Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Celular e Molecular

Rafael Ricardo de Castro Cuadrat

Exploração da diversidade de policetídeo sintases (PKSs) ambientais

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Ciências, com área de concentração em Biologia Celular e Molecular

Orientador (es): Prof. Dr. Alberto Martin Rivera Dávila

RIO DE JANEIRO, 2010

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

C961

Cuadrat, Rafael Ricardo de Castro.

Exploração da diversidade de policetídeo sintases (PKSs) ambientais / Rafael Ricardo de Castro Cuadrat. – Rio de Janeiro, 2010.
xiv, 83 f. : il. ; 30 cm.

Dissertação (mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Celular e Molecular, 2010.
Bibliografia: f. 72-75

1. Metagenômica. 2. Policetídeo Sintases. 3. PKS. 4. Arraial do Cabo. I. Título.

CDD 578.7

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Celular e Molecular

Rafael Ricardo de Castro Cuadrat

Exploração da diversidade de policetídeo sintases (PKSs) ambientais

ORIENTADOR: Prof. Dr. Alberto Martin Rivera Dávila

Aprovada em: 08/04/2010

EXAMINADORES:

Prof. Dr. Antônio Basílio de Miranda - Presidente
Prof. Dr. André Nóbrega Pitaluga
Profa. Dra. Valéria Maia de Oliveria
Prof. Dr. Juliano de Carvalho Cury
Prof. Dr. Fabio Faria da Mota

Rio de Janeiro, 08 de Abril de 2010

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Exploração da diversidade de PKSs ambientais

RESUMO

DISSERTAÇÃO DE MESTRADO

Rafael Ricardo de Castro Cuadrat

O uso abusivo de antibióticos nos últimos tempos tem causado o surgimento de cepas multi-resistentes, inclusive em relação às moléculas de última geração, como por exemplo as cepas de *Staphylococcus aureus* resistentes à Vancomicina. Isto torna importante a busca por novas moléculas com ação antimicrobiana. A indústria farmacêutica, durante décadas, tem trabalhado com a modificação química dos produtos naturais (produzidos a partir de microorganismos cultivados) na tentativa de aumentar a potência destes compostos, para torná-los eficazes contra as cepas resistentes aos atuais fármacos. Porém, a descoberta de novos compostos naturais se mostra mais eficiente e menos onerosa do que a modificação de compostos já conhecidos. Entretanto, estudos vêm demonstrando que somente uma pequena porcentagem (1%-10%) dos microrganismos presentes na natureza pode ser isolada e mantida em meios de cultura artificiais, fazendo com que estes isolados e suas respectivas moléculas sejam sempre "redescobertos", limitando o desenvolvimento de novos fármacos. Contudo, abordagens moleculares como a metagenômica e ferramentas de bioinformática têm sido utilizadas combinadamente para o acesso direto ao DNA dos microrganismos não cultiváveis. Através da clonagem de fragmentos de DNA ambiental e posterior triagem por hibridização, PCR e sequenciamento, podemos obter informações referentes a genes que codificam moléculas de interesse biotecnológico e farmacológico, como por exemplo: lipases, esterases, celulases, quitinases, policetídeo sintases, etc. As famílias de genes que codificam as enzimas PKSs (polyketides synthases) e halogenases flanqueadoras são consideradas de interesse biotecnológico pois são produzidas no metabolismo secundário de diversos organismos e são fundamentais para a síntese de compostos antimicrobianos e antitumor. Visando a identificação e análise da variabilidade das PKSs, é interessante que se utilize amostras de DNA de ambientes com alta diversidade genética, como é o caso dos ambientes marinhos costeiros. Trabalhos preliminares realizados por nosso grupo mostram considerável diversidade em águas superficiais marinhas, composta por fungos, dinoflagelados, cianobactérias e actinobactérias não cultivados. O objetivo deste trabalho é analisar a diversidade de PKSs em ambientes marinhos, realizando inferências sobre evolução e filogenia destas enzimas. Foi possível sequenciar 5 novas regiões KS de PKS tipo I iterativa e modular, além de constatar uma grande diversidade de PKS nos bancos de dados ambientais estudados.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Exploring the diversity of environmental PKSs

ABSTRACT

Rafael Ricardo de Castro Cuadrat

Overuse of antibiotics in recent times has caused the emergence multi-resistents strain, including some molecules of last generation, such as strains of *Staphylococcus aureus* resistant to Vancomycin. This makes it important to search for new molecules with anti-microbial activity. The pharmaceutical industry, for decades, has worked with the chemical modification of natural products (produced from cultivated microorganisms) in an attempt to increase the potency of these compounds to make them effective against strains resistant to current drugs. However, the discovery of new natural compounds is more efficient and less costly than the modification of compounds known. However, research has shown that only a small percentage (1% -10%) of microorganisms in nature can be isolated and kept in artificial culture medium, making these isolates and their molecules are always "rediscovered" by limiting the development of new drugs. However, molecular approaches such as metagenomic and bioinformatics tools have been used in combination for direct access to the DNA of non-cultivable microorganisms. By cloning DNA fragments and subsequent environmental screening by hybridization, PCR and sequencing, we can obtain information about the genes that encode molecules of pharmacological and biotechnological interest, for example, lipases, esterases, cellulases, chitinases, polyketide synthases, etc. The families of genes that encode enzymes PKSs (polyketides synthases) and flanking halogenases are considered of biotechnological interest because they are produced in the secondary metabolism of different organisms and are essential for the synthesis of antimicrobial compounds and antitumor. For the identification and analysis of variability of PKSs, it is interesting to use DNA samples from environments with high genetic diversity, as is the case of coastal marine environments. Preliminary work performed by our group show considerable diversity in marine surface waters, composed of fungi, dinoflagellates, cyanobacteria and uncultured actinobacteria. The objective of this study is to analyze the diversity of PKSs in marine environments, making inferences about evolution and phylogeny of these enzymes. It was possible to sequence 5 new regions of KS type I iterative and modular, and see a great diversity of PKSs in environmental databases studied.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACP – proteína carreadora de grupamento acil
AT - aciltransferas
ATP – adenosina trifosfato
BAC – cromossomo artificial de bactéria
CAMERA - Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis
CYC – claisen ciclase
CON – domínio de condensação
DH - deidratase
DMSO - dimetilsulfóxido
DNA – ácido desoxiribonucleico
dNTP – deoxinucleotideo trifosfato
ER – enoil redutase
FAS – ácido graxo sintase
Gfp – proteína verde fluorescente
KR – cetoredutase
KS - cetosintase
LB - meio luria-bertani
MT – metal transferase
NCBI – Centro Nacional para a Informação Biotecnológica
NR – não redundante
PCR – reação em cadeia da polimerase
pH – potencial hidrolítico
PKS – policetídeo sintase
rRNA – ácido ribonucleico ribossomal
SDS - dodecil sulfato de sódio
TE – tampão EDTA; tioesterase

LISTA DE FIGURAS

LISTA DE TABELAS

MATERIAL SUPLEMENTAR

S1 – Tabela com os 3 melhores hits entre as sequências KS tipo I iterativa e o banco ambiental do NCBI.

S2 – Tabela com os 10 melhores hits entre o modelo de KS tipo II e o banco ambiental do NCBI.

S3 – Tabela com os 3 melhores hits de cada sequência KS tipo I iterativa e o banco do CAMERA

S4 – Tabela com os 10 melhores hits entre o modelo de KS tipo II e o banco do CAMERA

Agradecimentos

Agradeço primeiramente aos meus pais e irmão, pois sem o apoio deles este trabalho não seria realizado. Agradeço a todos que de alguma forma participaram deste trabalho, a todos do LBCS e do LBMPV pela paciência e pela ajuda nos momentos em que mais precisei, principalmente Ana, Joana (muitíssimo obrigado por todo o suporte na bancada, pelas mini-preps, etc), Diogo, Adriana, Kary, André, Marina, Silvana, Erick, Tempone, Tatiana e João (desculpe se esqueci alguém, mas ando com memória de peixe). À Dra. Yara, por ceder o espaço e material de seu laboratório para a realização dos experimentos em bancada. Agradeço ao meu orientador Alberto Dávila pela confiança depositada para realização de um projeto dentro de uma linha de pesquisa nova em nosso grupo. Também não posso deixar de agradecer ao Juliano pela ajuda na bancada e pela revisão criteriosa do trabalho. Aos avaliadores, por terem aceitado o convite e pelo tempo dedicado à leitura deste trabalho. Agradeço muito a minha namorada Carol, pelo apoio incondicional em todos os momentos finais da realização deste trabalho, pelas ideias e sugestões tão valiosas. Aos amigos da Toca do Coelho, sempre presentes nos momentos de descontração. A Deus e todos os amigos espirituais que me auxiliam desde o começo e tenho certeza que até o fim. Muito obrigado.

1 - Introdução

1.1 – Biodiversidade marinha e a nova indústria da biotecnologia

Estudos demonstram que em ambientes marinhos existem aproximadamente 3,67x 10^{30} células microbianas (Whitman et al. 1998). Estima-se que a abundância de bactérias seja de até 10^{6} células por mililitro de água na zona pelágica marinha, representando a maior parte da biomassa oceânica (Azam et al. 1998). Esta gigantesca biodiversidade possui grande potencial biotecnológico, pois seu estudo permite a descoberta de novas enzimas de interesse para a indústria.

Os ambientes marinhos são extremamente diversos e os microorganismos que os habitam são expostos a extremos de pressão, temperatura, salinidade e disponibilidade de nutrientes. Os diferentes nichos marinhos possuem comunidades bacterianas únicas e muito distintas, adaptadas às mais diferentes situações. Isto leva a uma grande diversidade bioquímica, ainda pouco explorada (Kennedy et al. 2008).

Por este motivo, diversos estudos vêm sendo realizados para explorar estes ambientes de maneira mais eficiente e novas técnicas vêm surgindo para tornar possível o acesso a estes microorganismos, com o objetivo não apenas de prover informações sobre a sua importância central na cadeia alimentar ou para a reciclagem biogeoquímica nos ecossistemas marinhos, mas também para entender sua capacidade de produzir novas enzimas e metabólitos com possível aplicação biotecnológica (Azam et al. 1998).

Porém, assim como nos ambientes terrestres, não é possível cultivar a maioria dos microorganismos presentes nos ambientes marinhos. Estima-se que apenas cerca de 0,001 a 0,1% dos microorganismos presentes nos oceanos seja cultivável em laboratório (Amman et al. 1995). Para tentar superar esta limitação, foram desenvolvidas técnicas de estudo dos ácidos nucléicos independentes de cultivo dos organismos, como o estudo de biodiversidade baseado no gene codificador do rRNA da menor subunidade ribossomal (16S nos procariotos e 18S nos eucariotos) e a construção e sequenciamento de grandes bibliotecas de DNA ambiental.

No mundo inteiro, diversos estudos vêm utilizando estas novas técnicas para explorar este gigantesco potencial, como, por exemplo, o projeto Sargasso Sea (Mar de Sargasso) (Venter et al. 2004) e o gigantesco estudo ao redor do mundo, Global Ocean

Sampler Expedition (Expedição Oceânica Global) (Yooseph et al. 2007).

Apesar destas iniciativas internacionais, no Brasil poucos estudos têm enfocado a gigantesca massa oceânica do litoral. O Brasil possui 7.400 km de litoral, com grande diversidade geomorfológica, com uma grande variedade de estuários, recifes de corais, praias arenosas e rochosas, ilhas oceânicas e lagos de água salgada. A maior parte das águas litorâneas fica situada na região tropical, abrigando uma grande diversidade de espécies pobres em produção (poucas células de cada organismo) se comparadas com espécies de regiões frias, onde a diversidade de espécies é menor, mas com maior produtividade (Couto et al. 2003).

Nosso grupo vem realizando estudos de biodiversidade envolvendo o sequenciamento de fragmentos de genes de rRNA (16S e 18S) no litoral do estado do Rio de Janeiro, no município de Arraial do Cabo, situado na Região dos Lagos. Resultados preliminares vêm demonstrando grande biodiversidade nesta região, tornando-a um bom alvo para estudos mais profundos à procura de novos genes, enzimas e vias biossintéticas.

1.2 - Metagenômica

A metagenômica é uma abordagem que surgiu na década de 90, como forma de estudar os ácidos nucléicos de organismos não cultiváveis. Diversas estratégias vêm sendo desenvolvidas, com diferentes objetivos, como, por exemplo, o estudo da biodiversidade de um determinado ambiente, ou a localização de genes ou clusters metabólicos responsáveis pela síntese de compostos de interesse biotecnológico (figura 1.1).

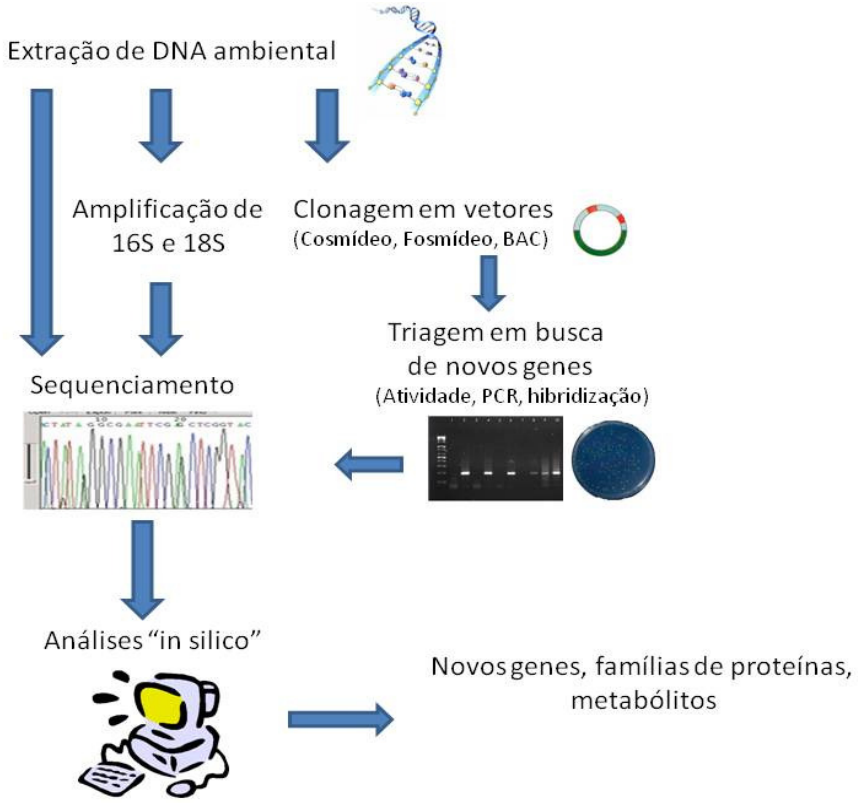


Figura 1.1: Fluxograma ilustrando algumas estratégias utilizadas em metagenômica.

