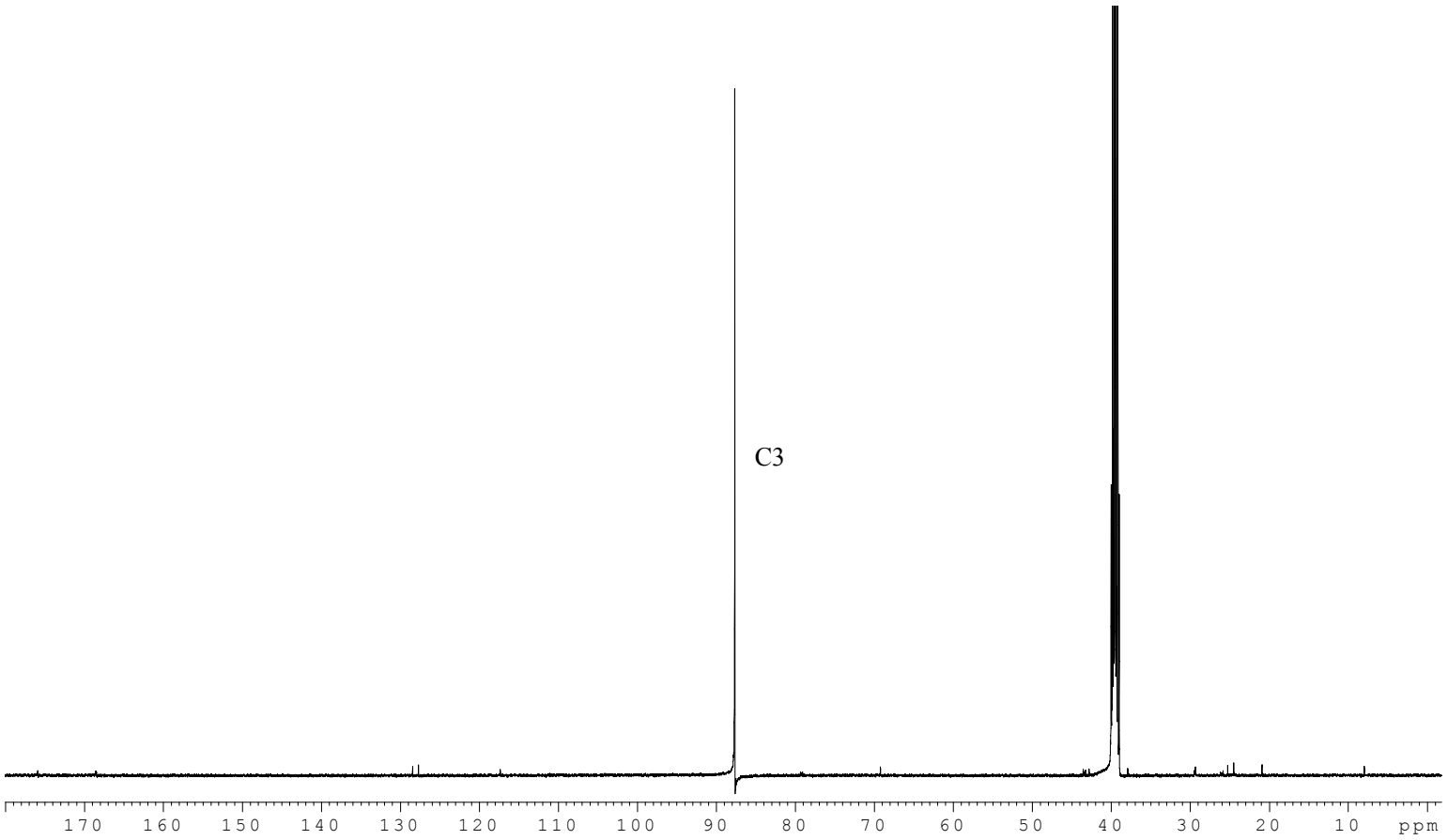
Apendix A, Figure 13. 1D ^{13}C-NMR spectrum in d_5-pyridine of [2-$^{13}C_1$]alanine-enriched **1**.



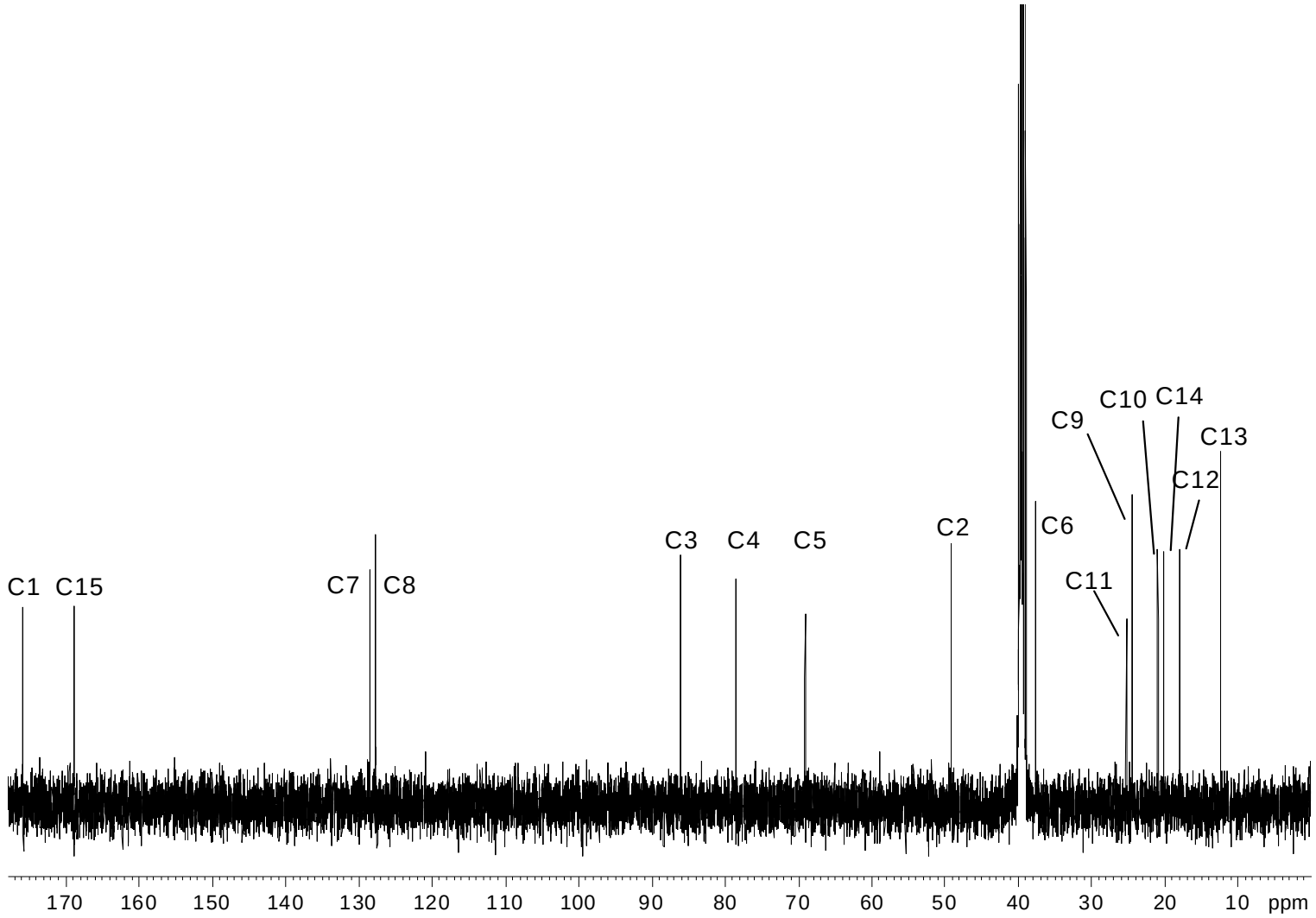
Apendix A, Figure 14. 1D ^{13}C-NMR spectrum in d_5-pyridine of 2-methyl-[2-$^{13}C_1$]butyrate-enriched **1**.



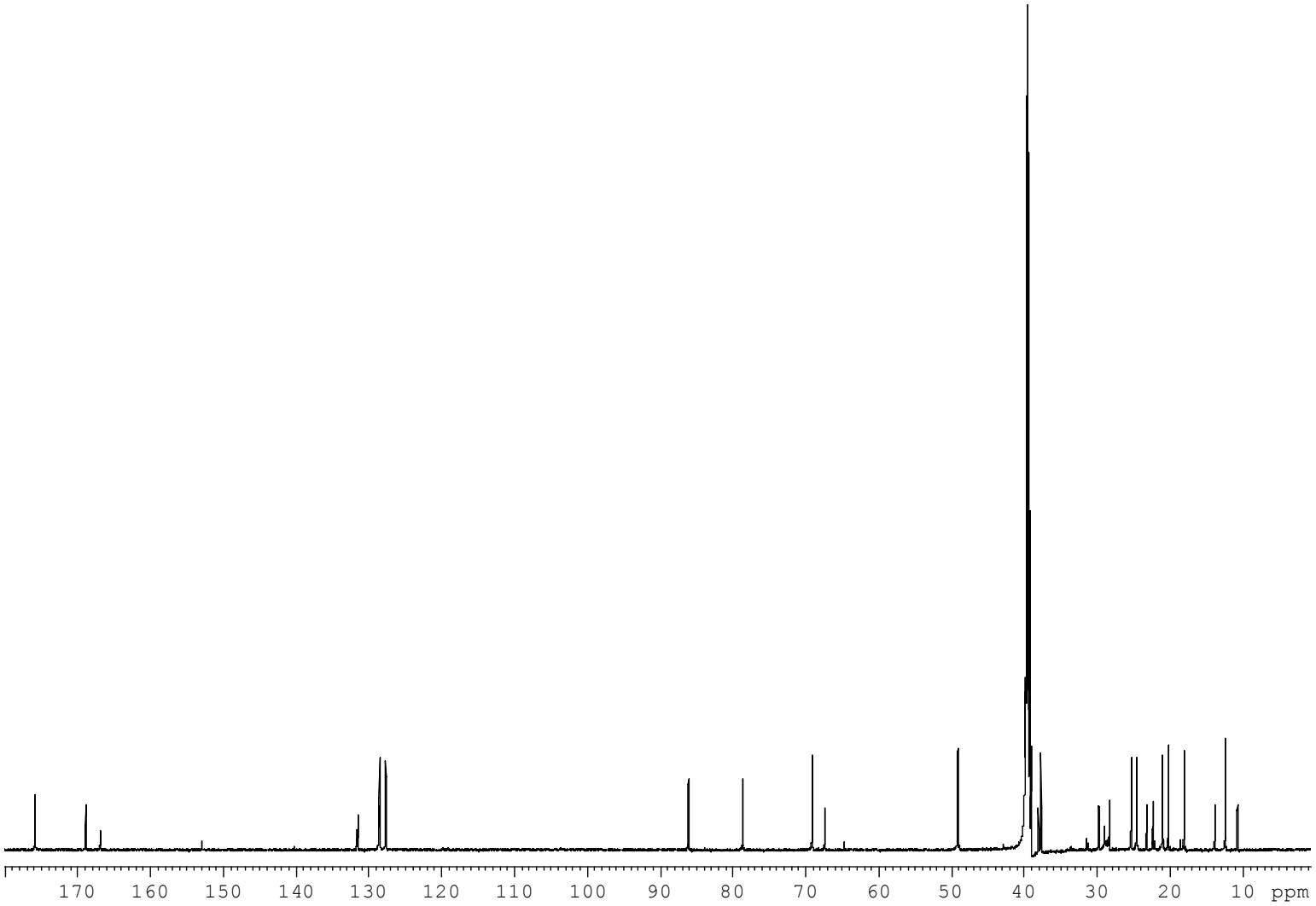
Appendix A, Figure 15. 1D ^{13}C-NMR in d_6-DMSO of [1-$^{13}C_1$]propionate enriched **5**.



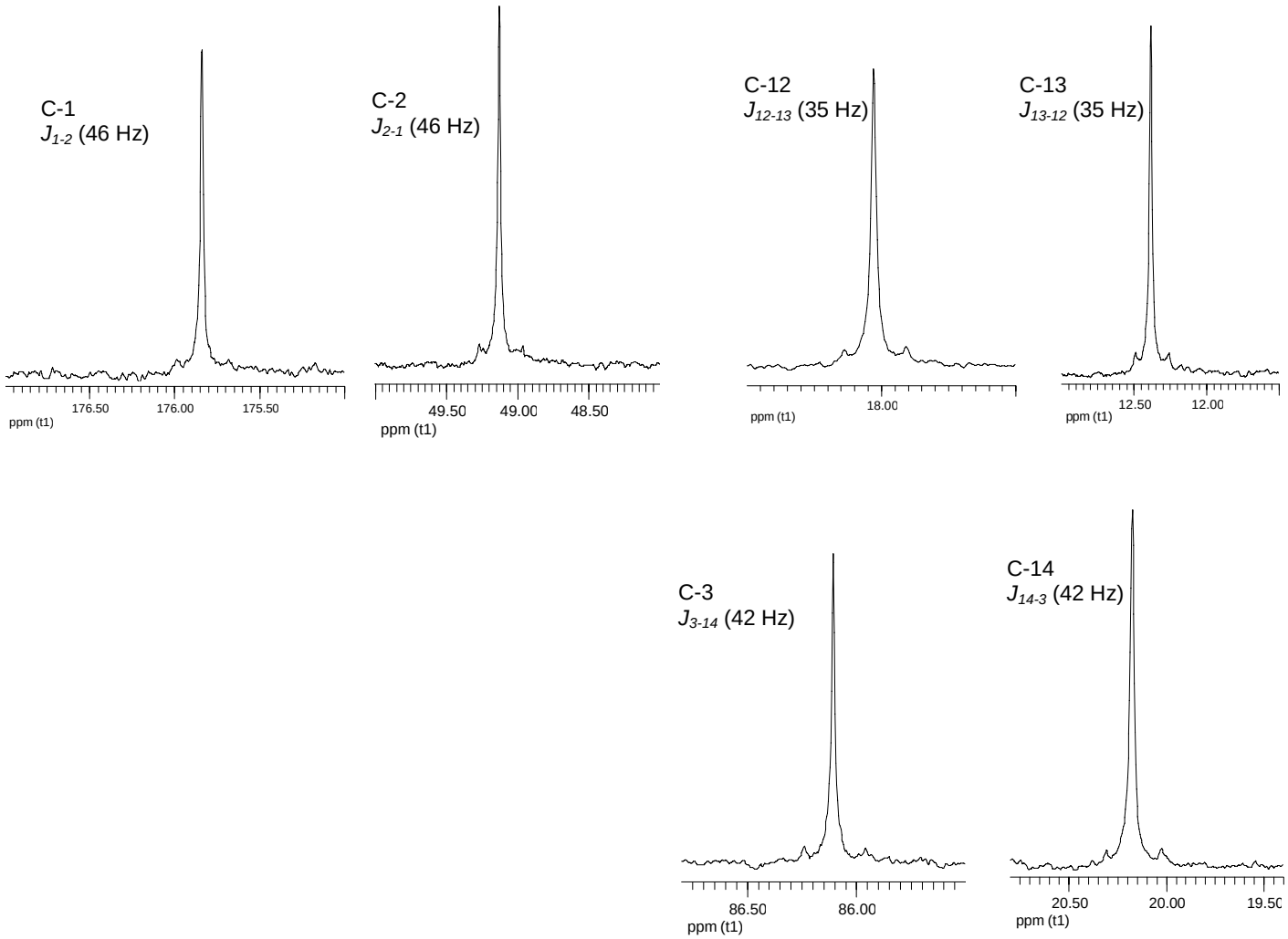
Appendix A, Figure 16. 1D ^{13}C-NMR in d_6-DMSO of [1-$^{13}C_1$]propionate enriched **6**.



Apendix A, Figure 17. 1D ^{13}C-NMR spectrum in d_6-DMSO of unlabeled **4**.



Apendix A, Figure 18. 1D ^{13}C-NMR spectrum in d_6-DMSO of [1,2-$^{13}C_2$]acetate-enriched **4**. Several minor peaks are present from minor contaminations in the sample.



Apendix A, Figure 19. Expanded 1D ^{13}C-NMR spectra of individual carbon resonances in d_6-DMSO of [1,2-$^{13}C_2$]acetate-enriched **4**.

APPENDIX B

Appendix B, Table 1. List of PCR Primers used to amplify probes for Southern hybridization

Primer name	Primer Sequence 5' - 3'
CCRF2	MGSGAGGCSHWISMIATGGACCCSCARCAIMG
CCRR2	SGGGTCSCCSGRIBDSGDSGTSCCIGTICCRTG
KSF1	CGSCACGACCTSCCSTACCACRUSAUCGGCTSSGAC
KSR1	SCGCTTSAGSSACATCCASAGGTASCGGTTGTCGTA
GN1_111_1	ACSGGSCGSCCSAAGGGSG
A8-2R	TCGATGTCSCCSAGKTCGATSCG
SAL101F1	GCSSTATCTBGGBCCBGARGG
SAL101R2	GGAASCCSGTBYGSGC
DAHPF1	CCTCACACGGCTTCAAG
DAHPR1	GTTCAACTCTGCC

Appendix B, Figure 1. Nucleotide sequences of the *sal* genes and amino acid translations.

```
>salA

1     ATG TCT GGT GAG CAA GCT GAG CAC GGT GTC CAC GCG ATC AGG GAA   45
1      M   S   G   E   Q   A   E   H   G   V   H   A   I   R   E    15

46    TTC CTG CTC ACC TGG CTT GTG GAT CTG CTC GGT GTC CCC CGG GAT   90
16     F   L   L   T   W   L   V   D   L   L   G   V   P   R   D    30

91    TCC GTG GAT CGG TGG GCC CCC CTG CAT CGG TAT GGC CTG GAC TCG   135
31     S   V   D   R   W   A   P   L   H   R   Y   G   L   D   S    45

136   CTG AAG TCG ACG GCA CTC GTC GCC GCC CTG TCC GAG TTT CTC GCC   180
46     L   K   S   T   A   L   V   A   A   L   S   E   F   L   A    60

181   CGG CCC GTA CCG ATC ACG TTC CTC TGG GAT CAC CCC ACA CTG GAA   225
61     R   P   V   P   I   T   F   L   W   D   H   P   T   L   E    75

226   GGC CTG TCC GTC GCA CTG GCG CGC GGA CGA CGA GTC GAG GAA TCC   270
76     G   L   S   V   A   L   A   R   G   R   R   V   E   E   S    90

271   GCC GGC GGC ACG GCC ACA CGA CAC CGG CCG GAA GAC GGT GCC CCA   315
91     A   G   G   T   A   T   R   H   R   P   E   D   G   A   P   105

316   CCC GAA CCG CTC GCA CTC GTG GGG ATC GGC TGC CGG CTA CCG GGG   360
106    P   E   P   L   A   L   V   G   I   G   C   R   L   P   G   120

361   GCG GCG GAC CCG TCG GCC TTC TGG CGC TTG CTC GTC AAC AGG CAC   405
121    A   A   D   P   S   A   F   W   R   L   L   V   N   R   H   135

406   GAC GCC ATA CGG CCG GTC CCG CAG CGC CGG CGT GAC CTG CTC GGC   450
136    D   A   I   R   P   V   P   Q   R   R   R   D   L   L   G   150

451   AAC GCG GAC ATC GAC GAT ACG TCC TTC CGC TGG GGC GGC TTC CTC   495
151    N   A   D   I   D   D   T   S   F   R   W   G   G   F   L   165

496   GAC GAC ATC GAC CTG TTC GAC CCT CTC TTC TTC GGG ATC TCA CCG   540
166    D   D   I   D   L   F   D   P   L   F   F   G   I   S   P   180

541   CGC GAA GCC CGT GTG ATG GAC CCG CAG CAG CGG TTG ATG CTG GAG   585
181    R   E   A   R   V   M   D   P   Q   Q   R   L   M   L   E   195

586   CTC GCC TGG GAG GCG CTG GAG GAT GCC GGC ATA GCG CCC GGC AAC   630
196    L   A   W   E   A   L   E   D   A   G   I   A   P   G   N   210

631   CTC GCC GGG ACC GAC ACC GGG GTC TTC GTG GGC ACC AGC TGG TGC   675
211    L   A   G   T   D   T   G   V   F   V   G   T   S   W   C   225

676   GAC TAC TCG GCC CTG GCG CAC GAG GCG GCA GCG CAG TTC GAA ATC   720
226    D   Y   S   A   L   A   H   E   A   A   A   Q   F   E   I   240

721   GGA CCG CAC GCG GCG ACC GGA ATG CAC GAC AGC ATC ATC GCC AAC   765
241    G   P   H   A   A   T   G   M   H   D   S   I   I   A   N   255
```

```
766   CGG GTC TCA TAC GCC CTC GCC CTG CAG GGC CCC AGC ATG GCG ATC   810
256    R   V   S   Y   A   L   A   L   Q   G   P   S   M   A   I   270

811   GAC ACC GCG TGC TCA TCC TCG CTG GTC GCG GTC CAT CTC GCG TGC   855
271    D   T   A   C   S   S   S   L   V   A   V   H   L   A   C   285

856   CAG TCA TTG TGG TCG GGC GAG AGC GAG CTC GCC CTC GCC GGC GGC   900
286    Q   S   L   W   S   G   E   S   E   L   A   L   A   G   G   300

901   GTC ACT CTC AAT ATC TTT GGG CCA CAC TAC CTG GCC ATG AAC GAG   945
301    V   T   L   N   I   F   G   P   H   Y   L   A   M   N   E   315

946   ACA GGC GCT CTC TCG CCG GAC GGC CGG TGC AAG ACC TTC GAC GCC   990
316    T   G   A   L   S   P   D   G   R   C   K   T   F   D   A   330

991   CGT GCC GAC GGC ACC GTA CGG GGC GAA GGA GCC GCC GTC GTC GTC  1035
331    R   A   D   G   T   V   R   G   E   G   A   A   V   V   V   345

1036  GTG ACG CCC CTG CGC GTG GCG CTG GAA CGC GGG CTG CCC GTC TAC  1080
346    V   T   P   L   R   V   A   L   E   R   G   L   P   V   Y   360

1081  TGC CTG ATC CGC GGC AGC GCC GTC AAC AAC GAT GGC CTG AGC AAC  1125
361    C   L   I   R   G   S   A   V   N   N   D   G   L   S   N   375

1126  GGG CTC ACC GCC CCG AAC CCA CGT GCT CAG GAG ACC CTA CTA CGT  1170
376    G   L   T   A   P   N   P   R   A   Q   E   T   L   L   R   390

1171  ACC GCG TAC CGC CGG GCC GGC TTG GCA CCA TGG CTC GTG GAC TAT  1215
391    T   A   Y   R   R   A   G   L   A   P   W   L   V   D   Y   405

1216  GTG GAG TGC CAC GGC ACA GGC ACA CCG CTG GGT GAC CCC ATC GAG  1260
406    V   E   C   H   G   T   G   T   P   L   G   D   P   I   E   420

1261  GCC AAG GCC CTC GGT ACG GTG CTC GGT CGC GAT CGC GAA CCA GGT  1305
421    A   K   A   L   G   T   V   L   G   R   D   R   E   P   G   435

1306  AAC GAG CTC CGC ATC GGT TCG GTG AAG ACG AAC ATC GGC CAC CTG  1350
436    N   E   L   R   I   G   S   V   K   T   N   I   G   H   L   450

1351  GAG CCG GCG GCC GGC GTC GCC GGT CTC GCC AAG GTC GCC CTG GCG  1395
451    E   P   A   A   G   V   A   G   L   A   K   V   A   L   A   465

1396  ATG CGT AAC GGC ATC CTG CCC GCC AGC CTG CAC TAC AGC CGC CCC  1440
466    M   R   N   G   I   L   P   A   S   L   H   Y   S   R   P   480

1441  AAC CCT CAG GTT CCC TTC GCC GAA GGT CAC CTG CGC GTC CAG GAT  1485
481    N   P   Q   V   P   F   A   E   G   H   L   R   V   Q   D   495

1486  GCG ACA ACC ACG TGG CAT CGC CGG TCA GAT CTA CGG CGC GCG GGA  1530
496    A   T   T   T   W   H   R   R   S   D   L   R   R   A   G   510

1531  ATC AGC GCC TTT GGC TTC GGC GGG ACG AAC TGC CAC CTG GTG ACA  1575
511    I   S   A   F   G   F   G   G   T   N   C   H   L   V   T   525

1576  GAG GAA CTG GCG CAG CCT GAG GCC GCC CTA TTG CTG CTA GCG GCG  1620
526    E   E   L   A   Q   P   E   A   A   L   L   L   L   A   A   540

1621  GAT TCC ACG CAG GAG CTC ACC AAC CAG GTG GCT GCG CTG CGC TCC  1665
541    D   S   T   Q   E   L   T   N   Q   V   A   A   L   R   S   555
```

```
1666  GCT CTC GCG GAG GAC CGA GTC GAC CTG GCT GCG GCC TGC CGG CAG   1710
556    A   L   A   E   D   R   V   D   L   A   A   A   C   R   Q    570

1711  GCG CTC GAC ACG CTG GGC ACC GGC CGT TTC CGG CTG TCG GCA GCC   1755
571    A   L   D   T   L   G   T   G   R   F   R   L   S   A   A    585

1756  GCG CGT ACC ACA GCC GAG CTG CGG GCG CAG CTG GAC GCG TAC GCC   1800
586    A   R   T   T   A   E   L   R   A   Q   L   D   A   Y   A    600

1801  GAG GGC CGG CCC CTA CCC GGC CTG TCC ACT GCG GAG ACA GCT CTT   1845
601    E   G   R   P   L   P   G   L   S   T   A   E   T   A   L    615

1846  CGT CGA CCC CGC GTC GCG TTT CTC TGC TCC GGA ACC GGC TCC CAG   1890
616    R   R   P   R   V   A   F   L   C   S   G   T   G   S   Q    630

1891  TGG TTC GGC ATG TGC CGA CAG CTG GTT GCT GGC ATG CCC GCG TTC   1935
631    W   F   G   M   C   R   Q   L   V   A   G   M   P   A   F    645

1936  CGG CGA TCG CTG GTG CGG GCG GCT GAG CGG ATA GAA CGA ATC CTC   1980
646    R   R   S   L   V   R   A   A   E   R   I   E   R   I   L    660

1981  GGC TTC TCA GTG CTC GAC CGG CTC TTC GAC GAC GAG TCG CAC GCC   2025
661    G   F   S   V   L   D   R   L   F   D   D   E   S   H   A    675

2026  CGG CTC GAC GAC ATG GAG ATC GTC CAG CCG ATG CTG TTC TGC GTC   2070
676    R   L   D   D   M   E   I   V   Q   P   M   L   F   C   V    690

2071  CAG GTC GCC CTG GCC GAC GCC TGG CGT TCG CTC GGT GTG GAG CCC   2115
691    Q   V   A   L   A   D   A   W   R   S   L   G   V   E   P    705

2116  GAT CTG GTA ATC GGC CAG AGT GTC GGC GAA ATC GCA GCC GCC CAC   2160
706    D   L   V   I   G   Q   S   V   G   E   I   A   A   A   H    720

2161  CTC ACG GGT GCG CTC AGC CTC GAC GAC GCG GCC TTG GTA GCG GCC   2205
721    L   T   G   A   L   S   L   D   D   A   A   L   V   A   A    735

2206  ACG CAC GCA CGG CTC GTG CAA CGC CTA GCC GTG GGT CGC GGC GAT   2250
736    T   H   A   R   L   V   Q   R   L   A   V   G   R   G   D    750

2251  AGC CTG GTC GTC GCG GCC GTC CCC GAA GTT GTC GCC GGG CAC CTA   2295
751    S   L   V   V   A   A   V   P   E   V   V   A   G   H   L    765

2296  GCA GCG GTC GAG GAG CGA CCC ACG CTA GCT GGA CGC CTC GGC CCG   2340
766    A   A   V   E   E   R   P   T   L   A   G   R   L   G   P    780

2341  GCC TCT ACG CTG GTC TCA GGG GAT ACC ACC GCG ATC CGG GCG CTG   2385
781    A   S   T   L   V   S   G   D   T   T   A   I   R   A   L    795

2386  ACC GAC CGG TTC GCG GCA GCG GAC ATC GCA GCC CAC CGG GTA CGC   2430
796    T   D   R   F   A   A   A   D   I   A   A   H   R   V   R    810

2431  ATG GGT TAC GCA TCC CAC TCG CCG CTG ATG GAC CCC GTG CTC ACA   2475
811    M   G   Y   A   S   H   S   P   L   M   D   P   V   L   T    825

2476  CCC CTG CGG TCG GAG ATC AAC GGC ATC GTC GCG CGC CCG GCT CAG   2520
826    P   L   R   S   E   I   N   G   I   V   A   R   P   A   Q    840

2521  CTG CCG ATG ATC TCG ACG GTC ACC GGG AAA GAA ATC GAC GGC GAA   2565
841    L   P   M   I   S   T   V   T   G   K   E   I   D   G   E    855
```

```
2566  TCC CTG GGC CCG GAC TAT TGG CCG GAC AAC GTC CGC CAG GAG AGC   2610
856    S   L   G   P   D   Y   W   P   D   N   V   R   Q   E   S    870

2611  CGG CTC ACG GCC GCC CTG GAC ACC ATG GCG GCG CAC GAC ATC GAC   2655
871    R   L   T   A   A   L   D   T   M   A   A   H   D   I   D    885

2656  GTC GTA CTG GAG CTC AGT CCG CAC CCG GTC CTA CTC AAG GGC GTA   2700
886    V   V   L   E   L   S   P   H   P   V   L   L   K   G   V    900

2701  CGG GCA AGC CTC GAC AGC GAA CGG GAG GGC GCG ACG AGG CGC CCC   2745
901    R   A   S   L   D   S   E   R   E   G   A   T   R   R   P    915

2746  GTC ATG TGC CTG GCG TCG CTT CAA CGA GGG ACG GAC GAG GCG TGG   2790
916    V   M   C   L   A   S   L   Q   R   G   T   D   E   A   W    930

2791  AGC CTA CTC GCC TCG CTG GGT GAA CTG CAT ACC GTG GGT CAA CCG   2835
931    S   L   L   A   S   L   G   E   L   H   T   V   G   Q   P    945

2836  GTC GCC GCC GGA TCC TTC CTG TGC GCA CCA CGC GGG CGT CAT ACC   2880
946    V   A   A   G   S   F   L   C   A   P   R   G   R   H   T    960

2881  GAG CCC CAG GCG GCC ATC GGT TCC GCG ACG GCG CGG GCC GAG GAG   2925
961    E   P   Q   A   A   I   G   S   A   T   A   R   A   E   E    975

2926  ACG CCC CTT TCC CTG CTG CCC GTA ACC GCC CAT TCC GCT GAC GCG   2970
976    T   P   L   S   L   L   P   V   T   A   H   S   A   D   A    990

2971  CTA CGG GAC ACC TGC CGA GAG CTG TCC AAC CAT GTC GAG CGA AAT   3015
991    L   R   D   T   C   R   E   L   S   N   H   V   E   R   N   1005

3016  GCC GCA CCA TGG CTG CCG GAC CTG GCG TAC ACC CTG GCG ACC CGG   3060
1006   A   A   P   W   L   P   D   L   A   Y   T   L   A   T   R   1020

3061  CGC ACC CCA CTT CCA CAC CGG ATC GCT TTC GTG GTG CGT GAT CGC   3105
1021   R   T   P   L   P   H   R   I   A   F   V   V   R   D   R   1035

3106  GAC GAT CTA CTC GAC GGC CTG GCG CAT ATT TCC GCT GGC AGG TCG   3150
1036   D   D   L   L   D   G   L   A   H   I   S   A   G   R   S   1050

3151  TAC CCC GGC GCC GTC AAG GGC ACC GTC GCT GGC GGC GGC GCC CGG   3195
1051   Y   P   G   A   V   K   G   T   V   A   G   G   G   A   R   1065

3196  CGC GTA GCA CTG GTG TTC TCC GGC GGA GGC ACG CAC TGG GCC GGC   3240
1066   R   V   A   L   V   F   S   G   G   G   T   H   W   A   G   1080

3241  ATG GGA CGT GCC TTG ATG GGC TGG CAC GCG GGC TTC CGG GCA TCG   3285
1081   M   G   R   A   L   M   G   W   H   A   G   F   R   A   S   1095

3286  ATG CAC GAG TGC GAC GCG GTG TTC CGG GAG CTG ATC GGA TGG TCG   3330
1096   M   H   E   C   D   A   V   F   R   E   L   I   G   W   S   1110

3331  GTA GTC GAC GAG CTC AGC CTG CCG GCG GAG CGG TCC CGG CTG GAC   3375
1111   V   V   D   E   L   S   L   P   A   E   R   S   R   L   D   1125

3376  GCG ACC GAT ATC CAG CAG CCG GTG CTG TTC ACC CTG CAG GTC TCA   3420
1126   A   T   D   I   Q   Q   P   V   L   F   T   L   Q   V   S   1140

3421  CTG GCC CGT CTC TGG ATG GAG CTT GGC ATC GAG CCG GAG GCA TTC   3465
1141   L   A   R   L   W   M   E   L   G   I   E   P   E   A   F   1155
```

```
3466  GTC GGA CAC AGC ATC GGC GAG GTC GCC GCC GTC TGT GTG GCG GGC   3510
1156   V   G   H   S   I   G   E   V   A   A   V   C   V   A   G    1170

3511  GGG CTG TCG GTG CGT GAC GCT GCC CGG GTC ACC ATC GCC CGC TCC   3555
1171   G   L   S   V   R   D   A   A   R   V   T   I   A   R   S    1185

3556  CAC CTT ATC CAG CAC CGC GCC GCG AAG GCC GCC ATG ATC GCC GTC   3600
1186   H   L   I   Q   H   R   A   A   K   A   A   M   I   A   V    1200

3601  CAA GCC GGA GAC GAA GAG ATC ATC CCG TTC CTC GCC CCG TAC GGC   3645
1201   Q   A   G   D   E   E   I   I   P   F   L   A   P   Y   G    1215

3646  GGG CGC GTC GCC ATC GCG GCG CTC AAT AGC CCA ACC AGC TCC GCG   3690
1216   G   R   V   A   I   A   A   L   N   S   P   T   S   S   A    1230

3691  GTT TCC GGC CCG CCC GAG GAG ATT CGC GCC CTG GAG GTC GCA CTA   3735
1231   V   S   G   P   P   E   E   I   R   A   L   E   V   A   L    1245

3736  AAT CGA GCC GGC ATC TCG AGC CGA GCC GTA CGC GTT GAT CGC CCC   3780
1246   N   R   A   G   I   S   S   R   A   V   R   V   D   R   P    1260

3781  GGT CAC AGC CCT GGT ATG GAT CCG CTG CTG TCA CCG CTG CGT GAA   3825
1261   G   H   S   P   G   M   D   P   L   L   S   P   L   R   E    1275

3826  GCG CTC ACC AAC ATT GAA CCT CGT GCA TTT TGG GCC AGA TTC CAC   3870
1276   A   L   T   N   I   E   P   R   A   F   W   A   R   F   H    1290

3871  TCG ACG GCG CTC GAC GGC GCG GTC GAT CCC GTG GTC AAC GCA GAC   3915
1291   S   T   A   L   D   G   A   V   D   P   V   V   N   A   D    1305

3916  TAC TGG GCG CAC AAC CTG CGC AAC CAG GTG CGC TTC GCG CCG ACG   3960
1306   Y   W   A   H   N   L   R   N   Q   V   R   F   A   P   T    1320

3961  GTG GCG GCT CTG GCC GAC GCG GGC ATT GAC ACG TTT GTT GAG ATC   4005
1321   V   A   A   L   A   D   A   G   I   D   T   F   V   E   I    1335

4006  AGC CCG CAT GGC ACC CTA CGC GGC GCG ATC GAG GAG ATC GCT CAG   4050
1336   S   P   H   G   T   L   R   G   A   I   E   E   I   A   Q    1350

4051  GCT CAG GGA GCC TCT GTT GTC GTG GCC GAC TCG ATC CGC CGC GGG   4095
1351   A   Q   G   A   S   V   V   V   A   D   S   I   R   R   G    1365

4096  GAA GAC GAC AAT CGC TGC TTC CTC AAC GCG GCT GCA TCG CTG TTC   4140
1366   E   D   D   N   R   C   F   L   N   A   A   A   S   L   F    1380

4141  GTA CAT GGC GTG CCC CTG TCG CTC GAG ACG CTG TTC TCG TCC GAC   4185
1381   V   H   G   V   P   L   S   L   E   T   L   F   S   S   D    1395

4186  GCG CAG GTT GTC GAG ACT CCG CTC GTG AGC TGG CAG AGG CAC CCC   4230
1396   A   Q   V   V   E   T   P   L   V   S   W   Q   R   H   P    1410

4231  TAC TGG CTG GAC GCC GCG CCC GTC CGC CAA CCA CGC ACG CCC GCT   4275
1411   Y   W   L   D   A   A   P   V   R   Q   P   R   T   P   A    1425

4276  GCG GCA TCC GTC GCC GGC CCG ACC ACC AGC TTC CAC AAT GAG AAC   4320
1426   A   A   S   V   A   G   P   T   T   S   F   H   N   E   N    1440

4321  GCC ACC ACC CCG ACA GCG GAA GCC ACC ATC ACC GGC GAG ATC GCG   4365
1441   A   T   T   P   T   A   E   A   T   I   T   G   E   I   A    1455
```

```
4366  CAC GTT CTG GGC ATG CCG GTC GAG GAG CTC GCC GAG GGC ACG CGG   4410
1456   H   V   L   G   M   P   V   E   E   L   A   E   G   T   R   1470

4411  CTG CGG GAC TTC GGG CTC GAC TCC ATG CTC GCC ATG CGG CTG GGC   4455
1471   L   R   D   F   G   L   D   S   M   L   A   M   R   L   G   1485

4456  AGC AGA ATC CAG ACC CTG TTC GGC CAT CGT GTC TCG CTG TTT GAG   4500
1486   S   R   I   Q   T   L   F   G   H   R   V   S   L   F   E   1500

4501  TTC TTC ACG GAC CGG ACG GTT GGC GAA CTC ACC GCG CGC ATC GCC   4545
1501   F   F   T   D   R   T   V   G   E   L   T   A   R   I   A   1515

4546  GAG CTC GTC GGT GCG CGC GCG TCC ACC GGC GGC GAG GAC AAC GCG   4590
1516   E   L   V   G   A   R   A   S   T   G   G   E   D   N   A   1530

4591  CCG CCG CGG CTG CGG AAT ACG GCT GTT GCG GAC AAT CTG TCC GAC   4635
1531   P   P   R   L   R   N   T   A   V   A   D   N   L   S   D   1545

4636  GCC GAC GCC GAG GAG TTG CTC GCG GAA CTG ACC GAT CGT GGG CTG   4680
1546   A   D   A   E   E   L   L   A   E   L   T   D   R   G   L   1560

4681  CTG GAG CCG GTT TCC CCG GAG GCG CCC GCG GCG CAA CCA CGC GAA   4725
1561   L   E   P   V   S   P   E   A   P   A   A   Q   P   R   E   1575

4726  GAG CTG CGG GCG GCG CTA GAA CGC ACC CAA TCC TTT GCG CTT GCG   4770
1576   E   L   R   A   A   L   E   R   T   Q   S   F   A   L   A   1590

4771  CCG GTG GCG CAC GGT CAG GCC GCC ATT TGG TTC ATG CAG CAG CTG   4815
1591   P   V   A   H   G   Q   A   A   I   W   F   M   Q   Q   L   1605

4816  TCG CTC GAC GGC GCA GCC TAC AAC CTC ATG TTC GGC GCT CGG GTG   4860
1606   S   L   D   G   A   A   Y   N   L   M   F   G   A   R   V   1620

4861  CCC GAT AAG ATC GAC GAG TCG GCC CTG CGC CGG GCC GCA GCC GCG   4905
1621   P   D   K   I   D   E   S   A   L   R   R   A   A   A   A   1635

4906  GTG GTG GAG CGA CAT CCC GCG CTG CGC ACC GTC TTC GTC GAA GCC   4950
1636   V   V   E   R   H   P   A   L   R   T   V   F   V   E   A   1650

4951  GGC GGC CAC CCC TAC CAG GTG ATA AAG GCG GAT CCC GGT TAC GAA   4995
1651   G   G   H   P   Y   Q   V   I   K   A   D   P   G   Y   E   1665

4996  TTC GAC ACG GTC GAC GGT ACG GGG CTC GAT GAC GCG GCA CTG ACC   5040
1666   F   D   T   V   D   G   T   G   L   D   D   A   A   L   T   1680

5041  GAC CTC CTC GCC GAG CAC GGC CAC CGG CCC TTC GAT CTC GAT CAG   5085
1681   D   L   L   A   E   H   G   H   R   P   F   D   L   D   Q   1695

5086  GGT CCG CTA TTG CGG TTG GTG TTG GTC AAC CGC GGT GCG GCA GAC   5130
1696   G   P   L   L   R   L   V   L   V   N   R   G   A   A   D   1710

5131  AAC TGC CTG CTG CTG GTC ATC CAC CAT GTG GCA GCT GAC GCG GCG   5175
1711   N   C   L   L   L   V   I   H   H   V   A   A   D   A   A   1725

5176  TCG GTG GAC ATC GTC GTC CGG GAC CTG CGC GAG TTC TAC GGC GAG   5220
1726   S   V   D   I   V   V   R   D   L   R   E   F   Y   G   E   1740

5221  GCA CAG CGC GGA ACG CTG CTC ACC CAG GCA CCC GAG ACA CCG TAC   5265
1741   A   Q   R   G   T   L   L   T   Q   A   P   E   T   P   Y   1755
```

```
5266  ACC GAG TTC GTC GAA TGG GAA CGC GAA TGG CTC GGC AGC CCG GCG   5310
1756   T   E   F   V   E   W   E   R   E   W   L   G   S   P   A    1770

5311  GCC GAG GCG GCG CTG CAC TGG TGG TCG CAA CAG TTG GCG CAG CCC   5355
1771   A   E   A   A   L   H   W   W   S   Q   Q   L   A   Q   P    1785

5356  CCG GCG CAC CTC GAC CTT CCA CGC CTG TCA GCG CCC GAC GCG ACC   5400
1786   P   A   H   L   D   L   P   R   L   S   A   P   D   A   T    1800

5401  GCC GAC AAG CGA CCA GCC AAA GAG CGG GAC GGT GTG GTG AGC TAC   5445
1801   A   D   K   R   P   A   K   E   R   D   G   V   V   S   Y    1815

5446  GCG GGG GAG GAT GCC AGC TTC CGG TGG GAC GCT GCC GAT ACC CGC   5490
1816   A   G   E   D   A   S   F   R   W   D   A   A   D   T   R    1830

5491  CTA CTG AAG GAC TTC GCC GTG CGC GCC GGC GTC TCG GTC AGC ACC   5535
1831   L   L   K   D   F   A   V   R   A   G   V   S   V   S   T    1845

5536  CTT GTC ATG GCA GGC TTC GTG GCC ACC CTG GGC CGG GCA ACC GGG   5580
1846   L   V   M   A   G   F   V   A   T   L   G   R   A   T   G    1860

5581  GCG AGG GAC ATC GTC CTG GGC ACG GCG GTC GCC CAG CGC AGT GAT   5625
1861   A   R   D   I   V   L   G   T   A   V   A   Q   R   S   D    1875

5626  GCG GGG CGC GAG TCC GCG GTG GGG TAC TAC CTC AAC ACC ATC CCA   5670
1876   A   G   R   E   S   A   V   G   Y   Y   L   N   T   I   P    1890

5671  GTC CGC GCC CGG CCC CGT CCG GAC ATC ACC TTC GAC GCG TTG CTG   5715
1891   V   R   A   R   P   R   P   D   I   T   F   D   A   L   L    1905

5716  CGT GAA GTC CAC GAC TTC GCG CTC GGG ATG TTG GAA CAC ATG AAC   5760
1906   R   E   V   H   D   F   A   L   G   M   L   E   H   M   N    1920

5761  TAT CCA CTG GAC CTG CTT GTC TCG GCG CTC AAA CCG CCG CGG CGG   5805
1921   Y   P   L   D   L   L   V   S   A   L   K   P   P   R   R    1935

5806  CAG GGC CGC ACG CCG TTG TTC GAC ATC GCT GTC AAC TGG CTC TCC   5850
1936   Q   G   R   T   P   L   F   D   I   A   V   N   W   L   S    1950

5851  GGT GAC GCG TTC ACC TAT GCG AAC ACA TTG TTC CAC GGT GCT GGC   5895
1951   G   D   A   F   T   Y   A   N   T   L   F   H   G   A   G    1965

5896  GTG GCG GCG TGG CCT GCC GGC CCC CTA CCA CTG GTC CCG CTG CCT   5940
1966   V   A   A   W   P   A   G   P   L   P   L   V   P   L   P    1980

5941  CTC CGG CGG CAC ATC GCA AAG TTC GAT CTC GAA ATC ACG ATG GGC   5985
1981   L   R   R   H   I   A   K   F   D   L   E   I   T   M   G    1995

5986  GAC GTC GCC GAC GAG GTG GTC GGT CAA GTG CAG TTC AAA CCC GCG   6030
1996   D   V   A   D   E   V   V   G   Q   V   Q   F   K   P   A    2010

6031  TTC CTG GAA AGA GAA ACC GTG ACG ACA CTG CTG GAG TAC TTC CAC   6075
2011   F   L   E   R   E   T   V   T   T   L   L   E   Y   F   H    2025

6076  CGC GTG CTG TTG CAG GCG ATC GAC CGG CCG GAT GTG CCG CTG AGT   6120
2026   R   V   L   L   Q   A   I   D   R   P   D   V   P   L   S    2040

6121  GAC CTC GTG CTG GAG ACC TCG GTG AAG CAG GCG GTC TCA TGA   6162
2041   D   L   V   L   E   T   S   V   K   Q   A   V   S   *
```

>salB

```
1     ATG ACG GCC GCA TCG ACA CTG CAC GGC GGC TTT GTC GCC CAC GCA   45
1      M   T   A   A   S   T   L   H   G   G   F   V   A   H   A    15

46    GCC GCG AAT CCG GAC ACG CTG GCC GTA GCG TCC GAC GCC GGC GTG   90
16     A   A   N   P   D   T   L   A   V   A   S   D   A   G   V    30

91    ATG ACC TAC GGT CGG CTT GAC GAG ACC TCG GCG GCG CTG GCC GAG   135
31     M   T   Y   G   R   L   D   E   T   S   A   A   L   A   E    45

136   CGG CTG TCT GCC TTG GGT GCT GGC CCG GGT GTT CCG ATC GGG GTC   180
46     R   L   S   A   L   G   A   G   P   G   V   P   I   G   V    60

181   TGT ATC GAA CGC ACG CCG GAC CTG CTC GTC GCT ATC CTC GGC GTG   225
61     C   I   E   R   T   P   D   L   L   V   A   I   L   G   V    75

226   CTG CGC GCG GGC GCC TGC TAT CTG CCG CTC GAT CCT CAA TAT TCA   270
76     L   R   A   G   A   C   Y   L   P   L   D   P   Q   Y   S    90

271   GCG CGC CAC CTC GGC TTC ATG GTG GCC GAC AGC GGG ACC CGC CTG   315
91     A   R   H   L   G   F   M   V   A   D   S   G   T   R   L   105

316   GTC GTT ACC ACA CGA TCC TCT CGG GAC GCG TGC CCG GAC GGC TGC   360
106    V   V   T   T   R   S   S   R   D   A   C   P   D   G   C   120

361   ACC GCG CTC GTC CTG GAG GAA TCC GAG GCG ATA GCC GAC CCG CCG   405
121    T   A   L   V   L   E   E   S   E   A   I   A   D   P   P   135

406   CCA GTG GCC GCG GTT CCG GAC GAT TCT GCC TAC GTC ATA TAT ACC   450
136    P   V   A   A   V   P   D   D   S   A   Y   V   I   Y   T   150

451   TCC GGC TCG ACC GGC ACG CCC AAG GGG GTG CCG ATC CGG CAC AGC   495
151    S   G   S   T   G   T   P   K   G   V   P   I   R   H   S   165

496   AGC TGC GCG GCG ATG CTT GCC GAG GCG GAC CGA ATT TTC GAG GGC   540
166    S   C   A   A   M   L   A   E   A   D   R   I   F   E   G   180

541   TGT GAC ATG AGC GGT ATC GCC GCC GTC ACC TCG GTC TGC TTC GAC   585
181    C   D   M   S   G   I   A   A   V   T   S   V   C   F   D   195

586   CTG TCA GTG CTG GAG ATC TTC TCC GCC CTC AGC CGT GGC CGG ACG   630
196    L   S   V   L   E   I   F   S   A   L   S   R   G   R   T   210

631   CTC GTC CTG GTG AAT AGT GCC AGC CAC CTT CCG GAG AGC TCC CAT   675
211    L   V   L   V   N   S   A   S   H   L   P   E   S   S   H   225

676   GTC GAA CGG GTG ACG CAC GTC AGC ACG GTC CCG TCC GCA ATG ACC   720
226    V   E   R   V   T   H   V   S   T   V   P   S   A   M   T   240

721   AGC CTG CTT GAC GCG CAA GCC GTT CCG GCC GGC CTG CGG AAC GTG   765
241    S   L   L   D   A   Q   A   V   P   A   G   L   R   N   V   255

766   GTG CTC GGC GGC GAA CCC GTA CGT CGG AGC CTG GTC GAC CGG ATC   810
256    V   L   G   G   E   P   V   R   R   S   L   V   D   R   I   270

811   TAC CGC GAG ACC AAC GTC GAC TTC GTC TTC AAC GGA TAC GGC CCG   855
271    Y   R   E   T   N   V   D   F   V   F   N   G   Y   G   P   285
```

```
856   ACG GAA GGC ACG GTC TTC TGT ACC TTC AAG CCC GTA TCC CGC GAC   900
286    T   E   G   T   V   F   C   T   F   K   P   V   S   R   D   300

901   GAG GCC GGC GAG CCG TCG ATC GGT ACG CCA TCC CTG ACC GCT CGC   945
301    E   A   G   E   P   S   I   G   T   P   S   L   T   A   R   315

946   GTC TAT GTG CTC GAC GAG AAG CTG CGG CCG TCG GCC GTC GGC GAG   990
316    V   Y   V   L   D   E   K   L   R   P   S   A   V   G   E   330

991   TCG GGT GAG CTG TAC CTC GGC GGT GCC GGA CTT ACC TGG GGC TAC   1035
331    S   G   E   L   Y   L   G   G   A   G   L   T   W   G   Y   345

1036  CTC AAC CGG CCC GGG CTG ACT GCG GAA CGG TTC GTA CCT GAT CCG   1080
346    L   N   R   P   G   L   T   A   E   R   F   V   P   D   P   360

1081  CAG GTG GCG GGT GAA CGC ATG TAT CGC ACC GGC GAC ATC GCT CGG   1125
361    Q   V   A   G   E   R   M   Y   R   T   G   D   I   A   R   375

1126  CTC AAC GAA GCA GGT GAA ATC GAG TTC GTG GGA CGC TCC GAC CTT   1170
376    L   N   E   A   G   E   I   E   F   V   G   R   S   D   L   390

1171  CAG GTG AAG GTC CGC GGG TAC CGC ATC GAG CTA GAA GAG GTC GAG   1215
391    Q   V   K   V   R   G   Y   R   I   E   L   E   E   V   E   405

1216  GCA CGA CTG ACC GAA TGC CCC GAG GTG CGG ACG GCT GCG GCC GTC   1260
406    A   R   L   T   E   C   P   E   V   R   T   A   A   A   V   420

1261  GTC CGT GAG CAG ACG CCG GGT ACG AGA GCC CTG ACC GCG TAC GCG   1305
421    V   R   E   Q   T   P   G   T   R   A   L   T   A   Y   A   435

1306  GTT CCG GCG AGT GGA GCA CCC GAC GGC GAC GGG CCC TGG CTC GAC   1350
436    V   P   A   S   G   A   P   D   G   D   G   P   W   L   D   450

1351  GCC GAC CTG CAG GCA ACG ATC AAG CAA CAG CTC GGT GCG CTG TTG   1395
451    A   D   L   Q   A   T   I   K   Q   Q   L   G   A   L   L   465

1396  CCC GGT TAC ATG GTT CCC GAA ACG ATC GTC TTC CTG CCC GCG CTC   1440
466    P   G   Y   M   V   P   E   T   I   V   F   L   P   A   L   480

1441  CCG CTG TCG CCA GTT GGG AAG CTG GAC CGC ACG GCG CTA CCG GCA   1485
481    P   L   S   P   V   G   K   L   D   R   T   A   L   P   A   495

1486  CCA CCC GTT GTC GAT GTG CTG CCC TCG GGG GAC TCC GCC ACC ACC   1530
496    P   P   V   V   D   V   L   P   S   G   D   S   A   T   T   510

1531  GAC ACC GAA CAA GCG CTT GCC GAG ATC TGG GGT GCA CTG CTG GAC   1575
511    D   T   E   Q   A   L   A   E   I   W   G   A   L   L   D   525

1576  CGG ACT CCG CAG TCC ATC GGC ATC CGC GAC ACA TTC TAC GAC CTC   1620
526    R   T   P   Q   S   I   G   I   R   D   T   F   Y   D   L   540

1621  GGC GGC AAC TCT TTG TTG TTG GTG CGG CTC GCG AAG CGA ATG GGT   1665
541    G   G   N   S   L   L   L   V   R   L   A   K   R   M   G   555

1666  CAG CGC TTT CAC CGC AAG GTC GGC GTG GCG GAC CTG TTC CGG TTC   1710
556    Q   R   F   H   R   K   V   G   V   A   D   L   F   R   F   570

1711  CGC GAC ATC GGC TCG CTC GCC AAG TGG CTG GAC GAC GAG AGC GGA   1755
571    R   D   I   G   S   L   A   K   W   L   D   D   E   S   G   585
```

```
1756  AAG AGT CCT GAG GAC ATC GAG CAG GCA CGA CGC CGT GCC AGC ACC   1800
586    K   S   P   E   D   I   E   Q   A   R   R   R   A   S   T    600

1801  AGA CGC TCG GTG GTG CGC GGC CAC AGC AGA TCA CCG AGC ACT CGA   1845
601    R   R   S   V   V   R   G   H   S   R   S   P   S   T   R    615

1846  ACC GAC CCG ACC GTC AAG AAC ACG CCC GCA TCA AAT GGA GGC CCA   1890
616    T   D   P   T   V   K   N   T   P   A   S   N   G   G   P    630

1891  CAT CCA TGA   1899
631    H   P   *
```

>salC

```
1     ATG AGC GAA GCC GTT GAG CGG AAC TAC GAT CCG GGA GAC GTC GCC   45
1      M   S   E   A   V   E   R   N   Y   D   P   G   D   V   A    15

46    ATC ATC GGC GCC TCG TGC AGA TTC CCT GGC GCA CGA AAC AAG GAA   90
16     I   I   G   A   S   C   R   F   P   G   A   R   N   K   E    30

91    CAG TAC TGG GAC AAC CTT CTC CAT GGG CGC GAG GGC GTG ACC TTT   135
31     Q   Y   W   D   N   L   L   H   G   R   E   G   V   T   F    45

136   TAC GCC AAG GAC GAG ATC GAG GTG GAC GAG ACG CTC ATC AAC AGT   180
46     Y   A   K   D   E   I   E   V   D   E   T   L   I   N   S    60

181   CCG GCG TAC GTG CGG GCC ACC GGA GCA CTG GAC GGA TAC GAC GAG   225
61     P   A   Y   V   R   A   T   G   A   L   D   G   Y   D   E    75

226   TTC GAT CCC GCC GTG TTC GGT GTT AGC GAT CGA ATG GCG GCG GCG   270
76     F   D   P   A   V   F   G   V   S   D   R   M   A   A   A    90

271   ATG ACT CCC GAG CAC CGG GTC TTC CTG GAG GCA TCC TGG GAG GTG   315
91     M   T   P   E   H   R   V   F   L   E   A   S   W   E   V   105

316   ATG GAG GAC GCC AGC TAC GAC CCC GAG CGG GTC CGC GGC GAG GTG   360
106    M   E   D   A   S   Y   D   P   E   R   V   R   G   E   V   120

361   GGC GTC TAC GCC TCG ACC AAC CCG CAG AGC GCG GCC CTC TAC AGC   405
121    G   V   Y   A   S   T   N   P   Q   S   A   A   L   Y   S   135

406   TCA CCA CCG GAC TGG GTA TCG GCC GGC CCG GAG GTG ATG GAC CGC   450
136    S   P   P   D   W   V   S   A   G   P   E   V   M   D   R   150

451   AGT AAC GCA TGG CTA CCG GAC ACG ATC ACC TCC AAC GTC TTG TAC   495
151    S   N   A   W   L   P   D   T   I   T   S   N   V   L   Y   165

496   TAC CTC GGC CTG ACC GGT GAG GCG GTC ACC GTG ACC GCC GTG TGC   540
166    Y   L   G   L   T   G   E   A   V   T   V   T   A   V   C   180

541   TCG GGC TTT CAC TAT GCG GTG CAC CTC GCG TGC CAG TCA CTT CTC   585
181    S   G   F   H   Y   A   V   H   L   A   C   Q   S   L   L   195

586   CTA GGC CAG ACT GAT ATG GCG ATC GCG GGC GGC GTC ATG GTT CGG   630
196    L   G   Q   T   D   M   A   I   A   G   G   V   M   V   R   210
```

```
631   CTG CCC CAG CGC CGC GGC TAT CTG TGG GAG GAG GGT CGA ATT CTG   675
211    L   P   Q   R   R   G   Y   L   W   E   E   G   R   I   L   225

676   TCC CGC GAC GGC CAC TGC CGC CCC TTT GAC GCG AAC GGC ACC GGA   720
226    S   R   D   G   H   C   R   P   F   D   A   N   G   T   G   240

721   TCG GTA CTG GCC AGC GGA GTC GCG GTG GTC CTA CTC AAG CCA CTG   765
241    S   V   L   A   S   G   V   A   V   V   L   L   K   P   L   255

766   CCA CAG GCC GTC GCC GAC CGC GAC CAC ATC TAT GCA ACG GTC AAG   810
256    P   Q   A   V   A   D   R   D   H   I   Y   A   T   V   K   270

811   GGA ACG GCC ATC AAC AAC AAC GGC ATC AGT GCC ATG GCG TAC GGC   855
271    G   T   A   I   N   N   N   G   I   S   A   M   A   Y   G   285

856   CTC GCG CAG CCT GAA CGG CTC AGC GCC TGC ATC GCC GGT GCG ATG   900
286    L   A   Q   P   E   R   L   S   A   C   I   A   G   A   M   300

901   CAG GCC GGC GAT GTG GCG CCC GAA ACG GTG TCC ATG TAC GAG GCC   945
301    Q   A   G   D   V   A   P   E   T   V   S   M   Y   E   A   315

946   AAT GGG TTC GGC ATG CCC ATC ACG GAC TCG CTC GAG GTG CAC GCC   990
316    N   G   F   G   M   P   I   T   D   S   L   E   V   H   A   330

991   GCC CAT CTG GCG TTC GGG AAG CAG TCC GGC ACC TGC TCG ATT GGG  1035
331    A   H   L   A   F   G   K   Q   S   G   T   C   S   I   G   345

1036  GCG GTC AAG GGC AAT ATC GGG CAC GCC GGC GTG GTT GCC GGT GGC  1080
346    A   V   K   G   N   I   G   H   A   G   V   V   A   G   G   360

1081  TCC GGC GCG GTC AAG GCG GCG TTC GCG CTG TAC CAC CGC AGT CTC  1125
361    S   G   A   V   K   A   A   F   A   L   Y   H   R   S   L   375

1126  CCG CCC ACC ATC AAC CTG ACC GAG CTC AAC GAA GAA ATC GAC TTC  1170
376    P   P   T   I   N   L   T   E   L   N   E   E   I   D   F   390

1171  CCG CGT ACT CCG TTC GTC CCG CAA CTG GAG CCC GCC GCC TGG CAG  1215
391    P   R   T   P   F   V   P   Q   L   E   P   A   A   W   Q   405

1216  CCG GAG TGC GGC ATC AGA CGG GCC GGC ATC ACC GCG CTC GGC GGC  1260
406    P   E   C   G   I   R   R   A   G   I   T   A   L   G   G   420

1261  GGC GGG TAC AAC GCG CAC CTG GTC CTC GAG GAG CCG CCG CGT CCG  1305
421    G   G   Y   N   A   H   L   V   L   E   E   P   P   R   P   435

1306  GTC GAA CGG GAG CCC GAG AAC CGC CGG CCC CGG ATC GTC ACG CTG  1350
436    V   E   R   E   P   E   N   R   R   P   R   I   V   T   L   450

1351  TCG GCG CTC GAC GAC GCC GCG GTC TCC CGC CAA CGC GCG GCC CTG  1395
451    S   A   L   D   D   A   A   V   S   R   Q   R   A   A   L   465

1396  TCC TCC TGG CTC GCT GAG CAT TCC GAT GCT CGG CTC GAC GAC ATC  1440
466    S   S   W   L   A   E   H   S   D   A   R   L   D   D   I   480

1441  GCC TAT AGC CTC AAC CTC GGG CGT AAG GCG CTG CCA TCC CGA TGG  1485
481    A   Y   S   L   N   L   G   R   K   A   L   P   S   R   W   495

1486  GCG GCC GTC ATC AGC ACC CGC GAC GAA CTG CTC GAG GTG CTG TCG  1530
496    A   A   V   I   S   T   R   D   E   L   L   E   V   L   S   510
```

```
1531  GGC GAC GGC AAG AGC GGT CGG GTG TCT CGC TTC GGC CAA GAG CGC   1575
511    G   D   G   K   S   G   R   V   S   R   F   G   Q   E   R    525

1576  CGC GCC GAC CTG GCG CGT TTT CGG CGC ACC GAA GAC GGA CTC GCT   1620
526    R   A   D   L   A   R   F   R   R   T   E   D   G   L   A    540

1621  CTC GGC AGC GGG GAC GAG ATG CGC GAC GTT CAG GCG CTG ACC GAG   1665
541    L   G   S   G   D   E   M   R   D   V   Q   A   L   T   E    555

1666  CTC GCC GCA GCA TGG GTG CAG GGG GAA CGC GTG AAC TTC GAG GTG   1710
556    L   A   A   A   W   V   Q   G   E   R   V   N   F   E   V    570

1711  TTG CAC GCC GAC GAA CGC AGC CAC CGC ATT TCG TTG CCC AAC TAT   1755
571    L   H   A   D   E   R   S   H   R   I   S   L   P   N   Y    585

1756  CCG TTC GCG CGC CGA CGG TTC TGG CGC ACC GAC TGG TGA   1794
586    P   F   A   R   R   R   F   W   R   T   D   W   *
```

```
>salD

1     ATG GGC GGC CGT CAT CAG CAC CCG CGA CGA ACT GCT CGA GGT GCT   45
1      M   G   G   R   H   Q   H   P   R   R   T   A   R   G   A    15

46    GTC GGG CGA CGG CAA GAG CGG TCG GGT GTC TCG CTT CGG CCA AGA   90
16     V   G   R   R   Q   E   R   S   G   V   S   L   R   P   R    30

91    GCG CCG CGC CGA CCT GGC GCG TTT TCG GCG CAC CGA AGA CGG ACT   135
31     A   P   R   R   P   G   A   F   S   A   H   R   R   R   T    45

136   CGC TCT CGG CAG CGG GGA CGA GAT GCG CGA CGT TCA GGC GCT GAC   180
46     R   S   R   Q   R   G   R   D   A   R   R   S   G   A   D    60

181   CGA GCT CGC CGC AGC ATG GGT GCA GGG GGA ACG CGT GAA CTT CGA   225
61     R   A   R   R   S   M   G   A   G   G   T   R   E   L   R    75

226   GGT GTT GCA CGC CGA CGA ACG CAG CCA CCG CAT TTC GTT GCC CAA   270
76     G   V   A   R   R   R   T   Q   P   P   H   F   V   A   Q    90

271   CTA TCC GTT CGC GCG CCG ACG GTT CTG GCG CAC CGA CTG GTG ACC   315
91     L   S   V   R   A   P   T   V   L   A   H   R   L   V   T   105

316   GCC GCG GAC ACA GGT CGC GTG GGT TCA GGC GAA AGC CAA ACG CCG   360
106    A   A   D   T   G   R   V   G   S   G   E   S   Q   T   P   120

361   GCC CAC ACC GTC GAC CTA GCC GAC CCC GCC ACC TTC GCC AAT CAC   405
121    A   H   T   V   D   L   A   D   P   A   T   F   A   N   H   135

406   GAC CTG ACC GGT TTC TGG CAA CAG CTG CGC GAC GAG GAG CCG ATC   450
136    D   L   T   G   F   W   Q   Q   L   R   D   E   E   P   I   150

451   CAC TGG AAT CCC CCG ACA GCC GGA CGT CGA GGC TTC TGG GTG GTG   495
151    H   W   N   P   P   T   A   G   R   R   G   F   W   V   V   165

496   TCG CGA TAC GCA GAC ATT CTC GAC GTC TAC CGC GAC GAC GTC ACC   540
166    S   R   Y   A   D   I   L   D   V   Y   R   D   D   V   T   180
```

```
541   TTC ACC TCA GAG CGC GGG AAT GTG CTC GTG ACG TTG CTG GCG GGC   585
181    F   T   S   E   R   G   N   V   L   V   T   L   L   A   G   195

586   GGT GAC GCG GGC GCT GGC CGG ATG CTG GCG GTC ACC GAC GGG CCG   630
196    G   D   A   G   A   G   R   M   L   A   V   T   D   G   P   210

631   CGT CAC GCG GAG TTG CGC AAG CTG CTG TTG CGG GCA TTG GGG CCG   675
211    R   H   A   E   L   R   K   L   L   L   R   A   L   G   P   225

676   CGC GTA CTC GCG CCA GTC TGC GCC GCG GTG CGG ACC AAC ACC CGG   720
226    R   V   L   A   P   V   C   A   A   V   R   T   N   T   R   240

721   CAG ATG ATC CGC GAG GCA GTC ACC AAG GGT GAG TGC GAC TTC GCG   765
241    Q   M   I   R   E   A   V   T   K   G   E   C   D   F   A   255

766   TCC GAC ATC GCC AGC CGC ATT CCG ATG ATG ACG ATC TCC AAC CTG   810
256    S   D   I   A   S   R   I   P   M   M   T   I   S   N   L   270

811   CTG GGG GTC CCC GAC GCG GAC CGT GCC TTC CTG CTG AGT CTG ACG   855
271    L   G   V   P   D   A   D   R   A   F   L   L   S   L   T   285

856   AAG ACG GCG CTG AGC GCC GAC GAC GAA AGC ATC AGC GAG ACC GAG   900
286    K   T   A   L   S   A   D   D   E   S   I   S   E   T   E   300

901   TCC GCG ATG GCA CGC AAC GAG ATC CTG CTC TAC TTC CAA GAC CTC   945
301    S   A   M   A   R   N   E   I   L   L   Y   F   Q   D   L   315

946   ATG GAG TTC CGG CGC GAC CAC CCC GGC GAG GAT GTC GTG AGC ATG   990
316    M   E   F   R   R   D   H   P   G   E   D   V   V   S   M   330

991   CTG GTG AAC AGC AGC ATC GAC GGA GCG CCC CTT TCG GAC GAC GAC   1035
331    L   V   N   S   S   I   D   G   A   P   L   S   D   D   D   345

1036  ATA GTG CTC AAC TGC TAC AGC CTG ATC ATC GGC GGC GAC GAA ACG   1080
346    I   V   L   N   C   Y   S   L   I   I   G   G   D   E   T   360

1081  AGT CGG CTA ACC ATG ATC GAT AGC GTG AAT ACG CTG GCT GCA CAC   1125
361    S   R   L   T   M   I   D   S   V   N   T   L   A   A   H   375

1126  CCC CAG CAG TGG AGG CGA CTC AAA GAT GGT CAA TGC GAG ATC GAC   1170
376    P   Q   Q   W   R   R   L   K   D   G   Q   C   E   I   D   390

1171  AAG GCC GTG GAC GAG GTA CTG CGC TGG GCC TCA CCC TCG ATG CAC   1215
391    K   A   V   D   E   V   L   R   W   A   S   P   S   M   H   405

1216  TTC GGA CGG GTG GCA GCG CGG GAC ACC ATT CTT CAT GGA GTG CGG   1260
406    F   G   R   V   A   A   R   D   T   I   L   H   G   V   R   420

1261  ATC CGG GCG GAC GAC ATA GTT ACC CTC TGG CAT GCC TCC GGC AAC   1305
421    I   R   A   D   D   I   V   T   L   W   H   A   S   G   N   435

1306  CGC GAT GAG CGG GTC TTC CAT CGG CCG GAG GTG TTC GAC CTG GGC   1350
436    R   D   E   R   V   F   H   R   P   E   V   F   D   L   G   450

1351  CGT ACA CCC AAC CGG CAC CTC TCG TTT GGG CAC GGG CCT CAC TAC   1395
451    R   T   P   N   R   H   L   S   F   G   H   G   P   H   Y   465

1396  TGC ATC GGT TCC TAC CTT GCC AAG GTG GAG ATC TCG GAG CTG CTG   1440
466    C   I   G   S   Y   L   A   K   V   E   I   S   E   L   L   480
```

```
1441  ATC GCT CTG CGG GAT CTG ACA TCG GGT TTT GAG ACA ACA GGC GAG   1485
481    I   A   L   R   D   L   T   S   G   F   E   T   T   G   E    495

1486  CCG CAA CGG ATC CGG TCG AAC CTG CTC ACT GGC TTC GCC ACC ATG   1530
496    P   Q   R   I   R   S   N   L   L   T   G   F   A   T   M    510

1531  CCA GTC CGC TTC GTT CCT GAT CGT GCA GGG CTC GCT CGC GAC GCG   1575
511    P   V   R   F   V   P   D   R   A   G   L   A   R   D   A    525

1576  CTG GAT GGG TGA   1587
526    L   D   G   *
```

>SalE

```
1     gtg aga caa atg att gcc aac ccc ctc gat ccg ggc gac ggc acg   45
1      V   R   Q   M   I   A   N   P   L   D   P   G   D   G   T    15

46    tgg ctg gtc gtg gtt aac gac gcc ggc caa cac gcg ctg tgg cgg   90
16     W   L   V   V   V   N   D   A   G   Q   H   A   L   W   R    30

91    ccc ttt ctg gat atg ccg ccc ggc tgg cgt gtc gtc cta cgc gac   135
31     P   F   L   D   M   P   P   G   W   R   V   V   L   R   D    45

136   gcc gac cac gac gcg gcc ctt gac tac gtc gag cgc aac tgg act   180
46     A   D   H   D   A   A   L   D   Y   V   E   R   N   W   T    60

181   gcc ctg caa ccc gcg ggc cac agc gaa ccg ggg gcg acg agg taa   225
61     A   L   Q   P   A   G   H   S   E   P   G   A   T   R   *
```

>salF

```
1     ATG CCG TTT GCG TTC GCC AAC GAC ATC AGG TTG CAC TAC GAA TTG   45
1      M   P   F   A   F   A   N   D   I   R   L   H   Y   E   L    15

46    CAT GGC ACG GGC GAT CCC GTG GTG CTG GTC AGT GGT GCC GGC GTA   90
16     H   G   T   G   D   P   V   V   L   V   S   G   A   G   V    30

91    TCC GGG AAA TCC TGG CTG ATC CAT CAG GGG CCG GCA CTG CTG GAG   135
31     S   G   K   S   W   L   I   H   Q   G   P   A   L   L   E    45

136   AAC GGC TAC CGG GTC TGC GTC TAC GAC AGC CGC GGC CAA CCA CCC   180
46     N   G   Y   R   V   C   V   Y   D   S   R   G   Q   P   P    60

181   AGC GAT GAG TGC GTC AGC GGC TTC GTC GTG GAG GAT CTG GTA GCC   225
61     S   D   E   C   V   S   G   F   V   V   E   D   L   V   A    75

226   GAC CTC GCA GCC CTA CTG GAG TTC CTC GAC GCT GGC CCG GCC CGC   270
76     D   L   A   A   L   L   E   F   L   D   A   G   P   A   R    90

271   CTC ATC GGC ACC TCG ACG GGT GCC TAC GTG GTG CAG GAG CTT GCC   315
91     L   I   G   T   S   T   G   A   Y   V   V   Q   E   L   A   105
```

```
316   CTG CGT CGA CCG GAG CTG GTT CGG CAG GCG GTG TTG ATG GCC AGC   360
106    L   R   R   P   E   L   V   R   Q   A   V   L   M   A   S   120

361   CGG GCC CGT CCC GAC GTG TTA CGG ACA CAG CTG GCG CGT GCC GAG   405
121    R   A   R   P   D   V   L   R   T   Q   L   A   R   A   E   135

406   ATC GAG TTG GGC GAG TCT GGG GTG ACG CTG CCG TCC TCG TAT CGC   450
136    I   E   L   G   E   S   G   V   T   L   P   S   S   Y   R   150

451   GCG GTC GTT CGC GCG CTC CAG ATG CTC TCG CCA CGA TCG ATG GAC   495
151    A   V   V   R   A   L   Q   M   L   S   P   R   S   M   D   165

496   GAC GAG TCG TCG ATC CGC GAC TGG CTG GCC CTG CTC GAG CTC GCC   540
166    D   E   S   S   I   R   D   W   L   A   L   L   E   L   A   180

541   CCG CCC GAC GGA GCG GGA GTA CGG CAT CAG CTC GCT CTC CAA CCG   585
181    P   P   D   G   A   G   V   R   H   Q   L   A   L   Q   P   195

586   ATG CCG GAC CGC CGC GCC GCC TAC GCC GAG ATC GCG GTG CCC TGC   630
196    M   P   D   R   R   A   A   Y   A   E   I   A   V   P   C   210

631   CAC GTC GTG TCG TTC GCG GAT GAC TTG ATC GCG CCG CCC GGG TAC   675
211    H   V   V   S   F   A   D   D   L   I   A   P   P   G   Y   225

676   GGC GAG GAG CTG GCG GGG TGT ATT CCC GGC GCC ACC TTC GAC TTG   720
226    G   E   E   L   A   G   C   I   P   G   A   T   F   D   L   240

721   GTC GGG GAC GCC GGC CAC TTC GGC TAC CTG GAA CGG CCC GAG GAG   765
241    V   G   D   A   G   H   F   G   Y   L   E   R   P   E   E   255

766   GTC AAC CAC ATC ATC GCC AAA CAT TTC GCG GGC ATC GGC GCC CGC   810
256    V   N   H   I   I   A   K   H   F   A   G   I   G   A   R   270

811   CTC ACC GGA CAC AAC CCA GAC ACC AGG AGG AGT GCG CCA TGG GCC   855
271    L   T   G   H   N   P   D   T   R   R   S   A   P   W   A   285

856   AGC TGA   861
286    S   *
```