As estratégias que mais vêm sendo utilizadas são as seguintes:

A) Estudo dos genes rRNA:

As análises de estrutura de comunidades microbianas utilizando as informações da sequência do gene que codifica a menor subunidade do RNA ribossômico (rRNA 16S para os procariotos e rRNA 18S para os eucariotos) tm se tornado cada vez mais difundidas. Este gene possui características fundamentais que possibilitam sua utilização em estudos de ecologia, tais como: presença de regiões com sequência de nucleotídeos hipervariáveis entre regiões conservadas; presença em todos os microrganismos; e tamanho considerado satisfatório, de cerca de 1500 nucleotídeos, para estudos filogenéticos (Amann & Ludwig, 2000). Nos últimos anos, o sequenciamento do rDNA 16S de procariotos e rDNA 18S de microeucariotos tem sido muito utilizado em estudos de diversidade, caracterização de comunidades complexas e taxonomia, levando à existência de um considerável volume de informações em bancos de dados públicos (Chelius & Triplett, 2001; Derakshani et al., 2001).

B) Clonagem de DNA ambiental em bibliotecas:

Para tornar possível o acesso ao DNA total dos ambientes estudados, a estratégia mais utilizada vem sendo a clonagem de pequenos ou grandes fragmentos em bibliotecas. O primeiro grande desafio a ser vencido para tornar possível a clonagem é isolar o DNA ambiental, em quantidade suficiente e com pureza adequada para que seja possível realizar as reações enzimáticas necessárias. Métodos físicos para lise dos microorganismos (como congelamento com nitrogênio líquido e sonicação) se mostram mais eficientes para a lise de todos os tipos de células, resultando em menor perda de biodiversidade do que os métodos químicos e enzimáticos. Porém, os métodos de lise física acabam por fragmentar o DNA da célula, gerando fragmentos de DNA com cerca de 10 kb (kilobases), tornando apenas possível a clonagem em pequenos vetores como plasmídeos. Já os métodos de lise química e enzimática que envolvem apenas utilização, por exemplo, de lisozima, proteinase K e SDS (dodecil sulfato de sódio), tornam possível a obtenção de fragmentos de 40 a 200 kilobases, passíveis de clonagem em vetores maiores como cosmídeos, fosmídeos e cromossomo artificial de bactérias

(BAC). A vantagem de se clonar fragmentos maiores é que dessa forma aumenta-se a chance de clonar genes ou até mesmo óperons inteiros, facilitando a detecção de novas enzimas ou até mesmo vias metabólicas completas. A clonagem de fragmentos pequenos tem ainda a desvantagem de poder levar a formação de indesejáveis quimeras. (Singh et al. 2009). Porém, a clonagem por si não leva à geração de novas informações. É preciso triar em busca de clones que possuam sequências de interesse, e isto pode ser realizado basicamente de três formas:

1. Triagem baseada em função: onde as colônias são testadas com relação a uma atividade específica (como, por exemplo, degradação de lipídios ou atividade antimicrobiana);

2. Triagem baseada em sequência: utilizando-se PCR (reação em cadeia da polimerase), hibridização molecular ou técnicas computacionais para localizar determinada sequência;

3. Triagem baseada em indução de expressão gênica por substrato (SIGEX): utilizada apenas para detectar atividade catabólica. São utilizados vetores de expressão com o sistema marcador *gfp* ("Green fluorencese protein"), que emite fluorescência quando há expressão do gene clonado (Uchiyama & Watanabe, 2008).

A vantagem da triagem baseada em função consiste no fato de que não é preciso conhecer previamente parte da sequência do gene, tornando possível a descoberta de novas famílias gênicas com atividade de interesse industrial. Porém, nem sempre é possível expressar facilmente os genes clonados em bibliotecas tradicionais, que utilizam algumas cepas da bactéria *Escherichia coli* como hospedeira. Muitas sequências dependem de fatores de regulação para serem expressas e nem sempre estes fatores estão presentes. Para superar esta limitação, diversos outros vetores além de outras bactérias vêm sendo testadas como alternativas para a expressão heteróloga das sequências clonadas, como, por exemplo, estirpes de *Pseudomonas putida* e *Streptomyces lividans* (Martinez et al. 2004), sendo a última especialmente útil como hospedeira para expressão heteróloga de policetídeo sintases (PKS) em bibliotecas metagenômicas, assim como para outros grupos de enzimas produtoras de metabólitos secundários com atividades interessantes para a indústria de biotecnologia (Courtois et al. 2003).

Entretanto, as técnicas baseadas em sequência têm a vantagem de não ser preciso conseguir expressar o gene, sendo apenas necessário conhecer um trecho conservado

entre os membros da família alvo, e assim desenhar iniciadores para reações de PCR ou sondas para hibridização. Esta técnica é muito útil para detectar novos membros de famílias gênicas já conhecidas.

C) Sequenciamento direto do DNA ambiental:

Adicionalmente, pode-se extrair e sequenciar diretamente o DNA total do ambiente, utilizando-se técnicas de pirosequenciamento, como por exemplo, o Global Ocean Sampling Expedition (GOS) (Rush et al. 2007) e o metagenoma do Mar de Sargasso (Venter et al. 2004), que gerou mais de 1 gpb de DNA, localizou aproximadamente 1,2 milhões de genes putativos, demonstrando o grande potencial desta técnica na identificação de novos genes. Porém, a inferência funcional destes genes continua sendo um desafio, por não existirem homólogos nos bancos curados para anotação baseada em similaridade.

Nestes casos, todas as análises são feitas *in silico*, e diversas questões podem ser respondidas, principalmente com relação à biodiversidade do ambiente estudado. A maior dificuldade desta abordagem é a montagem das sequências, sem formar quimeras entre os pequenos trechos de DNA sequenciados, oriundos dos diversos organismos ambientais. Em geral se obtém baixa cobertura, o que torna complicada a completa montagem dos trechos obtidos. Novas técnicas de montagem vêm sendo desenvolvidas, utilizando, por exemplo, análises de frequência de utilização de determinados códons e conteúdo de GC, porém a presença de rearranjos gênicos e transferência horizontal de genes complica este tipo de abordagem (Cowan et al. 2005). Outra grande desvantagem é que desta forma não se obtém a sequência clonada para posterior expressão. Porém esta técnica pode ser combinada com a construção de bibliotecas com o DNA extraído do mesmo ambiente, visando à construção de sondas para hibridização e/ou iniciadores para PCR, baseados em trechos de DNA sequenciados diretamente daquele ambiente e posterior triagem da biblioteca em busca dos clones que contenham as sequências alvo do estudo.

1.3 – Policetídeo Sintases (PKS)

PKSs são enzimas que produzem um grande grupo de metabólitos secundários chamados de policetídeos. Esses metabólitos possuem diversas aplicações na indústria, sendo que muitos possuem importância médica. Dentre os principais, podemos citar compostos com atividade antimicrobiana como, por exemplo, a Eritromicina, imunosupressora como a Rapamicina, antiparasitária como Avermectina, e até mesmo toxinas prejudiciais, como a Aflatoxina (Castoe et al., 2007). Na figura 1.2 pode-se observar a estrutura química de alguns policetídeos.

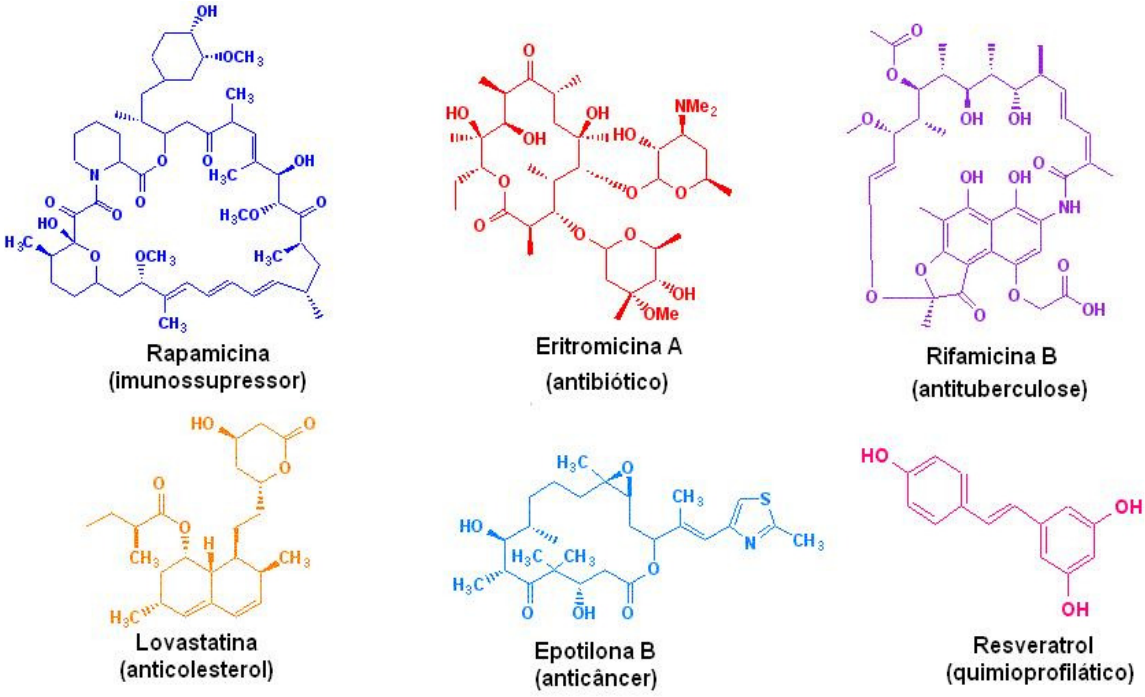


Figura 1.2: Estrutura química de alguns policetídeos com atividade de interesse médico (Fonte: http://linux1.nii.res.in/~pksdb/polyketide.html).

Estes metabólitos têm sido encontrados em diversos organismos como bactérias, fungos, plantas, insetos, dinoflagelados, moluscos e esponjas (Gokhale et al. 2007).

As PKSs possuem similaridade (tanto em sequência quanto em estrutura) com outras enzimas responsáveis pela produção de ácido graxo, denominadas Ácido Graxo Sintase (FAS). Ambas tipicamente catalisam sucessivas condensações de unidades simples de carbono (grupos acil-coA, geralmente acetil-coA e malonil-coA), para construir uma cadeia cetônica**.** Contudo, na biossíntese de ácidos graxos acontece a

completa redução dos grupos cetônicos, com a produção de cadeias de carbono completamente reduzidas, enquanto nos policetídios as cadeias permanecem parcialmente ou não reduzidas (Castoe et al. 2007).

Conforme a atuação do conjunto das enzimas (ou dos domínios) PKS durante a biossíntese, os policetídeos podem ser compostos aromáticos poliidroxilados (como a maioria dos pigmentos fúngicos), compostos alifáticos pouco oxigenados (ou policetídeos parcialmente reduzidos, como a lovastatina) e alifáticos altamente reduzidos (e.g. ácidos graxos) (Pastre et al. 2007).

Os domínios catalíticos (no caso das tipo I) ou enzimas (no caso das tipo II) presentes nas PKSs são: cetoacil sintase (KS), proteína carreadora do grupo ácido (ACP), acil transferase (AT), cetoredutase (KR), deidratase (DH), tioesterase (TE), enoil redutase (ER), metil trasferase (MT), Claisen ciclase (CYC) e domínio de condensação (CON). Os domínios essenciais para uma PKS modular mínima são KS, ACP e AT. Estes, além do domínio TE, realizam reações de condensação da cadeia. Já os domínios KR, DH e ER são responsáveis por reações de redução enquanto MT, CYC e CON realizam modificações pós-condensação (tabela 1.1) (www.rasmusfrandsen.dk/ny_side_8.htm).

Tabela 1.1: Domínios presentes nas PKSs e suas respectivas funções na síntese do policetídeo (fonte: http://www.rasmusfrandsen.dk/ny_side_8.htm).

Domínio	Função
Acil transferase (AT)	Carregamento de iniciadores, intermediários e extensor de unidades acil
Proteína Carreadora do grupo Acil (ACP)	Segura a cadeia policetídica em crescimento
β-cetoacil sintase (KS)	Realiza reação de condensação entre as unidades iniciadoras, intermediárias e extensoras
β-ceto redutase (KR)	Reduz β-cetonas a hidroxilas
Deidratase (DH)	Reduz hidroxilas a enoil (insaturado)
Enoil reductase (ER)	Reduz enoil a alcil (saturado)
Thioesterase (TE)	Facilita a liberação do produto final da enzima
Metiltransferase (MT)	Transfere grupos metil para o policetídeo em crescimento
Claisen ciclase (CYC)	Facilita a formação de anel através da reação de ciclização de Claisen
Domínio de condensação (CON)	Facilita a condensação do policetídeo sintetizado a outros policetídeos

As PKSs podem ser classificadas quanto ao tipo:

Tipo I – grandes enzimas multifuncionais, multidomínios, que possuem todas as atividades enzimáticas necessárias para o enlongamento e processamento da cadeia policetídica. Em alguns casos, a biossíntese de policetídeos por PKSs tipo I é realizada por mais de uma proteína, e os genes codificantes estão organizados em grupos (*clusters*), como por exemplo, o grupo de três genes responsáveis pela produção da PKS que sintetiza a Eritromicina. (Lal et al. 2000 e Cane et al. 1998).