>salG

```
1     GTG TAT TCC CGG CGC CAC CTT CGA CTT GGT CGG GGA CGC CGG CCA   45
1      V   Y   S   R   R   H   L   R   L   G   R   G   R   R   P   15

46    CTT CGG CTA CCT GGA ACG GCC CGA GGA GGT CAA CCA CAT CAT CGC   90
16     L   R   L   P   G   T   A   R   G   G   Q   P   H   H   R   30

91    CAA ACA TTT CGC GGG CAT CGG CGC CCG CCT CAC CGG ACA CAA CCC   135
31     Q   T   F   R   G   H   R   R   P   P   H   R   T   Q   P   45

136   AGA CAC CAG GAG GAG TGC GCC ATG GGC CAG CTG ACC GAC GCC GTG   180
46     R   H   Q   E   E   C   A   M   G   Q   L   T   D   A   V   60

181   CTC AGC AAC GCG ACG GCG GCA GAG ATC GCC GCC GCT CCG CTA CCC   225
61     L   S   N   A   T   A   A   E   I   A   A   A   P   L   P   75
```

```
226   GAG GAG TAC TTC GCC GCC CAC CTG CGG GTC GAA GAC GCC ACG ATG   270
76     E   E   Y   F   A   A   H   L   R   V   E   D   A   T   M    90

271   TTC GAG GGC GTC GTG GAC AAG GAC GTC CGC AAG TCG CTG CGG CTC   315
91     F   E   G   V   V   D   K   D   V   R   K   S   L   R   L   105

316   GGT CAC GTT CCA ATG CCC GAG CTG GCT CCG GAC GAG GTC CTC GTC   360
106    G   H   V   P   M   P   E   L   A   P   D   E   V   L   V   120

361   GCG GTG ATG GCG AGC GCG GTC AAC TAC AAC ACC GTG TGG TCG GCG   405
121    A   V   M   A   S   A   V   N   Y   N   T   V   W   S   A   135

406   ATG TTC GAG CCG ATG TCC AGC TTC CGC TAT CTG AAG GGG CTC GCC   450
136    M   F   E   P   M   S   S   F   R   Y   L   K   G   L   A   150

451   CGG CAG GGT GGC TGG GCG AAA CGG CAC GAC CAA CCC TTC CAC GTG   495
151    R   Q   G   G   W   A   K   R   H   D   Q   P   F   H   V   165

496   GTC GGC TCC GAC TGC GCT GGG GTG ATC GTG CGC ACC GGC TCC GGC   540
166    V   G   S   D   C   A   G   V   I   V   R   T   G   S   G   180

541   GTC CGC CGC TGG GAG ATC GGC CAG CAC GTC GTG GTG AGC CCG GCG   585
181    V   R   R   W   E   I   G   Q   H   V   V   V   S   P   A   195

586   TAC GTG GAC GAC CAG GAG CCC ACC ACG CAC GCC GAC GGA ATG CTC   630
196    Y   V   D   D   Q   E   P   T   T   H   A   D   G   M   L   210

631   GGC GGG GAC CAG AAG GCG TGG GGT TTC GAG ACA AAC TTC GGG GCG   675
211    G   G   D   Q   K   A   W   G   F   E   T   N   F   G   A   225

676   CTG GCC GAG TTC ACG GTG GTC CGG GCC ACC CAG CTG ATT CCC AAG   720
226    L   A   E   F   T   V   V   R   A   T   Q   L   I   P   K   240

721   CCC GCC CAC CTC ACC TGG GAG GAG GCC GCG AGC GTC CCG CTG TGC   765
241    P   A   H   L   T   W   E   E   A   A   S   V   P   L   C   255

766   GCG GGA ACC GCC TAC CGG ATG CTG GTC AGC GAG CGC GGC GCG AAC   810
256    A   G   T   A   Y   R   M   L   V   S   E   R   G   A   N   270

811   ATG AAG CAG GGA GAC GTC GTG CTC GTC TGG GGT GCG GCC GGC GGT   855
271    M   K   Q   G   D   V   V   L   V   W   G   A   A   G   G   285

856   CTC GGC GGC TAC GCA GTG CAG TAT GTG CGT AAC GGT GGC GGC ATC   900
286    L   G   G   Y   A   V   Q   Y   V   R   N   G   G   G   I   300

901   CCG GTG GCC GTC GTC GGG TCC GAG CGC AAG GCG GAG GCA ATG CGC   945
301    P   V   A   V   V   G   S   E   R   K   A   E   A   M   R   315

946   CGA CTC GGC TGC GAG GTG ATT CTC AAC CGG ACC GAG ATC GGC ATC   990
316    R   L   G   C   E   V   I   L   N   R   T   E   I   G   I   330

991   ACT GAC GAG ATC GCC GAC GAC CCG AAG CAG GTG ATC AGA GCC GGC  1035
331    T   D   E   I   A   D   D   P   K   Q   V   I   R   A   G   345

1036  GTG CGG CTC GGC GAA ATG ATC AAC GAA CAG GTG GGC CGG GCT CCG  1080
346    V   R   L   G   E   M   I   N   E   Q   V   G   R   A   P   360

1081  GAT ATC GTC TTC GAG CAC ATC GGA CGG CCC ACC TTC GGG ATA TCG  1125
361    D   I   V   F   E   H   I   G   R   P   T   F   G   I   S   375
```

```
1126  GTG TTC GTC GCA CAG CGC GGC GGC ACC GTA GTG ACG TGC GGC TCG   1170
376    V   F   V   A   Q   R   G   G   T   V   V   T   C   G   S    390

1171  AGC GCC GGG TAT CAG CAC AGC TTC GAC AAC CGC TAC CTG TGG ATG   1215
391    S   A   G   Y   Q   H   S   F   D   N   R   Y   L   W   M    405

1216  CGC CTC AAA CGG ATC ATC GGC AGC CAC ATC GCC AAC CTC CAA GAG   1260
406    R   L   K   R   I   I   G   S   H   I   A   N   L   Q   E    420

1261  TAC TGG GAT GCG CAC CGC CTG TTC GAA CTC GGT CAC ATC GTG CCG   1305
421    Y   W   D   A   H   R   L   F   E   L   G   H   I   V   P    435

1306  ACC ATG TCG GCG GTC TAT CCG TTG ACG GAC GTC GCG GCC GCC TGC   1350
436    T   M   S   A   V   Y   P   L   T   D   V   A   A   A   C    450

1351  CGT CTC GTG CAG GCC AAC CAG CAG ATC GGC AAA ATC GCC GTG CGC   1395
451    R   L   V   Q   A   N   Q   Q   I   G   K   I   A   V   R    465

1396  TGC CTG GCA CCC CGC GCC GGG CTC GGG GTG ACC AAC CCC GAG TTG   1440
466    C   L   A   P   R   A   G   L   G   V   T   N   P   E   L    480

1441  CGC GAG CGC ATC GGC GAG GAC CGC CTG AAC CCG TGG CGC GGC CTC   1485
481    R   E   R   I   G   E   D   R   L   N   P   W   R   G   L    495

1486  ACC ACC GGC TCT GAC GGT GCC GAG TTC GGA GGA CAG CGT GAC GAC   1530
496    T   T   G   S   D   G   A   E   F   G   G   Q   R   D   D    510

1531  CCG GCC TAA   1539
511    P   A   *
```

>SALH

```
1     ATG CGC GCC ACA GGT CTG TCG ACC GAG GAC CTG AAG AAG CCG ATG   45
1      M   R   A   T   G   L   S   T   E   D   L   K   K   P   M    15

46    ATC GGG GTA GCG CAC AGT TGG ATC GGC ACG ATG CCG TGC AAC CTC   90
16     I   G   V   A   H   S   W   I   G   T   M   P   C   N   L    30

91    AAC CAC CGG CGG CTC GCC CAG GAG GTG ATG GCC GGC GTG CGC GCC   135
31     N   H   R   R   L   A   Q   E   V   M   A   G   V   R   A    45

136   GCA GGT GGC ACC CCG ATC GAG ATC AAC ACC ATC GCG ATC TCG GAC   180
46     A   G   G   T   P   I   E   I   N   T   I   A   I   S   D    60

181   GTG ATC ACC ATG GGC ACG GAG GGA ATG CGG ACC TCA CTG GTG AGC   225
61     V   I   T   M   G   T   E   G   M   R   T   S   L   V   S    75

226   CGC GAG GTG ATC GCC GAC TCC ATC GAG TTG GTA TGC CGC GGC CAC   270
76     R   E   V   I   A   D   S   I   E   L   V   C   R   G   H    90

271   GGT CTC GAC GGC CTG GTG ACC CTG GCC GGC TGT GAC AAG ACC ATA   315
91     G   L   D   G   L   V   T   L   A   G   C   D   K   T   I   105

316   CCC GGT GCG GCG CTG GCC CAC GTG CGA CTC GAC ATC CCC GGA GCC   360
106    P   G   A   A   L   A   H   V   R   L   D   I   P   G   A   120
```

```
361   GTC ATC TAT TCC GGC ACG ATG ATG CCG GGC GAG CAC CTC GGG CGC   405
121    V   I   Y   S   G   T   M   M   P   G   E   H   L   G   R   135

406   GAC ATC ACC CTG CAG GAT GTG TTT GAG GCC GTC GGC AGC GCC ACC   450
136    D   I   T   L   Q   D   V   F   E   A   V   G   S   A   T   150

451   GCC ACC GGC TGC ACC GAC GAG CTC GAC AAG CTC GAG CGC GCG GCC   495
151    A   T   G   C   T   D   E   L   D   K   L   E   R   A   A   165

496   TGT CCC GGT ATC GGG GCC TGC GCG GGC CAC TAC ACG GCC AAT ACG   540
166    C   P   G   I   G   A   C   A   G   H   Y   T   A   N   T   180

541   ATG GCC GTG GTG CTG GAG TTC CTC GGG CTT TCC CCG TTC GGC TCG   585
181    M   A   V   V   L   E   F   L   G   L   S   P   F   G   S   195

586   ATG GAT CCG CCG GCA GTC GAC GCC CGC AAG GAC ACG GTC TGC CGC   630
196    M   D   P   P   A   V   D   A   R   K   D   T   V   C   R   210

631   CAG GCC GGC GAG CTG GTC ATG CGG GCG GTG GCA GAA GGG CTG CGG   675
211    Q   A   G   E   L   V   M   R   A   V   A   E   G   L   R   225

676   CCC AGC CGG TTC CTG ACG CCC TCC TCG TTG CGC AAC GCC ATC GCC   720
226    P   S   R   F   L   T   P   S   S   L   R   N   A   I   A   240

721   GCC GGG GTC GCC ACC GGC GGC TCG ACG AAC ATG GTG CTC CAC CTG   765
241    A   G   V   A   T   G   G   S   T   N   M   V   L   H   L   255

766   TTG GCG ATC GCT CGG GAG GCG GGC ATC CCG CTG GAC ATT GAC GAG   810
256    L   A   I   A   R   E   A   G   I   P   L   D   I   D   E   270

811   TTC GAC CGG ATC AGC TCG GTG ACC CCG ATC ATC GCT GAC CTG CGT   855
271    F   D   R   I   S   S   V   T   P   I   I   A   D   L   R   285

856   CCA AAC GGG ACG TAC ACC GCG GTG GAC CTC GAC CGG GCG GGC GGC   900
286    P   N   G   T   Y   T   A   V   D   L   D   R   A   G   G   300

901   ACC CGG GTG ATC GCC CGC CAC ATG GTC GAC GCG GGC CTG ATT GCT   945
301    T   R   V   I   A   R   H   M   V   D   A   G   L   I   A   315

946   GGC GAC GAA AGC ACC GTC ACC GGC CGC ACC GTC GCA CAC GAG GCC   990
316    G   D   E   S   T   V   T   G   R   T   V   A   H   E   A   330

991   GCG GAC GCG GCC GAG ACA CCC GGC CAA CGG GTC GTC ACC ACG GTG  1035
331    A   D   A   A   E   T   P   G   Q   R   V   V   T   T   V   345

1036  GAG GCA CCC CTA TCG CCT TCC GGT GCT CTC CTG ATC TTG CGG GGT  1080
346    E   A   P   L   S   P   S   G   A   L   L   I   L   R   G   360

1081  AAC CTA GCG CCC GAT GGC AGT GTG GTG AAG GCA CCC GGC GCG GTG  1125
361    N   L   A   P   D   G   S   V   V   K   A   P   G   A   V   375

1126  ACC CTG CGG ATG ACC GGC ACC GCA TTG GTG TTC AAC TGC GAG GAG  1170
376    T   L   R   M   T   G   T   A   L   V   F   N   C   E   E   390

1171  GAA GCG ATG GCC GCG GTC CAG ACT GGC CGC GTC CGG CCA GGC CAC  1215
391    E   A   M   A   A   V   Q   T   G   R   V   R   P   G   H   405

1216  GTC GTC GTC ATC CGC TAC GAG GGT CCG CGC GGG GGT CCA GGG ATG  1260
406    V   V   V   I   R   Y   E   G   P   R   G   G   P   G   M   420
```

```
1261  AGG GAA ATG CTC GGA GTG ACC TCG GCG CTC ATC GGC CGC GGC CTG   1305
421    R   E   M   L   G   V   T   S   A   L   I   G   R   G   L    435

1306  GGC ACG TCG GTC GGT CTG GTG ACC GAC GGC CGT TTC TCG GGC GCG   1350
436    G   T   S   V   G   L   V   T   D   G   R   F   S   G   A    450

1351  ACC AGG GGA CTG ATG GTG GGG CAC GTC GCT CCG GAG GCG GCG GAG   1395
451    T   R   G   L   M   V   G   H   V   A   P   E   A   A   E    465

1396  GGC GGA CCG ATC GCG GCG GTG TGC GAC GGT GAC CGG ATC ACC ATC   1440
466    G   G   P   I   A   A   V   C   D   G   D   R   I   T   I    480

1441  GAC CTG CAG CGG CGC GAA TGC TCT GTC GAC CTG GAT CCG GGC GAA   1485
481    D   L   Q   R   R   E   C   S   V   D   L   D   P   G   E    495

1486  CTG GCC GCG CGG ATG CGA GAC TGG TCG GCT CCG CCA CCG CGC TAC   1530
496    L   A   A   R   M   R   D   W   S   A   P   P   P   R   Y    510

1531  ACG ATC GGC GTC ATG GCC AAA TAC TGG TCG ACG GTC TCG TCG GCG   1575
511    T   I   G   V   M   A   K   Y   W   S   T   V   S   S   A    525

1576  GCC GTG GGC GCC GTG ACG ACC CCG CAC CCC ACC CAG GGC CCG GCG   1620
526    A   V   G   A   V   T   T   P   H   P   T   Q   G   P   A    540

1621  ACA GCG TCA GGT AAG GCC GAG GAG TGC CAG CAG GCG AGT GCG GTC   1665
541    T   A   S   G   K   A   E   E   C   Q   Q   A   S   A   V    555

1666  GAG GGC GTG ATG GCG GTT GGC GGC GGC GAT GTC GGT GCT GCC GGC   1710
556    E   G   V   M   A   V   G   G   G   D   V   G   A   A   G    570

1711  GGT TCG TAG   1719
571    G   S   *
```

>SALI

```
1     GTG TGT GAT GCC CGC GCT GTC ACC ATC GCT GCT TTC GTC CAT GTC   45
1      V   C   D   A   R   A   V   T   I   A   A   F   V   H   V    15

46    CTA CGC ACC GGC TGT CAC GGT GCC CGA AAC CAC TGC TGG ACC TGC   90
16     L   R   T   G   C   H   G   A   R   N   H   C   W   T   C    30

91    CGC CAC GTT GGT CAA CCT GCC GGA TCC GCA TGT CTG CCG GGT ATG   135
31     R   H   V   G   Q   P   A   G   S   A   C   L   P   G   M    45

136   CGG CAC CGG TTG ATC GTT GTG GTG ACT GCG GCG GGT GTG CGC TGT   180
46     R   H   R   L   I   V   V   V   T   A   A   G   V   R   C    60

181   GGT GGC TGG TTG CCG CTC CTA CAT CGC GTT ACC TAC AAC GAA GAC   225
61     G   G   W   L   P   L   L   H   R   V   T   Y   N   E   D    75

226   CAT TCA AGG GTC CGC ACC GGC ACC AGA CCG CAG GTC ATG GCC GTC   270
76     H   S   R   V   R   T   G   T   R   P   Q   V   M   A   V    90

271   CTC CGC AAC GCC GCC ATC AGC GCG CTA CGA ACC GCC GGC AGC ACC   315
91     L   R   N   A   A   I   S   A   L   R   T   A   G   S   T   105
```

```
316   GAC ATC GCC GCC GCC AAC CGC CAT CAC GCC CTC GAC CGC ACT CGC   360
106    D   I   A   A   A   N   R   H   H   A   L   D   R   T   R    120

361   CTG CTG GCA CTC CTC GGC CTT ACC TGA   387
121    L   L   A   L   L   G   L   T   *
```

>SALJ

```
1     GTG AAT CGG GGT CTG CCG TCC ACG TTC TTA ACA TTT CAT TCT TCC    45
1      V   N   R   G   L   P   S   T   F   L   T   F   H   S   S    15

46    TCG TTT GTC GAG TTG TTG AAG GAA TAC GAT CCC ACG CTG CTG CCG    90
16     S   F   V   E   L   L   K   E   Y   D   P   T   L   L   P    30

91    GCC AAG CGT GCC GCC GAG ACG TTC GGC GAG GCT CCG CAC GGC ACC   135
31     A   K   R   A   A   E   T   F   G   E   A   P   H   G   T    45

136   ACC ATA GTG GCC GCC ACG TTC GCT GGT GGC GTC CTG CTC GCT GGC   180
46     T   I   V   A   A   T   F   A   G   G   V   L   L   A   G    60

181   GAC CGT CGT ACG ACC ATG GGG AAC CTG ATC GCC GGC CGG GAC GTG   225
61     D   R   R   T   T   M   G   N   L   I   A   G   R   D   V    75

226   GAC AAA CTC ACG ATC ACG GAC GAC TAT TCG GCG GTC GGT TTC GCC   270
76     D   K   L   T   I   T   D   D   Y   S   A   V   G   F   A    90

271   GGC ACG GTG GGC ATC TCC ATT GAC ATG ACT CGG TTG TTT GTC GTC   315
91     G   T   V   G   I   S   I   D   M   T   R   L   F   V   V   105

316   GAG CTG GAG CAT TAT GAG AAG ATC GAG GGC GTC CCG CTC TCA CTG   360
106    E   L   E   H   Y   E   K   I   E   G   V   P   L   S   L   120

361   GAG GGC AAG AGC AAC CGG CTC GCT GCC ATA GTC AAG GGA AAC CTC   405
121    E   G   K   S   N   R   L   A   A   I   V   K   G   N   L   135

406   CCC ATG GCC AAC GCG GGC ATG GTT TCT ATC CCG CTG TTC ATC GGC   450
136    P   M   A   N   A   G   M   V   S   I   P   L   F   I   G   150

451   TAC GAT CTC GAG GCA TCC GAT CCT GCG AGG GCC GGC CGG ATC GTG   495
151    Y   D   L   E   A   S   D   P   A   R   A   G   R   I   V   165

496   TCG TTC GAC GCA ATC GGC GCC CGC TAT GAG GAA AAG TCC GGC TAC   540
166    S   F   D   A   I   G   A   R   Y   E   E   K   S   G   Y   180

541   CAG TCG ATC GGC TCG GGC TCC CCC TTC GCT AAA TCG TCT CTG AAG   585
181    Q   S   I   G   S   G   S   P   F   A   K   S   S   L   K   195

586   AAG CTA CAC GAC CCC GGT TCC GAT GAT GCG GCT ACC CTG AGT GCC   630
196    K   L   H   D   P   G   S   D   D   A   A   T   L   S   A   210

631   CTG ATC GAG GCG CTG TAC GAT GCC GCC GAT GAT GAC TCG GCC ACC   675
211    L   I   E   A   L   Y   D   A   A   D   D   D   S   A   T   225

676   GGC GGC CCC GAC ATG GTC CGC CGC ATT TTC CCC ACG GCC ATC CGC   720
226    G   G   P   D   M   V   R   R   I   F   P   T   A   I   R   240
```

```
721   ATC ACG GCC GAG GGC GCC TAC CGA TAC ACA GAG GAG GAA CTC GCG   765
241    I   T   A   E   G   A   Y   R   Y   T   E   E   E   L   A    255

766   GCA GTC GCC GAC GCC GTC GTC GCC CGG CGT AGT GCC GAC GAG CCC   810
256    A   V   A   D   A   V   V   A   R   R   S   A   D   E   P    270

811   GGC GTC CGA AGC TTT ACC GCG TCC TGA   837
271    G   V   R   S   F   T   A   S   *
```

>SALK

```
1     ATG TCT CGT GAA ATG TGG CTA CGG TTC GAG ATA TAC CAT GAT GTG    45
1      M   S   R   E   M   W   L   R   F   E   I   Y   H   D   V    15

46    GTA TAT GCG CCG GAG GCG AGC CTA GTG GCT AAA GAG CTG GGC TGC    90
16     V   Y   A   P   E   A   S   L   V   A   K   E   L   G   C    30

91    CGC GGC GGC TGG ATG GGC TAC TTC GGA ACC CGG TCG GCC CCG CTT   135
31     R   G   G   W   M   G   Y   F   G   T   R   S   A   P   L    45

136   GGT GCG GCG TCA CCC GAG GTG GTG ACC GCG GCG TTC TTC AAT TTC   180
46     G   A   A   S   P   E   V   V   T   A   A   F   F   N   F    60

181   CAC CCG CAG ATG GTT GCC AAG GCG GTT CCG CAA GTC TGG TCC GTG   225
61     H   P   Q   M   V   A   K   A   V   P   Q   V   W   S   V    75

226   GCC TCG CCA GAG CGT TTT CTA CAG GCG CGG CTG GAG ATC GTC GAC   270
76     A   S   P   E   R   F   L   Q   A   R   L   E   I   V   D    90

271   GCT GCG CTG CGC CGC ATG CTT GGC GCT GAT GTG GCC GGG CTC GAG   315
91     A   A   L   R   R   M   L   G   A   D   V   A   G   L   E   105

316   GAG ATG GTC ACG CTT GCC GCG TCC GCC GCC GCG GGC GTG TCT CCC   360
106    E   M   V   T   L   A   A   S   A   A   A   G   V   S   P   120

361   GGC GGC CGC GTA CTG GGT GCT GCC AAC CAG GCG CTT CCA CCG CCG   405
121    G   G   R   V   L   G   A   A   N   Q   A   L   P   P   P   135

406   GTC GAG CCG CAT CTC GCG CTC TGG CAG TCC TGC TCG ACG CTG CGA   450
136    V   E   P   H   L   A   L   W   Q   S   C   S   T   L   R   150

451   GAA TCG CGT GGT GAC GGC CAT GTG GCG GCG CTC GTG GCC GCG GGC   495
151    E   S   R   G   D   G   H   V   A   A   L   V   A   A   G   165

496   CTG GCA CCG TGC GAG GCG CTG GTC TTG TTC ACG GCC GAC AAG GGC   540
166    L   A   P   C   E   A   L   V   L   F   T   A   D   K   G   180

541   CTC GAC CCG GCA TAC ATG CGC CGG ATG CGC GGA TGG TCC GAA CAG   585
181    L   D   P   A   Y   M   R   R   M   R   G   W   S   E   Q   195

586   GAG TGG AAG GAG GCC GAA GCC GCG CTC GTC GAT CGT CGT CTG CTC   630
196    E   W   K   E   A   E   A   A   L   V   D   R   R   L   L   210


631   ACC GGC GGC GGC CTC ACC GTA GCC GGG CGG GAG CTT CGT GCG GAA   675
211    T   G   G   G   L   T   V   A   G   R   E   L   R   A   E   225
```

```
676   GTC GAA CGG TGG ACG GAC ATG GCA GCG GCG GCC CCG TGG CAG GCT   720
226    V   E   R   W   T   D   M   A   A   A   A   P   W   Q   A    240

721   CTG GGA GCC GGT CCC ACC GCC AGG CTC CGC GAC CTC ATG ACG CCG   765
241    L   G   A   G   P   T   A   R   L   R   D   L   M   T   P    255

766   GTG GTG CAG CGC CTC AGC GAG CTG AAC GAG TCC ATG AAG GTA AAC   810
256    V   V   Q   R   L   S   E   L   N   E   S   M   K   V   N    270

811   CCC ATG GCG CTG AAT CCG GCG GAG ATT GTG GCG TAG   846
271    P   M   A   L   N   P   A   E   I   V   A   *
```

>SALL

```
1     ATG TTC CTG TCT CGC TTC CGA CGG AGA AGC CTG CGG TTC GTG GTC   45
1      M   F   L   S   R   F   R   R   R   S   L   R   F   V   V    15

46    TGC ACA ACA ACC AAT CCT GAG GTC GGT TTC CGA CCG ATA AAA CGG   90
16     C   T   T   T   N   P   E   V   G   F   R   P   I   K   R    30

91    AGG TCA CTC ACC ATG CAG CAC AAT CTC ATT GCC TTC CTC TCC GAC   135
31     R   S   L   T   M   Q   H   N   L   I   A   F   L   S   D    45

136   GTC GGG TCC GCC GAC GAA GCC CAC GCA CTC TGC AAG GGA GTC ATG   180
46     V   G   S   A   D   E   A   H   A   L   C   K   G   V   M    60

181   TAC GGC GTC GCG CCG GCA GCG ACG ATC GTG GAC ATA ACC CAT GAC   225
61     Y   G   V   A   P   A   A   T   I   V   D   I   T   H   D    75

226   GTC GCG CCC TTT GAT GTC CGG GAA GGC GCA CTG TTT CTC GCT GAT   270
76     V   A   P   F   D   V   R   E   G   A   L   F   L   A   D    90

271   GTC CCG CAC TCC TTC CCG GCT CAT ACG GTC ATC TGC GCA TAC GTC   315
91     V   P   H   S   F   P   A   H   T   V   I   C   A   Y   V   105

316   TAT CCC GAG ACC GGC ACG GCG ACC CAC ACC ATT GCA GTG CGC AAT   360
106    Y   P   E   T   G   T   A   T   H   T   I   A   V   R   N   120

361   GAG AAG GGT CAA CTG CTC GTC GGA CCC AAC AAT GGG CTG CTG AGC   405
121    E   K   G   Q   L   L   V   G   P   N   N   G   L   L   S   135

406   TTC GCA CTC GAC GCC TCC CCT GCG GTC GAG TGT CAC GAG GTC CTC   450
136    F   A   L   D   A   S   P   A   V   E   C   H   E   V   L   150

451   TCC CCC GAT GTC ATG AAC CAG CCG GTG ACG CCA ACC TGG TAC GGC   495
151    S   P   D   V   M   N   Q   P   V   T   P   T   W   Y   G   165

496   AAG GAC ATC GTC GCT GCC TGC GCC GCC CAC CTC GCT GCC GGC ACC   540
166    K   D   I   V   A   A   C   A   A   H   L   A   A   G   T   180

541   GAC CTT GCC GCC GTC GGC CCC CGG ATC GAC CCG AAG CAG ATC GTT   585
181    D   L   A   A   V   G   P   R   I   D   P   K   Q   I   V   195

586   CGG CTG CCC TAT GCC TCC GCC AGC GAG GTG GAG GGC GGG ATC CGC   630
196    R   L   P   Y   A   S   A   S   E   V   E   G   G   I   R   210
```

```
631   GGG GAG GTG GTA CGC ATC GAC CGG GCT TTC GGC AAC GTA TGG ACC   675
211    G   E   V   V   R   I   D   R   A   F   G   N   V   W   T   225

676   AAC ATA CCC ACC CAC TTG ATT GGC TCG ATG CTG CAG GAC GGG GAG   720
226    N   I   P   T   H   L   I   G   S   M   L   Q   D   G   E   240

721   CGA CTG GAA GTC AAG ATC GAA GCC CTC AGC GAC ACG GTG CTC GAG   765
241    R   L   E   V   K   I   E   A   L   S   D   T   V   L   E   255

766   CTG CCC TTC TGC AAA ACC TTC GGC GAG GTC GAC GAG GGG CAG CCA   810
256    L   P   F   C   K   T   F   G   E   V   D   E   G   Q   P   270

811   CTG CTG TAC CTC AAC AGT CGG GGC CGG CTT GCC CTC GGG CTC AAC   855
271    L   L   Y   L   N   S   R   G   R   L   A   L   G   L   N   285

856   CAG TCC AAC TTC ATC GAG AAG TGG CCT GTC GTA CCC GGT GAC AGC   900
286    Q   S   N   F   I   E   K   W   P   V   V   P   G   D   S   300

901   ATC ACC GTC TCC CCA AGG GTG CCC GAT AGT AAC CTG GGC CCG GTG   945
301    I   T   V   S   P   R   V   P   D   S   N   L   G   P   V   315

946   CTC GGG TAG   954
316    L   G   *
```

>SALM

```
1     ATG ACG AAC GGT GGG CGC CTA TCG GGA AAA GTA AGT CTG ATC ACC   45
1      M   T   N   G   G   R   L   S   G   K   V   S   L   I   T   15

46    GGC GCC GCG CAT GGC ATA GGC CAT GCC ACG GCT GTG TGG TTC GCC   90
16     G   A   A   H   G   I   G   H   A   T   A   V   W   F   A   30

91    CGC GAG GGA GCC CGG CTG GTG GTC AGC GAC GTG GAC GGT GCC GCA   135
31     R   E   G   A   R   L   V   V   S   D   V   D   G   A   A   45

136   CTG GAG AAG TGC CAC GCC GAA CTG GCC GAG TCC GGA GCC GAG GTG   180
46     L   E   K   C   H   A   E   L   A   E   S   G   A   E   V   60

181   ACG ACG GTG ATC GCC GAC GTC TCA GAC GCC ACC CAG GCA CAC CGC   225
61     T   T   V   I   A   D   V   S   D   A   T   Q   A   H   R   75

226   ATG GTT CAG ACG GCC GTC GAT GTC TAC AAC CGG CTG GAC GTC GTG   270
76     M   V   Q   T   A   V   D   V   Y   N   R   L   D   V   V   90

271   GTC GCC AAC GCT GGC GTG ATC CCT CTG CAC GAG ATC ACC GAG GCG   315
91     V   A   N   A   G   V   I   P   L   H   E   I   T   E   A   105

316   ACC GAG CAG GAC TGG GAT GAG GTC ATG GCC GTC GAT GGC AAG GGC   360
106    T   E   Q   D   W   D   E   V   M   A   V   D   G   K   G   120

361   ATG TTC CTC ACG TGC AAG TAC GGC ATT GAC GCC ATG CTG GCC ACG   405
121    M   F   L   T   C   K   Y   G   I   D   A   M   L   A   T   135

406   GGC GGC TCG ATC GTG TGC CTT TCA TCT ATC TCC GGC GTC GCT GGC   450
136    G   G   S   I   V   C   L   S   S   I   S   G   V   A   G   150
```

```
451   CAG CGG GGG CAG AGC ACG TAC GGC CCG GCC AAG TTC GTG GCT TCC   495
151    Q   R   G   Q   S   T   Y   G   P   A   K   F   V   A   S    165

496   GGG CTC ACC AAG CAC TTG GCC GTC GAA TGC GCA CGA CAC TGC ATC   540
166    G   L   T   K   H   L   A   V   E   C   A   R   H   C   I    180

541   CGG GTC AAC GCG GTC GCT CCC GGG ACT ATC CGC ACC AAC CGG GTC   585
181    R   V   N   A   V   A   P   G   T   I   R   T   N   R   V    195

586   CGC CGG CTC GAA GAA GAA CCG GGC GGT CCG GAG TAC CTG GAG GAC   630
196    R   R   L   E   E   E   P   G   G   P   E   Y   L   E   D    210

631   ATC GTC CGG CTC CAT CCA GCC GGA CGC CTG GGC GAG CCG TCG GAG   675
211    I   V   R   L   H   P   A   G   R   L   G   E   P   S   E    225

676   GTG GCA GCC GCG ATC GGA TTC CTG GCC TCC GAC GAG GCC TCG TTC   720
226    V   A   A   A   I   G   F   L   A   S   D   E   A   S   F    240

721   ATC ACG GGT GTA GTA CTT CCC GTC GAC GGA GGT TAC CTC GCG CAG   765
241    I   T   G   V   V   L   P   V   D   G   G   Y   L   A   Q    255

766   TGA   768
256    *
```

```
>SALN

1     GTG TTT GTA CTG TTC ATC GGC GAC ATC GTG GGG GAC GAG GGG ACA   45
1      V   F   V   L   F   I   G   D   I   V   G   D   E   G   T    15

46    CGG TAC CTG GTA GAC CGG CTG CCC GAA CTG CGT GCC CGC CAC ACG   90
16     R   Y   L   V   D   R   L   P   E   L   R   A   R   H   T    30

91    GTG GAC CTC GTT GTG GCG AAT GCC GAG AAC TCG GCG GCC GAC GGC   135
31     V   D   L   V   V   A   N   A   E   N   S   A   A   D   G    45

136   TTG GGG ATG GGT CGT GTG CAG GTG GAG CAA CTG CTG TCC GGA GGG   180
46     L   G   M   G   R   V   Q   V   E   Q   L   L   S   G   G    60

181   GTT GAC GTC ATC ACC GGC GGA AAC CAC TCC TGG GAC AGC CCA GAG   225
61     V   D   V   I   T   G   G   N   H   S   W   D   S   P   E    75

226   TCG GTC GGG ATG CTC AAT CTG CCG CAA GTG GTT CGG CCC GCC AAC   270
76     S   V   G   M   L   N   L   P   Q   V   V   R   P   A   N    90

271   GTG GAC CCC GAG GTA CCC GGT CGG GGC TTT GTC CAT GTC CCG GTC   315
91     V   D   P   E   V   P   G   R   G   F   V   H   V   P   V   105

316   GGC GAG GAG ATG GTC ACC GTA CTG AAC CTG GCT GAT GGC TGC GCC   360
106    G   E   E   M   V   T   V   L   N   L   A   D   G   C   A   120

361   ATG GCG TCG ACC AAG GCG ACT GCG GGA CGG GTG GGT CCC GCC TAC   405
121    M   A   S   T   K   A   T   A   G   R   V   G   P   A   Y   135

406   TCG GCA TGG GAG GCG GCC GAC CGG CGC GGC ATC ACG ATC GTC GAC   450
136    S   A   W   E   A   A   D   R   R   G   I   T   I   V   D   150
```

```
451   TAC CAC GGC GAT CAC GTG CTG GAA AAG CAG ATC TTC GCC CAC ACC   495
151    Y   H   G   D   H   V   L   E   K   Q   I   F   A   H   T   165

496   GTC GAC GGC ACC GCC GCC GCG GTC GTC GGC ACG CAC ACC CAC GAA   540
166    V   D   G   T   A   A   A   V   V   G   T   H   T   H   E   180

541   GCC ACC GAC CTG CCA CGG CTG CTG CCG AAC GGC ACG GCG TTC GTC   585
181    A   T   D   L   P   R   L   L   P   N   G   T   A   F   V   195

586   ACA GAG GTG GGC ATG ACG GGA CCG GAC GGC GGC GTG CAG GGT TTC   630
196    T   E   V   G   M   T   G   P   D   G   G   V   Q   G   F   210

631   GCT CCC GCA AAC CTG GTG AAC AAC CTG CGG ACC ACC GGC AAC CCG   675
211    A   P   A   N   L   V   N   N   L   R   T   T   G   N   P   225

676   TTC GAC GGG CCG ATG CCC ACG GTG CAT CGA GCC CCG ATG GTG CTC   720
226    F   D   G   P   M   P   T   V   H   R   A   P   M   V   L   240

721   GGC GCG GTC CTG ATC GAG ATC GTC GGC GGT CTT GCC GCC GGC ATC   765
241    G   A   V   L   I   E   I   V   G   G   L   A   A   G   I   255

766   CAG CGG GTG CGG TGA   780
256    Q   R   V   R   *
```

>SALO

```
1     ATG TTG GAC GGC AGC GCC GTA CAC GGC CAA TCG GCG GTC CCC GCC   45
1      M   L   D   G   S   A   V   H   G   Q   S   A   V   P   A   15

46    GGC TAC CAG GTG GTT GAC CTT TCG GTG ACC ATC GCC GAG GAC TTG   90
16     G   Y   Q   V   V   D   L   S   V   T   I   A   E   D   L   30

91    CCA TGC TAC TGG GAG GGG CAT CAA CCC TTC CAG CAC AAG ACC TGG   135
31     P   C   Y   W   E   G   H   Q   P   F   Q   H   K   T   W   45

136   AAC TGG TTT GCC GAC CAT CGC AAC GCG ATG GCC TCC GTC TAC AGC   180
46     N   W   F   A   D   H   R   N   A   M   A   S   V   Y   S   60

181   CGC GGC GGC CCG TAC GCG ACC CGT TGG ATG GCC ATC GAC GAG CAC   225
61     R   G   G   P   Y   A   T   R   W   M   A   I   D   E   H   75

226   ACC GGA ACC CAC TTT GAT GCT CCT CTG CAC TTC ATC CCA CCT CCG   270
76     T   G   T   H   F   D   A   P   L   H   F   I   P   P   P   90

271   GAC TCC GGC CTG CCG GGA GCG GGT CCG GCG GGT GCC ATC TCG GCT   315
91     D   S   G   L   P   G   A   G   P   A   G   A   I   S   A   105

316   GAC CAG GTC GGG CTG GCG AGC CTC ATG GGA GCC GCC GCG GTA ATT   360
106    D   Q   V   G   L   A   S   L   M   G   A   A   A   V   I   120

361   GAC GTG GCA TCT CCC GGT GGA GGC GAA GGC GGG ATG AGC CCG CTG   405
121    D   V   A   S   P   G   G   G   E   G   G   M   S   P   L   135

406   ATC CAC CCG TCG GCG ATC GTG GCA TGG GAG GCC GCC AAC GGC ATG   450
136    I   H   P   S   A   I   V   A   W   E   A   A   N   G   M   150
```

```
451   CTG AAC CCC GGT GAG ATC GTG CTG CTG CGC ACC GGC TGG GAC CGG   495
151    L   N   P   G   E   I   V   L   L   R   T   G   W   D   R    165

496   CGC TAC CGA CGC GGT GCC GCC GGG GGC GGA TAT CTT CGT GAC GTC   540
166    R   Y   R   R   G   A   A   G   G   G   Y   L   R   D   V    180

541   CTC GCC GGG CGG AGC CCC GGA TGG CCG GCG CCG ACA GAG CCC ACA   585
181    L   A   G   R   S   P   G   W   P   A   P   T   E   P   T    195

586   GTG GAG CTA CTG CTG ACT CGG GGG GTG CAA TGC CTT GGC ATT GAC   630
196    V   E   L   L   L   T   R   G   V   Q   C   L   G   I   D    210

631   GCG CCG TCG ATC GGC CCG GCG CAC GAT CCG GTG GGC GTG CAT GTG   675
211    A   P   S   I   G   P   A   H   D   P   V   G   V   H   V    225

676   CGG GGC CTG GGA CAG GGA ATG GTC TAC GTC GAG TGT CTC GCC GCG   720
226    R   G   L   G   Q   G   M   V   Y   V   E   C   L   A   A    240

721   CTC GAC GCG CTG CCC CCT CGC GGC GCG TGG TTT TGT TTC CTT CCC   765
241    L   D   A   L   P   P   R   G   A   W   F   C   F   L   P    255

766   ATC AAG GTG GAG GGC AGC ACT GGG GCT CCT GGC CGG GCG ATC GCC   810
256    I   K   V   E   G   S   T   G   A   P   G   R   A   I   A    270

811   CTG ATG CCG GCC ACG GAG TCG TGA   834
271    L   M   P   A   T   E   S   *
```

```
>salP

1     ATG ACC AGC ATA GAA TCT AGT CCG AAA CGG ATT CTA CTC CTT GGC    45
1      M   T   S   I   E   S   S   P   K   R   I   L   L   L   G    15

46    GAC AGC ACT ATT TTT CGA GAG GCG ATG GCC GAG ATC TGC GAT CGC    90
16     D   S   T   I   F   R   E   A   M   A   E   I   C   D   R    30

91    GAG CCC GAC CTG CGG GTG GTC GGG CAG GCG GCG GCG GGG CCC ATG   135
31     E   P   D   L   R   V   V   G   Q   A   A   A   G   P   M    45

136   GCG GCG GCC CTG GTT CGA CGG ACG GCG CCA GAC GTG GTA CTG CTG   180
46     A   A   A   L   V   R   R   T   A   P   D   V   V   L   L    60

181   AGC GTC GGC CCG TCT GAC CAA GGG ATC ATG CGG CTG GTC GGC CAG   225
61     S   V   G   P   S   D   Q   G   I   M   R   L   V   G   Q    75

226   ATC CTG GCC GCG ACC TGC CAG CCC CGG CTG GTA ATC CTG GCG ATG   270
76     I   L   A   A   T   C   Q   P   R   L   V   I   L   A   M    90

271   CAC GAG AAC GTG CGC CTG GCG CGC CGG TAC CTC GCC ATG GGT ACG   315
91     H   E   N   V   R   L   A   R   R   Y   L   A   M   G   T   105

316   TGG GCG TAC GTC TCG CGG AAC TCG ACG CGT GAG GAA CTG CTG AGC   360
106    W   A   Y   V   S   R   N   S   T   R   E   E   L   L   S   120

361   GTC GTA CGG ATC GTG AGC CGC AAC CGC GAC AGC GCC GTG CTT TCC   405
121    V   V   R   I   V   S   R   N   R   D   S   A   V   L   S   135
```

```
406   GTG CAG AAG AGC ATG CTG CAA CAT CTG ACC GGG GTC GGC ACG CGG   450
136    V   Q   K   S   M   L   Q   H   L   T   G   V   G   T   R    150

451   TCA CTG TCG GGC CGG GAG TTC GAG GTG ATC GAG CTG GTC GCG GCG   495
151    S   L   S   G   R   E   F   E   V   I   E   L   V   A   A    165

496   GGC CTG AGC AAC GCC CAG ATC GCC GGC CGC CTC TCC ATC TCC GAG   540
166    G   L   S   N   A   Q   I   A   G   R   L   S   I   S   E    180

541   GGC ACG GTC AAG CGC CAC CTC ACC AAC ACG TAC GAC AAG CTC GAT   585
181    G   T   V   K   R   H   L   T   N   T   Y   D   K   L   D    195

586   GTG CGC TCA CGA GTG GCC GCC GTC AAC AGG CTC GCG GCG ATG GGT   630
196    V   R   S   R   V   A   A   V   N   R   L   A   A   M   G    210

631   TTG ATC GGC CCG AAC GCG ATG CTG GAG ACG AGC AGC CTG GTG GCC   675
211    L   I   G   P   N   A   M   L   E   T   S   S   L   V   A    225

676   TGA   678
226    *
```

>salQ

```
1     ATG TCC CCC GGC TTG CAG CCA AGC GAA TTC ACG TTG GAC CAG CGC    45
1      M   S   P   G   L   Q   P   S   E   F   T   L   D   Q   R    15

46    TAT CTG TGC CAG GAC GGC ACC GCC TAC CTC ACC GGC GTG CAG GCG    90
16     Y   L   C   Q   D   G   T   A   Y   L   T   G   V   Q   A    30

91    CTG GTG CGT CTC CTG CTT GAC CGT GTC CGC CAC GAC CGG CGT GCC   135
31     L   V   R   L   L   L   D   R   V   R   H   D   R   R   A    45

136   GGA AAC GAC GCG GCG GTC TTC GTC TCC GGC TAC GAA GGG TCG CCC   180
46     G   N   D   A   A   V   F   V   S   G   Y   E   G   S   P    60

181   CTG GCG GGC TAC GAC TTG GAG CTC GCC CGC CGG GGC CGA CTG CTG   225
61     L   A   G   Y   D   L   E   L   A   R   R   G   R   L   L    75

226   GCC GAG CAC GGC GTG GTG CAT CGG CCC GCA CTG AAC GAG GAG CTT   270
76     A   E   H   G   V   V   H   R   P   A   L   N   E   E   L    90

271   GCG GCG ACG GCC GTG CTC GGC AGC CAG CTC GCT GCG GGC GTC GGC   315
91     A   A   T   A   V   L   G   S   Q   L   A   A   G   V   G   105

316   AGA CAG CGC CCC GAC GGT GTC GTT GGT GTC TGG TAC GGC AAG GCG   360
106    R   Q   R   P   D   G   V   V   G   V   W   Y   G   K   A   120

361   CCG GGT CTC GAC CGC GCG GGC GAC GCG TTG CGC CAC GCG AAC TTC   405
121    P   G   L   D   R   A   G   D   A   L   R   H   A   N   F   135

406   GCG GGC ACC GAC CCG CGG GGC GGT GCC ATC GCG CTC GTC GGG GAC   450
136    A   G   T   D   P   R   G   G   A   I   A   L   V   G   D   150

451   GAT CCA GCG GCG AAG TCG TCG ACT GTG CCG AGC GCG TCG GAG TTC   495
151    D   P   A   A   K   S   S   T   V   P   S   A   S   E   F   165
```

```
496   GCC CTG GCG GAC CTC TCA ATG CCG GTC CTC GTG CCA GCA GAC GCC   540
166    A   L   A   D   L   S   M   P   V   L   V   P   A   D   A   180

541   GCC GAC CTG CTC CGC TAC GGC CTG CAC GCC ATC GAG ATG TCC CGG   585
181    A   D   L   L   R   Y   G   L   H   A   I   E   M   S   R   195

586   GCG TCC GGC CTA TGG TCA GCG CTC AAG ATC GTG ACC AAC GTC GCC   630
196    A   S   G   L   W   S   A   L   K   I   V   T   N   V   A   210

631   GAC GCG GCG AGC ACG GTG CCG GTC ACC CCG TCC TGG CGC TTC CCG   675
211    D   A   A   S   T   V   P   V   T   P   S   W   R   F   P   225

676   ACA GTG GCT GGT GGA TAT CGC CAC CGG CCC ACG TCC CGG CTG CTC   720
226    T   V   A   G   G   Y   R   H   R   P   T   S   R   L   L   240

721   GGC GCG GAA CTC GCC GCC TTG GAG GAG AGC CTT CAC CGG GTC CGC   765
241    G   A   E   L   A   A   L   E   E   S   L   H   R   V   R   255

766   CTC CCC TTG GCG GTG GAG TAT CTG CGC GCC GCC GGC CTG AAC GAG   810
256    L   P   L   A   V   E   Y   L   R   A   A   G   L   N   E   270

811   CTG TAC GGG CAC GCG GAT GAC CGG ATC GGG ATC GTG GCC GCC GGC   855
271    L   Y   G   H   A   D   D   R   I   G   I   V   A   A   G   285

856   AAG TCA TAT CTG GAC GTC CGC CAA GCG CTG CGG ACG CTT GGC ATG   900
286    K   S   Y   L   D   V   R   Q   A   L   R   T   L   G   M   300

901   GAC GAG GAA GCT CTG CGC CGC TAT CGC GTA CGG CTA CTC AAG CTG   945
301    D   E   E   A   L   R   R   Y   R   V   R   L   L   K   L   315

946   GCG GCG GTC TAC CCC CTT GAG TCA GCC GTG GTG CGG GAG TTC GCC   990
316    A   A   V   Y   P   L   E   S   A   V   V   R   E   F   A   330

991   GCG GGG CTG GAC GAG ATA GTG GTG GTG GAG GAG AAA CGG GGT TTC  1035
331    A   G   L   D   E   I   V   V   V   E   E   K   R   G   F   345

1036  ATC GAA TCC GCC GTC AAG GAT CTG CTC TAT GGC AGC GGC GAC GCG  1080
346    I   E   S   A   V   K   D   L   L   Y   G   S   G   D   A   360

1081  CCG GTG GTC TGC GGG AAG ACG GAT CCG GCG GGG CGG TCG CTG TTC  1125
361    P   V   V   C   G   K   T   D   P   A   G   R   S   L   F   375

1126  CGC CCC ACG GGT GAG CTG AAT CCC GAC GTG ATC GCG GCT GGC CTC  1170
376    R   P   T   G   E   L   N   P   D   V   I   A   A   G   L   390

1171  GCC CGG CGA CTG TCG GCG ACC GGC CGC TGC GAA CCC GTG GAA GCG  1215
391    A   R   R   L   S   A   T   G   R   C   E   P   V   E   A   405

1216  TGG CAG GGA CGT CGG CGC CGT GAC CGC ATC GAC CTG CCG GTG GTG  1260
406    W   Q   G   R   R   R   R   D   R   I   D   L   P   V   V   420

1261  GCG CGA ACC CCA TAC TTC TGC TCC GGC TGC CCG CAC AAC TCG TCG  1305
421    A   R   T   P   Y   F   C   S   G   C   P   H   N   S   S   435

1306  ACG AAG GCG CCT CCC GGT ACC CTG ACC GCC GGC GGC ACC GGA TGT  1350
436    T   K   A   P   P   G   T   L   T   A   G   G   T   G   C   450

1351  TCC ACC ATG ACC TTC TTC ATG GAT CCG AAG CGG GTC GGC GAC ATC  1395
451    S   T   M   T   F   F   M   D   P   K   R   V   G   D   I   465
```

```
1396  ATC GGC GTC ACG CAG ATG GGC GGC GAG GGT GCC CAG TGG ATC GGG   1440
466    I   G   V   T   Q   M   G   G   E   G   A   Q   W   I   G    480

1441  ATG GCA CCC TTC GTC GAC ATC GAT CAC TTC GTG CAG AAC GTA GGC   1485
481    M   A   P   F   V   D   I   D   H   F   V   Q   N   V   G    495

1486  GAC GGC ACC TTT CAC CAC TCG GCC AGC CTC GCG CTG CGA GCC GCA   1530
496    D   G   T   F   H   H   S   A   S   L   A   L   R   A   A    510

1531  GTG GCG GCC GGC GTC AAC ATT ACG TAC AAG GTG CTG TAC AAC TCT   1575
511    V   A   A   G   V   N   I   T   Y   K   V   L   Y   N   S    525

1576  GCC GTG GCC ATG ACC GGT GGT CAG GAC GCC GTG GGT GCC TTG CCG   1620
526    A   V   A   M   T   G   G   Q   D   A   V   G   A   L   P    540

1621  GTG CAT CGG CTG GCC GCT CTG CTT CTG GCC GAG GGC GCG ACC AAG   1665
541    V   H   R   L   A   A   L   L   L   A   E   G   A   T   K    555

1666  GTG ATC GTC ACC TCG GAG GAT CCA AAG CGG CTC GGC CGG CAT CCC   1710
556    V   I   V   T   S   E   D   P   K   R   L   G   R   H   P    570

1711  ATG CCG GCC GGC GTC GAG GTA CGC CAC CGC GAC GAT CTG CCG CGC   1755
571    M   P   A   G   V   E   V   R   H   R   D   D   L   P   R    585

1756  AGC CAG GCG GAA CTC GCC GTG ACT CCC GGC GTC ACC GTG CTC ATC   1800
586    S   Q   A   E   L   A   V   T   P   G   V   T   V   L   I    600

1801  CAC GAC CAG GAG TGC GCG GCG GTA CAA CGG CGT AAG CGG CGC CGC   1845
601    H   D   Q   E   C   A   A   V   Q   R   R   K   R   R   R    615

1846  GGC CTG CGG CCG CAG CCC GCC TAC AAA GTG CTG ATC AAC GAA CGG   1890
616    G   L   R   P   Q   P   A   Y   K   V   L   I   N   E   R    630

1891  TTG TGC GAG GGA TGC GGC GAC TGC GGG GCG CAG TCG AAC TGC CTG   1935
631    L   C   E   G   C   G   D   C   G   A   Q   S   N   C   L    645

1936  TCG GTC CAG CCG GTA GAA AGC GAC TTT GGC CGC AAG ACA CGT ATT   1980
646    S   V   Q   P   V   E   S   D   F   G   R   K   T   R   I    660

1981  CAC CAG TCC TCG TGT AAC GCC GAC TAC GCA TGC CTA CAG GGC CGG   2025
661    H   Q   S   S   C   N   A   D   Y   A   C   L   Q   G   R    675

2026  TGC CCA TCA TTC TTG ACC ATT GTC CCC GGC CGA TCC CCG CAC CGC   2070
676    C   P   S   F   L   T   I   V   P   G   R   S   P   H   R    690

2071  GAG CTG GCC TCG GTG CCC ACG GCT AAA CCA TCG ATG CCG CGG CAG   2115
691    E   L   A   S   V   P   T   A   K   P   S   M   P   R   Q    705

2116  CGC GTT GTC GGC GAC TTC TCA GTG CGG ATC ACC GGC ATC GGC GGC   2160
706    R   V   V   G   D   F   S   V   R   I   T   G   I   G   G    720

2161  AGC GGC GTG GTA ACC CTG GCT CAG ATC ATC GCC ACA GCC GCA GTG   2205
721    S   G   V   V   T   L   A   Q   I   I   A   T   A   A   V    735

2206  CTC GAC GGC AGG CAC GTG CGG ACG CTC GAC CAG ACC GGG CTG GCC   2250
736    L   D   G   R   H   V   R   T   L   D   Q   T   G   L   A    750

2251  CAG AAG GGC GGT GCG GTC GTC TCC GAT GTC CGC GTC ACC GAG CGT   2295
751    Q   K   G   G   A   V   V   S   D   V   R   V   T   E   R    765
```

```
2296  GCC GTC GAG TCT GCC GCA CGG CTA GCG GAC GGT GAA TGC GAT CTC   2340
766    A   V   E   S   A   A   R   L   A   D   G   E   C   D   L    780

2341  TAC CTA GCC ACC GAC CTG CTC GTC GCG GCT CAG CCC GCC AAC CTC   2385
781    Y   L   A   T   D   L   L   V   A   A   Q   P   A   N   L    795

2386  GCT GCA GCC GAC ACC GCC CGG ACC GTC GTC GTG GCA TCG ACC TCG   2430
796    A   A   A   D   T   A   R   T   V   V   V   A   S   T   S    810

2431  CCG GTG CCC ACC GGT CAG ATG GTT ACC GAG GTG TCG GGC CCA CGG   2475
811    P   V   P   T   G   Q   M   V   T   E   V   S   G   P   R    825

2476  CCC GAC GTA GAC GAA CTG TGC TCG GCT CTC GAT CGA TCG AGT GCG   2520
826    P   D   V   D   E   L   C   S   A   L   D   R   S   S   A    840

2521  CGC GCC ACC TTC TTC GAC GCC CGA TCC CTC GCC TTG CGC CTG CTG   2565
841    R   A   T   F   F   D   A   R   S   L   A   L   R   L   L    855

2566  GGT GAC GAC CAG TTC GCC AAT ATT CTG CAA CTA GGC GCC GCC TAC   2610
856    G   D   D   Q   F   A   N   I   L   Q   L   G   A   A   Y    870

2611  CAG GAG GGC GCC ATC CCG CTA GAA GCG GCC AGC ATC GAA CAG GCG   2655
871    Q   E   G   A   I   P   L   E   A   A   S   I   E   Q   A    885

2656  ATC CGG CTC AAC GGC GTC GGC GTG GAG CGC AAC CTC AGC GCC TTC   2700
886    I   R   L   N   G   V   G   V   E   R   N   L   S   A   F    900

2701  TGC CTC GGC CGC CTG GTG GTT GAC GAC CCT CAG ACG GTG TCG CCG   2745
901    C   L   G   R   L   V   V   D   D   P   Q   T   V   S   P    915

2746  GCA GCT CCC GTA GCC CAG GAG CAA CGA CCG TTG CAG GGC GAT GCA   2790
916    A   A   P   V   A   Q   E   Q   R   P   L   Q   G   D   A    930

2791  GCA TCG TTG CGG GCG ACA ATC GAT GCC GGC GAG GAC ACC CCA CTC   2835
931    A   S   L   R   A   T   I   D   A   G   E   D   T   P   L    945

2836  GCG CGC CGG CTG GAT GCG CTC GTG GCA GAT CTC GTC GCC TAC CAG   2880
946    A   R   R   L   D   A   L   V   A   D   L   V   A   Y   Q    960

2881  GAT CTG GCC TAT GCC CGC GAG TTT GTG GAG TTC GTG GAA CAC GTA   2925
961    D   L   A   Y   A   R   E   F   V   E   F   V   E   H   V    975

2926  CGC CGG GTG GAG GGA CGG GCG GGG CCG CTC ACC GCA GCC GTG GCG   2970
976    R   R   V   E   G   R   A   G   P   L   T   A   A   V   A    990

2971  GGC AAT CTA TAC AAA CTT ATG GCC TAC AAG GAT GAG TAC GAG GTG   3015
991    G   N   L   Y   K   L   M   A   Y   K   D   E   Y   E   V   1005

3016  GCC CGG CTC TCG CTC GAT GAA CGC CTG ACC CAC GAC ATA GCC GCG   3060
1006   A   R   L   S   L   D   E   R   L   T   H   D   I   A   A   1020

3061  CAG TTC GGA GCT GGG GCG CGA TAC CAC TAC CTA TTG CAG CCA CCG   3105
1021   Q   F   G   A   G   A   R   Y   H   Y   L   L   Q   P   P   1035

3106  GTT CTG AGG AAG CTC GTG CGC CGC AAG ATT CGC CTC GGG CAG TGG   3150
1036   V   L   R   K   L   V   R   R   K   I   R   L   G   Q   W   1050

3151  TTC CGA GTC GTG CTC CGG CTG CTG TAC GCA CTG CGC AGA CTG CGA   3195
1051   F   R   V   V   L   R   L   L   Y   A   L   R   R   L   R   1065
```

```
3196  GGG AGC AGG CTC GAT GTC TTC GGC TAC ACA CGC GTG CGC CGG ATC   3240
1066   G   S   R   L   D   V   F   G   Y   T   R   V   R   R   I   1080

3241  GAA CGG AGG CTT GTG GCG AAC TAT CGC AGG ACG ATC CTC GCC GCG   3285
1081   E   R   R   L   V   A   N   Y   R   R   T   I   L   A   A   1095

3286  TTG TCG GCG GCC AAC CCC GAC GAC CCG GCG CTG GTA GAG CTG GCG   3330
1096   L   S   A   A   N   P   D   D   P   A   L   V   E   L   A   1110

3331  TCT CTG CCA GAC ATG GTG CGT GGC TTC GAC GAG GTC AAG CTC GCC   3375
1111   S   L   P   D   M   V   R   G   F   D   E   V   K   L   A   1125

3376  AAT GTG ACC GAG TAC CAC CGG CGC CAG GAG CAG TTG CTG CAG TCG   3420
1126   N   V   T   E   Y   H   R   R   Q   E   Q   L   L   Q   S   1140

3421  CTG CCC CAC CCC GTC CAG CAG GTC GGG TAG   3450
1141   L   P   H   P   V   Q   Q   V   G   *
```

>salR

```
1     GTG GCC GAT CAC GCG GAC TTC AGC CTG GCT CAA CTC CGG TAC TTC   45
1      V   A   D   H   A   D   F   S   L   A   Q   L   R   Y   F   15

46    GTA ACT TCC GCC GAG ATG GGA AAC ATA TCC GGC GCG GCA CAG CAA   90
16     V   T   S   A   E   M   G   N   I   S   G   A   A   Q   Q   30

91    TTA CAT GCT TCG CAG TCG GCC GTC TCG ATG GCG ATC CAG CGC CTG   135
31     L   H   A   S   Q   S   A   V   S   M   A   I   Q   R   L   45

136   GAG AAA CAA CTG GAC ACC AAG CTG TTC TTC CGG CAC CGC ACC AAA   180
46     E   K   Q   L   D   T   K   L   F   F   R   H   R   T   K   60

181   GGA GTC AGC CTG ACT CCC AGC GGC CGC GTC CTG CTC GAC CCC GCG   225
61     G   V   S   L   T   P   S   G   R   V   L   L   D   P   A   75

226   AAG TCG TTG CTG GCC CAG GCG CAG GAG ATG CTC GCC CAC AGC CGC   270
76     K   S   L   L   A   Q   A   Q   E   M   L   A   H   S   R   90

271   GAG TGG CAG GGC GAG ACC TGC GGA ATG CTC AAT GTG GCC GTT CTC   315
91     E   W   Q   G   E   T   C   G   M   L   N   V   A   V   L   105

316   CGC TCA GTC GCG CCC GAT CTG GTG CCG GCT GTG ATG AGC CGA GTA   360
106    R   S   V   A   P   D   L   V   P   A   V   M   S   R   V   120

361   AAG AAG ATG CGC CCT GGG CTG GCG GTC ACC GTC CAC GAG GGA ACC   405
121    K   K   M   R   P   G   L   A   V   T   V   H   E   G   T   135

406   ATG GAC GAT GTT CTC GAA CTG CTG CGC ACG GGC ACG TGT GAA CTC   450
136    M   D   D   V   L   E   L   L   R   T   G   T   C   E   L   150

451   GCC GTA GTC CCG GAG CTG CCA GGT CAG GCT CTG TCG TTC GCC CGG   495
151    A   V   V   P   E   L   P   G   Q   A   L   S   F   A   R   165

496   CTG GCG AAG GTG CTC ATG GCA GCG GTC GTT GCA TGG TCC GAT CCG   540
166    L   A   K   V   L   M   A   A   V   V   A   W   S   D   P   180
```

```
541   CTG GCG CCC GTG GGG CGC GCA ACG CTT CGA CAG TTG GCC ACC CGG   585
181    L   A   P   V   G   R   A   T   L   R   Q   L   A   T   R   195

586   CGT CTC CTC GTG GCC GAC CTG GAC GTT GAT CGC CAA GTC GGC CTG   630
196    R   L   L   V   A   D   L   D   V   D   R   Q   V   G   L   210

631   GAC ACC CGG ACA CGG CTG TTC GCC GAC GCC GGC ACC CCC ACG CCA   675
211    D   T   R   T   R   L   F   A   D   A   G   T   P   T   P   225

676   GAG GTG GTG CTG ACA TCG AGC ATC GCC ACG ATG CTT GGC CTC GTC   720
226    E   V   V   L   T   S   S   I   A   T   M   L   G   L   V   240

721   GGG GTC GGC GAG GGC TTC GCC CTG CTA TAC CGG ACC GCA GCC AGC   765
241    G   V   G   E   G   F   A   L   L   Y   R   T   A   A   S   255

766   CAG TGC TAC CGA CAG GAA CTC GCC TGG ATC GAC ATC GTC TCC GAA   810
256    Q   C   Y   R   Q   E   L   A   W   I   D   I   V   S   E   270

811   CGC CCG TGC GCA TCC TGG CTT GGT GTG GCG ACC ATA CCC GGC CTG   855
271    R   P   C   A   S   W   L   G   V   A   T   I   P   G   L   285

856   GCT CTC AGC AGC CGT GCG GAG CTA TTC GTT TCC CTG CTC GAT GAC   900
286    A   L   S   S   R   A   E   L   F   V   S   L   L   D   D   300

901   GCT GTC CGG AGC GTC TAC TGC ACG ACG GCG GAC GAC GGG GTG AGT   945
301    A   V   R   S   V   Y   C   T   T   A   D   D   G   V   S   315

946   TCC CCT GGC AGC ACT GAT ATC TTC GAT CAT GGA GCC TCT TTT GCT   990
316    S   P   G   S   T   D   I   F   D   H   G   A   S   F   A   330

991   ACG GAC CCC GCC GCA CCC GCC TGG TCG ACC TAG   1023
331    T   D   P   A   A   P   A   W   S   T   *
```

>salS

```
1     ATG AGG CCG TTC ATC ACT TCG CCC GCG GTG TTG CTG GGG AAG TCG   45
1      M   R   P   F   I   T   S   P   A   V   L   L   G   K   S   15

46    GCG TCG TCC CAG CGT AGG GGC GAG AAG CTC GAT CGG CTG ACC GGA   90
16     A   S   S   Q   R   R   G   E   K   L   D   R   L   T   G   30

91    CTT TCA CCC ATC AGA CGG GGT CGG ACT CGT CGC CGC GCG CAT CAG   135
31     L   S   P   I   R   R   G   R   T   R   R   R   A   H   Q   45

136   ATT CTG CTA CGC CGA GCT GGT CGG CTT CGG AAA GAG GCA ATG ACC   180
46     I   L   L   R   R   A   G   R   L   R   K   E   A   M   T   60

181   ACC GGC CCG TAC CCA GCC TCG AGC CCG AGG GCC GCG GTG TGG CTG   225
61     T   G   P   Y   P   A   S   S   P   R   A   A   V   W   L   75

226   CCC GCG ATC GGT CTG CTG GTG GCG AGC GAT CAG CCC AGG AAG CGA   270
76     P   A   I   G   L   L   V   A   S   D   Q   P   R   K   R   90

271   AAG ACG GGT TCG GCG ACA CAG ACG GGC TTG TCG TCG CCC TGT CGT   315
91     K   T   G   S   A   T   Q   T   G   L   S   S   P   C   R   105
```

```
316   TCA ACC GAG GCG GCG GCC GTC ACC TGC AGT CCG CCC CTA ACC GGT   360
106    S   T   E   A   A   A   V   T   C   S   P   P   L   T   G    120

361   TCG ATA TCG ACG AGG GTG GCG GAC AGG CGT AGC TTG GCA CCG ACC   405
121    S   I   S   T   R   V   A   D   R   R   S   L   A   P   T    135

406   GGA ACA GCA GCA GGG AAG CGG ACC CTG TTG AGC CCG TAG   444
136    G   T   A   A   G   K   R   T   L   L   S   P   *
```

>SALT

```
1     ATG AGC GCT CCG AAT CTG AAC GCG GAC ATC GGC GTC ATC GGC GGC    45
1      M   S   A   P   N   L   N   A   D   I   G   V   I   G   G    15

46    ACG GGT CTG TAC GAG TTG CTG GGC GAG ACC GAA GAG GTG AAA ATC    90
16     T   G   L   Y   E   L   L   G   E   T   E   E   V   K   I    30

91    TCC ACG CCG TAT GGC GCG CCG TCG GAT CTG ATC ACG GTG GCA GAG   135
31     S   T   P   Y   G   A   P   S   D   L   I   T   V   A   E    45

136   TTG AAC AGC CGC CAG GTC GCC TTC CTA CCG CGG CAC GGG CGG GAC   180
46     L   N   S   R   Q   V   A   F   L   P   R   H   G   R   D    60

181   CAC CGT TTC CCG CCG CAC AGA ATC CCG TAC CGG GCC AAC ATC TGG   225
61     H   R   F   P   P   H   R   I   P   Y   R   A   N   I   W    75

226   GCG CTG CGT TCG CTG GGT GTG CGG CAG ATC CTT GCA CCG TGC GCA   270
76     A   L   R   S   L   G   V   R   Q   I   L   A   P   C   A    90

271   GTG GGC GGC CTT CGG CCG GAC TTC ATT GCC GGC ACG TTC GTC GTG   315
91     V   G   G   L   R   P   D   F   I   A   G   T   F   V   V   105

316   CCG GAT CAG CTG ATC GAC CGC ACG AGT GGG CGG GAG CAG ACC TTC   360
106    P   D   Q   L   I   D   R   T   S   G   R   E   Q   T   F   120

361   TGC GAC ACC GGC GCA GTG CAT ACG ACC TTC GCA GAG CCG TAT TGC   405
121    C   D   T   G   A   V   H   T   T   F   A   E   P   Y   C   135

406   CCG GAC GGG CGC CGC GCC GCG TTG GCC GCG GCG GAA CGG GTA AAC   450
136    P   D   G   R   R   A   A   L   A   A   A   E   R   V   N   150

451   CGC GTC GAC GGC GGC ACC ATG GTC GTA ATC GAC GGG CCA CGG TTC   495
151    R   V   D   G   G   T   M   V   V   I   D   G   P   R   F   165

496   TCG ACC CGT GCC GAG TCA CGG TGG CAC GCG GCC GCC GGC GGC ACC   540
166    S   T   R   A   E   S   R   W   H   A   A   A   G   G   T   180

541   ATC ATC AAC ATG ACC GGC CAG CCC GAG GCC GCG CTC GCC CGC GAG   585
181    I   I   N   M   T   G   Q   P   E   A   A   L   A   R   E   195

586   CTC GCA CTC TGC TAC ACC AGC ATC GCT TTG GTA ACC GAC CTC GAC   630
196    L   A   L   C   Y   T   S   I   A   L   V   T   D   L   D   210

631   GCC GGC GTG GAA GGC AGC CAG GGC GTG ACC CAC GAC GAG GTG CTG   675
211    A   G   V   E   G   S   Q   G   V   T   H   D   E   V   L   225
```

```
676   CGG GCC TTT GCG AAG AAC GTG ACC GTC CTG CGA CGC GTA CTC TTC   720
226    R   A   F   A   K   N   V   T   V   L   R   R   V   L   F   240

721   GAT GCC GTC GCT GCG CTA CCG GTG GAA CGC ACC TGT CCC TGC TCG   765
241    D   A   V   A   A   L   P   V   E   R   T   C   P   C   S   255

766   CGC GCC TTC GCC GTT CCC GTG GAC CTC GAC GCC GGT   801
256    R   A   F   A   V   P   V   D   L   D   A   G
```