As PKS tipo I podem ser modulares (geralmente em bactérias) (figura 1.3) ou iterativas (geralmente em fungos, porém presente em algumas bactérias) (figura 1.4).

Nas modulares, cada polipeptídio inclui um ou mais módulos, e cada módulo é responsável por um turno de condensação e processamento da cadeia. Cada domínio catalítico nas PKS modulares é utilizado apenas uma vez na biossíntese do policetídeo. Estas PKSs sintetizam policetídeos macrocíclicos, através da condensação de acetatos, propionatos e butiratos. O nível de redução dos grupos beta-carbonil realizado em cada ciclo de condensação é variável. Os policetídeos macrocíclicos são os de maior importância clínica. Avanços nos estudos sobre as PKS modulares realizados nas últimas décadas vem demonstrando ser possível realizar recombinações entre os módulos destas enzimas, modificando a estrutura e função do policetídeo, o que gera um grande potencial de produção de novos compostos (Rup et al, 2000).

As PKSs tipo I iterativas possuem um único módulo, que realiza vários turnos de enlongamento da cadeia, utilizando cada domínio várias vezes durante a biossíntese. Estas PKSs catalisam a formação de policetídeos aromáticos, como por exemplo, o ácido 6-metilsalicíclico (Shen, 2003).

As PKSs e FAS compartilham uma estrutura conservada, que inclui domínios funcionais homólogos entre elas.

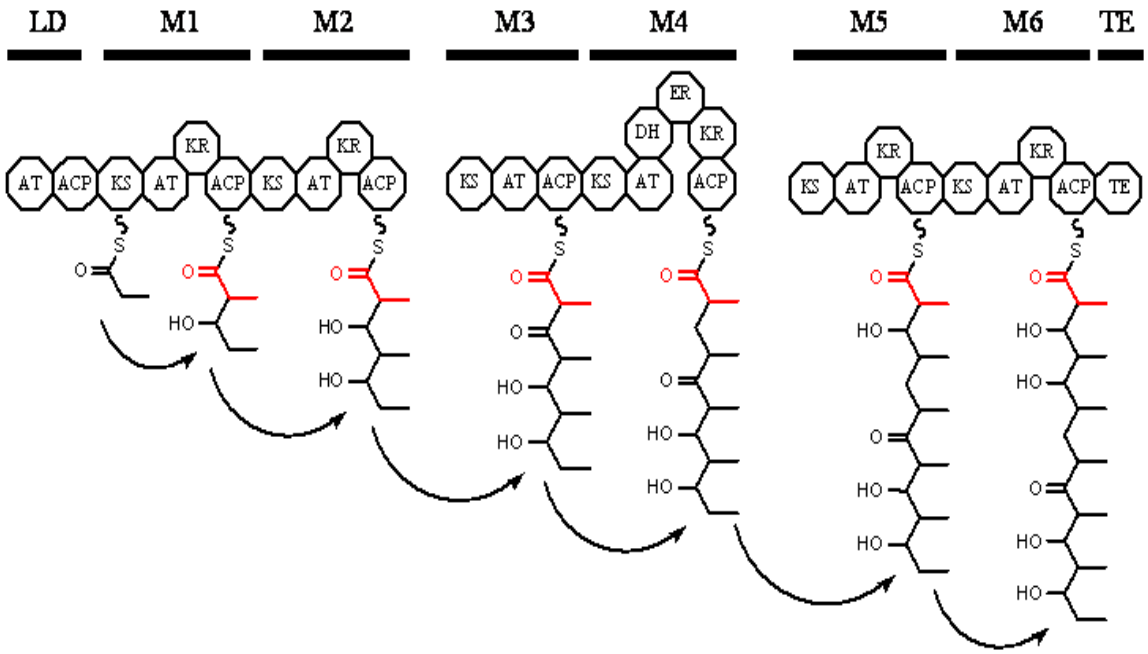


Figura 1.3: Estrutura das três PKSs modulares tipo I, responsáveis pela produção da Eritromicina, mostrando o crescimento da cadeia policetídica. Siglas: AT – Acil transferase; ACP – Proteína Carreadora do Grupo Acil; KS – Cetoacil Sintase; KR – Ceto redutase; DH – Deidratase; ER – Enoil redutase; TE – Tioesterase. LD – Módulo iniciador; M1 a M6 – Módulos 1 a 6. (Fonte: http://www.rasmusfrandsen.dk/ny_side_8.htm).

Figura 1.4: Esquema demonstrando o funcionamento de PKS tipo I iterativa, produtora de Lovastatina. As seguintes siglas são referentes aos domínios cataliticos da enzima: ACP – Proteína Carreadora do Grupo Acil; KS – Cetoacil Sintase; DH – Deidratase; ER – Enoil redutase; TE – Tioesterase. MAT – Malonil coenzima A:ACP transaciclase. (Fonte: http://linux1.nii.res.in/~pksdb/polyketide.html).

Tipo II – Ao contrário das tipo I, a atividade enzimática para o enlongamento e processamento da cadeia é realizada por enzimas separadas e cada domínio é utilizado de maneira iterativa (Castoe et al, 2007) (Figura 1.5).

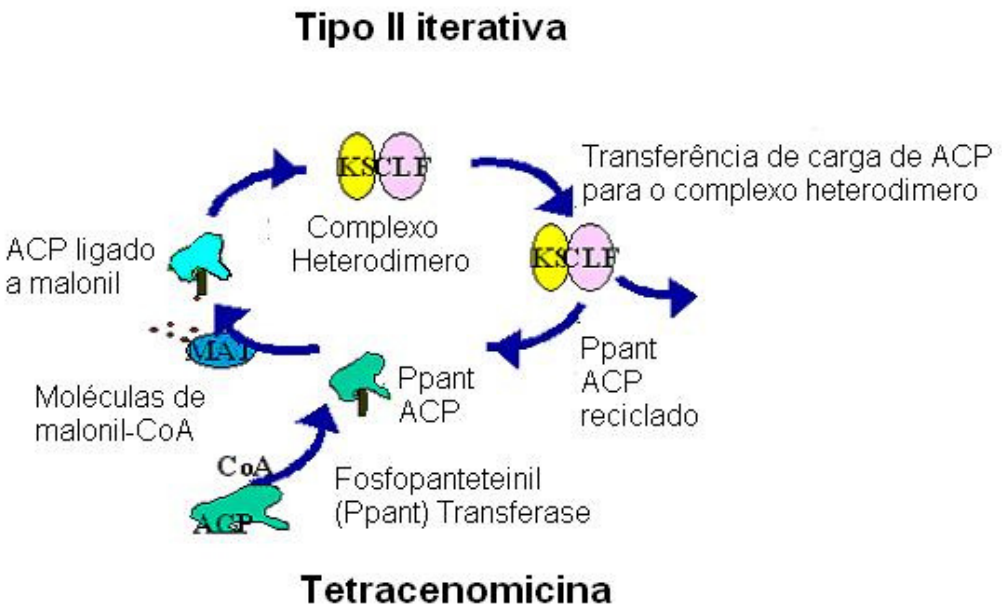


Figura 1.5: Esquema demonstrando o funcionamento de PKS tipo II produtora de Tetracenomicina. As seguintes siglas se referem aos domínios catalíticos: ACP – Proteína Carreadora do Grupo Acil; KS – Cetoacil Sintase; MAT – Malonil coenzima A:ACP transaciclase. CFL – Fator limitante de condensação (Fonte: http://linux1.nii.res.in/~pksdb/polyketide.html).

Tipo III – Responsáveis pela produção de Chalcona em plantas e polihidroxi-fenois em bactérias. Diferentemente dos outros tipos de PKS, na tipo III a cadeia é enlongada e processada em um único e multifuncional sítio ativo. Não existe domínio ACP neste tipo de PKS, atuando a mesma diretamente nos grupos acil-coA (Castoe et al, 2007) (figura 1.6).

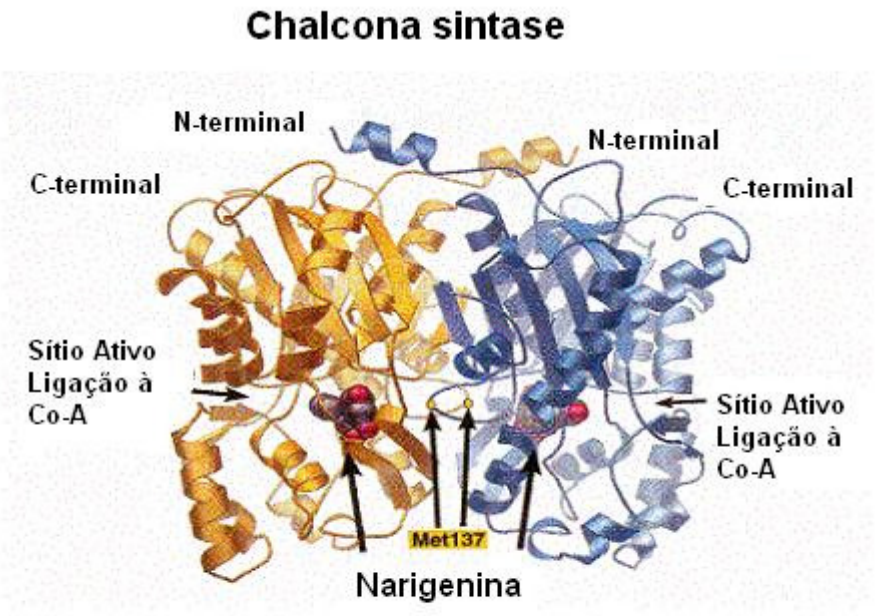


Figura 1.6: Representação estrutural de Chalcona Sintase (PKS tipo III). Sigla CoA – Coenzima A (Fonte: http://linux1.nii.res.in/~pksdb/polyketide.html).

A tabela 1.2 resume os tipos de PKS existentes, suas funções e estrutura, além dos organismos nos quais podem ser encontrados.

Tabela 1.2: Resumo dos tipos de PKSs existentes, com suas respectivas estruturas, mecanismos e distribuição nos organismos. Adaptado de Watanabe & Ebizuka 2004.

Grupo	Estrutura protéica	Mecanismo de síntese do metabólito	Encontrada em
Tipo I modular	Proteína única com múltiplos módulos e múltiplos domínios	Linear, cada sítio ativo utilizado uma única vez	Bactéria
Tipo I iterativa	Proteína única com único módulo e múltiplos domínios	Iterativo, cada sítio ativo utilizado várias vezes	Fungos e bactérias
Tipo II	Múltiplas proteínas, cada uma com um domínio ativo	Iterativo, cada sítio ativo utilizado uma ou mais vezes	Bactéria
Tipo III	Proteína única com múltiplos módulos	Iterativo, cada sítio ativo utilizado várias vezes	Plantas e bactérias

Apesar disso, diversos estudos sugerem que a diversidade de PKSs é muito maior em termos de mecanismo e estrutura do que se pode classificar com o sistema de classificação em tipos. Um exemplo é a atividade não iterativa exibida durante a catálise de um policetídeo por enzimas codificadas por um grupo de genes para PKS clonado a

partir de *Streptomyces griseus*. Esta, além de não possuir atividade iterativa, como as PKS tipo II (apesar de ser composta por múltiplas enzimas de domínio único), não utiliza domínio ACP, sendo, desta forma, similar às PKS tipo III. Todavia, seus domínios KS são claramente homólogos aos de tipo I e II, não podendo assim ser classificada como tipo III (Shen, 2003).

Diversos estudos vêm sendo realizados buscando novas PKSs em diversos ambientes com ajuda da metagenômica, como, por exemplo em esponjas marinhas (Kennedy et al. 2008, Schirmer et al. 2005) e em solos (Courtois et al. 2003, Wawrik et al. 2005). Entretanto, não foram encontrados estudos na literatura, realizados em busca de PKSs na zona pelágica e na superfície marinha, sendo este um campo a ser explorado.

2 – Objetivos

2.1 – Objetivo Geral

- Identificar, sequenciar e analisar domínios de PKSs de origem ambiental.

2.2 – Objetivos específicos

- Triagem *in silico* em busca de PKSs nos bancos públicos de sequências ambientais, principalmente de ambientes marinhos, para comprovar a existência das mesmas nesses ambientes.
- Mapeamento parcial da biodiversidade de microorganismos baseado em estudo de rDNA da região portuária de Arraial do Cabo, para comprovar a existência de organismos produtores de PKSs neste ambiente.
- Construção de bibliotecas metagenômicas a partir de água do mar da região portuária de Arraial do Cabo para posterior triagem por PCR em busca de PKSs.
- Realizar inferências sobre filogenia de PKSs.

3 - Material e Métodos

3.1 - Triagem "in silico" de bibliotecas metagenômicas obtidas em bancos de dados públicos

Foram utilizados dois bancos de sequências ambientais: o banco ambiental do NCBI (Environmental) (http://www.ncbi.nlm.nih.gov/) com um total de 6.028.192 proteínas, e o banco de proteínas de todos os projetos do CAMERA (http://camera.calit2.net/) com um total de 43.240.119 sequências. Todas as sequências em formato FASTA de três domínios de PKSs modulares (KS, AT e ACP) foram baixadas do banco PKSDB (http://linux1.nii.res.in/~pksdb/DBASE/pageALL4.html). Além destas, as sequências da região KS de todas as PKSs iterativas foram obtidas no IterDB (http://202.54.226.229/~pksdb/ITRDB/pageALL4.html) (Ansari et al. 2004).

O programa MAFFT (versão 6.240) (Katoh et al. 2002) foi utilizado para gerar 60 alinhamentos múltiplos (utilizando-se os parâmetros padrões), sendo três alinhamentos (um por domínio) para cada metabólito do banco PKSDB. Cada alinhamento múltiplo foi utilizado com o programa hmmbuild do pacote HMMER (versão 2.3.2), para gerar modelos ocultos de markov (*hmm*) utilizando os parâmetros padrões do programa. Estes modelos foram calibrados com o programa hmmcalibrate (com os parâmetros padrão) e posteriormente utilizados para buscas utilizando o programa hmmsearch (com parâmetro de corte de *e-value* em $10e^{-5}$) contra o banco ambiental do NCBI e contra o banco de proteínas do CAMERA.

Foram realizadas comparações entre as tabelas de hits dos 3 modelos para localizar quais sequências ambientais apresentaram hit contra pelo menos 2 modelos, e contra os 3 modelos. As sequências ambientais que apresentaram hit contra os 3 domínios foram submetidas ao SEARCHPKS (http://www.nii.res.in/searchpks.html) (Yadav et al. 2003).

O programa BLAST (versão 2.2.17) foi utilizado com o algoritmo BLASTP para buscar as regiões KS do IterDB nos bancos ambientais (CAMERA e NCBI), utilizando os parâmetros padrão.

Para buscar PKSs do tipo II, foram obtidas sequências dos genes KS do NCBI (gi|12744820, gi|12744821, gi|153497, gi|153496 e gi|161335626). As mesmas foram alinhadas com o MAFFT e modelos foram gerados para a busca contra os mesmos

bancos ambientais (utilizando o HMMER com a mesma metodologia utilizada para a busca de PKSs tipo I). As sequências dos bancos ambientais que apresentaram similaridade com os modelos foram extraídas com o FASTACMD do pacote BLAST.

Foram geradas listas com os identificadores das sequências ambientais que apresentaram maior similaridade com cada modelo de KS (modulares) e com cada KS iterativa (5 melhores para cada). Com o programa FASTACMD do pacote BLAST, as sequências foram extraídas e foram gerados arquivos no formato FASTA. A redundância foi removida com o programa CD-HIT (com parâmetro de corte em 100% de identidade).

Posteriormente cada sequência foi submetida ao SEARCHPKS para extração dos domínios KS e posterior análise filogenética.

A figura 3.1 ilustra a metodologia utilizada para a triagem de PKSs nos bancos ambientais.

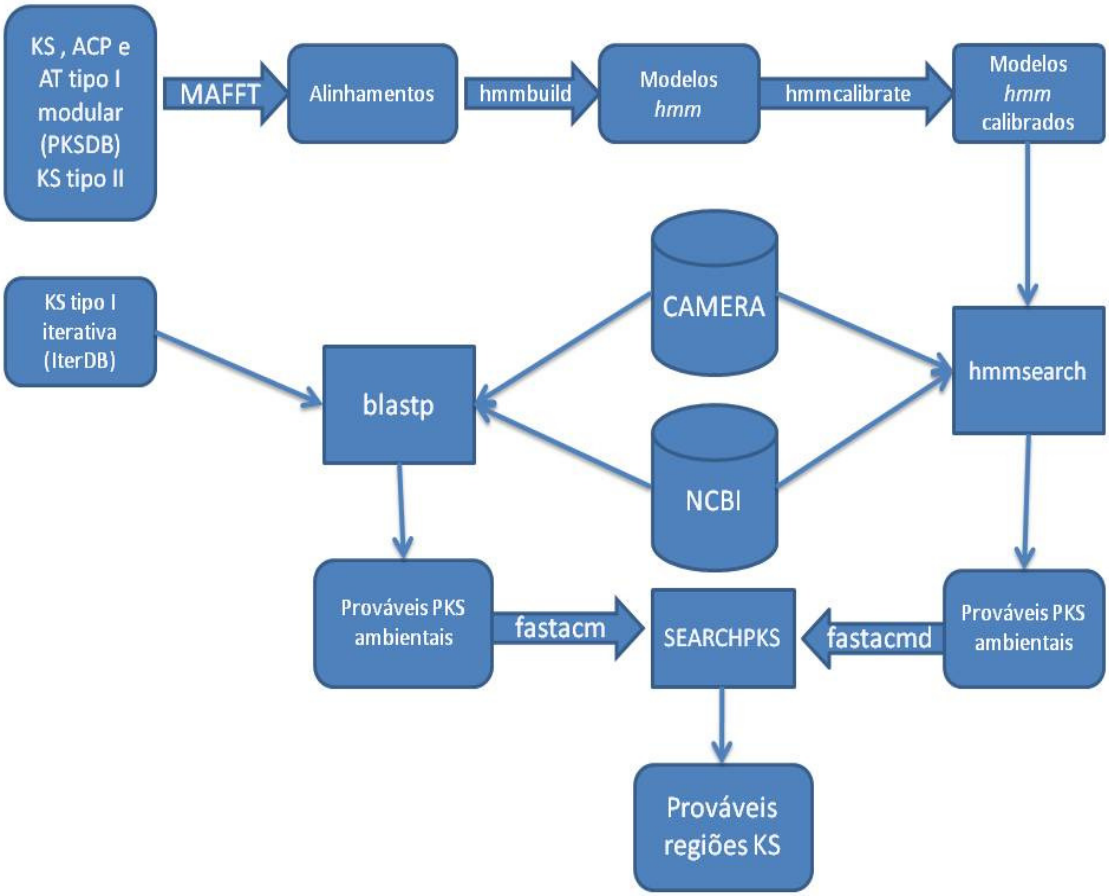


Figura 3.1: Fluxograma ilustrando a metodologia "in silico" para triagem de PKSs em bancos ambientais. Siglas: KS – Cetoacil sintase; ACP – Proteína Carreadora do grupamento acil; AT – Acil transferase; *hmm* – Modelos ocultos de markov.

3.2 - Coletas de água em Arraial do Cabo – RJ

Foram realizadas 8 coletas no município de Arraial do Cabo – RJ. A tabela 3.1 mostra as datas de cada coleta e o volume de água coletado.

Tabela 3.1: Data e horário das 8 coletas realizadas em Arraial do Cabo – RJ, com seus respectivos volumes.

	Data e hora da coleta	Volume coletado
Coleta 1	9/01/2008 – 14 h	10 litros
Coleta 2	12/03/2008 – 13 h	20 litros
Coleta 3	07/05/2008 – 13 h	40 litros
Coleta 4	09/07/2008 – 12 h	60 litros
Coleta 5	03/09/2008 – 13 h	80 litros
Coleta 6	11/12/2008 – 12:30 h	80 litros
Coleta 7	7/08/2009 – 16 h	140 litros
Coleta 8	6/12/2009 – 14 h	40 litros

Todas as coletas foram realizadas na superfície (máximo de 1 metro de profundidade) do porto da Praia dos Anjos (22°58'18.40"S; 42° 1'5.14"O). Em todas as coletas, a temperatura, o pH e a salinidade foram medidos. Os galões foram transportados para o Rio de Janeiro, armazenados a 4º C (graus Celsius) até a filtragem.

3.3 - Filtração

Coletas 1 a 7: As amostras foram filtradas em membranas Millipore de 0,22 micrômetros de diâmetro dos poros. As mesmas foram armazenadas a -80º C imediatamente após a filtragem.

Coleta 8: A amostra foi pré-filtrada em membrana, com 0,5 mm de diâmetro dos poros, para remover a sujeira. Apenas 1 litro não foi pré-filtrado, pois o mesmo foi separado para análises de biodiversidade baseada em genes ribossomais. Os restantes 39 litros foram filtrados logo em seguida em membranas com poros de 0.22 micrômeros (variando de 2 a 6 litros por membrana), sendo 4 delas lavadas imediatamente para extração do DNA, enquanto as demais foram armazenadas em "ultra-freezer" a -80º C.

3.4 - Extração do DNA

As membranas utilizadas no processo de filtragem das coletas 1 a 7 foram cortadas em pequenas fatias, distribuídas em tubos de 50 mililitros. A seguir, foram acrescentados 30 ml de tampão TE (Tris-HCl 10 mM, EDTA 1 mM) por tubo. Os tubos foram agitados com ajuda de um homogeneizador tipo vortex por 15 minutos. Posteriormente o líquido foi coletado e o processo foi repetido mais duas vezes para cada tubo, totalizando 45 minutos de agitação para cada tubo. O líquido coletado foi centrifugado a 10.000 g para se obter um material centrifugado com as células. O sobrenadante foi descartado e todos os centrifugados foram ressuspendidos em 1 ml de TE, para posteriormente serem transferidos para microtubos de 1,5 ml.

O material centrifugado foi lavado duas vezes com PBS (8 g NaCl, 0.2 KCl, 1.44 g Na_2HPO_4, 0.24 g KH_2PO_4 em 1 litro de água mili-Q, pH 7,2) gelado (o dobro do volume do centrifugado), centrifugando-se a 12.000 g por 10 minutos. Ressuspendeu-se o centrifugado em "Set Buffer" (Sucrose 20%, EDTA 50 mM, Tris-HCl 50 mM pH 7,6) (2 vezes o volume do centrifugado). Adicionou-se lisozima (concentração final 1 mg/ml), incubou-se por 45 minutos a 37º C e adicionou-se proteinase K (concentração final 0.2 mg/ml e SDS 10%), incubando-se a 55º C por 2 horas. Adicionou-se igual volume de fenol equilibrado, homogeneizando-se os tubos por inversão vagarosamente durante 10 minutos e centrifugou-se a 12.000 g por mais 10 minutos. A fase aquosa foi transferida para um novo tubo, onde se adicionou igual volume de fenol: clorofórmio (1:1), homogeneizando-se os tubos por inversão vagarosamente durante 10 minutos. Centrifugou-se a 12.000 g por 10 minutos (a 4ºC) e transferiu-se a fase aquosa para um novo tubo, adicionando-se igual volume de clorofórmio, homogeneizando-se os tubos por inversão vagarosamente durante 10 minutos. Centrifugou-se a 12.000 g por 10 minutos (a 4 ºC) e transferiu-se a fase aquosa para um novo tubo. Foi adicionado 1/10 do volume em solução de acetato de sódio 3 M pH 5.2 e 2.5 vezes o volume em etanol 100% gelado. Incubando-se por 1 hora a -20º C. Posteriormente, centrifugou-se a 13000 g por 30 minutos (a 4 ºC). Descartou-se o sobrenadante e lavou-se o DNA precipitado duas vezes com etanol 70% gelado. O DNA foi seco invertendo-se o tubo em papel absorvente por 10 minutos. Posteriormente ressuspendeu-se o DNA de cada tubo em 25 µl de TE por 2 horas em temperatura ambiente. Adicionou-se RNAse A (20 mg/ml) e incubou-se em banho-maria por 1 hora a 37º C. Uma alíquota de 2 µl foi

verificada por eletroforese em gel de agarose 1%, a 100 volts por 45 minutos. Ao final, os tubos foram estocados a -20º C.

Para extrair o DNA do material da coleta 8, foi utilizado o kit EPICENTRE "Metagenomic DNA Isolation Kit for Water" seguindo as recomendações do fabricante. Foram feitas 3 extrações, sendo a primeira a partir de 2 membranas e a segunda e a terceira a partir de uma única membrana cada. Após a lavagem e centrifugação das células na primeira extração, o precipitado foi armazenado a -20º C e a extração foi realizada no dia seguinte, enquanto as demais foram realizadas imediatamente após a filtragem. O resultado das 3 extrações foi verificado em eletroforese em gel de agarose 1%.

3.5 – Amplificação e clonagem de genes ribossomais (rDNA 16S e 18S)

Com as amostras de DNA extraídas a partir da coleta 8, foram realizadas 5 reações de PCR para os genes 16S e 18S , utilizando-se os iniciadores BAC27F (5' – AGAGTTTGATCMTGGCTCAG 3') e BAC518R (3' - ATTACCGCGGCTGCTGG – 5") para rDNA 16S de bactéria, e EK7F (5' – ACCTGGTTGATCCTGCCAG – 3') e EK516R (5'- ACCAGACTTGCCCTCC – 3') para rDNA 18S (eucariotos). Cada reação de 25 µl foi realizada com 10 ng de DNA ambiental, 2,5 µl de tampão de PCR, 2 µl de $MgCl_2$ (a 25 mM), 0,5 µl de dNTP mix (a 10 mM), 1 pmol de cada iniciador (senso e anti-senso), 18 µl de água mili-Q e uma unidade de Taq DNA polimerase recombinante (Fermentas). O ciclo utilizado no termociclador (Applied Biosystems) para 16S de bactéria foi: 5 minutos de desnaturação inicial a 95º C; 30 ciclos de 1 minuto desnaturando a 92°C, 1 minuto de anelamento a 55°C e 1 minuto de enlongamento a 72°C; 10 minutos a 72°C de enlongamento final. Para 18S o ciclo foi: 5 minutos de desnaturação inicial a 95º C; 30 ciclos de 1 minuto a 95º C, 1 minuto a 55º C e 1 minuto a 72º C; 10 minutos de extensão final a 72º C.

Os produtos de PCR foram purificados com o kit Qiagen (*QIAquick Gel Extraction*), seguindo instruções do fabricante e clonados com o kit "pGEM-T Easy Vector System", seguindo as instruções do fabricante, transformados em células competentes *E. coli* DH5α, plaqueados em LB ágar com ampicilina (100 µg/ml), x-gal (80 µg/ml) e IPTG (0,5 mM). As colônias brancas foram crescidas em meio LB líquido

com ampicilina (100 µg/ml) e submetidas à extração do DNA por lise alcalina para posterior sequenciamento, utilizando o primer M13, pela plataforma PDTIS-IOC-FIOCRUZ. Os cromatogramas foram submetidos ao sistema Stingray (stingray.biowebdb.org) para análise da qualidade e de similaridade.

Posteriormente as sequências foram todas modificadas para o sentido 5´ - 3´ da fita senso e as sequências de rDNA 16S foram submetidas ao RDP classifier (http://rdp.cme.msu.edu/classifier/classifier.jsp) (Release 10) com valor de confiança de 70%.