>salU

```
1     ATG GCA ACG AAT CAA CTC CTG CCC CAT ACG AGA GGC GAT TTT GGT   45
1      M   A   T   N   Q   L   L   P   H   T   R   G   D   F   G   15

46    GTG TAT GGA ACG CGC ATG CGA CCC GGC GTG GAG CGG TCG ATC CGC   90
16     V   Y   G   T   R   M   R   P   G   V   E   R   S   I   R   30

91    CAT TCG CCA CCA GCG ACG AAC GAT GGC GAC AGA GAA CTC TGC CTA   135
31     H   S   P   P   A   T   N   D   G   D   R   E   L   C   L   45

136   GAC CGG TGG CGC GAG CTA CCC CGG CGG CAG GTG CCA CCC TGG CCC   180
46     D   R   W   R   E   L   P   R   R   Q   V   P   P   W   P   60

181   GAT CCC GCC GAG GTG GCA GCA GTG TGC GCG ACA CTC GGC AAA ATG   225
61     D   P   A   E   V   A   A   V   C   A   T   L   G   K   M   75

226   CCG CCG ATC GTC ACA CCC TAC GAG GTC GAT GAA CTG CGT CAC CGC   270
76     P   P   I   V   T   P   Y   E   V   D   E   L   R   H   R   90

271   CTG GCC GAG GTA TGC GAA GGA CGC GCC TTC CTG TTA CAG GGC GGC   315
91     L   A   E   V   C   E   G   R   A   F   L   L   Q   G   G   105

316   GAC TGC GCC GAG ACG TTC ACC GGA AAC ACT GAG AGC CAC CTG CTG   360
106    D   C   A   E   T   F   T   G   N   T   E   S   H   L   L   120

361   GGC ACC ACG CGG ACG CTG CTG CAG ATG GCG ATG GCG ATA ACG TAC   405
121    G   T   T   R   T   L   L   Q   M   A   M   A   I   T   Y   135

406   GGC GGT TCC GTT CCG GTG GTG AAG GTG GCG CGC CTC GCC GGC CAG   450
136    G   G   S   V   P   V   V   K   V   A   R   L   A   G   Q   150

451   TAC GGG AAG CCC CGA TCC TCG GCG ACC GAT TCC CTG GGG TTG CCC   495
151    Y   G   K   P   R   S   S   A   T   D   S   L   G   L   P   165

496   GCG TAC CGT GGC GAT ATT ATC AAT GCG CGG CAC CCG GCC GAG TCG   540
166    A   Y   R   G   D   I   I   N   A   R   H   P   A   E   S   180

541   GCG CGG GCC GCC GAT CCG CAG CGC ATG ATC GAT GCG TAC GCG AAT   585
181    A   R   A   A   D   P   Q   R   M   I   D   A   Y   A   N   195

586   TCC GCG GTC GCG ATG AAC CTC ATC CGC GCC TAT CCG CCG GAC GAT   630
196    S   A   V   A   M   N   L   I   R   A   Y   P   P   D   D   210

631   CTG ACC GAC CTC GAA GAG CTG TAC GAC GAT ACC TAC GAC CTC ATC   675
211    L   T   D   L   E   E   L   Y   D   D   T   Y   D   L   I   225
```

```
676   CGC GCT TCC CCG GCC GGA GCC CGG TAC CAG GTC ATC TCC GGC GAG   720
226    R   A   S   P   A   G   A   R   Y   Q   V   I   S   G   E   240

721   ATC GAC CGG GCG CGC GGC TTC GTC CGC GCG TGG GGG CCG AGC GAG   765
241    I   D   R   A   R   G   F   V   R   A   W   G   P   S   E   255

766   CGC CAC GCG TTG CGG GAA TCA AAG GTG TAC TGC TCG CAC GAG GCG   810
256    R   H   A   L   R   E   S   K   V   Y   C   S   H   E   A   270

811   CTG GTG CTC GAG TAC GAC CGG GCG CTG ACC CGA ATC AAC GAC GGC   855
271    L   V   L   E   Y   D   R   A   L   T   R   I   N   D   G   285

856   CGG GCA TAC GCG TTG TCG GGT CAC TTT CTG TGG GTC GGC GAA CGC   900
286    R   A   Y   A   L   S   G   H   F   L   W   V   G   E   R   300

901   ACC CGC CAG CTT GAC CAC GCG CAC GTC GAC TTT GTC GCG CGT ATC   945
301    T   R   Q   L   D   H   A   H   V   D   F   V   A   R   I   315

946   GCC AAC CCA ATC GGC GTG AAG CTC GGC CCG GCC GCC AGC CCG CAT   990
316    A   N   P   I   G   V   K   L   G   P   A   A   S   P   H   330

991   GCT GCG ATC GAG CTG TGC GAG CGG CTC AAC CCC GAG AAC CTG CCC  1035
331    A   A   I   E   L   C   E   R   L   N   P   E   N   L   P   345

1036  GGG CGG CTC ACC CTG ATC AGC CGC ATG GGC AAC CGC CAG GTA CGT  1080
346    G   R   L   T   L   I   S   R   M   G   N   R   Q   V   R   360

1081  GAC GTG TTT CCG GCG ATC GTC GAC AAG GTC ACG GCC GCG GGC GCC  1125
361    D   V   F   P   A   I   V   D   K   V   T   A   A   G   A   375

1126  AAG GTC GTT TGG CAG TGC GAT CCG ATG CAC GGC AAC ACC GAG CAG  1170
376    K   V   V   W   Q   C   D   P   M   H   G   N   T   E   Q   390

1171  TCC TCA CAC GGC TTC AAG ACC CGG CGC CTC GAC CGG GTC GTT GAC  1215
391    S   S   H   G   F   K   T   R   R   L   D   R   V   V   D   405

1216  GAG TTG CTG GGC TAC TTC GAC GTG CAC CGC AGT CTG GGC ACC CAT  1260
406    E   L   L   G   Y   F   D   V   H   R   S   L   G   T   H   420

1261  CCG GGA GGC GTC CAC GTC GAG CTC ACC GGT GAG AAC GTC ACC GAG  1305
421    P   G   G   V   H   V   E   L   T   G   E   N   V   T   E   435

1306  TGC CTC GAC GGG ATA CGT GGC GTC GAG GAC CAA CAC CTA CCG GAT  1350
436    C   L   D   G   I   R   G   V   E   D   Q   H   L   P   D   450

1351  CGC TAC GAG ACT GCC TGC GAC CCG CGG CTG AAC ATG CGG CAG AGC  1395
451    R   Y   E   T   A   C   D   P   R   L   N   M   R   Q   S   465

1396  TTA GAA CTC GCG TTG CTG GTC GCC GAG ATT CTA CGC GGC   1434
466    L   E   L   A   L   L   V   A   E   I   L   R   G
```

>salV

```
1     ATG TTC ACA CAG TAC TAT CCG GAA GAC GCC GCT CTC ACA GTC GAG   45
1      M   F   T   Q   Y   Y   P   E   D   A   A   L   T   V   E   15

46    CTG CGT AAT CAG GTT GAG GCG TTC ATT GCC GGC AAG TGG GAC GCG   90
16     L   R   N   Q   V   E   A   F   I   A   G   K   W   D   A   30

91    GCG ATG CTG GCA GCG CAC CAG ACC CGC AAG CTG CGT GAG GTC GTT   135
31     A   M   L   A   A   H   Q   T   R   K   L   R   E   V   V   45

136   GCT TTT GCC CGC GAC AAC TCG CCG TTT TAC GGG GAG CGC CTG GTC   180
46     A   F   A   R   D   N   S   P   F   Y   G   E   R   L   V   60

181   GGG GTG ACC GAC GAG GTG ATC GCG GCC CTG CAC CCC GGT AAC CTG   225
61     G   V   T   D   E   V   I   A   A   L   H   P   G   N   L   75

226   CGC GAC CTC CCG TTC ACG ACG AAG GAT GAC CTG CGG CGG GCG CAG   270
76     R   D   L   P   F   T   T   K   D   D   L   R   R   A   Q   90

271   TTC GAC ATG TTG TCG CGG CCG CTC TCC GAC GCG TGG ATC TTC TAC   315
91     F   D   M   L   S   R   P   L   S   D   A   W   I   F   Y   105

316   GAA ACC ACG GGC ACC ACT GGT GTT GCG ACA CCG TGT CCA CGT ACC   360
106    E   T   T   G   T   T   G   V   A   T   P   C   P   R   T   120

361   AAT GTG GAC TCT TTA CAC AAC AAC ACC ATT CTC ACG GCG TAC TAC   405
121    N   V   D   S   L   H   N   N   T   I   L   T   A   Y   Y   135

406   CGC GAC ATT TTC GGC GCG CGT GGC AGC GGC CAG GTC ATC GGT ATC   450
136    R   D   I   F   G   A   R   G   S   G   Q   V   I   G   I   150

451   TCG GGC CCG AGT GAG CTG CAC TCT ACC GGC GAC ACC TTC GGC GAC   495
151    S   G   P   S   E   L   H   S   T   G   D   T   F   G   D   165

496   GTT TGC CGC AGC CTG GGC CAC GCG GTC GCC AAG ATG TGG CCA CAC   540
166    V   C   R   S   L   G   H   A   V   A   K   M   W   P   H   180

541   TCG CCG ATG GTC GGC TTC GAC CGG GCC CTG GAA CTG ATG CGT CGG   585
181    S   P   M   V   G   F   D   R   A   L   E   L   M   R   R   195

586   CTA CCC GTG ACC GGC CTG TTC TGC ACG CCC GGC ATG GCC ATG CAC   630
196    L   P   V   T   G   L   F   C   T   P   G   M   A   M   H   210

631   CTG GCA CAG CGA GCC ATC GAC GCC GGG CTG CGG CCG CGG CTG GAC   675
211    L   A   Q   R   A   I   D   A   G   L   R   P   R   L   D   225

676   TTC TCC CTC GAT CTG CTG CTG CTC ACC GGC GAG CTG CTC TCG CCT   720
226    F   S   L   D   L   L   L   L   T   G   E   L   L   S   P   240

721   AGT CTG CTG ACC AAC ATC GGC CTG CTC TGG GAC GCC GAG GTG CAC   765
241    S   L   L   T   N   I   G   L   L   W   D   A   E   V   H   255

766   AAC GTG CTC TAC GCC TCA CAG GAG GCC TCC GTG CTA GCG GCG GCG   810
256    N   V   L   Y   A   S   Q   E   A   S   V   L   A   A   A   270

811   ACT TCG GAC GGT CGG CTG CGA ACC GCC CCG CTG ATC AAC CTC TAC   855
271    T   S   D   G   R   L   R   T   A   P   L   I   N   L   Y   285


856   GAG GTC ATC GAT CCC GAC AGC GGG GAA CCG GTC GAT TCC GAC GGT   900
286    E   V   I   D   P   D   S   G   E   P   V   D   S   D   G   300
```

```
901   GCC GGC CAG TGC TAC GGC GAG CTC GTG GTC ACC AAC CTC TAC TTG   945
301    A   G   Q   C   Y   G   E   L   V   V   T   N   L   Y   L    315

946   GGC GCC AAG CCT CTC ATC AGA TAC CGC ACC GGA GAC CTG GTC CGG   990
316    G   A   K   P   L   I   R   Y   R   T   G   D   L   V   R    330

991   ATG ACC GAG TCT GGC GTG TCT GCC GCC GTG CCC GCA CCC TCG GTC   1035
331    M   T   E   S   G   V   S   A   A   V   P   A   P   S   V    345

1036  GAG GTC CTC GGC CGG GCA CAC GAT CGG TTG ACG CTC AAC GGA CAC   1080
346    E   V   L   G   R   A   H   D   R   L   T   L   N   G   H    360

1081  CTC ATC AGC GGA TAT GAC TTC GAG AAT CTT TTG CTC GGC CAG TTC   1125
361    L   I   S   G   Y   D   F   E   N   L   L   L   G   Q   F    375

1126  CGT GGC TTC TGT GAC TAC CAG GTC ACA ATC AAG CGC GAC GAT GCT   1170
376    R   G   F   C   D   Y   Q   V   T   I   K   R   D   D   A    390

1171  GGC GTC GAT CGA GTC GCT CTG ACG CTG CGG TCC GAT CCC ACC ATC   1215
391    G   V   D   R   V   A   L   T   L   R   S   D   P   T   I    405

1216  GGC TTT CCG AGC GTC GAG GGT GCC GTG CAG CAG TGC CTC GAC GAG   1260
406    G   F   P   S   V   E   G   A   V   Q   Q   C   L   D   E    420

1261  CTC GAC ACG CCG CTG ACC GTA CGG TTC GGC ACT CCC GGC GCG ATC   1305
421    L   D   T   P   L   T   V   R   F   G   T   P   G   A   I    435

1306  ACC GGG ACT GGC GCG ATG GTG AGC TGG AAG GCT GCC CGC GTG ATC   1350
436    T   G   T   G   A   M   V   S   W   K   A   A   R   V   I    450

1351  GAC CAG CGG ACC GGC CAG TCG GAC GAG GTC GAA GCG GCC GAG AAC   1395
451    D   Q   R   T   G   Q   S   D   E   V   E   A   A   E   N    465

1396  ATC GCG GCG TCG AGG TCC TGA   1416
466    I   A   A   S   R   S   *
```

>salW

```
1     ATG GCG GGC TTG GAC CGG GCA GCG TCG CCC GGA CTG CGG CTC GGA   45
1      M   A   G   L   D   R   A   A   S   P   G   L   R   L   G    15

46    CGC TGG GCC TAC CAC CGC GGC GAG TTC GTG CCG GCG TCT GAT CCG   90
16     R   W   A   Y   H   R   G   E   F   V   P   A   S   D   P    30

91    CAG TTG CCG CTC AGC ACG CAG GCT TTG CAC TAT GGG ATC GGT GTA   135
31     Q   L   P   L   S   T   Q   A   L   H   Y   G   I   G   V    45

136   TTC GAG GGG ATT CGG GCG TAC CGC AGC GCT GAC GGC CTG TTT CTG   180
46     F   E   G   I   R   A   Y   R   S   A   D   G   L   F   L    60

181   TTC CGT GCG TAC GAC CAC TAC GAG CGC ATG CTG CGT GGC TGC CGA   225
61     F   R   A   Y   D   H   Y   E   R   M   L   R   G   C   R    75

226   ACG CTG AGG ATT CCG CTG CCG GGG AAG CCC GGC GAT CTG GTC GAC   270
76     T   L   R   I   P   L   P   G   K   P   G   D   L   V   D    90
```

```
271   ATC ACC GTC GAG CTG CTG CGG CGC AAT GCT CAC GAT GAA GAC GTG   315
91     I   T   V   E   L   L   R   R   N   A   H   D   E   D   V   105

316   TAC GTG CGG CCG GTG GCC TAC AAG CTC TCG TTG CTG CCC GGC ATG   360
106    Y   V   R   P   V   A   Y   K   L   S   L   L   P   G   M   120

361   CCA CCT GGC GTA TTC CTG AGC GGG CTG TCG GAC GCC ATG TCG ATC   405
121    P   P   G   V   F   L   S   G   L   S   D   A   M   S   I   135

406   ATC AGC TAC AGC CTG CCG GTT GAG CGG CTC GGC CAG GGC GTG CGG   450
136    I   S   Y   S   L   P   V   E   R   L   G   Q   G   V   R   150

451   TGT GGC ATC AGC TCC TGG CGG CGC CCG CCA CGC GAC ACA CTG CCC   495
151    C   G   I   S   S   W   R   R   P   P   R   D   T   L   P   165

496   GCC CAG GCG AAG ATC ACC GGA GGC TAT GTG ACG AGC GCT TTC GCT   540
166    A   Q   A   K   I   T   G   G   Y   V   T   S   A   F   A   180

541   ACC GAC GAG GCA CGA GCA GGC GGG CAG GAC GAC GCC ATC CTG CTT   585
181    T   D   E   A   R   A   G   G   Q   D   D   A   I   L   L   195

586   GAC CGC TCG GGC AAC GTG GCC GAG GCG ACA ACC GCC AAC GTC TTC   630
196    D   R   S   G   N   V   A   E   A   T   T   A   N   V   F   210

631   ACG GTC CGC GAC GGT TGC CTG GTC ACC CCC CCG ACA ACC GGA GAC   675
211    T   V   R   D   G   C   L   V   T   P   P   T   T   G   D   225

676   CTG CTG CCC GGC ATT ACC AGG GAC ACG CTC CTC ACG CTC ATC CGC   720
226    L   L   P   G   I   T   R   D   T   L   L   T   L   I   R   240

721   GAG GTG GGC CTG CCC GTC GCT GAG CGG TCG GTC AGC CCC GCC GAG   765
241    E   V   G   L   P   V   A   E   R   S   V   S   P   A   E   255

766   CTG TTC TCC GCA GAC GAG GTT TTC CTC TGT TCC ACC GGC AAA GGT   810
256    L   F   S   A   D   E   V   F   L   C   S   T   G   K   G   270

811   GTG GTG CCG GTC ATC GCC GTG GCA GGC CGT GAC GTC GGC ACC GGT   855
271    V   V   P   V   I   A   V   A   G   R   D   V   G   T   G   285

856   GCG ATC GGC CCA GTT ACC GCC AAG GTC CGT GCG CTC TAC GCG GCC   900
286    A   I   G   P   V   T   A   K   V   R   A   L   Y   A   A   300

901   GCG ACG ACC ATG CCG GGC GGT ATT CAC GCG GAC TGG CTT ACT CCC   945
301    A   T   T   M   P   G   G   I   H   A   D   W   L   T   P   315

946   GTC ATG GAG ACA CCG TGA   963
316    V   M   E   T   P   *
```

```
>salX

1     GTG ACC AAA GCA ACG ACG TCG TTA CCC TGT TTC GTC ACC CTC ACA   45
1      V   T   K   A   T   T   S   L   P   C   F   V   T   L   T   15

46    CCC GCC GCT GCG GGT GGG ACG CCG ACC GTG GGC ACG CTG GGT CCG   90
16     P   A   A   A   G   G   T   P   T   V   G   T   L   G   P   30
```

```
91    CCC GGC ACG AGC AGC GAG CTG GCC GCG ACA CTG CTC GCT GGC CGG   135
31     P   G   T   S   S   E   L   A   A   T   L   L   A   G   R    45

136   CTC GGA GCG GCC GAT CCC GGG ACG GTC GAG GGC GGG CCG ACC AGA   180
46     L   G   A   A   D   P   G   T   V   E   G   G   P   T   R    60

181   GTG CGG CTC TTC GAT ACC TAT GAG AGA GCC GGG AGC GCA CTG CGC   225
61     V   R   L   F   D   T   Y   E   R   A   G   S   A   L   R    75

226   GCG CGC TTG GTG ACG CAC GTG GTC GTG GCC AAC GCC TAC AAG GCC   270
76     A   R   L   V   T   H   V   V   V   A   N   A   Y   K   A    90

271   GTG CAC GAG TTC TAC ATG GAC CCG GAT TTC GAG CTG GCC GTG GTG   315
91     V   H   E   F   Y   M   D   P   D   F   E   L   A   V   V   105

316   TTC GTT CTG GAC ACC CCG ATG TAC GGC ATC GCC AAG CGG CGT GGT   360
106    F   V   L   D   T   P   M   Y   G   I   A   K   R   R   G   120

361   GCC GGT GCC TCG CCG CCT TCA CCC ACC ATC GCG AGC CAC CCG TCA   405
121    A   G   A   S   P   P   S   P   T   I   A   S   H   P   S   135

406   CCC GTA CCC ATC ATC GGG CAG TTG CTG AAC GGG TCC GGC GCC CAC   450
136    P   V   P   I   I   G   Q   L   L   N   G   S   G   A   H   150

451   GTG GTG ATG CAC GCC GAC TCG ACC AGC TCG GCC GCA CTC GCC GCC   495
151    V   V   M   H   A   D   S   T   S   S   A   A   L   A   A   165

496   GCG GAG GGT CGG GTC GAC CTG GCG CTC ACC ACG GCA CCC GCC GCG   540
166    A   E   G   R   V   D   L   A   L   T   T   A   P   A   A   180

541   CAG CTG TAC GAC CTG GAG TTC ATC ACG CGC ACA AGA TGC ATT CGT   585
181    Q   L   Y   D   L   E   F   I   T   R   T   R   C   I   R   195

586   ATG GTC TGG TCT GCT TTT GTC CTA TCG GAT CCA GCT CAG GCC ACT   630
196    M   V   W   S   A   F   V   L   S   D   P   A   Q   A   T   210

631   ACG GAG TCA GAC GGA GTC GCC TGA   654
211    T   E   S   D   G   V   A   *
```

>salY

```
1     ATG CTG CTG TTG AGC GGT AAG CAA GTG CTG GAG CTG TAT TCG CCG   45
1      M   L   L   L   S   G   K   Q   V   L   E   L   Y   S   P    15

46    GCC GAC GCG TAC CAG GTC ATC GCG GAC GTC ATG CGT CGT TAC AGC   90
16     A   D   A   Y   Q   V   I   A   D   V   M   R   R   Y   S    30

91    GGC GGC GGC GTC GTG CAG CCG GTG CGT ACG GTG CTC AGC TCC GAC   135
31     G   G   G   V   V   Q   P   V   R   T   V   L   S   S   D    45

136   CGT GAC GCG TCA CTC TTC GCG TCC ATG CCC TGT CAT GTT TCC GGC   180
46     R   D   A   S   L   F   A   S   M   P   C   H   V   S   G    60

181   GAG GTC GAC TCG GGT TTC GGA CTC AAG GCC GTA CTT CAC GTG CCC   225
61     E   V   D   S   G   F   G   L   K   A   V   L   H   V   P    75
```

```
226   GGC AAC GCC GAG CGC GGC CTC GTG TCG CAT GTG GGT GCC ATC CTC   270
76     G   N   A   E   R   G   L   V   S   H   V   G   A   I   L    90

271   GTC TTC CAC CCT GAC ACC GGG CAA CCG CTG GCC GTC ATC GAC GCC   315
91     V   F   H   P   D   T   G   Q   P   L   A   V   I   D   A   105

316   GCA GCG GTC ACG AGT CTG CGT ACG GCC GCA GCT TCG GCC GTA GCA   360
106    A   A   V   T   S   L   R   T   A   A   A   S   A   V   A   120

361   ACC GAC GCA TTG GCC CGC GCG GAT GCC GGG GAC CTA GCC CTG CTG   405
121    T   D   A   L   A   R   A   D   A   G   D   L   A   L   L   135

406   GGC GCC GGA GTG CAG GCC CGC AGT CAC CTG GTG GCC ATG AAC CGG   450
136    G   A   G   V   Q   A   R   S   H   L   V   A   M   N   R   150

451   ATC CGG ACC CTG CGC CGG GTT CGC GTG TGG AAC CGC ACT CCC GCC   495
151    I   R   T   L   R   R   V   R   V   W   N   R   T   P   A   165

496   CAC GCC GAG CAG TTT CGG GCC TGG GCC CGC TCC CTG GTA GAG GTC   540
166    H   A   E   Q   F   R   A   W   A   R   S   L   V   E   V   180

541   GAC GTC GAG CTG ATG CCG ACG CCG GCC GCT GCG CTG CGC GGC GCC   585
181    D   V   E   L   M   P   T   P   A   A   A   L   R   G   A   195

586   GAC GTG GTA TGC ACC ACG CTG GGC AGC AGC GAT CCG GTG GTC TCC   630
196    D   V   V   C   T   T   L   G   S   S   D   P   V   V   S   210

631   GCA GCC GAC ATC GCC CCA GGT TGC CAC GTT AAC GCG GTC GGG GCA   675
211    A   A   D   I   A   P   G   C   H   V   N   A   V   G   A   225

676   TCG GTG CCC AGC AAG CGG GAG CTC GAC TCG GAG GCG GTC GCG TCG   720
226    S   V   P   S   K   R   E   L   D   S   E   A   V   A   S   240

721   TGT TCT GTC TTC GTC GAC AGC CGC GAG GCG GCC ATG CGC GAG TCC   765
241    C   S   V   F   V   D   S   R   E   A   A   M   R   E   S   255

766   TCC GAG ATC ACC ATG CCG CTC GCG GAG CGA CTC ATC CCA GGT GAC   810
256    S   E   I   T   M   P   L   A   E   R   L   I   P   G   D   270

811   TGG TAC TTC CCC GAG ATC GGC GAG GTA CTC CTC GGC CGG CAC CCG   855
271    W   Y   F   P   E   I   G   E   V   L   L   G   R   H   P   285

856   GGC CGC ACA GAC CGC GAC GAA CGA ACC CTC TAC AAG TCG CTG GGC   900
286    G   R   T   D   R   D   E   R   T   L   Y   K   S   L   G   300

901   ATG GCG GCG CAG GAT GTC GCG TCG GGC TTT GCC ATC GTA CGA GCT   945
301    M   A   A   Q   D   V   A   S   G   F   A   I   V   R   A   315

946   GCC CAG CGA CTC GGC GTC GGC ACG GTG ACC GAC ATC TTC AGC TGA   990
316    A   Q   R   L   G   V   G   T   V   T   D   I   F   S   *


>salZ

1     ATG GCC GGA TTG CGG CGG CGG CGC TTC GCA CAG GTG GAC GTC TTC   45
1      M   A   G   L   R   R   R   R   F   A   Q   V   D   V   F   15
```

```
46    GCC GAG GAG GCA TAC CTC GGC AAC CCG GTC GCG GTG GTG CTC GAC   90
16     A   E   E   A   Y   L   G   N   P   V   A   V   V   L   D    30

91    GGC GAC GGA ATT ACC GAG CGG GAC ATG GCG CAG TTC GCG AGG TGG   135
31     G   D   G   I   T   E   R   D   M   A   Q   F   A   R   W    45

136   ACG AAT CTG TCC GAG ACG ACC TAC GTT CTC CCG TCG GCC GTG GCC   180
46     T   N   L   S   E   T   T   Y   V   L   P   S   A   V   A    60

181   GAC TAC CGC TTA CGG ATC TTC ACA CCC GGT GGT GAG ATC CCC TTT   225
61     D   Y   R   L   R   I   F   T   P   G   G   E   I   P   F    75

226   GCG GGA CAC CCG ACG CTG GGC TCG GCA CAC GCG TGG CTG GAG GCG   270
76     A   G   H   P   T   L   G   S   A   H   A   W   L   E   A    90

271   GGT GGT ATC CCA CAC TCC GAC GTG CTC GTT CAG GAG TGC GAC CTA   315
91     G   G   I   P   H   S   D   V   L   V   Q   E   C   D   L   105

316   GGA TTG GTG ACG GTC CGC CGC GCC GTC ACC GGC CTG AGC TTT GCC   360
106    G   L   V   T   V   R   R   A   V   T   G   L   S   F   A   120

361   GCT CCG CCC CTA CGG CGC GAT GGC CCC CTG GCC TCC GAT CAC CTC   405
121    A   P   P   L   R   R   D   G   P   L   A   S   D   H   L   135

406   GCC CGC ATC GTG CGG GCT CTG CGG ATT GCC CCC GAG CAG GTC GTC   450
136    A   R   I   V   R   A   L   R   I   A   P   E   Q   V   V   150

451   GCG CAC GGC TGG GTG GAC AAC GGC CCG GGC TGG GCA GCC GTG CGG   495
151    A   H   G   W   V   D   N   G   P   G   W   A   A   V   R   165

496   CTC GCG TCC GCG GAC GAG GTG CTG GCG GTG GAT CCC GAT GAA CAG   540
166    L   A   S   A   D   E   V   L   A   V   D   P   D   E   Q   180

541   CAG ATG TAT GAC CTC ATG CTC GGC GTT GTG GGC GCC TAT CCC GCA   585
181    Q   M   Y   D   L   M   L   G   V   V   G   A   Y   P   A   195

586   GGA TCG CCG CTG CAG TTC GAG GTG CGC GCG TTC GCC GCG CCC GCC   630
196    G   S   P   L   Q   F   E   V   R   A   F   A   A   P   A   210

631   GGT GTG CGT GAG GAT CCG GTC ACT GGG AGC CTC AAC GCC GGC ATT   675
211    G   V   R   E   D   P   V   T   G   S   L   N   A   G   I   225

676   GCC CAG TGG CTC ATC AGC ACG GGT GCC GCG CCA GGC AGC TAC CGG   720
226    A   Q   W   L   I   S   T   G   A   A   P   G   S   Y   R   240

721   GCG GGT CAG GGT GCG CGC GTC GGG CGT CGT GGT GTG CTG ACC GTC   765
241    A   G   Q   G   A   R   V   G   R   R   G   V   L   T   V   255

766   AGC GCG GAC GGC GAC GTC GTG TGG GTG GGG GGA CGG AGC ACG ACT   810
256    S   A   D   G   D   V   V   W   V   G   G   R   S   T   T   270

811   TAC GTT CGC GGC GTG GTG GAC CTG TGA   837
271    Y   V   R   G   V   V   D   L   *
```

Appendix B, Figure 2. Complete DNA sequence of the *sal* gene cluster from *S. tropica* CNB440.

cggtctcggtgtccttccccgtgaacggcgagcagccaacacccgagggcggaccgggcggtggtcctcg
gctaccaggcaggcgatagcaaccactccgagctaaaccgacattagccccatcaaggcatgaacacgaag
aacatgcctgaacgaggcatcgtcaactgctaccgggttcgcagccctacccccaggagaaacccgatgac
ccacgactatctgatctcactgatccgcaccgccgtcggcgccgccctggcgtggctggcgtccgaggcgg
ggatcgtgctcgactcgggcaccgcgctgaccgcgggcgtggtcgcgctggcgatggccggctactacggc
ctcgtgcgggccacactcccgggctcacaccgggctcacaacgaccccataaatcgcgtcaatttggatga
atgggatcgcccgaagctccggaaaccccagttcaaggggccctcttggtggtgggccgccgggggctcga
accccgaacctactgattaaaagtcagcagctctaccattgagctagcggcccgccgagccacaagactac
ccgactccctggcgagcaagcacgcggatctccggcaaccggtcccccacagtccatccttgccaagaatg
cgtctgctgctgcgcccgcatcccctcgaggcgcatgctgcgccgggtctacacccgcctgccaacaccct
gtcgggactacctgtacaggtccacgccccgtcggggctactccttcttgatcaaccagccgtcgtccacg
tgagcagacacgtcacccagccgttcagatcggcctgccggctaccgcttcgcaggccagcagagaacgcc
gaaccggcgtggaccaagccgcaggtagagcatccgggctgacacgaatggctgtggtcctactccagcgc
aagggacgccgccaaggtaccgtcgatgttcaccgggcagtgcccggcagccttctcgttcaggggctgag
tagatcggccggatactgcagggcctcctcgcgtaccggcagcagcgcagtgcgtatgacactcatgtcga
cgcaggacccgcgaggccgctcacaggtccaccacgccgcgaacgtaagtcgtgctccgtccccccaccca
cacgacgtcgccgtccgcgctgacggtcagcacaccacgacgcccgacgcgcgcaccctgacccgcccggt
agctgcctggcgcggcacccgtgctgatgagccactgggcaatgccggcgttgaggctcccagtgaccgga
tcctcacgcacaccggcgggcgcggcgaacgcgcgcacctcgaactgcagcggcgatcctgcgggataggc
gcccacaacgccgagcatgaggtcatacatctgctgttcatcgggatccaccgccagcacctcgtccgcgg
acgcgagccgcacggctgcccagcccgggccgttgtccacccagccgtgcgcgacgacctgctcgggggca
atccgcagagcccgcacgatgcgggcgaggtgatcggaggccaggggccatcgcgccgtaggggcggagc
ggcaaagctcaggccggtgacggcgcggcggaccgtcaccaatcctaggtcgcactcctgaacgagcacgt
cggagtgtgggataccacccgcctccagccacgcgtgtgccgagcccagcgtcgggtgtcccgcaaagggg
atctcaccaccgggtgtgaagatccgtaagcggtagtcggccacggccgacgggagaacgtaggtcgtctc
ggacagattcgtccacctcgcgaactgcgccatgtcccgctcggtaattccgtcgccgtcgagcaccaccg
cgaccgggttgccgaggtatgcctcctcggcgaagacgtccacctgtgcgaagcgccgccgccgcaatccg
gccatagtggcgttccctcccggtacggttctgggtgcctagactgcttggccctcgacgaggatcaggtc
catgtgtgtaccggcgcggagcggcacgatctcgcggtactcgggcgagcggtaccaccggtgggcttgct
cggccgaggggaagcccatcaccgagaggtggtccggccgccagattccctccagcacctcgaacgtgccg
ttgcggatgacaaacccgccaccgtgcagggcgagcgtcgaacccacctttgcccggtactcggccatggc
ctccgggtcaccgagcgtacccacgttgccgatcagaagcgccgccatggtgatgactccgttctcgttcg
tggtgatgacaggacgagcgggctcagctgaagatgtcggtcaccgtgccgacgccgagtcgctgggcagc
tcgtacgatggcaaagcccgacgcgacatcctgcgccgccatgcccagcgacttgtagagggttcgttcgt
cgcggtctgtgcggcccgggtgccggccgaggagtacctcgccgatctcggggaagtaccagtcacctggg
atgagtcgctccgcgagcggcatggtgatctcggaggactcgcgcatggccgcctcgcggctgtcgacgaa
gacagaacacgacgcgaccgcctccgagtcgagctcccgcttgctgggcaccgatgccccgaccgcgttaa
cgtggcaacctggggcgatgtcggctgcggagaccaccggatcgctgctgcccagcgtggtgcataccacg
tcggcgccgcgcagcgcagcggccggcgtcggcatcagctcgacgtcgacctctaccagggagcgggccca
ggcccgaaactgctcggcgtgggcgggagtgcggttccacacgcgaacccggcgcagggtccggatccggt
tcatggccaccaggtgactgcgggcctgcactccggcgcccagcagggctaggtccccggcatccgcgcgg
gccaatgcgtcggttgctacggccgaagctgcggccgtacgcagactcgtgaccgctgcggcgtcgatgac
ggccagcggttgcccggtgtcagggtggaagacgaggatggcacccacatgcgacacgaggccgcgctcgg
cgttgccgggcacgtgaagtacggccttgagtccgaaacccgagtcgacctcgccggaaacatgacagggc
atggacgcgaagagtgacgcgtcacggtcggagctgagcaccgtacgcaccggctgcacgacgccgccgcc
gctgtaacgacgcatgacgtccgcgatgacctggtacgcgtcggccggcgaatacagctccagcacttgct
taccgctcaacagcagcattggctcgctcccttcaggcgactccgtctgactccgtagtggcctgagctgg
atccgataggacaaaagcagaccagaccatacgaatgcatcttgtgcgcgtgatgaactccaggtcgtaca
gctgcgcggcgggtgccgtggtgagcgccaggtcgacccgaccctccgcggcggcgagtgcggccgagctg
gtcgagtcggcgtgcatcaccacgtgggcgccggacccgttcagcaactgcccgatgatgggtacgggtga

cgggtggctcgcgatggtgggtgaaggcggcgaggcaccggcaccacgccgcttggcgatgccgtacatcg
gggtgtccagaacgaacaccacggccagctcgaaatccgggtccatgtagaactcgtgcacggccttgtag
gcgttggccacgaccacgtgcgtcaccaagcgcgcgcgcagtgcgctcccggctctctcataggtatcgaa
gagccgcactctggtcggcccgccctcgaccgtcccgggatcggccgctccgagccggccagcgagcagtg
tcgcggccagctcgctgctcgtgccgggcggacccagcgtgcccacggtcggcgtcccacccgcagcggcg
ggtgtgagggtgacgaaacagggtaacgacgtcgttgctttggtcacggtgtctccatgacgggagtaagc
cagtccgcgtgaataccgcccggcatggtcgtcgcggccgcgtagagcgcacggaccttggcggtaactgg
gccgatcgcaccggtgccgacgtcacggcctgccacggcgatgaccggcaccacaccttgccggtggaac
agaggaaaacctcgtctgcggagaacagctcggcggggctgaccgaccgctcagcgacgggcaggcccacc
tcgcggatgagcgtgaggagcgtgtccctggtaatgccgggcagcaggtctccggttgtcgggggggtgac
caggcaaccgtcgcggaccgtgaagacgttggcggttgtcgcctcggccacgttgcccgagcggtcaagca
ggatggcgtcgtcctgcccgcctgctcgtgcctcgtcggtagcgaaagcgctcgtcacatagcctccggtg
atcttcgcctgggcgggcagtgtgtcgcgtggcgggcgccgccaggagctgatgccacaccgcacgccctg
gccgagccgctcaaccggcaggctgtagctgatgatcgacatggcgtccgacagcccgctcaggaatacgc
caggtggcatgccgggcagcaacgagagcttgtaggccaccggccgcacgtacacgtcttcatcgtgagca
ttgcgccgcagcagctcgacggtgatgtcgaccagatcgccgggcttccccggcagcggaatcctcagcgt
tcggcagccacgcagcatgcgctcgtagtggtcgtacgcacggaacagaaacaggccgtcagcgctgcggt
acgcccgaatcccctcgaatacaccgatcccatagtgcaaagcctgcgtgctgagcggcaactgcggatca
gacgccggcacgaactcgccgcggtggtaggcccagcgtccgagccgcagtccgggcgacgctgcccggtc
caagcccgccatcaggacctcgacgccgcgatgttctcggccgcttcgacctcgtccgactggccggtccg
ctggtcgatcacgcgggcagccttccagctcaccatcgcgccagtcccggtgatcgcgccgggagtgccga
accgtacggtcagcggcgtgtcgagctcgtcgaggcactgctgcacggcaccctcgacgctcggaaagccg
atggtgggatcggaccgcagcgtcagagcgactcgatcgacgccagcatcgtcgcgcttgattgtgacctg
gtagtcacagaagccacggaactggccgagcaaaagattctcgaagtcatatccgctgatgaggtgtccgt
tgagcgtcaaccgatcgtgtgcccggccgaggacctcgaccgagggtgcgggcacggcggcagacacgcca
gactcggtcatccggaccaggtctccggtgcggtatctgatgagaggcttggcgcccaagtagaggttggt
gaccacgagctcgccgtagcactggccggcaccgtcggaatcgaccggttccccgctgtcgggatcgatga
cctcgtagaggttgatcagcggggcggttcgcagccgaccgtccgaagtcgccgccgctagcacggaggcc
tcctgtgaggcgtagagcacgttgtgcacctcggcgtcccagagcaggccgatgttggtcagcagactagg
cgagagcagctcgccggtgagcagcagcagatcgagggagaagtccagccgcggccgcagcccggcgtcga
tggctcgctgtgccaggtgcatggccatgccgggcgtgcagaacaggccggtcacgggtagccgacgcatc
agttccagggcccggtcgaagccgaccatcggcgagtgtggccacatcttggcgaccgcgtggcccaggct
gcggcaaacgtcgccgaaggtgtcgccggtagagtgcagctcactcgggcccgagataccgatgacctggc
cgctgccacgcgcgccgaaaatgtcgcggtagtacgccgtgagaatggtgttgttgtgtaaagagtccaca
ttggtacgtggacacggtgtcgcaacaccagtggtgcccgtggtttcgtagaagatccacgcgtcggagag
cggccgcgacaacatgtcgaactgcgcccgccgcaggtcatccttcgtcgtgaacgggaggtcgcgcaggt
taccggggtgcagggccgcgatcacctcgtcggtcaccccgaccaggcgctccccgtaaaacggcgagttg
tcgcgggcaaaagcaacgacctcacgcagcttgcgggtctggtgcgctgccagcatcgccgcgtcccactt
gccggcaatgaacgcctcaacctgattacgcagctcgactgtgagagcggcgtcttccggatagtactgtg
tgaacattccgagcccccacatgttagggcgattcgacggtatggtcgcgcaggtcaatcagggtcagcca
ggtcggccacgacccagtcggatatcagtctccacttctcgatggcggctcggtggaatgcattttcctgg
tagcgctgcagtgccaccagatcagccagcacgccaatggcggcgaagtcgtaggcgatgtcccggtcgac
cgtgttccagccggtgcgccagtcgacgagttccggtacgtggcggccgacctcacgggcgagctccaccg
cctcggtgacctgccgggcgtcgcggtctacgccatccttgagcttgaagagtacgatatgcgtgatcatt
tagcccttcccgaaatcagccatacctgaagaacgggctgaacatgcgctactgaccctagccgggaatct
ccagggccgagtctgtccgatgccccgcagcccgttcaatggaagtttcttgatatcctctatcgattgaa
ctgatgataaaggatgctcatattcattaaactccggtgatatgattggctcggctgtttccagcctgtca
tccatggccgacaaggaatggcaacgaatcaactcctgccccatacgagaggcgattttggtgtgtatgga
acgcgcatgcgacccggcgtggagcggtcgatccgccattcgccaccagcgacgaacgatggcgacagaga
actctgcctagaccggtggcgcgagctaccccggcggcaggtgccaccctggcccgatcccgccgaggtgg
cagcagtgtgcgcgacactcggcaaaatgccgccgatcgtcacaccctacgaggtcgatgaactgcgtcac
cgcctggccgaggtatgcgaaggacgcgccttcctgttacagggcggcgactgcgccgagacgttcaccgg
aaacactgagagccacctgctgggcaccacgcggacgctgctgcagatggcgatggcgataacgtacggcg
gttccgttccggtggtgaaggtggcgcgcctcgccggccagtacgggaagccccgatcctcggcgaccgat