Posteriormente as sequências de rDNA 16S foram alinhadas com vizinhos obtidos no Greengenes (http://greengenes.lbl.gov/ - versão de março/2009) (DeSantis et al. 2006). utilizando o CLUSTALW, a região coberta pelas sequências obtidas neste estudo foi extraída do alinhamento e o mesmo foi convertido para o formato do MEGA (versão 4.0). As sequências de rDNA 18S foram alinhadas com vizinhos obtidos no SILVA (release 102) (http://www.arb-silva.de/) (Pruesse et al. 2007). Duas árvores foram construídas (uma para 16S e outra para 18S) através de agrupamento com vizinhos com valor de *bootstrap* 1000. No caso de 16S foi utilizada uma sequência da arqueia *Caldisphaera lagunensis* como raiz. No caso de 18S, além dos vizinhos obtidos no SILVA, foi feita uma busca com o BLASTN contra o NR e os 3 melhores hits de cada clone foram adicionados ao alinhamento para a construção da árvore.

3.6 – Amplificação e clonagem de regiões conservadas de PKSs a partir do DNA ambiental

Com cerca de 50 ng do DNA ambiental obtido com as coletas 7 e 8 foram realizadas reações de PCR, utilizando o par de iniciadores degKS2F.i 5'–GCIATGGAYCCICARCARMGIVT-3' e degKS5R.i 5'-GTICCIGTICCRTGISCYTCIAC-3' - específico para as regiões KS de PKSs tipo I (Schirmer et al. 2005). As reações foram feitas com volume de 25 µl, sendo 1 µl de cada primer (5 pmol) (senso e anti-senso), 2,5 µl de tampão de Taq DNA polimerase, 2,0 µl de $MgCl_2$ (do estoque a 25mM), 15,5 µl de água mili-Q, 2,5ul de dimetilsulfóxido (DMSO), 0,5 µl de dNTP mix (do estoque a 10mM), 1 unidade de Taq DNA polimerase recombinante (Fermentas). O ciclo utilizado no termociclador (Applied Biosystems) foi:

95º C por 5 minutos; 40 ciclos de 40 segundos a 94º C seguido de 1 minuto a 45º C e 45 segundos a 72ºC; 10 minutos a 72º C. Como controle positivo para a reação, foi utilizada uma espécie não caracterizada de *Streptomyces* cedida pela Dra Rosalie Reed Rodrigues Coelho do Laboratório de Biotecnologia de Actinomicetos, CCS, UFRJ.

O produto de PCR foi purificado com o kit Qiagen (*QIAquick Gel Extraction*), seguindo instruções do fabricante e clonado utilizando 3 µl da reação, com o kit "pGEM-T Easy Vector System", seguindo as instruções do fabricante, transformados em células competentes *E. coli* DH5α, plaqueados em LB ágar com ampicilina (100 µg/ml), x-gal (80 µg/ml) e IPTG (0,5 mM). As colônias brancas (um total de 96) foram crescidas em meio LB líquido com ampicilina (100 µg/ml) e submetidas à extração do DNA por lise alcalina para posterior sequenciamento, utilizando o primer M13, pela plataforma PDTIS-IOC-FIOCRUZ. Os cromatogramas foram submetidos ao sistema Stingray (stingray.biowebdb.org) para análise da qualidade e de similaridade. As sequências com qualidade foram submetidas à análise com o BLAST (blastn) contra o banco de nucleotídeos não redundante do NCBI (NR). As que obtiveram similaridade com PKSs foram traduzidas nas 6 quadros de leitura utilizando o programa TRANSEQ, e as sequências traduzidas foram submetidas ao BLASTP contra o banco NR para que fosse possível verificar o quadro de leitura correta. As sequências de aminoácidos que apresentaram similaridade com PKSs foram utilizadas para as análises subsequentes.

3.7 – Clonagem do DNA ambiental em fosmídeo.

Foi utilizado o kit EPICENTRE EpiFOS™ Fosmid Library Production (Figura 3.2)

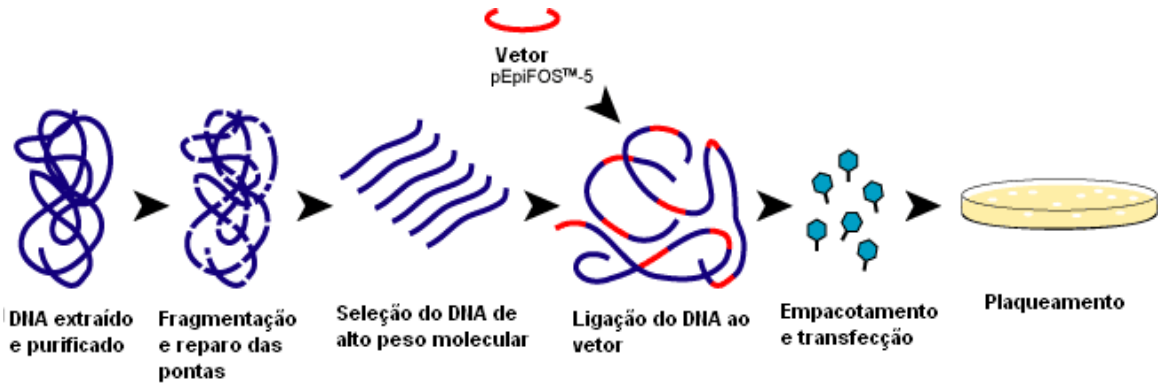


Figura 3.2: Esquema demonstrando o processo de clonagem de DNA com o kit EPICENTRE EpiFOS™ Fosmid Library Production.

Utilizou-se o DNA obtido com a coleta 7, extraído por fenol-clorofórmio e a amostra de DNA da coleta 8 extraído com o kit da EPICENTRE, para a construção de duas bibliotecas.

3.7.1 - Seleção de DNA de alto peso molecular (apenas para a coleta 7)

O DNA de alto peso molecular (entre 25 e 40 kb) foi selecionado através de eletroforese em gel de agarose de baixo ponto de fusão, com voltagem de 30 V por 16 horas a 4° C. Todo o DNA foi aplicado em um poço grande. O DNA de alto peso molecular fornecido com o kit foi colocado em um poço vizinho, enquanto o marcador HI-RANGE (Fermentas) foi colocado em um terceiro poço. Após a corrida, o gel foi corado em banho com SYBR GOLD por 40 minutos. Uma parte do gel contendo o DNA entre 25 e 40 Kb foi cortada com um bisturí, visualizando as bandas com ajuda de uma lâmpada emissora de luz negra. As partes de gel cortadas foram armazenadas em microtubos de 1,5 ml a -20° C. Posteriormente os tubos foram aquecidos em banho-maria a 70º C por 15 minutos para derreter o gel. O tubo contendo "GELase 50X

Buffer" foi pré-aquecido a 45° C e posteriormente acrescentou-se a GELase ao gel derretido, para uma concentração de 1X. Acrescentou-se 1 µl de GELase Enzyme Preparation na proporção de 1 µl para cada 100 µl de gel derretido. Incubou-se a 45º C por duas horas. Posteriormente os tubos foram incubados a 70° C por 10 minutos para inativar a enzima. Após este processo, foram removidas alíquotas de 500 µl em microtubos de 1,5 ml, e os mesmos foram resfriados no gelo por 5 minutos. Centrifugou-se a 10.000 rpm para precipitar os oligossacarídeos insolúveis. Removeu-se o sobrenadante cuidadosamente para novos tubos. Adicionou-se 1/10 do volume em acetato de sódio 3 molar pH 7.0 e 2.5 vezes o volume em etanol 100% gelado. Armazenou-se a -20° C por 18 h para potencializar a precipitação de DNA. Centrifugou-se a 14.000 rpm para precipitar o DNA. O sobrenadante foi descartado e o precipitado foi lavado duas vezes com etanol 70% gelado. Secou-se o precipitado invertendo os tubos por 10 minutos e o DNA foi ressuspendido em 54 µl de tampão TE. Uma alíquota de 2 µl foi utilizada para verificação da qualidade e quantidade em eletroforese em gel de agarose 1%, por 45 minutos a 100 volts.

3.7.2 - Reação de reparo das pontas

Seguindo instruções do fabricante do kit EPICENTRE, foi feita a reação de reparo das pontas do DNA, misturando-se os componentes da seguinte forma: 8 µl End-Repair 10X Buffer, 8 µl 2.5 mM dNTP Mix, 8 µl 10 mM ATP, 52 µl do DNA ressuspendido em TE, 4 µl End-Repair Enzyme Mix. Incubou-se por 45 minutos em temperatura ambiente e transferiu-se para 70º C para inativar a enzima. Adicionou-se 1/10 do volume em acetato de sódio 3 M pH 7,0 e 2,5 vezes o volume em etanol 100% gelado. Armazenou-se a -20º C por 18h para potencializar a precipitação de DNA. Centrifugou-se a 14.000 rpm para precipitar o DNA. O sobrenadante foi descartado e o precipitado foi lavado duas vezes com etanol 70% gelado. Secou-se o precipitado invertendo-se os tubos por 10 minutos e o DNA foi ressuspendido em 10 µl de tampão TE. Uma alíquota de 1 µl foi utilizada para verificação da qualidade e quantidade em gel de agarose 1%.

3.7.3 - Ligação do DNA ao vetor

Os seguintes componentes do kit EPICENTRE foram misturados na seguinte ordem: 1 µl "10X Fast-Link Ligation Buffer", 1 µl 10 mM ATP, 1 µl "pEpiFOS-5 Vector" (0.5 mg/ml) (figura 3.3) , 6 µl da amostra de DNA com as pontas reparadas e 1 µl "Fast-Link DNA Ligase". Incubou-se a temperatura ambiente por 2 horas. Transferiu-se para 70º C por 10 minutos para desativar a enzima.

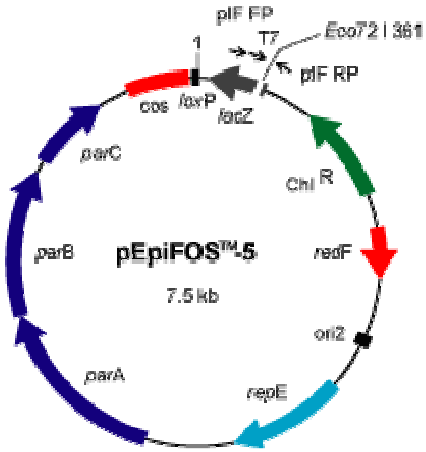


Figura 3.3: Vetor pEpiFOS-5, utilizado para a clonagem da amostra de DNA ambiental. (Fonte: EPICENTRE)

3.7.4 - Empacotamento do DNA em fago

Para cada reação de ligação realizada, um "MaxPlax Lambda Packaging Extract" foi retirado do freezer -80º C e descongelado no gelo. Foram acrescentados 25 µl (metade do extrato) para cada reação de ligação, congelando-se a -80 º C a outra metade. Transferiu-se o tubo para 30º C por 90 minutos. Após este período, descongelou-se novamente a metade do extrato restante e acrescentou-se à reação, incubando-se por mais 90 minutos a 30 º C. Ao término, foi acrescentado tampão de diluição de fagos para o volume final de 1 ml, e 25 µl de clorofórmio. Armazenou-se a 4º C por até 1 mês.

3.7.5 - Transfecção dos fagos

Uma única colônia de *E. coli* **EPI100-T1R** fornecida com o kit EPICENTRE foi retirada de uma placa de ágar e adicionada em 50 ml de meio de cultura Luria-Bertani (LB) suplementado com $MgSO_4$ (25 mM) em um recipiente autoclavado de 200 ml com tampa. O recipiente foi colocado em agitação de 200 rpm a 37º C por 16 horas. Foram retirados, em condições estéreis em cabine de segurança biológica, 5 ml da cultura e adicionados a um recipiente com mais 50 ml de LB suplementado. Incubou-se a 200 rpm, 37º C por 1h e 30 min. (acompanhando a densidade ótica com o espectrofotômetro "Biophotometer" fabricado pela Eppendorf até atingir o valor de 1,0). Após a cultura atingir o valor de densidade óptica desejado, foram aliquotados 100 µl da cultura em diversos microtubos de 1.5ml. Alíquotas de 10 µl do DNA empacotado no passo anterior foram adicionadas a cada tubo com a cultura. A mistura foi incubada a 37º C em estufa sem agitar por 20 minutos. Posteriormente alíquotas de 220 µl da mistura foram espalhadas em placas de ágar com cloranfenicol na concentração de 12,5 µg/ml. Uma placa foi utilizada como controle negativo, utilizando-se 220 µl da cultura de célula que não foi transfectada. As placas foram incubadas em estufa a 37º C por 24 horas.

3.7.6 - Estoque das colônias em glicerol

As colônias foram retiradas com palitos de madeira autoclavados e postas individualmente em poços de placas de 96 poços profundos com 1 ml de LB com cloranfenicol a 12.5 µg/ml. As placas foram incubadas com agitação a 200 rpm com temperatura de 37º C por 18 h. Foram colocados 100 µl de cada clone cultivado com 100 µl de glicerol 50% autoclavado em placas de poço raso, e as mesmas foram armazenadas a -80º C em "ultra-freezer".

3.7.7 - Extração de fosmídeos dos clones

Os fosmídeos foram extraídos pelo método de lise alcalina com algumas modificações: cada clone foi crescido em meio "Cicle-Grow" com cloranfenicol (12,5 µg/ml) por 16 horas a 37º C com agitação a 200 rpm. Transferiu-se 1,5 ml de cultura para tubos de 1,5 ml e centrifugou-se por 10 minutos a 13000 g, descartando-se o sobrenadante. Repetiu-se o procedimento por mais 3 vezes, acrescentando-se mais 1,5 ml de cultura ao mesmo tubo e centrifugando-o novamente. O processo foi realizado com um total de 48 ml de cultura para cada clone, em 8 tubos de 1,5 ml. O material de cada tubo centrifugado foi ressuspendido em 100 µl de TE, e agitado por 5 vezes por suaves inversões. Foram acrescentados 100 µl de tampão de lise (2 µl de NaOH 2 N , 5 µl de SDS 10% e 93 µl de água mili-Q) e agitou-se por 5 minutos por suaves inversões. Incubou-se por 5 minutos em temperatura ambiente. Posteriormente, foram adicionados 300 µl de solução de neutralização (acetato de potássio 3 M pH 4,8) e agitou-se por suaves inversões durante 5 minutos. Centrifugou-se por 5 minutos a 12000 g e transferiu-se por inversão (sem usar pipeta) o sobrenadante para novos tubos. Adicionou-se 1 µl de RNAse a 10 ng/ µl em cada tubo, incubando por 20 minutos a 37º C. Posteriormente foi adicionado 1 ml de etanol 95% gelado e centrifugou-se por 15 minutos a 12000 g, descartando-se o sobrenadante. Adicionou-se 1 ml de etanol 70% ao sobrenadante e agitou-se por suaves inversões 5 vezes. Centrifugou-se por 5 minutos a 12000 g, descartando novamente o sobrenadante. Drenou-se o resíduo de líquido dos tubos invertendo-os em papel absorvente por 20 minutos. Ressuspendeu-se o DNA precipitado de um tubo em 50 µl de água mili-Q, por 5 minutos a 37º C, a mesma alíquota de água foi utilizada para ressuspensão dos outros tubos, resultando em uma maior concentração de DNA no final do processo. Foi verificada uma alíquota de 2 µl em eletroforese em gel de agarose 1%, 100 volts por 20 minutos. Quantificou-se também o DNA utilizando um espectrofotômetro (NanoDrop). Por fim armazenou-se em freezer a -20º C.

3.8 – Triagem em busca de PKSs nos clones das bibliotecas construídas

Foram realizadas extrações a partir de "pools" com 96 clones cada placa de crescimento, pelo método de lise alcalina descrito em 3.7.7. Posteriormente foram realizadas reações de PCR com o par de iniciadores e as condições descritas em 3.6. As placas que apresentaram amplificação foram novamente cultivadas para que novos "pools" fossem construídos, com 12 clones de cada linha da placa. Mais uma vez foram realizadas extrações pelo método descrito em 3.7.7 e novas reações de PCR nas mesmas condições descritas em 3.6. Os "pools" de linhas positivos foram selecionados para cultivo e extração de fosmídeos de todos os clones dos mesmos, para que fosse possível localizar o clone contendo sequência de PKS.

3.9 – Análises filogenéticas das sequências de PKSs

As sequências de regiões KS obtidas nos bancos PKSDB e IterDB foram alinhadas com sequências de fabF e fabB (proteínas de ligação a ácido graxo) de *Escherichia coli* K12 (grupo externo), utilizando o programa MAFFT. O alinhamento foi submetido ao programa MODELGENERATOR para inferir o melhor modelo evolutivo a ser utilizado. O programa READSEQ foi utilizado para conversão do formato de alinhamento para formato PHYLIP. Posteriormente uma árvore filogenética foi construída através do método de máxima verossimilhança utilizando o programa PHYML (versão 2.4.4), utilizando o modelo sugerido pelo MODELGENERATOR e análise de *bootstrap* com valor 100.

Para situar as regiões KS amplificadas a partir das amostras ambientais de Arraial do Cabo, as mesmas foram alinhadas com as regiões KS do PKSDB, IterDB além da fabF e fabB. O alinhamento foi editado manualmente para retirar apenas as regiões cobertas pelas sequências amplificadas. Posteriormente o alinhamento foi submetido ao MODELGENERATOR e convertido para o formato PHYLIP com o READSEQ. Uma árvore foi construída com o PHYML com os mesmos parâmetros utilizados na árvore anterior.

As regiões KS extraídas do banco ambiental do NCBI e do CAMERA com o SEARCHPKS foram utilizadas para a construção de uma árvore para cada caso. Porém, nestas árvores foram incluídas 5 sequências de KS tipo II obtidas no GENBANK

(gi|12744820, gi|12744821, gi|153497, gi|153496 e gi|161335626), além de fabH de *E. coli e Mycobacterium bovis*. As árvores foram construídas com os mesmos parâmetros das anteriores.

Outra árvore foi construída utilizando as sequências KS tipo II curadas, as fabH da árvore anterior e as sequências ambientais do CAMERA e do NCBI que apresentaram hit com o modelo de KS tipo II construído, seguindo os mesmos parâmetros utilizados nas outras árvores.

4 - Resultados

4.1 - Triagem "in silico" por PKSs ambientais:

4.1.1 – Busca por PKSs tipo I modular no banco ambiental do NCBI:

Utilizando os 20 modelos *hmm* construídos com o alinhamento dos domínios KS do PKSDB, obtivemos um total de 13467 hits contra um total de 1445 proteínas do banco de sequências ambientais do GenBank, sendo 9,93% destes hits similares ao modelo de PKSs produtoras de mixalamida, 8,12% relacionados à epotilona e 7,84% similares às PKSs produtoras de tilactona. Os demais percentuais podem ser observados na figura 4.1.

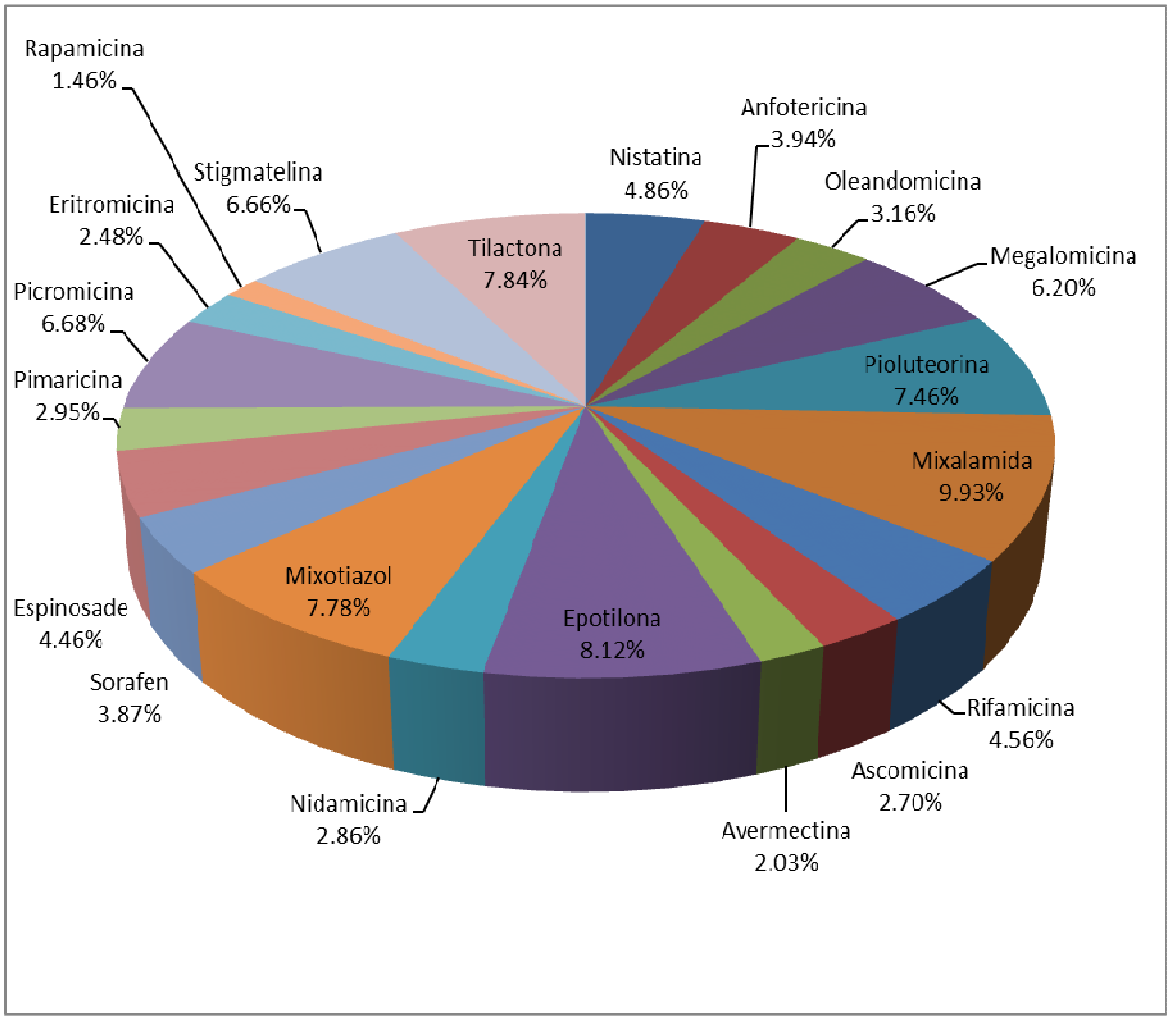


Figura 4.1: Distribuição dos hits obtidos com a busca por similaridade entre os domínios KS de PKSs tipo I modulares (separados por metabólito produzido) e o banco ambiental do NCBI, utilizando o pacote HMMER.

Já com os modelos *hmm* gerados a partir dos domínios AT, foram obtidos 12145 hits (com 1167 sequências), sendo 8,57%, 8,29% e 7,94% similares à Anfotericina, Estigmatelina e Rifamicina respectivamente, o que pode ser observado na figura 4.2.

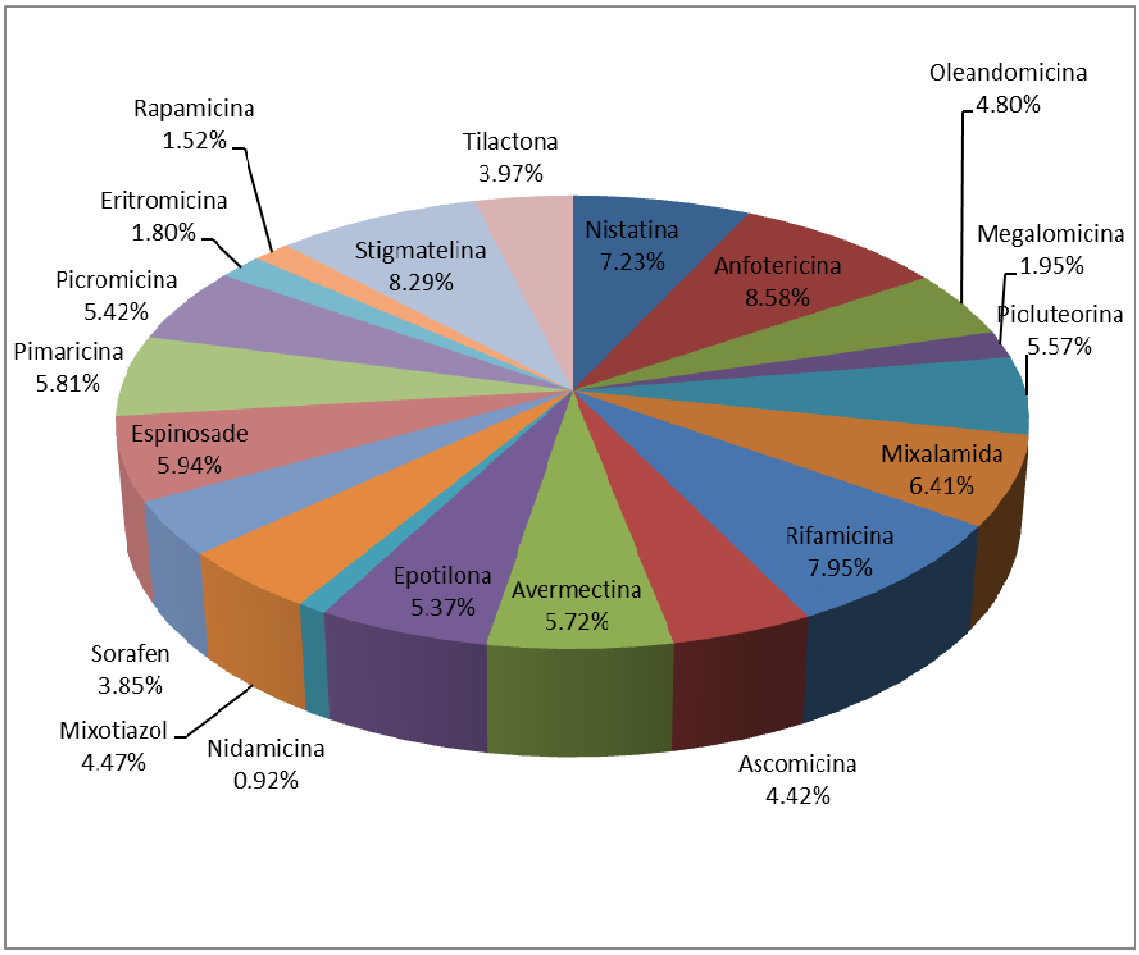


Figura 4.2: Distribuição dos hits obtidos com a busca por similaridade entre os domínios AT de PKSs tipo I modulares (separados por metabólito produzido) e o banco ambiental do NCBI, utilizando o pacote HMMER.

Com os domínios ACP, o número de hits foi de 889 (totalizando 117 sequências similares a um ou mais modelos de ACP), sendo 10,14% relacionados à Sorafen, 8,84% relacionados à Nidamicina e 8,84% similares à Epotilona, como demonstrado na figura 4.3.