tccctggggttgcccgcgtaccgtggcgatattatcaatgcgcggcacccggccgagtcggcgcgggccgc
cgatccgcagcgcatgatcgatgcgtacgcgaattccgcggtcgcgatgaacctcatccgcgcctatccgc
cggacgatctgaccgacctcgaagagctgtacgacgatacctacgacctcatccgcgcttccccggccgga
gcccggtaccaggtcatctccggcgagatcgaccgggcgcgcggcttcgtccgcgcgtgggggccgagcga
gcgccacgcgttgcgggaatcaaaggtgtactgctcgcacgaggcgctggtgctcgagtacgaccgggcgc
tgacccgaatcaacgacggccgggcatacgcgttgtcgggtcactttctgtgggtcggcgaacgcacccgc
cagcttgaccacgcgcacgtcgactttgtcgcgcgtatcgccaacccaatcggcgtgaagctcggcccggc
cgccagcccgcatgctgcgatcgagctgtgcgagcggctcaaccccgagaacctgcccgggcggctcaccc
tgatcagccgcatgggcaaccgccaggtacgtgacgtgtttccggcgatcgtcgacaaggtcacggccgcg
ggcgccaaggtcgtttggcagtgcgatccgatgcacggcaacaccgagcagtcctcacacggcttcaagac
ccggcgcctcgaccgggtcgttgacgagttgctgggctacttcgacgtgcaccgcagtctgggcacccatc
cgggaggcgtccacgtcgagctcaccggtgagaacgtcaccgagtgcctcgacgggatacgtggcgtcgag
gaccaacacctaccggatcgctacgagactgcctgcgacccgcggctgaacatgcggcagagcttagaact
cgcgttgctggtcgccgagattctacgcggctgatcccttcgacacatcccaaggcacatcggcatccgca
gggggaaagaatgagcgctccgaatctgaacgcggacatcggcgtcatcggcggcacgggtctgtacgagt
tgctgggcgagaccgaagaggtgaaaatctccacgccgtatggcgcgccgtcggatctgatcacggtggca
gagttgaacagccgccaggtcgccttcctaccgcggcacgggcgggaccaccgtttcccgccgcacagaat
cccgtaccgggccaacatctgggcgctgcgttcgctgggtgtgcggcagatccttgcaccgtgcgcagtgg
gcggccttcggccggacttcattgccggcacgttcgtcgtgccggatcagctgatcgaccgcacgagtggg
cgggagcagaccttctgcgacaccggcgcagtgcatacgaccttcgcagagccgtattgcccggacgggcg
ccgcgccgcgttggccgcggcggaacgggtaaaccgcgtcgacggcggcaccatggtcgtaatcgacgggc
cacggttctcgacccgtgccgagtcacggtggcacgcggccgccggcggcaccatcatcaacatgaccggc
cagcccgaggccgcgctcgcccgcgagctcgcactctgctacaccagcatcgctttggtaaccgacctcga
cgccggcgtggaaggcagccagggcgtgacccacgacgaggtgctgcgggcctttgcgaagaacgtgaccg
tcctgcgacgcgtactcttcgatgccgtcgctgcgctaccggtggaacgcacctgtccctgctcgcgcgcc
ttcgccgttcccgtggacctcgacgccggttgagcggctgctccccagttgacgaggtgccgttatggctt
cacggcaatggcgatattcccgctcgttctggagatggttgcgctgtcagcgttcgccgtgggaagcggcg
aacggcggctgcggcagccaccagtaccatccccgcctccgggcagcccgaggccgttgggcgccgaccgc
tacccactatcaacggaggaagtgaatgaccactacgaccaccctggccggactgccctcgctgaagggat
gtgtgctcggcccaagctcctggagtgagattacgcaggagcgcgtcggcaccttcgccgacgccaccgac
gaccaccagtggatccacgtcgatgccgatcgggcggtgagagagagccccttcggcggtccgatcggcca
cggcttcctgaccctgtcgttgatgatcccaatgtggagcgagattctcaccgtctcggatgccggcatgg
tggtgaactacgggctcaacagggtccgcttccctgctgctgttccggtcggtgccaagctacgcctgtcc
gccaccctcgtcgatatcgaaccggttaggggcggactgcaggtgacggccgccgcctcggttgaacgaca
gggcgacgacaagcccgtctgtgtcgccgaacccgtctttcgcttcctgggctgatcgctcgccaccagca
gaccgatcgcgggcagccacaccgcggccctcgggctcgaggctgggtacgggccggtggtcattgcctct
ttccgaagccgaccagctcggcgtagcagaatctgatgcgcgcggcgacgagtccgaccccgtctgatggg
tgaaagtccggtcagccgatcgagcttctcgcccctacgctgggacgacgccgacttccccagcaacaccg
cgggcgaagtgatgaacggcctcattcgcgaccaggtgcaccacgcatggtggcggttaggttgttgaggc
agatttgattctgtgatgatgatcgtgtagccgacgtcgcgggtgtgctgagtttctcatgttttcgttcg
tggaccggtccgggtccggcctctgcgttctctgagagtaacgttgatcatcgcttgtaggtctacggagt
cgctgtggtggcccgcaggcccggccgggtgtggaagcgatcaggttggtcgcgggccgtgtgcaaagcgg
cggtaggagatctcgtcgggatcgatgctgaggtcgaccgcggcggtcacggcgtggtcgatgagtgcgtc
gtagacggtcagcagttccccgattccttgttctgccgggtccggggataccacaggatcacctgcacacc
atcccgcaagcagccaccacgggcggcgtccagatcacccaccgcacccagcgatcgctacccgaccgaca
ccaacctcgcccgatatcactgagtagtcagacaattacagtaggttctttgcgtagtggacttgacacat
atcgcggaagtcgatttagaatagacgtatattagatcggttcactttcagtacacattcgtgtcgcactg
caagggagttgacccacccccccctcaacgaacggaccatccgtgtggcatcagctcacggtaatcacga
caatagtgattaccgtaatgacgcctgcggttaccacggagtcggccgaacccaaagaaaagaccgacccc
tcgttgttctagacgaagcgaaaccgaggatagctgcgaaccccctgacctgggaacgcgactacagcacc
gtcatctaagaagggaggcgttgcagtggccgatcacgcggacttcagcctggctcaactccggtacttcg
taacttccgccgagatgggaaacatatccggcgcggcacagcaattacatgcttcgcagtcggccgtctcg
atggcgatccagcgcctggagaaacaactggacaccaagctgttcttccggcaccgcaccaaaggagtcag
cctgactcccagcggccgcgtcctgctcgaccccgcgaagtcgttgctggcccaggcgcaggagatgctcg

cccacagccgcgagtggcagggcgagacctgcggaatgctcaatgtggccgttctccgctcagtcgcgccc
gatctggtgccggctgtgatgagccgagtaaagaagatgcgccctgggctggcggtcaccgtccacgaggg
aaccatggacgatgttctcgaactgctgcgcacgggcacgtgtgaactcgccgtagtcccggagctgccag
gtcaggctctgtcgttcgcccggctggcgaaggtgctcatggcagcggtcgttgcatggtccgatccgctg
gcgcccgtggggcgcgcaacgcttcgacagttggccacccggcgtctcctcgtggccgacctggacgttga
tcgccaagtcggcctggacacccggacacggctgttcgccgacgccggcacccccacgccagaggtggtgc
tgacatcgagcatcgccacgatgcttggcctcgtcggggtcggcgagggcttcgccctgctataccggacc
gcagccagccagtgctaccgacaggaactcgcctggatcgacatcgtctccgaacgcccgtgcgcatcctg
gcttggtgtggcgaccatacccggcctggctctcagcagccgtgcggagctattcgtttccctgctcgatg
acgctgtccggagcgtctactgcacgacggcggacgacggggtgagttcccctggcagcactgatatcttc
gatcatggagcctcttttgctacggaccccgccgcacccgcctggtcgacctagcagtgctgtgtcaggtc
gacgtgtcgttgtgtcgtctgagtggttcggggtagttcgatggtggggtgacccctgacggggctttccg
ttgcgcggcaacggcttgaggcgttcgccgcggatgtgttcaccccactgatgcggccggatcagcgggtg
aagagtgggacgtatctgcggggtctgctgctggatgcgcggcgtacgtcgatgcagccgacgacggcatg
atgatcggcgactggaacagcaacccacgcccgttcgtctggaccatgaccgctgaagaaatcctcgactc
actcgcccgactccgtaggcgaattccccggcgcagaacacgagggcctgatcggcgcggtaggagccttc
gcgccttcgcttgcccacgcgatccgaacgcttcagtcggacgcccggcctcccacctacacggatccacc
gctggcactcaccgcatccccatcaactgtcgaactcggggttgaggatgcgatctccgcctcgcttgaga
cggtgtcggacgaggcggccggagccggggtcgggtagccggagggcatttcaccctccggctaccacagg
tcccggcgtgacgttgggtaggccgccgacgaggcacctgtggtcaggtctcttccgacggccccaagcca
gatcccggtcgcctcagagtccattgaccagcacagaagcggttcaagcgcggatgtctaagatgtctacg
aactgcggaacccggtccctcggtcgccgcccaccctgtatcgaaggaccagcccagtaacgagagagtcg
ggctaccctacccgacctgctggacggggtggggcagcgactgcagcaactgctcctggcgccggtggtac
tcggtcacattggcgagcttgacctcgtcgaagccacgcaccatgtctggcagagacgccagctctaccag
cgccgggtcgtcggggttggccgccgacaacgcggcgaggatcgtcctgcgatagttcgccacaagcctcc
gttcgatccggcgcacgcgtgtgtagccgaagacatcgagcctgctccctcgcagtctgcgcagtgcgtac
agcagccggagcacgactcggaaccactgcccgaggcgaatcttgcggcgcacgagcttcctcagaaccgg
tggctgcaataggtagtggtatcgcgccccagctccgaactgcgcggctatgtcgtgggtcaggcgttcat
cgagcgagagccgggccacctcgtactcatccttgtaggccataagtttgtatagattgcccgccacggct
gcggtgagcggccccgcccgtccctccacccggcgtacgtgttccacgaactccacaaactcgcgggcata
ggccagatcctggtaggcgacgagatctgccacgagcgcatccagccggcgcgcgagtggggtgtcctcgc
cggcatcgattgtcgcccgcaacgatgctgcatcgccctgcaacggtcgttgctcctgggctacgggagct
gccggcgacaccgtctgagggtcgtcaaccaccaggcggccgaggcagaaggcgctgaggttgcgctccac
gccgacgccgttgagccggatcgcctgttcgatgctggccgcttctagcgggatggcgccctcctggtagg
cggcgcctagttgcagaatattggcgaactggtcgtcacccagcaggcgcaaggcgagggatcgggcgtcg
aagaaggtggcgcgcgcactcgatcgatcgagagccgagcacagttcgtctacgtcgggccgtgggcccga
cacctcggtaaccatctgaccggtgggcaccggcgaggtcgatgccacgacgacggtccgggcggtgtcgg
ctgcagcgaggttggcgggctgagccgcgacgagcaggtcggtggctaggtagagatcgcattcaccgtcc
gctagccgtgcggcagactcgacggcacgctcggtgacgcggacatcggagacgaccgcaccgcccttctg
ggccagcccggtctggtcgagcgtccgcacgtgcctgccgtcgagcactgcggctgtggcgatgatctgag
ccagggttaccacgccgctgccgccgatgccggtgatccgcactgagaagtcgccgacaacgcgctgccgc
ggcatcgatggtttagccgtgggcaccgaggccagctcgcggtgcggggatcggccggggacaatggtcaa
gaatgatgggcaccggccctgtaggcatgcgtagtcggcgttacacgaggactggtgaatacgtgtcttgc
ggccaaagtcgctttctaccggctggaccgacaggcagttcgactgcgccccgcagtcgccgcatccctcg
cacaaccgttcgttgatcagcactttgtaggcgggctgcggccgcaggccgcggcgccgcttacgccgttg
taccgccgcgcactcctggtcgtggatgagcacggtgacgccgggagtcacggcgagttccgcctggctgc
gcggcagatcgtcgcggtggcgtacctcgacgccggccggcatgggatgccggccgagccgctttggatcc
tccgaggtgacgatcaccttggtcgcgccctcggccagaagcagagcggccagccgatgcaccggcaaggc
acccacggcgtcctgaccaccggtcatggccacggcagagttgtacagcaccttgtacgtaatgttgacgc
cggccgccactgcggctcgcagcgcgaggctggccgagtggtgaaaggtgccgtcgcctacgttctgcacg
aagtgatcgatgtcgacgaagggtgccatcccgatccactgggcaccctcgccgcccatctgcgtgacgcc
gatgatgtcgccgacccgcttcggatccatgaagaaggtcatggtggaacatccggtgccgccggcggtca
gggtaccgggaggcgccttcgtcgacgagttgtgcgggcagccggagcagaagtatggggttcgcgccacc
accggcaggtcgatgcggtcacggcgccgacgtccctgccacgcttccacgggttcgcagcggccggtcgc

cgacagtcgccgggcgaggccagccgcgatcacgtcgggattcagctcacccgtggggcggaacagcgacc
gccccgccggatccgtcttcccgcagaccaccggcgcgtcgccgctgccatagagcagatccttgacggcg
gattcgatgaaaccccgtttctcctccaccaccactatctcgtccagccccgcggcgaactcccgcaccac
ggctgactcaaggggtagaccgccgccagcttgagtagccgtacgcgatagcggcgcagagcttcctcgt
ccatgccaagcgtccgcagcgcttggcggacgtccagatatgacttgccggcggccacgatcccgatccgg
tcatccgcgtgcccgtacagctcgttcaggccggcggcgcgcagatactccaccgccaaggggaggcggac
ccggtgaaggctctcctccaaggcggcgagttccgcgccgagcagccgggacgtgggccggtggcgatatc
caccagccactgtcgggaagcgccaggacggggtgaccggcaccgtgctcgccgcgtcggcgacgttggtc
acgatcttgagcgctgaccataggccggacgcccgggacatctcgatggcgtgcaggccgtagcggagcag
gtcggcggcgtctgctggcacgaggaccggcattgagaggtccgccagggcgaactccgacgcgctcggca
cagtcgacgacttcgccgctggatcgtccccgacgagcgcgatggcaccgccccgcgggtcggtgcccgcg
aagttcgcgtggcgcaacgcgtcgcccgcgcggtcgagacccggcgccttgccgtaccagacaccaacgac
accgtcggggcgctgtctgccgacgcccgcagcgagctggctgccgagcacggccgtcgccgcaagctcct
cgttcagtgcgggccgatgcaccacgccgtgctcggccagcagtcggccccggcgggcgagctccaagtcg
tagcccgccaggggcgacccttcgtagccggagacgaagaccgccgcgtcgtttccggcacgccggtcgtg
gcggacacggtcaagcaggagacgcaccagcgcctgcacgccggtgaggtaggcggtgccgtcctggcaca
gatagcgctggtccaacgtgaattcgcttggctgcaagccgggggacatgcgactccttctcggcaattcg
tctgaaccgatcatggtgaggtagcccacgcgaattaccgtccggagcggaatgcctggacgagaaatcca
cagcccgcccactagtcgacttattcgggccaatcttcccgctgcggcagcaatcaatcaattgcataagt
tcgatgatgtggattaatcatgccgatgacgcacgttgggcaggccggccaacaggatgataaagtgctgc
tacttgccaagtggccagagtccaccagatcgctcaagccttaagcgaaaggcgaacataggtgcacccaa
ctaccttcggcggcgtgtcaaacattcttgaagccgcttcgaccagagccactccgattgcggatcgagtc
gactaatgaccagcatagaatctagtccgaaacggattctactccttggcgacagcactatttttcgagag
gcgatggccgagatctgcgatcgcgagcccgacctgcgggtggtcgggcaggcggcggcggggcccatggc
ggcggccctggttcgacggacggcgccagacgtggtactgctgagcgtcggcccgtctgaccaagggatca
tgcggctggtcggccagatcctggccgcgacctgccagccccggctggtaatcctggcgatgcacgagaac
gtgcgcctggcgcgccggtacctcgccatgggtacgtgggcgtacgtctcgcggaactcgacgcgtgagga
actgctgagcgtcgtacggatcgtgagccgcaaccgcgacagcgccgtgctttccgtgcagaagagcatgc
tgcaacatctgaccggggtcggcacgcggtcactgtcgggccgggagttcgaggtgatcgagctggtcgcg
gcgggcctgagcaacgcccagatcgccggccgcctctccatctccgagggcacggtcaagcgccacctcac
caacacgtacgacaagctcgatgtgcgctcacgagtggccgccgtcaacaggctcgcggcgatgggtttga
tcggcccgaacgcgatgctggagacgagcagcctggtggcctgaccgtcgtcgccgcggcctgcggacgag
gacctcacgactccgtggccggcatcagggcgatcgcccggccaggagccccagtgctgccctccaccttg
atgggaaggaaacaaaaccacgcgccgcgaggggcagcgcgtcgagcgcggcgagacactcgacgtagac
cattccctgtcccaggccccgcacatgcacgcccaccggatcgtgcgccgggccgatcgacggcgcgtcaa
tgccaaggcattgcacccccgagtcagcagtagctccactgtgggctctgtcggcgccggccatccgggg
ctccgcccggcgaggacgtcacgaagatatccgcccccggcggcaccgcgtcggtagcgccggtcccagcc
ggtgcgcagcagcacgatctcaccggggttcagcatgccgttggcggcctcccatgccacgatcgccgacg
ggtggatcagcgggctcatcccgccttcgcctccaccgggagatgccacgtcaattaccgcggcggctccc
atgaggctcgccagcccgacctggtcagccgagatggcacccgccggacccgctcccggcaggccggagtc
cggaggtgggatgaagtgcagaggagcatcaaagtgggttccggtgtgctcgtcgatggccatccaacggg
tcgcgtacgggccgccgcggctgtagacggaggccatcgcgttgcgatggtcggcaaaccagttccaggtc
ttgtgctggaagggttgatgcccctcccagtagcatggcaagtcctcggcgatggtcaccgaaaggtcaac
cacctggtagccggcggggaccgccgattggccgtgtacggcgctgccgtccaacatacgatcaccgcacc
cgctggatgccggcggcaagaccgccgacgatctcgatcaggaccgcgccgagcaccatcggggctcgatg
caccgtgggcatcggcccgtcgaacgggttgccggtggtccgcaggttgttcaccaggtttgcgggagcga
aaccctgcacgccgccgtccggtcccgtcatgcccacctctgtgacgaacgccgtgccgttcggcagcagc
cgtggcaggtcggtggcttcgtgggtgtgcgtgccgacgaccgcggcggcggtgccgtcgacggtgtgggc
gaagatctgcttttccagcacgtgatcgccgtggtagtcgacgatcgtgatgccgcgccggtcggccgcct
cccatgccgagtaggcgggacccacccgtcccgcagtcgccttggtcgacgccatggcgcagccatcagcc
aggttcagtacggtgaccatctcctcgccgaccgggacatggacaaagccccgaccgggtacctcggggtc
cacgttggcgggccgaaccacttgcggcagattgagcatcccgaccgactctgggctgtcccaggagtggt
ttccgccggtgatgacgtcaacccctccggacagcagttgctccacctgcacacgacccatccccaagccg
tcggccgccgagttctcggcattcgccacaacgaggtccaccgtgtggcgggcacgcagttcgggcagccg

gtctaccaggtaccgtgtcccctcgtcccccacgatgtcgccgatgaacagtacaaacacggtcgtgtacc
tccctgggggggcggctttctcgccgacgcccggcgatgcagcccagagtctcggtcttcctcgccggcggc
atctgctgaaggtcacagttcgcctacaactttggtctgtcgcggcgcccgaaacgatctcatagcgtggg
ctagcgatcggccggccagggcagccgtcggacgaggtcgttgcagagtggtgaagggctgccgaacaggt
caggaggtggttgatgacgaacggtgggcgcctatcgggaaaagtaagtctgatcaccggcgccgcgcatg
gcataggccatgccacggctgtgtggttcgcccgcgagggagcccggctggtggtcagcgacgtggacggt
gccgcactggagaagtgccacgccgaactggccgagtccggagccgaggtgacgacggtgatcgccgacgt
ctcagacgccacccaggcacaccgcatggttcagacggccgtcgatgtctacaaccggctggacgtcgtgg
tcgccaacgctggcgtgatccctctgcacgagatcaccgaggcgaccgagcaggactgggatgaggtcatg
gccgtcgatggcaagggcatgttcctcacgtgcaagtacggcattgacgccatgctggccacgggcggctc
gatcgtgtgcctttcatctatctccggcgtcgctggccagcgggggcagagcacgtacggcccggccaagt
tcgtggcttccgggctcaccaagcacttggccgtcgaatgcgcacgacactgcatccgggtcaacgcggtc
gctcccgggactatccgcaccaaccgggtccgccggctcgaagaagaaccgggcggtccggagtacctgga
ggacatcgtccggctccatccagccggacgcctgggcgagccgtcggaggtggcagccgcgatcggattcc
tggcctccgacgaggcctcgttcatcacgggtgtagtacttcccgtcgacggaggttacctcgcgcagtga
cacgagcgccgtcgaccatcgcatagatggtccgacgatgttcctgtctcgcttccgacggagaagcctgc
ggttcgtggtctgcacaacaaccaatcctgaggtcggtttccgaccgataaaacggaggtcactcaccatg
cagcacaatctcattgccttcctctccgacgtcgggtccgccgacgaagcccacgcactctgcaagggagt
catgtacggcgtcgcgccggcagcgacgatcgtggacataacccatgacgtcgcgccctttgatgtccggg
aaggcgcactgtttctcgctgatgtcccgcactccttcccggctcatacggtcatctgcgcatacgtctat
cccgagaccggcacggcgacccacaccattgcagtgcgcaatgagaagggtcaactgctcgtcggacccaa
caatgggctgctgagcttcgcactcgacgcctcccctgcggtcgagtgtcacgaggtcctctcccccgatg
tcatgaaccagccggtgacgccaacctggtacggcaaggacatcgtcgctgcctgcgccgcccacctcgct
gccggcaccgaccttgccgccgtcggcccccggatcgacccgaagcagatcgttcggctgccctatgcctc
cgccagcgaggtggagggcgggatccgcggggaggtggtacgcatcgaccgggctttcggcaacgtatgga
ccaacatacccacccacttgattggctcgatgctgcaggacggggagcgactggaagtcaagatcgaagcc
ctcagcgacacggtgctcgagctgcccttctgcaaaaccttcggcgaggtcgacgaggggcagccactgct
gtacctcaacagtcggggccggcttgccctcgggctcaaccagtccaacttcatcgagaagtggcctgtcg
tacccggtgacagcatcaccgtctccccaagggtgcccgatagtaacctgggcccggtgctcgggtagctg
acccaggggcggcatgctgctgggtagctcctgtccctaggtcgagtagccgggaggaatttcacctccc
ggctctcgccgatctgggcgtctctcccgtcacccggctctgccgcgccaggtaggcgtagacgtgggcaa
gaagtggacggccacgccgagcggcgtaccggcccgtagtaccggatccagccgcgaacagtggccgatcc
gggccttggtcgggtgcagttgcagcccgacctgggccgttgcagcccgacctgggccaccctgcccgcga
tccgcgagacaagcccgaccagacatcaatcgggcaccacgtcgtaggacgacctacgccacaatctccgc
cggattcagcgccatggggtttaccttcatggactcgttcagctcgctgaggcgctgcaccaccggcgtca
tgaggtcgcggagcctggcggtgggaccggctcccagagcctgccacggggccgccgctgccatgtccgtc
caccgttcgacttccgcacgaagctcccgcccggctacggtgaggccgccgccggtgagcagacgacgatc
gacgagcgcggcttcggcctccttccactcctgttcggaccatccgcgcatccggcgcatgtatgccgggt
cgaggcccttgtcggccgtgaacaagaccagcgcctcgcacggtgccaggcccgcggccacgagcgccgcc
acatggccgtcaccacgcgattctcgcagcgtcgagcaggactgccagagcgcgagatgcggctcgaccgg
cggtggaagcgcctggttggcagcacccagtacgcggccgccgggagacacgcccgcggcggcggacgcgg
caagcgtgaccatctcctcgagcccggccacatcagcgccaagcatgcggcgcagcgcagcgtcgacgatc
tccagccgcgcctgtagaaaacgctctggcgaggccacggaccagacttgcggaaccgccttggcaaccat
ctgcgggtggaaattgaagaacgccgcggtcaccacctcgggtgacgccgcaccaagcggggccgaccggg
ttccgaagtagcccatccagccgccgcggcagcccagctctttagccactaggctcgcctccggcgcatat
accacatcatggtatatctcgaaccgtagccacatttcacgagacatgtcaagctcctgccggtcgtcgtc
catcgtcggaacacgcgcccccatcgcgtccgtattgaggattcaaccctcacggcccactgtcgtgtcaa
tcccctggaatcgtggagggctcggttatgcagcctggttctctaggccgcgggtggttgaacgttcacag
tcgatcggtgagtgtccgccgcaggccgagcttgccgccgaccggctttgccatcagattcagcgctgtat
tgcatcaacaaattgggtttgttgttgatgcgagggtgccgaatagtaggtacacgtcgagcttcgggagc
cggagcagcgttcccactatcctcttgaccggctgagcacgcttccagccggctctcaccgatgagggatt
cacatgtctggtgagcaagctgagcacggtgtccacgcgatcagggaattcctgctcacctggcttgtgga
tctgctcggtgtccccgggattccgtggatcggtgggcccccctgcatcggtatggcctggactcgctga
agtcgacggcactcgtcgccgccctgtccgagtttctcgcccggcccgtaccgatcacgttcctctgggat

caccccacactggaaggcctgtccgtcgcactggcgcgcggacgacgagtcgaggaatccgccggcggcac
ggccacacgacaccggccggaagacggtgccccacccgaaccgctcgcactcgtggggatcggctgccggc
taccgggggcggcggacccgtcggccttctggcgcttgctcgtcaacaggcacgacgccatacggccggtc
ccgcagcgccggcgtgacctgctcggcaacgcggacatcgacgatacgtccttccgctggggcggcttcct
cgacgacatcgacctgttcgaccctctcttcttcgggatctcaccgcgcgaagcccgtgtgatggacccgc
agcagcggttgatgctggagctcgcctgggaggcgctggaggatgccggcatagcgcccggcaacctcgcc
gggaccgacaccggggtcttcgtgggcaccagctggtgcgactactcggccctggcgcacgaggcggcagc
gcagttcgaaatcggaccgcacgcggcgaccggaatgcacgacagcatcatcgccaaccgggtctcatacg
ccctcgccctgcagggccccagcatggcgatcgacaccgcgtgctcatcctcgctggtcgcggtccatctc
gcgtgccagtcattgtggtcgggcgagagcgagctcgccctcgccggcggcgtcactctcaatatctttgg
gccacactacctggccatgaacgagacaggcgctctctcgccggacggccggtgcaagaccttcgacgccc
gtgccgacggcaccgtacggggcgaaggagccgccgtcgtcgtcgtgacgcccctgcgcgtggcgctggaa
cgcgggctgcccgtctactgcctgatccgcggcagcgccgtcaacaacgatggcctgagcaacgggctcac
cgccccgaacccacgtgctcaggagaccctactacgtaccgcgtaccgccgggccggcttggcaccatggc
tcgtggactatgtggagtgccacggcacaggcacaccgctgggtgaccccatcgaggccaaggccctcggt
acggtgctcggtcgcgatcgcgaaccaggtaacgagctccgcatcggttcggtgaagacgaacatcggcca
cctggagccggcggccggcgtcgccggtctcgccaaggtcgccctggcgatgcgtaacggcatcctgcccg
ccagcctgcactacagccgccccaaccctcaggttcccttcgccgaaggtcacctgcgcgtccaggatgcg
acaaccacgtggcatcgccggtcagatctacggcgcgcgggaatcagcgcctttggcttcggcgggacgaa
ctgccacctggtgacagaggaactggcgcagcctgaggccgccctattgctgctagcggcggattccacgc
aggagctcaccaaccaggtggctgcgctgcgctccgctctcgcggaggaccgagtcgacctggctgcggcc
tgccggcaggcgctcgacacgctgggcaccggccgtttccggctgtcggcagccgcgcgtaccacagccga
gctgcgggcgcagctggacgcgtacgccgagggccggcccctacccggcctgtccactgcggagacagctc
ttcgtcgaccccgcgtcgcgtttctctgctccggaaccggctcccagtggttcggcatgtgccgacagctg
gttgctggcatgcccgcgttccggcgatcgctggtgcgggcggctgagcggatagaacgaatcctcggctt
ctcagtgctcgaccggctcttcgacgacgagtcgcacgcccggctcgacgacatggagatcgtccagccga
tgctgttctgcgtccaggtcgccctggccgacgcctggcgttcgctcggtgtggagcccgatctggtaatc
ggccagagtgtcggcgaaatcgcagccgcccacctcacgggtgcgctcagcctcgacgacgcggccttggt
agcggccacgcacgcacggctcgtgcaacgcctagccgtgggtcgcggcgatagcctggtcgtcgcggccg
tccccgaagttgtcgccgggcacctagcagcggtcgaggagcgacccacgctagctggacgcctcggcccg
gcctctacgctggtctcaggggataccaccgcgatccgggcgctgaccgaccggttcgcggcagcggacat
cgcagcccaccgggtacgcatgggttacgcatcccactcgccgctgatggaccccgtgctcacacccctgc
ggtcggagatcaacggcatcgtcgcgcgcccggctcagctgccgatgatctcgacggtcaccgggaaagaa
atcgacggcgaatccctgggcccggactattggccggacaacgtccgccaggagagccggctcacggccgc
cctggacaccatggcggcgcacgacatcgacgtcgtactggagctcagtccgcacccggtcctactcaagg
gcgtacgggcaagcctcgacagcgaacgggagggcgcgacgaggcgccccgtcatgtgcctggcgtcgctt
caacgagggacggacgaggcgtggagcctactcgcctcgctgggtgaactgcataccgtgggtcaaccggt
cgccgccggatccttcctgtgcgcaccacgcgggcgtcataccgagccccaggcggccatcggttccgcga
cggcgcgggccgaggagacgccccttccctgctgcccgtaaccgcccattccgctgacgcgctacgggac
acctgccgagagctgtccaaccatgtcgagcgaaatgccgcaccatggctgccggacctggcgtacaccct
ggcgacccggcgcaccccacttccacaccggatcgctttcgtggtgcgtgatcgcgacgatctactcgacg
gcctggcgcatatttccgctggcaggtcgtaccccggcgccgtcaagggcaccgtcgctggcggcggcgcc
cggcgcgtagcactggtgttctccggcggaggcacgcactgggccggcatgggacgtgccttgatgggctg
gcacgcgggcttccgggcatcgatgcacgagtgcgacgcggtgttccgggagctgatcggatggtcggtag
tcgacgagctcagcctgccggcggagcggtcccggctggacgcgaccgatatccagcagccggtgctgttc
accctgcaggtctcactggcccgtctctggatggagcttggcatcgagccggaggcattcgtcggacacag
catcggcgaggtcgccgccgtctgtgtggcgggcgggctgtcggtgcgtgacgctgcccgggtcaccatcg
cccgctcccaccttatccagcaccgcgccgcgaaggccgccatgatcgccgtccaagccggagacgaagag
atcatcccgttcctcgccccgtacggcgggcgcgtcgccatcgcggcgctcaatagcccaaccagctccgc
ggtttccggcccgcccgaggagattcgcgccctggaggtcgcactaaatcgagccggcatctcgagccgag
ccgtacgcgttgatcgcccccggtcacagccctggtatggatccgctgctgtcaccgctgcgtgaagcgctc
accaacattgaacctcgtgcattttgggccagattccactcgacggcgctcgacggcgcggtcgatcccgt
ggtcaacgcagactactgggcgcacaacctgcgcaaccaggtgcgcttcgcgccgacggtggcggctctgg
ccgacgcgggcattgacacgtttgttgagatcagcccgcatggcaccctacgcggcgcgatcgaggagatc