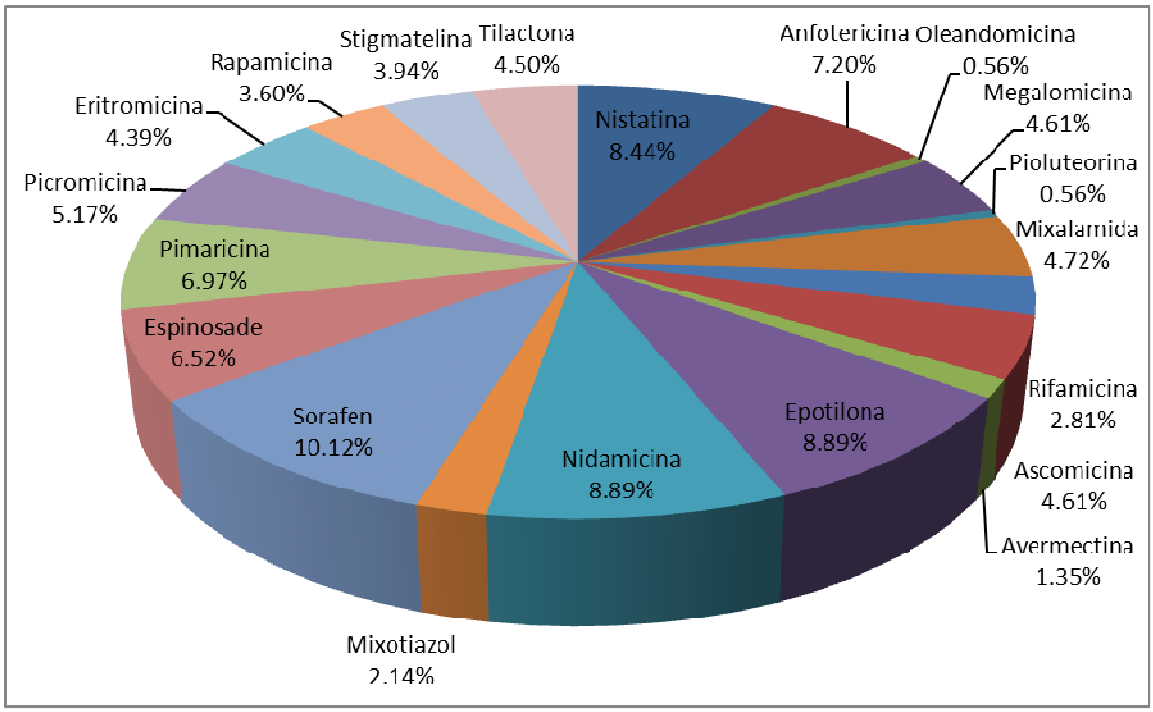


Figura 4.3: Distribuição dos hits obtidos com a busca por similaridade entre os domínios ACP de PKSs tipo I modulares (separados por metabólito produzido) e o banco ambiental do NCBI, utilizando o pacote HMMER.

A tabela 4.1 mostra o número de hits relativos a cada domínio.

Tabela 4.1: Número de hits obtidos com o HMMER entre os modelos de PKSs tipo I iterativa e o banco ambiental do NCBI

Metabólito produzido pela PKS	Domínio KS	Domínio AT	Domínio ACP
Nistatina	654	878	75
Anfotericina	531	1042	64
Oleandomicina	426	583	5
Megalomicina	835	237	41
Pioluteorina	1005	676	5
Mixalamida	1337	778	42
Rifamicina	614	965	25
Ascomicina	363	537	41
Avermectina	274	695	12
Epotilona	1094	652	79
Nidamicina	385	112	79
Mixotiazol	1048	543	19
Sorafen	521	468	90
Espinosade	600	722	58
Pimaricina	397	706	62
Picromicina	900	658	46
Eritromicina	334	219	39
Rapamicina	196	185	32
Stigmatelina	897	1007	35
Tilactona	1056	482	40
TOTAL	13467	12145	889

Apenas 3 sequências do banco ambiental do NCBI apresentaram similaridade aos 3 domínios, estas estão representadas na tabela 4.2.

Tabela 4.2: Sequências do banco ambiental do NCBI que apresentaram similaridade com os 3 domínios (AT, ACP e KS).

Identificador	Descrição	Sequência
gi|136007265| gb|EBL28474.1	Hypothetical protein GOS_8596358 [marine metagenome]	GVLHAVVLPQAPPVSSLVPVQWDRMLGNDNAPAFLSSMSSSVRRRT STVPVEPGATCAISLETVLDMVRRTAGGGVDADAPPMEAGVDSLGA VELRNQLQRAVGDSISLSSTLMFDHPTARQVATHLGGSAVAAADKR TVNAQLASTGTHVEIVGTVVSLPMGGSAVRGVSHCARDLLCVIPLT RWDVEAAARDLLGSPPAVASRVRHGGFLRDAELFEHRFFTMSAAEA AAMDPQQRQLLEHGYTAVHAAGRSKASLLGDIIAVNVGQWQSEFGA VLLGTPAGRSVYASTGFSCSVTCGRVSFVLGLQGPCASFDTACSAS LVANHGSMRALQRKECVAALSAGVNMILDPATMRGNAVAGFTSVRG RSHTFDARADGYARGEAIGAVVSRLREVGVRSAAEMRGSAVRQDGR SASLTAPNGQAQQGVLGASLVDAQADAGEVATLEAHGTGTALGDPI EAGAVAAIFLARRVSMGQSL
gi|142022786| gb|ECU92436.1	hypothetical protein GOS_2992782 [marine metagenome]	LACRQRRRKDRGARAQRALPRGRENPAHARRAVRSRRPVASRGRRG CDQRNRAFDAAAQGSLSPRRRTRRCAADTPGRRLLPSRRQRQGRRR VVSARADGRPAARPFRLVLVDGRADHHARPGQLRGREQFPRRARPA SARAGETGAQRQLGAVGGDRPRRHRLRTARARTTRRARRRHAAART GHRDAGTADGVRRHPVRGRANRLADPVPGRCAGRRVRAVFRAGATG RAAGAAGDGVAAPAACARAARTGRTHHRHARGDAGRNLAPFQPRCH RARAIAARSRPGFAGRARTDGSPHQGVRKTVSRDVVLFLSEPADAR PVRAQRTVAIAPRAGRRRSIRRPRRGRPFRTDRPGDRRPMNAKATH ALKAALDELRLRRAEIAALRSNRNEPIAVIGMACRFPGRSDTPDAF WQLLDGARDAVTEVPGERWDIDRYYDPDPSTPGKMATRHGAFLERV DQFDAAFFGIAPREATYLDPQQRLLLEVAWEALENAHLAPERFRQS ATGVYVGITCFDHAIQVSNASMPSSSYAGTGSALNMAAGRLSFVLG LTGPSMAIDTACSSSLVCLHLACESLRSRESNMALAGGVNLMLSPE VMVSFSQARMLSPDGRCKTFDAAADGYVRGEGCGIVVLKRLADALV DGDRVLGIVRGTAVDQGGAGGGLTVPSRDSQERVIRRALHQAGLAP GDVSYVEAHGTGTSLGDPIEVEALAGVYGPGRAANEPLVIGSVKTN IGHLESASGIAGLIKVLLSFEHDRIPAHLHFTQPNPHTPWQDIPIR VAADPVAWQRGERRRIAGVSAFGFSGTNAHAIVEEPPVAPARAAQR ALLLLSARSEAALAALVQRYERAIAGATPQELAAICRAAATGRSHY PFRAAYVSGVPASSAAAPRTGKALRMGFRFGVPDSGVAHALHASEP LFRDAFARCSVPLDALETDAGRFAIQFAWAELWKGWGIRPAVVSGH GIGEYVAACVAGVVSVADALRLVAARSNAEALRAVLRDMPLARPSV RLISGCLGADVTDEVTHPQYWLQLAGASDQADASHPPEGLADGWLP PPCAGDALERALAALYVQGAQFDWRALFPAPAQPATTLPNYPFERQ RFSLEKIPSPIVGMDAGSIDAALRHLKSSGKYPEDMLNAFPDLLRT AFAPAETVAPHAHPLYHVVWEQQAALPTAQVAADASPWLIFADASG VGERLAVLLRARGASCSLVRPGPDYVAGAEAGWQVAPERPDDFVRL LNETAAPGQRIVFLWALDEAVGETRMSTALLHLVHALVGSEREWTP STRPRISVVTRDAVEAGEAPHVSGLAQAALSGLARGAMIEHPEWFG IAIDLDPAAPEDETHALLQEMLGESREEQVALRHGARHVARLSPLA QAETAALPVDPDAAYLITGGFGALGLHTARWLAARGAGTLILVGRQ GAASDESQRAIAELRERNVTLRCERLDIADPAAVAAFFAALRRDGV PLKGIVHAAGIVGYKPIMQVERDELDAVLQPKVAGAWLLHQQSEHF PLDFFLLFSSIASAWGSREQAHYSAANRFLDALAHHRRGQGLPALS VNWGPWAEGGMTFPEAEALLRRVGIRSLAADRALDVLNRLPAVPQV AVVDIDLALFQGSYEARGPKPFLDRVRVAKSAPSAPAMPALSDASP RERKRLLADSIDRAVAQVLGYDAGTLDRDLGFFEMGMDSLMALDVR THLENALGIPLSVALLFDHPTVNALADFLAEQASGTAPDAHVAPAQ AQSVPPPQQPRPVAPAIDAREAGTPEPIAIVGMSCRFPGAAHDLDA YWQLLNDGVDAISEVPRERWDVDAYYDPDPEAPGRMYSRFGGFLDD VDQFDPAFFRITPREAAAMDPQQRLLLEVSHEALEHAGIPVDSLKG SRTGVFVGITTNDYANLQLRNGGGSGIDGYFFTGNPLNTAAGRISY GLGVQGPSMAIDTACSSSLTAIHTASQNLRSGECDVAIAGGVNLIL SPDNSIAVSRTRALAPDGRCKTFDAAADGFVRSEGCGALVLKRLSD

		ALAAGDRVLAVLRGSAVNHDGASSGFTAPNGRAQEAVIRQALGGLP AASIDYVEAHGTGTPLGDPVELQALATVFGAGRDASRRLRVGSVKT NIGHTESAAGIAGVIKVVLSLNHDRLPAHLHFRQPSPLVQWDALPL EICAEASAWPRGERPRRAGVSAFGASGTNAHLVLEEAPAPALQATP SRHKVHPLVLSAKTPAALRELAGRYQRRLEAEPGLDIAAVAFSAAT GRSHFAHRLAWPVTSLDDAIDKLRAFHAKEPAGAAQPAPRVKMAFL FTGQGSQYAGMGRRLYDAYPVFRDAIDRCRAVADPLLDKPLLEVLS AQGEDIHQTGYSQPALFSLQYALTTLLASFGVVPDAVMGHSVGEYA AACAAGVFSPEDGLRLIAERGRLMQALPRDGEMAAIFADLATVERA IDAWPHEVAVAAVNGPASIVISGKRERIAMLVDTFAARDIRSVPLN TSHAFHSPLLEPMLDSFQLAAKTVPVARPAIPFYSNLTGAVMDEAP TDTYWRRHCREPVQFASSVERLAEAGFNVLVEIGPKPVLVNLARAC CAPDAGIQFLALQRPQVEQQALIETLSSLYARGVDVDWAPTETPAP ARIALPSYPFQRSRTWFQKADTSMTQTSASPIAAAPTHNRSGEILE WLRGKIGELIQADPATINIELPFLEMGADSIVLIEAIRHIEAEYGV KLAMRRFFEDLATVQALAEYVADNLPAAAAPSGAEAVAVAVAVSEP STPAVAVAPSAAGLAPLAAAPAEWVAAEGDSTVERVLREQNQLLSH VMSQQMELLRTSLTGQPGVRPATAAVQAVASTASVAPQAASAAPAA APAAKPAPAAAAAPAADNPPPKPMMPWGSPVQQRARGLSAAQQEHL EALIVRYTTRTRKSKDSVQASRPVLADSRATVGFRFSTKEMLYPIV GDRAAGSRLWDIDGNEYIDFTMGFGVHLFGHTPDFIQQQVTREWQR PLELGARSSLVGEVAARFARVTGLDRVAFSNTGTEAVMTAMRLARA VTERDKIVMFTHSYHGHADGTLAAANAEGVTETIAPGVPFGSVENM ILLDYGSDAALEAIRGMAPTLAAVMVEPVQSRNPSLQPVAFLKELR RITEEAGVALIFDEMITGFRVHPGGSQAMFGIRADLATYGKIIGGG LPLGVIAGTSRFMDAIDGGMWTYGDHSFPAADRTAFGGTFCQYPLA MAAALAVLEKIEQEGPALQAALNERTAQIAGTLNAFFAEAEAPIKV TWFGSMFRFEFTENLDLFFYHMLEKGIYIWEWRTCFLSTAHTDADI DRFIRAVKDSVADLRRGGFIRPHSKHGTVAALSEAQRQLWVLSEID PEGSLAYNVNTTLELNGRLDEAAMRAAVQSLVDRHEALRTTMMADG SGQIVHPSLTLEIPLIDTDPNAWREQESRQPFDLVNGPLFRAALVR LGSERHLLVMTAHHIICDGSTFGVLLEDLARAYAGAAPADAPLQFR AYLKQLDGQRHSPETKANREYWLAQCARQAAPLNLPLDYPRPAVKT FHGERVSLHLDAAAAATLRTAARQNGCTLYMVLLAGFNLFLHRVAG QQEIVTGIPVTGRSVAGSDRLAGYCTHLLPLHSTLPEQATVASFLA GTRQNLLDALEHQDYPFAELVREIGAQRDLNAAPLVSAVFNLEPVS ALPELRGLTVGLVAPLIRHTAFDLNVNVLDAGQALLIDCDYNTDLF DASTVQRFLDIYRTLLTHLAEDASAAVARLPLSSDAERKLLTVEWN RTDTDFGDAAAQPLHRLFEQQVERTPDAVAVVFDDTALTYAELNLR ANRLAHHLIALGVGPDALVGVAMERSLDMSVALLAILKAGGAYVPV DPDYPAERVRFMIDHAQLRWLLTQQHLRDALPDTDAHVIVVDRDAL DLDAAATSNPAPALNGDNLAYMIYTSGSTGRPKGALNTHRAITNRI LWMQHAYALGADDAVLQKTPFSFDVSVWELFWPLVTGARLVFARPG GQRETDYLVELIERERITTIHFVPSMLRAFLDHPDLDAHCASLRRV VCSGEALPHDLQQRCLERLDVELYNLYGPTEAAVDVTAWECRRDDP HRIVPIGRPIANTRLYIVDAQMQPTPIGVAGELLIGGTPVGRGYHG EPELSAEKFIADPFSADPLARLYRTGDLARYRHDGNIEFLGRIDHQ IKLRGLRIEPGEIEAALTSHPLVDAAVVALRGVDDGARLVGWLCSS HPEAELVEAVRGHLRQRLPDYMVPSAFVVVSAFEHLPNGKLDRARL PEPGDGLDHVAPVNALEAQLAAIWQEVLGQARISTTANFFELGGNS LLATKVVARIRRDLHAKLEIRSLFALPTISSLAKRIADTQPIDYAP VTPLPAQASYALSPAQTRLWVQDRLHAAQAEGPLPTSLLFEGVLDV DALVRAFRALSERHEILRTRFVLEGNQPVQHVLPPGEAAFPVEIVD LQDAEDRDAQAASIQASERLVPMDLATGPLFRVKLLRLSEVRHVCI CTMHHVVSDGWSTEVLLDDLSALYDAFVQRRDDPLPALPIQYKDYA GWLNRLLAGPEGARMKDYWMTKLGGGLRALELPGDVEQPAAPSWKS WRFDLPAAETAALESLGKRHGATLFIALLSAIKALFYRRSGQEDIV VGTPVAGRELPELESQVGPYLNVLALRDRVAGDDRFDTLLTRVRDT TLEAFSHPLYPLDRLLDELHIKRVAGRNPLFDIGLTLQNQRHGPVD RYAGQVHIAELPDHDPQRADTEAATDFWFLAEPHAEGLAIRVVYHA GRFSEALVQGLANELTSVIGEVLANPGVRVRNLTLGQRALRAEARQ PTVELSAF
gi|142606603|gb|ECZ28301.1	hypothetical protein GOS_2210759	TAKDTSDVAMDPSTVVTRAVAGALGRDVERDAPLMEEGLDSLSAVE LGNTLQTATGIEMPATLVFDYPSQDAIIEYLDGAISKHRTQSRTTE KRSEVTKRMAIVGDADNRGSEVAQRTFKLETCEARGLVLNALGEYP GMSATSYLSGLFHSAVDGAPISGKISMSHFGEDQPAQPGLSLSLAV