gctcaggctcagggagcctctgttgtcgtggccgactcgatccgccgcggggaagacgacaatcgctgctt
cctcaacgcggctgcatcgctgttcgtacatggcgtgcccctgtcgctcgagacgctgttctcgtccgacg
cgcaggttgtcgagactccgctcgtgagctggcagaggcacccctactggctggacgccgcgcccgtccgc
caaccacgcacgcccgctgcggcatccgtcgccggcccgaccaccagcttccacaatgagaacgccaccac
cccgacagcggaagccaccatcaccggcgagatcgcgcacgttctgggcatgccggtcgaggagctcgccg
agggcacgcggctgcgggacttcgggctcgactccatgctcgccatgcggctgggcagcagaatccagacc
ctgttcggccatcgtgtctcgctgtttgagttcttcacggaccggacggttggcgaactcaccgcgcgcat
cgccgagctcgtcggtgcgcgcgcgtccaccggcggcgaggacaacgcgccgccgcggctgcggaatacgg
ctgttgcggacaatctgtccgacgccgacgccgaggagttgctcgcggaactgaccgatcgtgggctgctg
gagccggtttccccggaggcgcccgcggcgcaaccacgcgaagagctgcgggcggcgctagaacgcaccca
atcctttgcgcttgcgccggtggcgcacggtcaggccgccatttggttcatgcagcagctgtcgctcgacg
gcgcagcctacaacctcatgttcggcgctcgggtgcccgataagatcgacgagtcggccctgcgccgggcc
gcagccgcggtggtggagcgacatcccgcgctgcgcaccgtcttcgtcgaagccggcggccacccctacca
ggtgataaaggcggatcccggttacgaattcgacacggtcgacggtacggggctcgatgacgcggcactga
ccgacctcctcgccgagcacggccaccggcccttcgatctcgatcagggtccgctattgcggttggtgttg
gtcaaccgcggtgcggcagacaactgcctgctgctggtcatccaccatgtggcagctgacgcggcgtcggt
ggacatcgtcgtccgggacctgcgcgagttctacggcgaggcacagcgcggaacgctgctcacccaggcac
ccgagacaccgtacaccgagttcgtcgaatgggaacgcgaatggctcggcagcccggcggccgaggcggcg
ctgcactggtggtcgcaacagttggcgcagcccccggcgcacctcgaccttccacgcctgtcagcgcccga
cgcgaccgccgacaagcgaccagccaaagagcgggacggtgtggtgagctacgcgggggaggatgccagct
tccggtgggacgctgccgatacccgcctactgaaggacttcgccgtgcgcgccggcgtctcggtcagcacc
cttgtcatggcaggcttcgtggccaccctgggccgggcaaccggggcgagggacatcgtcctgggcacggc
ggtcgcccagcgcagtgatgcggggcgcgagtccgcggtggggtactacctcaacaccatcccagtccgcg
cccggccccgtccggacatcaccttcgacgcgttgctgcgtgaagtccacgacttcgcgctcgggatgttg
gaacacatgaactatccactggacctgcttgtctcggcgctcaaaccgccgcggcggcagggccgcacgcc
gttgttcgacatcgctgtcaactggctctccggtgacgcgttcacctatgcgaacacattgttccacggtg
ctggcgtggcggcgtggcctgccggccccctaccactggtcccgctgcctctccggcggcacatcgcaaag
ttcgatctcgaaatcacgatgggcgacgtcgccgacgaggtggtcggtcaagtgcagttcaaacccgcgtt
cctggaaagagaaaccgtgacgacactgctggagtacttccaccgcgtgctgttgcaggcgatcgaccggc
cggatgtgccgctgagtgacctcgtgctggagacctcggtgaagcaggcggtctcatgacggccgcatcga
cactgcacggcggctttgtcgcccacgcagccgcgaatccggacacgctggccgtagcgtccgacgccggc
gtgatgacctacggtcggcttgacgagacctcggcggcgctggccgagcggctgtctgccttgggtgctgg
cccgggtgttccgatcggggtctgtatcgaacgcacgccggacctgctcgtcgctatcctcggcgtgctgc
gcgcgggcgcctgctatctgccgctcgatcctcaatattcagcgcgccacctcggcttcatggtggccgac
agcgggacccgcctggtcgttaccacacgatcctctcgggacgcgtgcccggacggctgcaccgcgctcgt
cctggaggaatccgaggcgatagccgacccgccgccagtggccgcggttccggacgattctgcctacgtca
tatatacctccggctcgaccggcacgcccaaggggtgccgatccggcacagcagctgcgcggcgatgctt
gccgaggcggaccgaattttcgagggctgtgacatgagcggtatcgccgccgtcacctcggtctgcttcga
cctgtcagtgctggagatcttctccgccctcagccgtggccggacgctcgtcctggtgaatagtgccagcc
accttccggagagctcccatgtcgaacgggtgacgcacgtcagcacggtcccgtccgcaatgaccagcctg
cttgacgcgcaagccgttccggccggcctgcggaacgtggtgctcggcggcgaacccgtacgtcggagcct
ggtcgaccggatctaccgcgagaccaacgtcgacttcgtcttcaacggatacggcccgacggaaggcacgg
tcttctgtaccttcaagcccgtatcccgcgacgaggccggcgagccgtcgatcggtacgccatccctgacc
gctcgcgtctatgtgctcgacgagaagctgcggccgtcggccgtcggcgagtcgggtgagctgtacctcgg
cggtgccggacttacctggggctacctcaaccggcccgggctgactgcggaacggttcgtacctgatccgc
aggtggcgggtgaacgcatgtatcgcaccggcgacatcgctcggctcaacgaagcaggtgaaatcgagttc
gtgggacgctccgaccttcaggtgaaggtccgcgggtaccgcatcgagctagaagaggtcgaggcacgact
gaccgaatgccccgaggtgcggacggctgcggccgtcgtccgtgagcagacgccgggtacgagagccctga
ccgcgtacgcggttccggcgagtggagcacccgacggcgacgggccctggctcgacgccgacctgcaggca
acgatcaagcaacagctcggtgcgctgttgcccggttacatggttcccgaaacgatcgtcttcctgcccgc
gctcccgctgtcgccagttgggaagctggaccgcacggcgctaccggcaccacccgttgtcgatgtgctgc
cctcgggggactccgccaccaccgacaccgaacaagcgcttgccgagatctggggtgcactgctggaccgg
actccgcagtccatcggcatccgcgacacattctacgacctcggcggcaactctttgttgttggtgcggct
cgcgaagcgaatgggtcagcgctttcaccgcaaggtcggcgtggcggacctgttccggttccgcgacatcg

gctcgctcgccaagtggctggacgacgagagcggaaagagtcctgaggacatcgagcaggcacgacgccgt
gccagcaccagacgctcggtggtgcgcggccacagcagatcaccgagcactcgaaccgacccgaccgtcaa
gaacacgcccgcatcaaatggaggcccacatccatgagcgaagccgttgagcggaactacgatccgggaga
cgtcgccatcatcggcgcctcgtgcagattccctggcgcacgaaacaaggaacagtactgggacaaccttc
tccatgggcgcgagggcgtgacctttacgccaaggacgagatcgaggtggacgagacgctcatcaacagt
ccggcgtacgtgcgggccaccggagcactggacggatacgacgagttcgatcccgccgtgttcggtgttag
cgatcgaatggcggcggcgatgactcccgagcaccgggtcttcctggaggcatcctgggaggtgatggagg
acgccagctacgaccccgagcgggtccgcggcgaggtgggcgtctacgcctcgaccaacccgcagagcgcg
gccctctacagctcaccaccggactgggtatcggccggcccggaggtgatggaccgcagtaacgcatggct
accggacacgatcacctccaacgtcttgtactacctcggcctgaccggtgaggcggtcaccgtgaccgccg
tgtgctcgggctttcactatgcggtgcacctcgcgtgccagtcacttctcctaggccagactgatatggcg
atcgcgggcggcgtcatggttcggctgccccagcgccgcggctatctgtgggaggagggtcgaattctgtc
ccgcgacggccactgccgccccttgacgcgaacggcaccggatcggtactggccagcggagtcgcggtgg
tcctactcaagccactgccacaggccgtcgccgaccgcgaccacatctatgcaacggtcaagggaacggcc
atcaacaacaacggcatcagtgccatggcgtacggcctcgcgcagcctgaacggctcagcgcctgcatcgc
cggtgcgatgcaggccggcgatgtggcgcccgaaacggtgtccatgtacgaggccaatgggttcggcatgc
ccatcacggactcgctcgaggtgcacgccgcccatctggcgttcgggaagcagtccggcacctgctcgatt
ggggcggtcaagggcaatatcgggcacgccggcgtggttgccggtggctccggcgcggtcaaggcggcgtt
cgcgctgtaccaccgcagtctcccgcccaccatcaacctgaccgagctcaacgaagaaatcgacttcccgc
gtactccgttcgtcccgcaactggagcccgccgcctggcagccggagtgcggcatcagacgggccggcatc
accgcgctcggcggcggcgggtacaacgcgcacctggtcctcgaggagccgccgcgtccggtcgaacggga
gcccgagaaccgccggccccggatcgtcacgctgtcggcgctcgacgacgccgcggtctcccgccaacgcg
cggccctgtcctcctggctcgctgagcattccgatgctcggctcgacgacatcgcctatagcctcaacctc
gggcgtaaggcgctgccatcccgatgggcggccgtcatcagcacccgcgacgaactgctcgaggtgctgtc
gggcgacggcaagagcggtcgggtgtctcgcttcggccaagagcgccgcgccgacctggcgcgttttcggc
gcaccgaagacggactcgctctcggcagcggggacgagatgcgcgacgttcaggcgctgaccgagctcgcc
gcagcatgggtgcaggggaacgcgtgaacttcgaggtgttgcacgccgacgaacgcagccaccgcatttc
gttgcccaactatccgttcgcgcgccgacggttctggcgcaccgactggtgaccgccgcggacacaggtcg
cgtgggttcaggcgaaagccaaacgccggcccacaccgtcgacctagccgaccccgccaccttcgccaatc
acgacctgaccggtttctggcaacagctgcgcgacgaggagccgatccactggaatcccccgacagccgga
cgtcgaggcttctgggtggtgtcgcgatacgcagacattctcgacgtctaccgcgacgacgtcaccttcac
ctcagagcgcgggaatgtgctcgtgacgttgctggcgggcggtgacgcgggcgctggccggatgctggcgg
tcaccgacgggccgcgtcacgcggagttgcgcaagctgctgttgcgggcattggggccgcgcgtactcgcg
ccagtctgcgccgcggtgcggaccaacacccggcagatgatccgcgaggcagtcaccaagggtgagtgcga
cttcgcgtccgacatcgccagccgcattccgatgatgacgatctccaacctgctgggggtccccgacgcgg
accgtgccttcctgctgagtctgacgaagacggcgctgagcgccgacgacgaaagcatcagcgagaccgag
tccgcgatggcacgcaacgagatcctgctctacttccaagacctcatggagttccggcgcgaccaccccgg
cgaggatgtcgtgagcatgctggtgaacagcagcatcgacggagcgccccttcggacgacgacatagtgc
tcaactgctacagcctgatcatcggcggcgacgaaacgagtcggctaaccatgatcgatagcgtgaatacg
ctggctgcacacccccagcagtggaggcgactcaaagatggtcaatgcgagatcgacaaggccgtggacga
ggtactgcgctgggcctcaccctcgatgcacttcggacgggtggcagcgcgggacaccattcttcatggag
tgcggatccgggcggacgacatagttaccctctggcatgcctccggcaaccgcgatgagcgggtcttccat
cggccggaggtgttcgacctgggccgtacacccaaccggcacctctcgtttgggcacgggcctcactactg
catcggttcctaccttgccaaggtggagatctcggagctgctgatcgctctgcgggatctgacatcgggtt
ttgagacaacaggcgagccgcaacggatccggtcgaacctgctcactggcttcgccaccatgccagtccgc
ttcgttcctgatcgtgcagggctcgctcgcgacgcgctggatgggtgagacaaatgattgccaacccccctc
gatccgggcgacggcacgtggctggtcgtggttaacgacgccggccaacacgcgctgtggcggccctttct
ggatatgccgcccggctggcgtgtcgtcctacgcgacgccgaccacgacgcggcccttgactacgtcgagc
gcaactggactgccctgcaacccgcgggccacagcgaaccgggggcgacgaggtaatgccgtttgcgttcg
ccaacgacatcaggttgcactacgaattgcatggcacgggcgatcccgtggtgctggtcagtggtgccggc
gtatccgggaaatcctggctgatccatcaggggccggcactgctggagaacggctaccgggtctgcgtcta
cgacagccgcggccaaccacccagcgatgagtgcgtcagcggcttcgtcgtggaggatctggtagccgacc
tcgcagccctactggagttcctcgacgctggcccggcccgcctcatcggcacctcgacgggtgcctacgtg
gtgcaggagcttgccctgcgtcgaccggagctggttcggcaggcggtgttgatggccagccgggcccgtcc

cgacgtgttacggacacagctggcgcgtgccgagatcgagttgggcgagtctggggtgacgctgccgtcct
cgtatcgcgcggtcgttcgcgcgctccagatgctctcgccacgatcgatggacgacgagtcgtcgatccgc
gactggctggccctgctcgagctcgccccgcccgacggagcgggagtacggcatcagctcgctctccaacc
gatgccggaccgccgcgccgcctacgccgagatcgcggtgccctgccacgtcgtgtcgttcgcggatgact
tgatcgcgccgcccgggtacggcgaggagctggcggggtgtattcccggcgccaccttcgacttggtcggg
gacgccggccacttcggctacctggaacggcccgaggaggtcaaccacatcatcgccaaacatttcgcggg
catcggcgcccgcctcaccggacacaacccagacaccaggaggagtgcgccatgggccagctgaccgacgc
cgtgctcagcaacgcgacggcggcagagatcgccgccgctccgctacccgaggagtacttcgccgcccacc
tgcgggtcgaagacgccacgatgttcgagggcgtcgtggacaaggacgtccgcaagtcgctgcggctcggt
cacgttccaatgcccgagctggctccggacgaggtcctcgtcgcggtgatggcgagcgcggtcaactacaa
caccgtgtggtcggcgatgttcgagccgatgtccagcttccgctatctgaaggggctcgcccggcagggtg
gctgggcgaaacggcacgaccaacccttccacgtggtcggctccgactgcgctggggtgatcgtgcgcacc
ggctccggcgtccgccgctgggagatcggccagcacgtcgtggtgagcccggcgtacgtggacgaccagga
gcccaccacgcacgccgacggaatgctcggcggggaccagaaggcgtggggtttcgagacaaacttcgggg
cgctggccgagttcacggtggtccgggccacccagctgattcccaagcccgcccacctcacctgggaggag
gccgcgagcgtcccgctgtgcgcgggaaccgcctaccggatgctggtcagcgagcgcggcgcgaacatgaa
gcagggagacgtcgtgctcgtctggggtgcggccggcggtctcggcggctacgcagtgcagtatgtgcgta
acggtggcggcatcccggtggccgtcgtcgggtccgagcgcaaggcggaggcaatgcgccgactcggctgc
gaggtgattctcaaccggaccgagatcggcatcactgacgagatcgccgacgacccgaagcaggtgatcag
agccggcgtgcggctcggcgaaatgatcaacgaacaggtgggccgggctccggatatcgtcttcgagcaca
tcggacggcccaccttcgggatatcggtgttcgtcgcacagcgcggcggcaccgtagtgacgtgcggctcg
agcgccgggtatcagcacagcttcgacaaccgctacctgtggatgcgcctcaaacggatcatcggcagcca
catcgccaacctccaagagtactgggatgcgcaccgcctgttcgaactcggtcacatcgtgccgaccatgt
cggcggtctatccgttgacggacgtcgcggccgcctgccgtctcgtgcaggccaaccagcagatcggcaaa
atcgccgtgcgctgcctggcaccccgcgccgggctcggggtgaccaaccccgagttgcgcgagcgcatcgg
cgaggaccgcctgaacccgtggcgcggcctcaccaccggctctgacggtgccgagttcggaggacagcgtg
acgacccggcctaacccactcttcgtcggcatcgagcgcgccccagcgcgcgcctccatgcgcgccacagg
tctgtcgaccgaggacctgaagaagccgatgatcggggtagcgcacagttggatcggcacgatgccgtgca
acctcaaccaccggcggctcgcccaggaggtgatggccggcgtgcgcgccgcaggtggcaccccgatcgag
atcaacaccatcgcgatctcggacgtgatcaccatgggcacggagggaatgcggacctcactggtgagccg
cgaggtgatcgccgactccatcgagttggtatgccgcggccacggtctcgacggcctggtgaccctggccg
gctgtgacaagaccatacccggtgcggcgctggcccacgtgcgactcgacatccccggagccgtcatctat
tccggcacgatgatgccgggcgagcacctcgggcgcgacatcaccctgcaggatgtgtttgaggccgtcgg
cagcgccaccgccaccggctgcaccgacgagctcgacaagctcgagcgcgcggcctgtcccggtatcgggg
cctgcgcgggccactacacggccaatacgatggccgtggtgctggagttcctcgggctttccccgttcggc
tcgatggatccgccggcagtcgacgcccgcaaggacacggtctgccgccaggccggcgagctggtcatgcg
ggcggtggcagaagggctgcggcccagccggttcctgacgccctcctcgttgcgcaacgccatcgccgccg
gggtcgccaccggcggctcgacgaacatggtgctccacctgttggcgatcgctcgggaggcgggcatcccg
ctggacattgacgagttcgaccggatcagctcggtgaccccgatcatcgctgacctgcgtccaaacgggac
gtacaccgcggtggacctcgaccgggcgggcggcacccgggtgatcgcccgccacatggtcgacgcgggcc
tgattgctggcgacgaaagcaccgtcaccggccgcaccgtcgcacacgaggccgcggacgcggccgagaca
cccggccaacgggtcgtcaccacggtggaggcacccctatcgccttccggtgctctcctgatcttgcgggg
taacctagcgcccgatggcagtgtggtgaaggcacccggcgcggtgaccctgcggatgaccggcaccgcat
tggtgttcaactgcgaggaggaagcgatggccgcggtccagactggccgcgtccggccaggccacgtcgtc
gtcatccgctacgagggtccgcgcgggggtccagggatgagggaaatgctcggagtgacctcggcgctcat
cggccgcggcctgggcacgtcggtcggtctggtgaccgacggccgtttctcgggcgcgaccaggggactga
tggtggggcacgtcgctccggaggcggcggagggcggaccgatcgcggcggtgtgcgacggtgaccggatc
accatcgacctgcagcggcgcgaatgctctgtcgacctggatccgggcgaactggccgcgcggatgcgaga
ctggtcggctccgccaccgcgctacacgatcggcgtcatggccaaatactggtcgacggtctcgtcggcgg
ccgtgggcgccgtgacgaccccgcaccccacccagggcccggcgacagcgtcaggtaaggccgaggagtgc
cagcaggcgagtgcggtcgagggcgtgatggcggttggcggcggcgatgtcggtgctgccggcggttcgta
gcgcgctgatggcggcgttgcggaggacggccatgacctgcggtctggtgccggtgcggacccttgaatgg
tcttcgttgtaggtaacgcgatgtaggagcggcaaccagccaccacagcgcacacccgccgcagtcaccac
aacgatcaaccggtgccgcatacccggcagacatgcggatccggcaggttgaccaacgtggcggcaggtcc

agcagtggtttcgggcaccgtgacagccggtgcgtaggacatggacgaaagcagcgatggtgacagcgcgg
gcatcacacacttcagacgatcaaggcatcgataccctgaggatcatccaaaccgaaaatgcccgccgtac
tgcccgtcactcgggcgcttctttaccgactgtctcgagttgaacgagtccacaaggtcaccggggcactg
gcactccaccggccctcgatcaggacgcggtaaagcttcggacgccgggctcgtcggcactacgccgggcg
acgacggcgtcggcgactgccgcgagttcctcctctgtgtatcggtaggcgccctcggccgtgatgcggat
ggccgtggggaaaatgcggcggaccatgtcggggccgccggtggccgagtcatcatcggcggcatcgtaca
gcgcctcgatcagggcactcagggtagccgcatcatcggaaccggggtcgtgtagcttcttcagagacgat
ttagcgaaggggggagcccgagccgatcgactggtagccggacttttcctcatagcgggcgccgattgcgtc
gaacgacacgatccggccggccctcgcaggatcggatgcctcgagatcgtagccgatgaacagcgggatag
aaaccatgcccgcgttggccatggggaggtttcccttgactatggcagcgagccggttgctcttgccctcc
agtgagagcgggacgccctcgatcttctcataatgctccagctcgacgacaaacaaccgagtcatgtcaat
ggagatgcccaccgtgccggcgaaaccgaccgccgaatagtcgtccgtgatcgtgagtttgtccacgtccc
ggccggcgatcaggttccccatggtcgtacgacggtcgccagcgagcaggacgccaccagcgaacgtggcg
gccactatggtggtgccgtgcggagcctcgccgaacgtctcggcggcacgcttggccggcagcagcgtggg
atcgtattccttcaacaactcgacaaacgaggaagaatgaaatgttaagaacgtggacggcagaccccgat
tcacctcttgttccatagtcgccagtatagcggctctccgaagttcagcagggttcgcggctggggagtcc
ggatgcgaggggcccacatctcatcgaacaccatctctttgtcgggtttgggtgggcaccggtgaagatct
acgagcatggtcaacgatacgacccggctactggggctggacggcctgacggtcgagcgggtcgagctaga
cgccatcggtgtccggtggaggacctgtccaccggttgtgagcaggcgcggggcagcccgaagtgcgggca
gcggacggtccgggtacagcaatgggcccctgacgccaccgagacgttgccgcatcgtcgatgccgctcag
ttcccgaccctcgaaggaccaggtcggcgcgctgcagagtcacgaaccgggccgcccggggttgggacgtg
cagagtgagcgacaccgccctatgtcaagtggcagccacagaagagtctccaacttgcctccgtgtcgagc
gcctaccccatcaactgtgatcacggcacgagcaggggcatggcacccgtggccgtgccatggacgtccg
ctgctggaactgagtgaggggacattggtcgggtagcgggagcgcattgctgcctcccagccgtgcatgcg
gcgccaaccgcacacggctcaagcgatgacctcgacgcgacccggaacggcggtatgcggtcgctgttcgg
cggcgtgtacgcgccgtcaacactcggctcgttcctgcgcaccttcacccacgggcacgtccggcagttgc
aggccgccgcccgcgacaccctgatcggcctcacagggcaccgggtcggtcacgccacagtctcgctcacc
gacttgcctctgcgtcgagcgccttcccatcgacgacaaatcccggcaggagcggacggcattgcactcac
gaccgctccgttgacgccccttgcgggagctggccgatgagacgtgggtgctccgtgggtccgcctagact
cacgggcaagaatctgggaaaagtgcttcacatccagtagcggacgcacgaagaacagctgggagtcaaag
tcctcataccattgagctaacggcccgctgtggcacccggctaccgatcgctggctccgccccagggtccg
ccggggcggcacgctgccggctaggccggaaatccctcggcgggctgcccgcggctccgcccatctggttg
gtagcctccggtgctcgacctcgtgaatcccgaaggataccg

REFERENCES

Adams, J., M. Behnke, S. Chen, A. A. Cruickshank, L. R. Dick, L. Grenier, J. M. Klunder, Y. T. Ma, L. Plamondon and R. L. Stein (1998). "Potent and selective inhibitors of the proteasome: dipeptidyl boronic acids." Bioorg Med Chem Lett **8**(4): 333-8.

Adams, J., V. J. Palombella, E. A. Sausville, J. Johnson, A. Destree, D. D. Lazarus, J. Maas, C. S. Pien, S. Prakash and P. J. Elliott (1999). "Proteasome inhibitors: a novel class of potent and effective antitumor agents." Cancer Res **59**(11): 2615-22.

Alo, P. L., P. Visca, A. Marci, A. Mangoni, C. Botti and U. Di Tondo (1996). "Expression of fatty acid synthase (FAS) as a predictor of recurrence in stage I breast carcinoma patients." Cancer **77**(3): 474-82.

Alo, P. L., P. Visca, G. Trombetta, A. Mangoni, L. Lenti, S. Monaco, C. Botti, D. E. Serpieri and U. Di Tondo (1999). "Fatty acid synthase (FAS) predictive strength in poorly differentiated early breast carcinomas." Tumori **85**(1): 35-40.

Ansari, M. Z., G. Yadav, R. S. Gokhale and D. Mohanty (2004). "NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases." Nucleic Acids Res **32**(Web Server Issue): W405–W13.

Arakawa, K., R. Muller, T. Mahmud, T. W. Yu and H. G. Floss (2002). "Characterization of the early stage aminoshikimate pathway in the formation of 3-amino-5-hydroxybenzoic acid: the RifN protein specifically converts kanosamine into kanosamine 6-phosphate." J Am Chem Soc **124**(36): 10644-5.

Arcamone, F., G. Cassinelli, G. Fantini, A. Grein, P. Orezzi, C. Pol and C. Spalla (1969). "Adriamycin, 14-hydroxydaunomycin, a new antitumor antibiotic from *S. peucetius* var. caesius." Biotechnol Bioeng **11**(6): 1101-10.

Arvigo, R. and M. Balick (1993). Rainforest Remedies. Twin Lakes, Lotus Press.

Bentley, R. (1990). "The shikimate pathway-a metabolic tree with many branches." Crit Rev Biochem Mol Biol **25**(5): 307-84.

Bentley, S. D., K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Niel, E. Rabbinowitsch, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill and D. A. Hopwood (2002). "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)." Nature **417**: 141-47.

Beran, M. and J. Zima (1993). "Determination of monensins A and B in the fermentation broth of *Streptomyces cinnamonensis* by high performance liquid chromotography." Chromatographia **35**: 206-08.

Bergendahl, V., U. Linne and M. A. Marahiel (2002). "Mutational analysis of the C-domain in nonribosomal peptide synthesis." Eur J Biochem **269**(2): 620-9.

Bergmann, W. and D. C. Burke (1956). "Contributions to the study of marine products. XL. The nucleosides of sponges.1 IV. Spongosine." J. Org. Chem. **21**(2): 226-28.

Bergmann, W. and R. J. Feeney (1950). "The Isolation of a new thymine pentoside from sponges." J. Am. Chem. Soc **72**(6): 2809-10.

Bergmann, W. and R. J. Feeney (1951). "Contributions to the study of marine products. XXXII. The nucleosides of sponges. I." J Org Chem **16**(6): 981-87.

Beyer, S., B. Kunze, B. Silakowski and R. Muller (1999). "Metabolic diversity in myxobacteria: identification of the myxalamid and the stigmatellin biosynthetic gene cluster of *Stigmatella aurantiaca* Sg a15 and a combined polyketide-(poly)peptide gene cluster from the epothilone producing strain *Sorangium cellulosum* So ce90." Bioch. et Biophys. Acta **1445**: 185-95.

Bierman, M., R. Logan, K. O'Brien, E. T. Seno, R. N. Rao and B. E. Schoner (1992). "Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces* spp." Gene **116**(1): 43-9.

Bisang, C., P. F. Long, J. Cortes, J. Westcott, J. Crosby, A. L. Matharu, R. J. Cox, T. J. Simpson, J. Staunton and P. F. Leadlay (1999). "A chain initiation factor common to both modular and aromatic polyketide synthases." Nature **401**(6752): 502-5.

Blanc, V., P. Gil, N. Bamas-Jacques, S. Lorenzon, M. Zagorec, J. Schleuniger, D. Bisch, F. Blanche, L. Debussche, J. Crouzet and D. Thibaut (1997). "Identification and analysis of genes from *Streptomyces pristinaespiralis* encoding enzymes involved in the biosynthesis of the 4-dimethylamino-L-phenylalanine precursor of pristinamycin I." Mol Microbiol **23**(2): 191-202.

Bollag, D., P. McQueney, J. Zhu, O. Hensens, L. Koupal, J. Leisch, M. Goetz, E. Lazarides and C. Woods (1995). "Epothilones, a new class of microtubule-stabilizing agents with a taxol-like mechanism of action." Cancer Res **55**(11): 2325-33.

Brakhage, A. A. (1998). "Molecular regulation of beta-lactam biosynthesis in filamentous fungi." Microbiol Mol Biol Rev **62**(3): 547-85.

Brakhage, A. A. and M. L. Caruso (2004). Biotechnical genetics of antibiotic biosynthesis. Genetics and Biotechnology. U. Kuck. Berlin, Springer Verlag**:** 317–53.

Bross, P. F., R. Kane, A. T. Farrell, S. Abraham, K. Benson, M. E. Brower, S. Bradley, J. V. Gobburu, A. Goheer, S. L. Lee, J. Leighton, C. Y. Liang, R. T. Lostritto, W. D. McGuinn, D. E. Morse, A. Rahman, L. A. Rosario, S. L. Verbois, G. Williams, Y. C. Wang and R. Pazdur (2004). "Approval summary for bortezomib for injection in the treatment of multiple myeloma." Clin Cancer Res **10**(12 Pt 1): 3954-64.

Buchanan, G. O., P. G. Williams, R. H. Feling, C. A. Kauffman, P. R. Jensen and W. Fenical (2005). "Sporolides A and B: structurally unprecedented halogenated macrolides from the marine actinomycete *Salinispora tropica*." Org Lett **7**(13): 2731-4.

Bull, J. H., G. Ellison, A. Patel, G. Muir, M. Walker, M. Underwood, F. Khan and L. Paskins (2001). "Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray." Br J Cancer **84**(11): 1512-9.

Burja, A. M., B. Banaigs, E. Abou-Mansour, J. G. Burgess and P. C. Wright (2001). "Marine cyanobacteria - a prolific source of natural products." Tetrahedron **57**: 9347-77.

Byrne, K., A. Shaffiee, J. Nielsen, R. Arison, L. Monoghan and L. Kaplan (1993). "The biosynthesis and enzymology of an immunosuprressant, immunomycin, produced by *Streptomyces hygroscopicus* var. *ascomyceticus*." Dev. Ind. Microbiol **32**: 29-44.

Cane, D. (1997). "Polyketide and nonribosomal polypeptide biosynthesis." Chem. Rev. **97**: 2463-706.

Cane, D. E. (1997). "A special thematic issue on polyketide and non-ribosomal polypeptide biosynthesis." Chem Biol **97**: 2463-706.

Cane, D. E. and C. T. Walsh (1999). "The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases." Chem Biol **6**: R319-R25.

Cane, D. E., C. T. Walsh and C. Khosla (1998). "Harnessing the biosynthetic code: Combinations, permutations, and mutations." Science **282**(5386): 63-68.

Challis, G. L. and D. A. Hopwood (2003). "Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species." PNAS **100**(90002): 14555-61.

Challis, G. L., J. Ravel and C. A. Townsend (2000). "Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains." Chem Biol **7**(3): 211-24.

Chang, Z., P. Flatt, W. H. Gerwick, V. A. Nguyen, C. L. Willis and D. H. Sherman (2002). "The barbamide biosynthetic gene cluster: a novel marine cyanobacterial system of mixed polyketide synthase (PKS)-non-ribosomal peptide synthetase (NRPS) origin involving an unusual trichloroleucyl starter unit." Gene **296**(1-2): 235-47.

Chater, K. F. and S. Horinouchi (2003). "Signalling early developmental events in two highly diverged *Streptomyces* species." Mol Microbiol **48**(1): 9-15.

Chauhan, D., L. Catley, G. Li, K. Podar, T. Hideshima, M. Velankar, C. Mitsiades, N. Mitsiades, H. Yasui, A. Letai, H. Ovaa, C. Berkers, B. Nicholson, T. H. Chao, S. T. Neuteboom, P. Richardson, M. A. Palladino and K. C. Anderson (2005). "A novel orally active proteasome inhibitor induces apoptosis in multiple myeloma cells with mechanisms distinct from Bortezomib." Cancer Cell **8**(5): 407-19.

Chen, H. and C. T. Walsh (2001). "Coumarin formation in novobiocin biosynthesis: beta-hydroxylation of the aminoacyl enzyme tyrosyl-S-NovH by a cytochrome P450 NovI." Chem Biol **8**(4): 301-12.

Chen, P. and M. Hochstrasser (1996). "Autocatalytic subunit processing couples active site formation in the 20S proteasome to completion of assembly." Cell **86**(6): 961-72.

Chirala, S. S. and S. J. Wakil (2004). "Structure and function of animal fatty acid synthase." Lipids **39**(11): 1045-53.

Cimino, G., S. De Rosa and S. De Sttefano (1984). "Antiviral agents from a gorgonian, *Eunicella cavolini*." Experientia **40**: 339.

Conti, E., T. Stachelhaus, M. A. Marahiel and P. Brick (1997). "Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S." Embo J **16**(14): 4174-83.

Copeland, A., S. Lucas, A. Lapidus, K. Barry, T. Glavina del Rio, E. Dalin, H. Tice, D. Bruce, S. Pitluck and P. Richardson (2006). Sequencing of the draft genome and assembly of *Salinispora tropica* CNB-440, NCBI.

Corey, E. J. and W. D. Li (1999). "Total synthesis and biological activity of lactacystin, omuralide and analogs." Chem Pharm Bull (Tokyo) **47**(1): 1-10.

Cortes, J., S. Haydock, G. Roberts, D. Bevitt and P. Leadlay (1990). "An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*." Nature **346**: 176-78.

Costilow, R. N. and L. Laycock (1971). "Ornithine cyclase (deaminating). Purification of a protein that converts ornithine to proline and definition of the optimal assay conditions." J Biol Chem **246**(21): 6655-60.

Courtois, S., C. M. Cappellano, M. Ball, F. X. Francou, P. Normand, G. Helynck, A. Martinez, S. J. Kolvek, J. Hopke, M. S. Osburne, P. R. August, R. Nalin, M. Guerineau, P. Jeannin, P. Simonet and J. L. Pernodet (2003). "Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products." Appl Environ Microbiol **69**(1): 49-55.

Coux, O., K. Tanaka and A. L. Goldberg (1996). "Structure and functions of the 20S and 26S proteasomes." Annu Rev Biochem **65**: 801-47.

Cragg, G. M. and D. J. Newman (2001). "Medicinals for the millenia - the historical record." Ann. NY Acad. Sci. **953**: 3-25.

Davidson, B. S. (1995). "New dimensions in natural products research: cultured marine microorganisms." Curr. Op. Biotechnol. **6**: 284-91.

Deng, H., S. Cobb, A. McEwan, R. McGlinchey, J. Naismith, D. O'Hagan, D. Robinson and J. Spencer (2006). "The fluorinase from *Streptomyces cattleya* is also a chlorinase." Angew Chem Int Ed Engl. **45**(5): 759-62.

Dieckmann, R., Y. O. Lee, H. van Liempt, H. von Dohren and H. Kleinkauf (1995). "Expression of an active adenylate-forming domain of peptide synthetases corresponding to acyl-CoA-synthetases." FEBS Lett **357**(2): 212-6.

Donadio, S., M. Staver, J. McAlpine, S. Swanson and L. Katz (1991). "Modular organization of genes required for complex polyketide biosynthesis." Science **252**: 675–79.

Dong, C., H. Deng, M. Dorward, C. Schaffrath, D. O'Hagan and J. H. Naismith (2003). "Crystallization and X-ray diffraction of 5'-fluoro-5'-deoxyadenosine synthase, a fluorination enzyme from *Streptomyces cattleya*." Acta Crystallogr D Biol Crystallogr **59**(Pt 12): 2292-3.

Dong, C., F. Huang, H. Deng, C. Schaffrath, J. B. Spencer, D. O'Hagan and J. H. Naismith (2004). "Crystal structure and mechanism of a bacterial fluorinating enzyme." Nature **427**(6974): 561-5.

Drent, M. L., I. Larsson, T. William-Olsson, F. Quaade, F. Czubayko, K. von Bergmann, W. Strobel, L. Sjostrom and E. A. van der Veen (1995). "Orlistat (Ro 18-0647), a lipase inhibitor, in the treatment of human obesity: a multiple dose study." Int J Obes Relat Metab Disord **19**(4): 221-6.

Du, L., M. Chen, C. Sanchez and B. Shen (2000). "An oxidation domain in the BlmIII non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in *Streptomyces verticillus* ATCC15003." FEMS Microbiology Letters **189**(2): 171-75.

Du, L., Y. Q. Cheng, G. Ingenhorst, G. L. Tang, Y. Huang and B. Shen (2003). "Hybrid peptide-polyketide natural products: biosynthesis and prospects towards engineering novel molecules." Genet Eng (NY) **25**: 227-67.

Du, L., C. Sanchez, M. Chen, D. Edwards and B. Shen (2000). "The biosynthetic gene cluster for the antitumor drug bleomycin from *Streptomyces verticillus* ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase." Chem Biol. **7**(8): 623-42.

Du, L., C. Sanchez, M. Chen, D. Edwards and B. Shen (2003). "BlmIII and BlmIV nonribosomal peptide synthetase-catalyzed biosynthesis of the bleomycin bithiazole moiety involving both in *cis* and in *trans* aminoacylation." Biochemistry **42**: 9731-49.

Du, L., C. Sanchez and B. Shen (2001). "Hybrid peptide-polyketide natural products: Biosynthesis and prospects toward engineering novel molecules." Metabolic Engineering **3**(1): 78-95.

Duchaud, E., C. Rusniok, L. Frangeul, C. Buchrieser, A. Givaudan, S. Taourit, S. Bocs, C. Boursaux-Eude, M. Chandler, J. F. Charles, E. Dassa, R. Derose, S. Derzelle, G. Freyssinet, S. Gaudriault, C. Medigue, A. Lanois, K. Powell, P. Siguier, R. Vincent, V. Wingate, M. Zouine, P. Glaser, N. Boemare, A. Danchin and F. Kunst (2003). "The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*." Nat Biotechnol **21**(11): 1307-13.

Duitman, E. H., L. W. Hamoen, M. Rembold, G. Venema, H. Seitz, W. Saenger, F. Bernhard, R. Reinhardt, M. Schmidt, C. Ullrich, T. Stein, F. Leenders and J. Vater (1999). "The mycosubtilin synthetase of *Bacillus subtilis* ATCC6633: A multifunctional hybrid between a peptide synthetase, an amino transferase, and a fatty acid synthase." PNAS **96**(23): 13294-99.

Eberhard, A., A. L. Burlingame, C. Eberhard, G. L. Kenyon, K. H. Nealson and N. J. Oppenheimer (1981). "Structural identification of autoinducer of *Photobacterium fischeri* luciferase." Biochemistry **20**(9): 2444-9.

Ehmann, D. E., A. M. Gehring and C. T. Walsh (1999). "Lysine biosynthesis in *Saccharomyces cerevisiae*: mechanism of alpha-aminoadipate reductase (Lys2) involves posttranslational phosphopantetheinylation by Lys5." Biochemistry **38**(19): 6171-7.

Ehmann, D. E., C. A. Shaw-Reid, H. C. Losey and C. T. Walsh (2000). "The EntF and EntE adenylation domains of Escherichia coli enterobactin synthetase: sequestration and selectivity in acyl-AMP transfers to thiolation domain cosubstrates." Proc Natl Acad Sci U S A **97**(6): 2509-14.

Elofsson, M., U. Splittgerber, J. Myung, R. Mohan and C. M. Crews (1999). "Towards subunit-specific proteasome inhibitors: synthesis and evaluation of peptide alpha',beta'-epoxyketones." Chem Biol **6**(11): 811-22.

Endo, A. and S. J. Danishefsky (2005). "Total synthesis of salinosporamide A." J Am Chem Soc **127**(23): 8298-9.

Feling, R. H., G. O. Buchanan, T. J. Mincer, C. A. Kauffman, P. R. Jensen and W. Fenical (2003). "Salinosporamide A: A highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinospora*." Angew Chem Int Ed Engl. **42**(3): 355-7.

Fenical, W. (1993). "Chemical studies of marine bacteria: Developing a new resource." Chem. Rev. **93**: 1673-83.

Fenical, W. and P. R. Jensen (1993). Marine microorganisms: A new biomedical resource. Marine Biotechnology, Volume 1: Pharmaceutical and Bioactive

Natural Products. D. H. Attaway and O. R. Zaborsky. New York, Plenum Press**:** 419-57.

Fenteany, G., R. Standaert, W. Lane, S. Choi, E. Corey and S. Schreiber (1995). "Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin." Science **268**(5211): 726-31.

Fenteany, G., R. F. Standaert, G. A. Reichard, E. J. Corey and S. L. Schreiber (1994). "A beta-lactone related to lactacystin induces neurite outgrowth in a neuroblastoma cell line and inhibits cell cycle progression in an osteosarcoma cell line." PNAS **91**(8): 3358-62.

Finking, R. and M. A. Marahiel (2004). "Biosynthesis of nonribosomal peptides." Annu Rev Microbiol **58**: 453-88.

Floss, H. G. (1997). "Natural products derived from unusual variants of the shikimate pathway." Nat Prod Rep **14**(5): 433-52.

Fruh, K., M. Gossen, K. Wang, H. Bujard, P. A. Peterson and Y. Yang (1994). "Displacement of housekeeping proteasome subunits by MHC-encoded LMPs: a newly discovered mechanism for modulating the multicatalytic proteinase complex." Embo J **13**(14): 3236-44.

Furuya, Y., S. Akimoto, K. Yasuda and H. Ito (1997). "Apoptosis of androgen-independent prostate cell line induced by inhibition of fatty acid synthesis." Anticancer Res **17**(6D): 4589-93.

Gandecha, A. R., S. L. Large and E. Cundliffe (1997). "Analysis of four tylosin biosynthetic genes from the tylLM region of the *Streptomyces fradiae* genome." Gene **184**(2): 197-203.

Garcia-Carbonero, R., J. I. Mayordomo, M. V. Tornamira, M. Lopez-Brea, A. Rueda, V. Guillem, A. Arcediano, A. Yubero, F. Ribera, C. Gomez, A. Tres, J. L. Perez-Gracia, C. Lumbreras, J. Hornedo, H. Cortes-Funes and L. Paz-Ares (2001). "Granulocyte colony-stimulating factor in the treatment of high-risk febrile neutropenia: a multicenter randomized trial." J Natl Cancer Inst **93**(1): 31-8.

Gehring, A. M., E. DeMoll, J. D. Fetherston, I. Mori, G. F. Mayhew, F. R. Blattner, C. T. Walsh and R. D. Perry (1998). "Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by *Yersinia pestis*." Chem Biol **5**(10): 573-86.

Gibson, M. I. and F. Gibson (1962). "A new intermediate in aromatic biosynthesis." Biochim Biophys Acta **65**: 160-3.

Gibson, M. I. and F. Gibson (1964). "Preliminary studies on the isolation and metabolism of an intermediate in aromatic biosynthesis: chorismic acid." Biochem J **90**(2): 248-56.

Glickman, M. H., D. M. Rubin, O. Coux, I. Wefes, G. Pfeifer, Z. Cjeka, W. Baumeister, V. A. Fried and D. Finley (1998). "A subcomplex of the proteasome regulatory particle required for ubiquitin-conjugate degradation and related to the COP9-signalosome and eIF3." Cell **94**(5): 615-23.

Gocht, M. and M. A. Marahiel (1994). "Analysis of core sequences in the D-Phe activating domain of the multifunctional peptide synthetase TycA by site-directed mutagenesis." J Bacteriol **176**(9): 2654-62.

Goodfellow, M. and J. A. Haynes (1984). Actinomycetes in marine sediments. Biological, Biochemical, and Biomedical Aspects of Actinomycetes. L. Ortiz-Ortiz, L. F. Bojalil and V. Yakoleff. Orlando, Academic Press**:** 453-72.

Grifo, F., D. J. Newman, A. Fairfield, B. Bhattacharya and J. Grupenhoff (1997). The Origins of Prescription Drugs. Washington D.C., Island Press.

Groettrup, M., R. Kraft, S. Kostka, S. Standera, R. Stohwasser and P. M. Kloetzel (1996). "A third interferon-gamma-induced subunit exchange in the 20S proteasome." Eur J Immunol **26**(4): 863-9.

Groll, M., L. Ditzel, J. Lowe, D. Stock, M. Bochtler and H. D. Bartunik (1997). "Structure of 20S proteasome from yeast at 2.4A resolution." Nature **386**(3): 463-71.

Groll, M., R. Huber and B. C. Potts (2006). "Crystal structures of salinosporamide A (NPI-0052) and B (NPI-0047) in complex with the 20S proteasome reveal important consequences of beta-lactone ring opening and a mechanism for irreversible binding." J Am Chem Soc **128**(15): 5136-41.

Groves, J. T. (2003). "The bioinorganic chemistry of iron in oxygenases and supramolecular assemblies." Proc Natl Acad Sci U S A **100**(7): 3569-74.

Grunewald, J. and M. A. Marahiel (2006). "Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides." Microbiol Mol Biol Rev **70**(1): 121-46.

Han, S. K., O. I. Nedashkovskaya, V. V. Mikhailov, S. B. Kim and K. S. Bae (2003). "*Salinibacterium amurskyense* gen. nov., sp. nov., a novel genus of the family Microbacteriaceae from the marine environment." Int J Syst Evol Microbiol **53**(Pt 6): 2061-6.

Hanada, M., K. Sugawara, K. Kaneta, S. Toda, Y. Nishiyama, K. Tomita, H. Yamamoto, M. Konishi and T. Oki (1992). "Epoxomicin, a new antitumor agent of microbial origin." J Antibiot (Tokyo) **45**(11): 1746-52.

Haslam, E. (1993). Shikimic Acid: Metabolism and Metabolites. Wiley. New York.

Hata, T., Y. Sano, R. Sugawara, A. Matsumae, K. Kanomori, T. Shima and T. Hoshi (1956). "Mitomycin, a new antibiotic from *Streptomyces*." J. Antibiot. **Ser. A**(9): 141–46.

He, J., N. Magarvey, M. Piraee and L. C. Vining (2001). "The gene cluster for chloramphenicol biosynthesis in *Streptomyces venezuelae* ISP5230 includes novel shikimate pathway homologues and a monomodular non-ribosomal peptide synthetase gene." Microbiology **147**(10): 2817-29.

Hecht, H. J., H. Sobek, T. Haag, O. Pfeifer and K. H. van Pee (1994). "The metal-ion-free oxidoreductase from *Streptomyces aureofaciens* has an alpha/beta hydrolase fold." Nat Struct Biol **1**(8): 532-7.

Helmke, E. and H. Weyland (1984). "*Rhodococcus marinonascens* sp. nov., an actinomycete from the sea." Int. J. Syst. Bacteriol. **34**: 127-38.

Helmke, E., H. Weyland and. (1984). "*Rhodococcus marinonascens* sp. nov., an actinomycete from the sea." Int J Syst Bacteriol **34**: 127–38.

Hochhut, B., Y. Lotfi, D. Mazel, S. M. Faruque, R. Woodgate and M. K. Waldor (2001). "Molecular analysis of antibiotic resistance gene clusters in vibrio cholerae O139 and O1 SXT constins." Antimicrob Agents Chemother **45**(11): 2991-3000.

Hochstrasser, M. (1995). "Ubiquitin, proteasomes, and the regulation of intracellular protein degradation." Curr Opin Cell Biol **7**(2): 215-23.

Hopwood, D. A. (1997). "Genetic contributions to understanding polyketide synthases." Chem Rev **97**: 465-97.

Huang, F., S. F. Haydock, D. Spiteller, T. Mironenko, T.-L. Li, D. O'Hagan, P. F. Leadlay and J. B. Spencer (2006). "The gene cluster for fluorometabolite biosynthesis in *Streptomyces cattleya:* A thioesterase confers resistance to fluoroacetyl-Coenzyme A." Chemistry & Biology **13**(5): 475-84.

Ikiel, I., I. C. P. Smith and D. Sprott (1984). "Biosynthesis of isoleucine in methanogenic bacteria: A 13C NMR study." Biochemistry **23**: 1683-87.

Ishikawa, J. and K. Hotta (1999). "FramePlot: a new implementation of the Frame analysis for predicting protein-coding regions in bacterial DNA with a high G+C content." FEMS Microbiol. Lett. **174**: 251-53.

Jensen, P. R., R. Dwight and W. Fenical (1991). "Distribution of actinomycetes in near-shore tropical marine sediments." Appl. Environ. Microbiol. **57**: 1102-08.

Jensen, P. R. and W. Fenical (1996). "Marine bacterial diversity as a resource for novel microbial products." J. Indust. Microbiol. **17**: 346-51.

Kakavas, S. J., L. Katz and D. Stassi (1997). "Identification and characterization of the niddamycin polyketide synthase genes from *Streptomyces caelestis*." J Bacteriol **179**(23): 7515-22.

Kalaitzis, J. A., M. Izumikawa, L. Xiang, C. Hertweck and B. S. Moore (2003). "Mutasynthesis of enterocin and wailupemycin analogues." J Am Chem Soc **125**(31): 9290-1.

Kato, T., H. Hinoo, Y. Terui, J. Kikuchi and J. Shoji (1988). "The structures of katanosins A and B." J Antibiot (Tokyo) **41**(6): 719-25.

Katz, L. (1997). "Manipulation of Modular Polyketide Synthases." Chem Rev **97**(7): 2557-76.

Keating, T. A. and C. T. Walsh (1999). "Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis." Curr Opin Chem Biol **3**(5): 598-606.

Kieser, T., M. Bibb, M. Buttner, K. Chater and D. A. Hopwood (2000). Practical Streptomyces Genetics. United Kingdom, John Innes Foundation.

Kim, B. S., T. A. Cropp, B. J. Beck, D. H. Sherman and K. A. Reynolds (2002). "Biochemical evidence for an editing role of thioesterase II in the biosynthesis of the polyketide pikromycin." J. Biol. Chem. **277**(50): 48028-34.

Kim, K. and C. M. Crews (2004). Natural product and synthetic proteasome inhibitors. Totowa, NJ, Humana.

Kino, T., H. Hatanaka, S. Miyata, N. Inamura, N. Nishiyama, T. Yajima, T. Goto, M. Okuhara, M. Kohasaka, H. Aoki and T. Ochiai (1987). "FK506, a novel immunosuppressant isolated from a *Streptomyces*." J. Antibiot **40**(1256-1265).

Kletzin, A. and M. W. Adams (1996). "Molecular and phylogenetic characterization of pyruvate and 2-ketoisovalerate ferredoxin oxidoreductases from *Pyrococcus furiosus* and pyruvate ferredoxin oxidoreductase from *Thermotoga maritima*." J. Bacteriol. **178**(1): 248-57.

Knight, V., J. J. Sanglier, D. DiTullio, S. Braccili, P. Bonner, J. Waters, D. Hughes and L. Zhang (2003). "Diversifying microbial natural products for drug discovery." Appl. Microbiol. Biotechnol. **62**: 446-58.

Kohli, R. M., C. T. Walsh and M. D. Burkart (2002). "Biomimetic synthesis and optimization of cyclic peptide antibiotics." Nature **418**(6898): 658-61.

Konz, D., A. Klens, K. Schorgendorfer and M. A. Marahiel (1997). "The bacitracin biosynthesis operon of Bacillus licheniformis ATCC 10716: molecular characterization of three multi-modular peptide synthetases." Chem Biol **4**(12): 927-37.

Konz, D. and M. A. Marahiel (1999). "How do peptide synthetase generate structural diversity?" Chem Biol **6**: R39-R48.

Kridel, S. J., F. Axelrod, N. Rozenkrantz and J. W. Smith (2004). "Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity." Cancer Res **64**(6): 2070-5.

Lall, M. S., C. Karvellas and J. C. Vederas (1999). "Beta-lactones as a new class of cysteine proteinase inhibitors: inhibition of hepatitis A virus 3C proteinase by N-Cbz-serine beta-lactone." Org Lett **1**(5): 803-6.

Lambalot, R. H., A. M. Gehring, R. S. Flugel, P. Zuber, M. LaCelle, M. A. Marahiel, R. Reid, C. Khosla and C. T. Walsh (1996). "A new enzyme superfamily - the phosphopantetheinyl transferases." Chem Biol **3**(11): 923-36.

Lee, W. W., A. Benitez, L. Goodman and B. R. Baker (1960). "Potential anticancer agents.1 XL. synthesis of the -anomer of 9-(Darabinofuranosyl)-adenine." J Am Chem Soc **82**(10): 2648-49.

Li, L., J. P. Bannantine, Q. Zhang, A. Amonsin, B. J. May, D. Alt, N. Banerji, S. Kanjilal and V. Kapur (2005). "The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis." Proc Natl Acad Sci U S A **102**(35): 12344-9.

Li, L., J. P. Bannantine, Q. Zhang, A. Amonsin, B. J. May, D. Alt, N. Banerji, S. Kanjilal and V. Kapur (2005). "The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis." PNAS **102**(35): 12344-49.

Liu, H. and K. A. Reynolds (1999). "Role of crotonyl coenzyme A reductase in determining the ratio of polyketides monensin A and monensin B produced by Streptomyces cinnamonensis." J Bacteriol **181**(21): 6806-13.

Liu, H. and K. A. Reynolds (2001). "Precursor supply for polyketide biosynthesis: the role of crotonyl-CoA reductase." Metab Eng **3**(1): 40-8.

Lombo, F., A. Velasco, A. Castro, F. de la Calle, A. F. Brana, J. M. Sanches-Puesses, C. Mendez and J. A. Salas (2005). Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in *Streptomyces* hosts. unpublished. EMBL, NCBI.

Lomovskaya, N., Y. Doi-Katayama, S. Filippini, C. Nastro, L. Fonstein, M. Gallo, A. L. Colombo and C. R. Hutchinson (1998). "The *Streptomyces peucetius* dpsY and dnrX genes govern early and late steps of daunorubicin and doxorubicin biosynthesis." J Bacteriol **180**(9): 2379-86.

Long, P. F., C. J. Wilkinson, C. P. Bisang, J. Cortes, N. Dunster, M. Oliynyk, E. McCormick, H. McArthur, C. Mendez, J. A. Salas, J. Staunton and P. F. Leadlay (2002). "Engineering specificity of starter unit selection by the erythromycin-producing polyketide synthase." Mol Microbiol **43**(5): 1215-25.

Mai, X. and M. W. Adams (1994). "Indolepyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*. A new enzyme involved in peptide fermentation." J. Biol. Chem. **269**(24): 16726-32.

Maldonado, L. A., W. Fenical, P. R. Jensen, C. A. Kauffman, T. J. Mincer, A. C. Ward, A. T. Bull and M. Goodfellow (2005). "Salinispora arenicola gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae." Int J Syst Evol Microbiol **55**(5): 1759-66.

Marahiel, M. A., T. Stachelhaus and H. D. Mootz (1997). "Modular peptide Synthetases involved in nonribosomal peptide synthesis." Chem Rev **97**(7): 2651-74.

Mavrodi, D. V., V. N. Ksenzenko, R. F. Bonsal, R. J. Cook, A. M. Boronin and L. S. Thomoshow (1998). "A seven-gene locus for synthesis of phenazine-1-carboxylic acid by *Pseudomonas*." J Bacteriol **180**(9): 2541-9.

May, J. J., N. Kessler, M. A. Marahiel and M. T. Stubbs (2002). "Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases." Proc Natl Acad Sci U S A **99**(19): 12120-5.

McDonald, M., D. V. Mavrodi, L. S. Thomashow and H. G. Floss (2001). "Phenazine biosynthesis in *Pseudomonas fluorescens*: branchpoint from the primary shikimate biosynthetic pathway and role of phenazine-1,6-dicarboxylic acid." J Am Chem Soc **123**(38): 9459-60.

Mincer, T. J., W. Fenical and P. R. Jensen (2005). "Culture-dependent and culture-independent diversity within the obligate marine actinomycete genus *Salinispora*." Appl Environ Microbiol **71**(11): 7019-28.

Mincer, T. J., P. R. Jensen, C. A. Kauffman and W. Fenical (2002). "Widespread and persisitent populations of a major new marine actinomycete taxon in ocean sediments." Appl Envi Micr **68**(10): 5005-11.

Moffitt, M. C. and B. A. Neilan (2003). "Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations." J Mol Evol **56**(4): 446-57.

Molnar, I., T. Schupp, M. Ono, R. Zirkle, M. Milnamow, B. Nowak-Thompson, N. Engel, C. Toupet, A. Stratmann, D. D. Cyr, J. Gorlach, J. M. Mayo, A. Hu, S. Goff, J. Schmid and J. M. Ligon (2000). "The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from *Sorangium cellulosum* So ce90." Chem Biol **7**(2): 97-109.

Moore, B., J. Kalaitzis and L. Xiang (2005). "Exploiting marine actinomycete biosynthetic pathways for drug discovery." Antonie van Leeuwenhoek **87**(1): 49-57.

Moore, B. S. and C. Hertweck (2002). "Biosynthesis and attachment of novel bacterial polyketide synthase starter units." Nat Prod Rep **19**(1): 70-99.

Moore, B. S., K. Poralla and H. G. Floss (1993). "Biosynthesis of the yclohexanecarboxylic acid starter unit of ω-cyclohexyl fatty acids in Alicyclobacillus acidocaldarius." J Am Chem Soc **115**: 5267-74.

Mootz, H. D. and M. A. Marahiel (1997). "The tyrocidine biosynthesis operon of Bacillus brevis: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains." J Bacteriol **179**(21): 6843-50.

Mootz, H. D., D. Schwarzer and M. A. Marahiel (2002). "Ways of assembling complex natural products on modular nonribosomal peptide synthetases." Chembiochem **3**(6): 490-504.

Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut and N. Ward (2004). "Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment." Nature **432**(7019): 910.

Moran, M. A., L. T. Rutherford and R. E. Hodson (1995). "Evidence for indigenous Streptomyces populations in a marine environment determined with a 16S rRNA probe." Appl Environ Microbiol **61**(10): 3695-700.

Nakagawa, A., S. Takahashi and K. Uchida (1994). "Biosynthesis of lactacystin. Origin of the carbons and stereospecific NMR assignment of the two diastereotopic methyl groups." Tetrahedron Letters **35**(28): 5009-12.

Nakamura, Y., T. Gojobori and T. Ikemura (2000). "Codon usage tabulated from the international DNA sequence databases: status for the year 2000." Nucl. Acids Res. **292**.

Newman, D. J. and G. M. Cragg (2006). "Natural products from marine invertebrates and microbes as modulators of anitumor targets." Current Drug Targets **7**: 279-304.

Newman, D. J., G. M. Cragg and K. M. Snader (2000). "The influence of natural products upon drug discovery." Nat. Prod. Rep. **17**: 215-34.

Ni, X. and L. P. Hager (1999). "Expression of *Batis maritima* methyl chloride transferase in *Escherichia coli*." Proc Natl Acad Sci U S A **96**(7): 3611-5.

O'Hagan, D., C. Schaffrath, S. Cobb, J. Hamilton and C. Murphy (2002). "Biochemistry: biosynthesis of an organofluorine molecule." nature **21**(416): 279.

Omura, S. (1984). Production, structure, and biological properties of macrolide-like antibiotics. New York, Academic Press.

Omura, S., H. Ikeda, J. Ishikawa, A. Hanamoto, C. Takahashi, M. Shinose, Y. Takahashi, H. Horikawa, H. Nakazawa, T. Osonoe, H. Kikuchi, T. Shiba, Y. Sakaki and M. Hattori (2001). "Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites." Proc Natl Acad Sci U S A **98**(21): 12215-20.

Omura, S., K. Matsuzaki, F. T., Kosuge K, F. T., F. S. and A. Nakagawa (1991). "Structure of lactacystin, a new microbial metabolite which induces differentiation of neuroblastoma cells." J Antibiot (Tokyo) **44**(1): 117-8.

Piel, J., D. Butzke, N. Fusetani, D. Hui, M. Platzer, G. Wen and S. Matsunaga (2005). "Exploring the chemistry of uncultivated bacterial symbionts: antitumor polyketides of the pederin family." J Nat Prod **68**(3): 472-9.

Piel, J., I. Hofer and D. Hui (2004). "Evidence for a symbiosis island involved in horizontal acquisition of pederin biosynthetic capabilities by the bacterial symbiont of *Paederus fuscipes* beetles." J Bacteriol **186**(5): 1280-6.

Piel, J., D. Hui, G. Wen, D. Butzke, M. Platzer, N. Fusetani and S. Matsunaga (2004). "Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*." Proc Natl Acad Sci U S A **101**(46): 16222-7.

Pospiech, A., B. Cluzel, J. Bietenhader and T. Schupp (1995). "A new *Myxococcus xanthus* gene cluster for the biosynthesis of the antibiotic saframycin Mx1 encoding a peptide synthetase." Microbiology **141**(8): 1793-803.

Rainey, F. A., S. Klatte, R. M. Kroppenstedt and E. Stackebrandt (1995). "*Dietzia*, a new genus including *Dietzia maris* comb. nov., formerly *Rhodococcus maris*." Int J Syst Bacteriol **45**(1): 32-6.

Rajagopalan, J. S., L.-C. Chen and E. K. Jaffe (1992). "Preparation, ^{13}C NMR characterization, and enzymatic transformation of [U-^{13}C]chorismate to [U-^{13}C]prephenate and [U-^{13}C]hydroxyphenyl pyruvate." Bioorganic Chemistry **20**(2): 115-23.

Rangaswamy, V., R. Mitche, M. Ullrich and C. Bender (1998). "Analysis of genes involved in biosynthesis of coronafacic acid, the polyketide component of the phytotoxin coronatine." J Bacteriol **180**: 3330-38.

Rao, K. V., N. Kasanah, S. Wahyuono, B. L. Tekwani, R. F. Schinazi and M. T. Hamann (2004). "Three new manzamine alkaloids from a common Indonesian sponge and their activity against infectious and tropical parasitic diseases." J Nat Prod **67**(8): 1314-8.

Rawlings, B. J. (2001). "Type I polyketide biosynthesis in bacteria (Part A - erythromycin biosynthesis)." Nat Prod Rep **18**(2): 190-227.

Reddy, L. R., J. F. Fournier, B. V. Reddy and E. J. Corey (2005). "New synthetic route for the enantioselective total synthesis of salinosporamide A and biologically active analogues." Org Lett **7**(13): 2699-701.

Reddy, L. R., J. F. Fournier, B. V. Subba Reddy and E. J. Corey (2005). "An efficient, stereocontrolled synthesis of a potent omuralide-salinosporin hybrid for selective proteasome inhibition." J Am Chem Soc **127**(25): 8974-6.

Reddy, L. R., P. Saravanan and E. J. Corey (2004). "A simple stereocontrolled synthesis of salinosporamide A." J Am Chem Soc **126**(20): 6230-1.

Redenbach, M., H. M. Kieser, D. Denapaite, A. Eichner, J. Cullum, H. Kinashi and D. A. Hopwood (1996). "A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome." Mol Microbiol **21**(1): 77-96.

Reimmann, C., H. M. Patel, L. Serino, M. Barone, C. T. Walsh and D. Haas (2001). "Essential PchG-dependent reduction in pyochelin biosynthesis of *Pseudomonas aeruginosa*." J Bacteriol **183**(3): 813-20.

Reynolds, K. A., D. O'Hagan, D. Gani and J. A. Robinson (1988). "Butyrate metabolism in streptomycetes. Characterization of a vicinal interchange rearrangement linking isobutyrate and butyrate in Streptomyces cinnamonensis." J. Chem Soc. Perkin Trans. 1: 3195-207.

Rieger, C. E. and J. L. Turnbull (1996). "Small scale biosynthesis and purification of gram quantities of chorismic acid." Prep Biochem Biotechnol **26**(1): 67-76.

Robinson, I. M. and M. J. Allison (1969). "Isoleucine biosynthesis from 2-methylbutyric acid by anaerobic bacteria from the rumen." J Bacteriol **97**(3): 1220-6.

Sambrook, J., E. Fritsch and T. Maniatis (1989). Molecular vloning: a laboratory manual, 2nd edition. Cold Springs Harbor, Cold Springs Harbor Laboratory.

Saxena, D., S. Aouad, J. Attieh and H. S. Saini (1998). "Biochemical characterization of chloromethane emission from the wood-rotting fungus phellinus pomaceus." Appl Environ Microbiol **64**(8): 2831-5.

Schaffrath, C., S. L. Cobb and D. O'Hagan (2002). "Cell-free biosynthesis of fluoroacetate and 4-fluorothreonine in *Streptomyces cattleya*." Angew Chem Int Ed Engl **41**(20): 3913-5.

Schiwara, H., W. Domschk and G. Domagk (1968). "Sugar dehydrogenases in mammalian liver. Differentiation of various sugar dehydrogenases from pig liver by disc electrophoresis and ion exchange chromatography." Hoppe-Syler's Z. Physiol. Chem. **349**(11): 1574-81.

Schlumbohm, W., T. Stein, C. Ullrich, J. Vater, M. Krause, M. A. Marahiel, V. Kruft and B. Wittmann-Liebold (1991). "An active serine is involved in covalent substrate amino acid binding at each reaction center of gramicidin S synthetase." J Biol Chem **266**(34): 23135-41.

Schuhr, C., W. Eisenreich, M. Goese, P. Stohler, W. W, E. Kupfer and A. Bacher (2002). " Biosynthetic precursors of the lipase inhibitor lipstatin." J Org Chem **67**(7): 2257-62.

Schupp, T., C. Toupet, B. Cluzel, S. Neff, S. Hill, J. J. Beck and J. M. Ligon (1995). "A Sorangium cellulosum (myxobacterium) gene cluster for the biosynthesis of the macrolide antibiotic soraphen A: cloning, characterization, and homology to polyketide synthase genes from actinomycetes." J Bacteriol **177**(13): 3673-9.

Schwarzer, D. and M. A. Marahiel (2001). "Multimodular biocatalysts for natural product assembly." Naturwissenschaften **88**(3): 93-101.

Schwarzer, D., H. D. Mootz, U. Linne and M. A. Marahiel (2002). "Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases." PNAS **99**(22): 14083-88.

Sekowska, A., V. Denervaud, H. Ashida, K. Michoud, D. Haas, A. Yokota and A. Danchin (2004). "Bacterial variations on the methionine salvage pathway." BMC Microbio. **4**(9).

Sheldrick, G. M., P. G. Jones, O. Kennard, D. H. Williams and G. A. Smith (1978). "Structure of vancomycin and its complex with acetyl-D-alanyl-D-alanine." Nature **271**(5642): 223-5.

Shen, B. (2000). "The biosynthesis of aromatic polyketides." Topics in Current Chemistry **209**: 1-51.

Shimada, K., D. J. Hook, G. F. Warner and H. G. Floss (1978). "Biosynthesis of the antibiotic 2,5-dihydrophenylalanine by *Streptomyces arenae*." Biochemistry **17**(15): 3054-8.

Shine, J. and L. Dalgarno (1975). "Determinant of cistron specificity in bacterial ribosomes." Nature **254**(5495): 34-8.

Silakowski, B., B. Kunze and R. Muller (2001). "Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*." Gene **275**(2): 233-40.

Silakowski, B., G. Nordsiek, B. Kunze, H. Blocker and R. Muller (2001). "Novel features in a combined polyketide synthase/non-ribosomal peptide synthetase: the myxalamid biosynthetic gene cluster of the myxobacterium *Stigmatella aurantiaca* Sga15." Chem Biol **8**(1): 59-69.

Silakowski, B., H. U. Schairer, H. Ehret, B. Kunze, S. Weinig, G. Nordsiek, P. Brandt, H. Blocker, G. Hofle, S. Beyer and R. Muller (1999). "New Lessons for Combinatorial Biosynthesis from Myxobacteria. The myxothiazol biosynthetic gene cluster of *Stigmatella aurantiaca* DW4/3-1." J. Biol. Chem. **274**(52): 37391-99.

Smith, N., M. Mayhew, H. Robinson, A. Heroux, D. Charlton, M. J. Holden and D. T. Gallagher (2003). "Crystallization and phasing of alanine dehydrogenase from *Archaeoglobus fulgidus*." Acta Crystallographica Section D **59**(12): 2328-31.

Sosio, M., E. Bossi, A. Bianchi and S. Donadio (2000). "Multiple peptide synthetase gene cluster in Actinomycetes." Mol Gen Genet **264**(3): 213-21.

Stach, J. E., L. A. Maldonado, A. C. Ward, A. T. Bull and M. Goodfellow (2004). "*Williamsia maris* sp. nov., a novel actinomycete isolated from the Sea of Japan." Int J Syst Evol Microbiol **54**(Pt 1): 191-4.

Stachelhaus, T., A. Huser and M. A. Marahiel (1996). "Biochemical characterization of peptidyl carrier protein (PCP), the thiolation domain of multifunctional peptide synthetases." Chem Biol **3**(11): 913-21.

Stachelhaus, T. and M. A. Marahiel (1995). "Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA." J Biol Chem **270**(11): 6163-9.

Stachelhaus, T., H. D. Mootz, V. Bergendahl and M. A. Marahiel (1998). "Peptide bond formation in nonribosomal peptide biosynthesis. Catalytic role of the condensation domain." J Biol Chem **273**(35): 22773-81.

Stachelhaus, T., H. D. Mootz and M. A. Marahiel (1999). "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases." Chem Biol **6**(8): 493-505.

Stachelhaus, T., A. Schneider and M. A. Marahiel (1995). "Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains." Science **269**(5220): 69-72.

Stassi, D. L., S. J. Kakavas, K. A. Reynolds, G. Gunawardana, S. Swanson, D. Zeidner, M. Jackson, H. Liu, A. Buko and L. Katz (1998). "Ethyl-substituted erythromycin derivatives produced by directed metabolic engineering." Proc Natl Acad Sci U S A **95**(13): 7305-9.

Staunton, J. and B. Wilkinson (1998). "The biosynthesis of aliphatic polyketides." Topics in Current Chemistry **195**: 49-92.

Steffensky, M., A. Muhlenweg, Z. X. Wang, S. M. Li and L. Heide (2000). "Identification of the novobiocin biosynthetic gene cluster of *Streptomyces spheroides* NCIB 11891." Antimicrob Agents Chemother **44**(5): 1214-22.

Stegmann, E., C. Rausch, S. Stockert, D. Burkert and W. Wohlleben (2006). "The balhimycin biosynthetic gene cluster of the glycopeptide producer *Amycolatopsis balhimycina* includes a gene (orf1) with unknown function." FEMS Microbiol Lett **262**(1): 85-92.

Strohl, W. (1992). "Compilation and analysis of DNA sequences associated with apparent streptomycete promoters." Nucl. Acids Res. **20**(5): 961-74.

Suckling, C. J. (1991). "Chemical approaches to the discovery of new drugs." Sci Prog **75**(298 Pt 3-4): 323-59.

Sugio, S., G. A. Petsko, J. M. Manning, K. Soda and D. Ringe (1995). "Crystal structure of a D-amino acid aminotransferase: how the protein controls stereoselectivity." Biochemistry **34**(30): 9661-9.

Takahashi, S., K. Uchida, A. Nakagawa, Y. Miyake, M. Kainosho, K. Matsuzaki and S. Omura (1995). "Biosynthesis of lactacystin." J Antibiot (Tokyo) **48**(9): 1015-20.

Tamura, T., I. Nagy, A. Lupas, F. Lottspeich, Z. Cejka, G. Schoofs, K. Tanaka, R. De Mot and W. Baumeister (1995). "The first characterization of a eubacterial proteasome: the 20S complex of *Rhodococcus*." Curr Biol **5**(7): 766-74.

Tang, G. and S. H. Leppla (1999). "Proteasome activity is required for anthrax lethal toxin to kill macrophages." Infect. Immun. **67**: 3055-60.

Teoh, C. Y. and K. J. Davies (2004). "Potential roles of protein oxidation and the immunoproteasome in MHC class I antigen presentation: the 'PrOxI' hypothesis." Arch Biochem Biophys **423**(1): 88-96.

Tillett, D., M. Dittmann, M. Erhard, H. von Doren, t. Burner and A. Neilan (2000). "Structural organization of microcytin biosynthesis in *M. aeruginosa* PCC7806: An integrated peptide-polyketide synthetase system." Chem Biol **7**): 753-64.

Turgay, K. and M. A. Marahiel (1994). "A general approach for identifying and cloning peptide synthetase genes." Pept Res **7**(5): 238-41.

Umezawa, H., T. Aoyagi, K. Uotani, M. Hamada, T. Takeuchi and S. Takahashi (1980). "Ebelactone, an inhibitor of esterase, produced by actinomycetes." J Antibiot (Tokyo) **33**(12): 1594-6.

Uotani, K., H. Naganawa, A. T and H. Umezawa (1982). "Biosynthetic studies of ebelactone A and B by ^{13}C NMR spectrometry." J Antibiot (Tokyo) **35**(12): 1670-4.

Vaillancourt, F. H., E. Yeh, D. A. Vosburg, S. E. O'Connor and C. T. Walsh (2005). "Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis." Nature **436**(7054): 1191-4.

van Pee, K. H. (2001). "Microbial biosynthesis of halometabolites." Arch Microbiol **175**(4): 250-8.

Velasco, J. A., J. Cansado, M. C. Pena, T. Kawakami, J. Laborda and V. Notario (1993). "Cloning of the dihydroxyacid dehydratase-encoding gene (ILV3) from Saccharomyces cerevisiae." Gene **137**(2): 179-85.

Vinitsky, A., C. Cardozo, L. Sepp-Lorenzino, C. Michaud and M. Orlowski (1994). "Inhibition of the proteolytic activity of the multicatalytic proteinase complex (proteasome) by substrate-related peptidyl aldehydes." J Biol Chem **269**(47): 29860-6.

Voorhees, P. M. and R. Z. Orlowski (2006). "The proteasome and proteasome inhibitors in cancer therapy." Annu Rev Pharmacol Toxicol **46**: 189-213.

Wallace, K. K., Z. Y. Bao, H. Dai, R. Digate, G. Schuler, M. K. Speedie and K. A. Reynolds (1995). "Purification of crotonyl-CoA reductase from *Streptomyces collinus* and cloning, sequencing and expression of the corresponding gene in *Escherichia coli*." Eur J Biochem **233**(3): 954-62.

Walsh, C. T., H. Chen, T. A. Keating, B. Hubbard and H. C. Losey (2001). "Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines." Curr Opin Chem Biol **5**: 525-34.

Walsh, C. T., A. M. Gehring, P. H. Weinreb, L. E. Quadri and R. S. Flugel (1997). "Post-translational modification of polyketide and nonribosomal peptide synthases." Curr Opin Chem Biol **1**(3): 309-15.

Weibel, E. K., P. Hadvary, E. Hochuli, E. Kupfer and H. Lengsfeld (1987). "Lipstatin, an inhibitor of pancreatic lipase, produced by *Streptomyces toxytricini*. I. Producing organism, fermentation, isolation and biological activity." J Antibiot (Tokyo) **40**(8): 1081-5.

WHO. (2003, May 2003). "Traditional Medicine." 2006.

Wiesmann, K. E., J. Cortes, M. J. Brown, A. L. Cutter, J. Staunton and P. F. Leadlay (1995). "Polyketide synthesis in vitro on a modular polyketide synthase." Chem Biol **2**(9): 583-9.

Wilk, S. and M. Orlowski (1983). "Evidence that pituitary cation-sensitive neutral endopeptidase is a multicatalytic protease complex." J Neurochem **40**(3): 842-9.

Wu, K., L. Chung, W. P. Revill, L. Katz and C. D. Reeves (2000). "The FK520 gene cluster of *Streptomyces hygroscopicus* var. ascomyceticus (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units." Gene **251**(1): 81.

Xiang, L., J. A. Kalaitzis, G. Nilsen, L. Chen and B. S. Moore (2002). "Mutational analysis of the enterocin favorskii biosynthetic rearrangement." Org Lett **4**(6): 957-60.

Xie, G., M. Uttamchandani, G. Y. Chen, X. Bu, S. S. Lin, K. M. Wong, W. Yan, S. Q. Yao and Z. Guo (2002). "Substrate spectrum of tyrocidine thioesterase probed with randomized peptide N-acetylcysteamine thioesters." Bioorg Med Chem Lett **12**(6): 989-92.

Zerbe-Burkhardt, K., A. Ratnatilleke, N. Philippon, A. Birch, A. Leiser, J. W. Vrijbloed, D. Hess, P. Hunziker and J. A. Robinson (1998). "Cloning, sequencing, expression, and insertional inactivation of the gene for the large subunit of the coenzyme B_{12}-dependent isobutyryl-CoA mutase from *Streptomyces cinnamonensis*." J Biol Chem **273**(11): 6508-17.

Zubrod, C. G. (1984). "Origins and development of chemotherapy research at the National Cancer Institute." Cancer Treat Rep **68**(1): 9-19.