	[marine metagenome]	CRSSESDVDTSSISDRCGILPLGRFALDATADSPPSEIVNRMMHTM ENVELFDASAFRISPAESEAMDPQQRILLESVLHARRGIDGRDLIH RSGVVVGASGSQYFESQSIGPHSAVGSQQSVLCGRVSYAFGLKGPS VCVDTACSSSLVATHISAESVKTAACGNSIAAGIMVSSGMVAHSTL SAARMLSADGRCKTLDISADGYGRGECCGVVHIDRVSDGDASTSTL LVGTGVNQDGRSSSLTAPNGPSQQMLIADTMRVAGVSGDSVVQLEM HGTGTSLGDPIEVGAASTVLCGASKATASDSLILQAAKSHIGHCEP GAGIIGIVSAMSRLGAVTVSSLQHLRTLNPHVEAIVKRL

Estas sequências foram submetidas ao SEARCHPKS, e o resultado pode ser visto na figura 4.4.

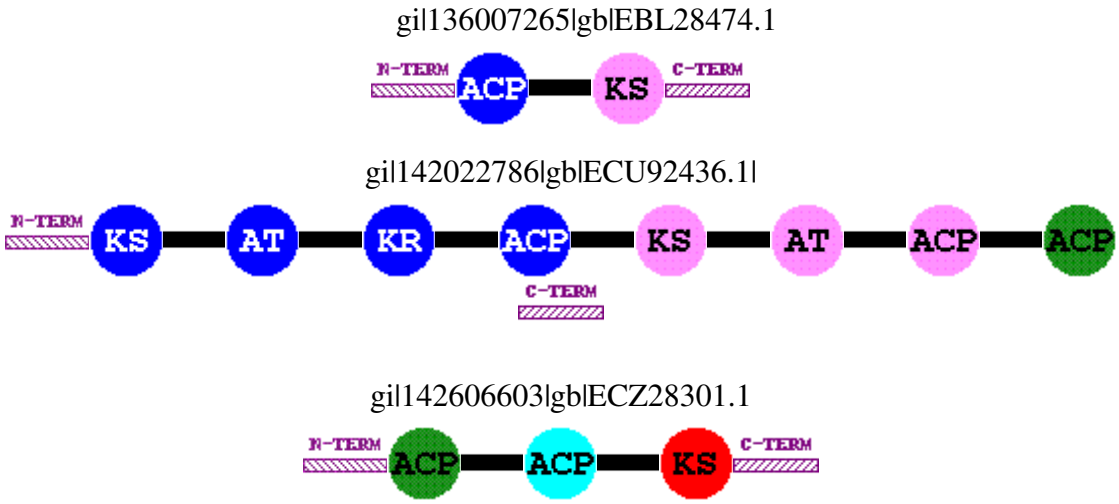


Figura 4.4: Domínios encontrados pelo programa SEARCHPKS nas três sequências similares aos 3 modelos utilizados no HMMER.

A sequência gi|142022786|gb|ECU92436.1, apresentou a maior similaridade com todos os modelos de PKSs tipo I modulares e da maior parte dos modelos de PKSs tipo I iterativas.

4.1.2 – Busca por PKSs tipo I Iterativas no banco ambiental do NCBI:

A lista com os 250 melhores hits entre cada sequência de região KS do IterDB e o banco ambiental do NCBI encontra-se no material suplementar (S1). Os 5 melhores hits de cada sequência foram extraídos com o FASTACMD e incluídos nas análises filogenéticas.

4.1.3 – Busca por PKSs tipo II no banco ambiental do NCBI:

Com o modelo gerado a partir do alinhamento de genes KS de PKSs tipo II, obtivemos um total de 2390 hits contra o banco ambiental do GenBank. As 10 sequências de maior similaridade (tabela S2) foram extraídas do banco com o FASTACMD e incluídas nas análises filogenéticas.

4.1.4 – Busca por PKSs tipo I modular no banco de proteínas do CAMERA:

Obtivemos um total de 14691 hits quando comparamos os 20 modelos da região KS do PKSDB contra o banco de proteínas do CAMERA, totalizando 1996 sequências com similaridade à PKS modular tipo I. Destes, 12,78% foram similares ao modelo de Mixalamida, 10,24% ao de Epotilona e 9,93% ao modelo de Mixotiazol. A figura 4.5 mostra a distribuição dos hits com os 20 modelos utilizados.

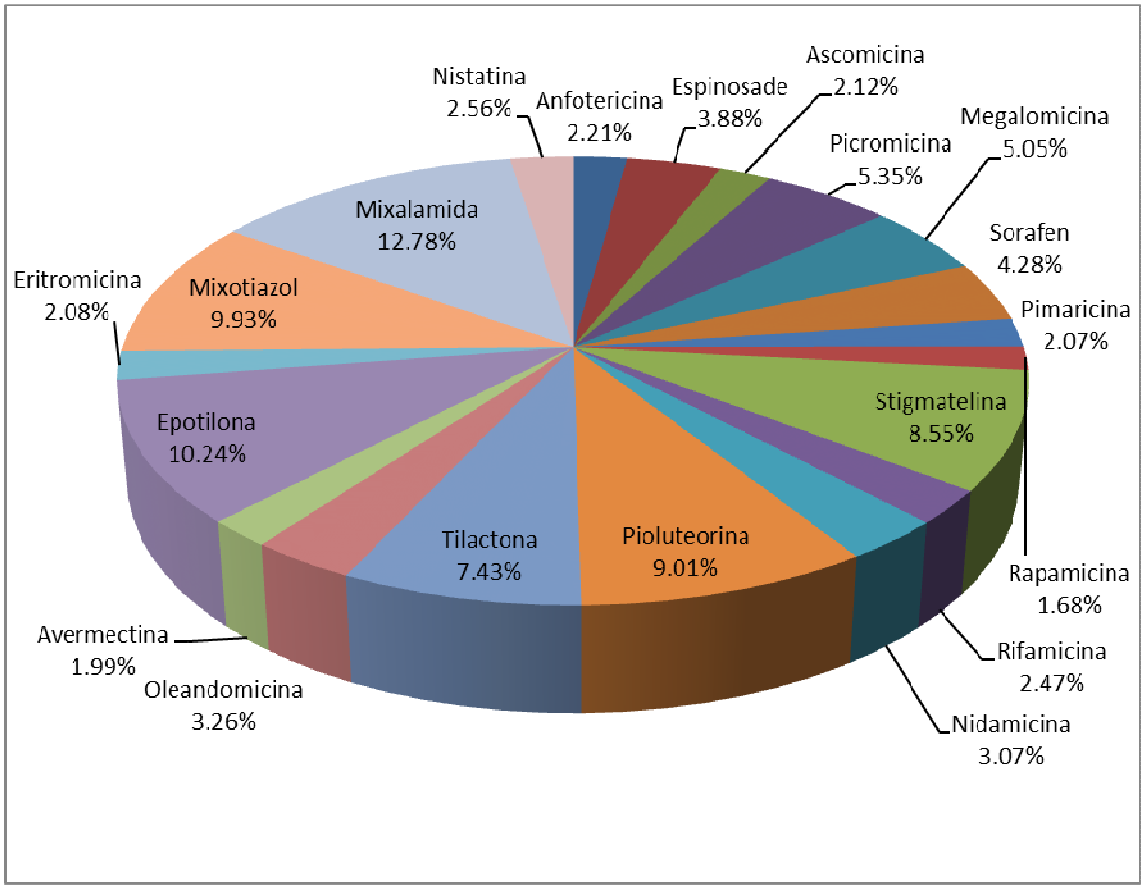


Figura 4.5: Distribuição dos hits obtidos na comparação entre os modelos da região KS de PKSs tipo I modulares (separados por metabólito produzido) contra o banco de proteínas do CAMERA.

Com os modelos da região AT, obtivemos um total de 12752 hits contra o banco do CAMERA, dos quais 12,71% similares ao modelo de PKS produtora de Anfotericina, 10,25% ao de Rifamicina e 10,00% ao modelo de Stigmatelina. A figura 4.6 exibe a distribuição dos hits contra os 20 modelos de PKSs do banco PKSDB.

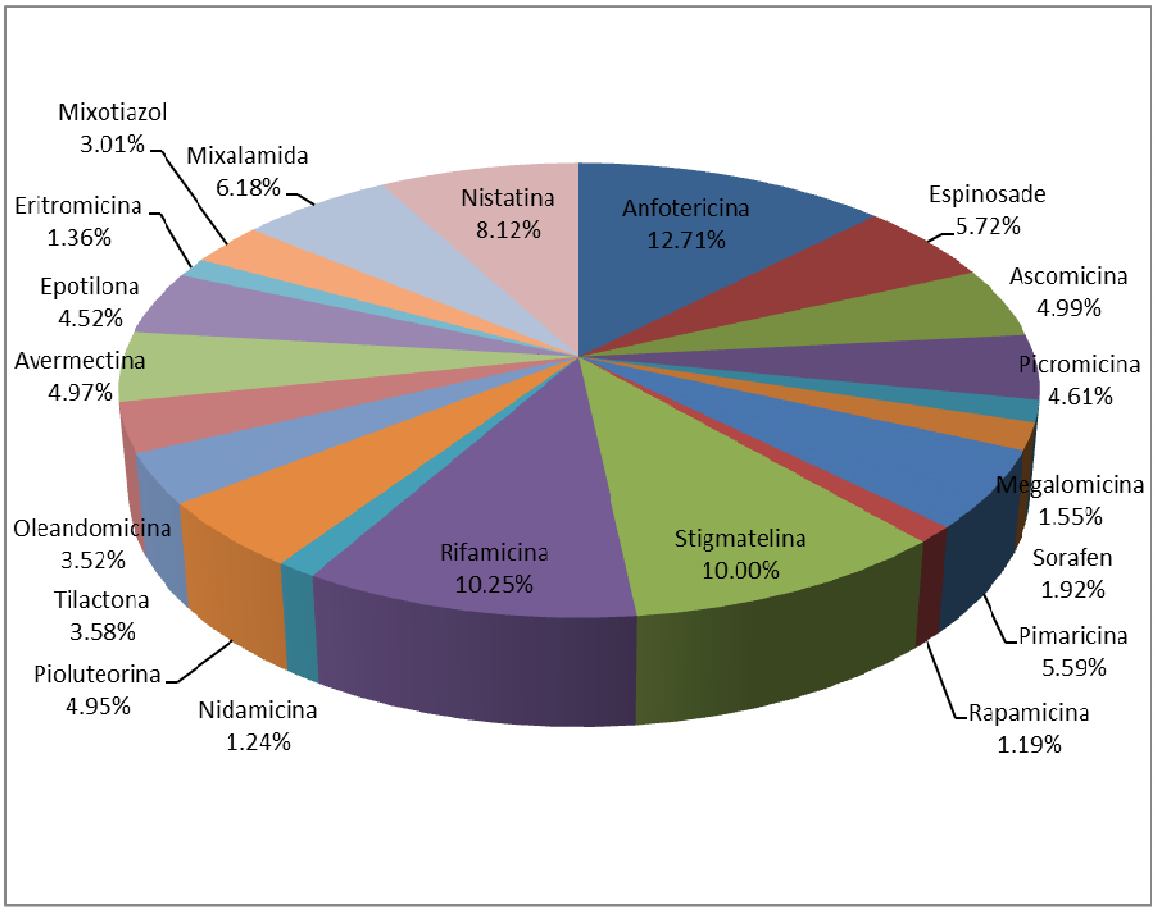


Figura 4.6: Distribuição dos hits obtidos na comparação entre os modelos da região AT de PKSs tipo I modulares (separados por metabólito produzido) contra o banco de proteínas do CAMERA.

A busca realizada com os modelos construídos a partir dos domínios ACP contra o banco de proteínas do CAMERA gerou um total de 1193 hits. Destes, 11.48% apresentaram similaridade ao modelo de PKS produtora de Epotilona, 10,81% são similares à Sorafen e 10,14% à Nidamicina. A figura 4.7 mostra a distribuição dos hits contra os 20 modelos.

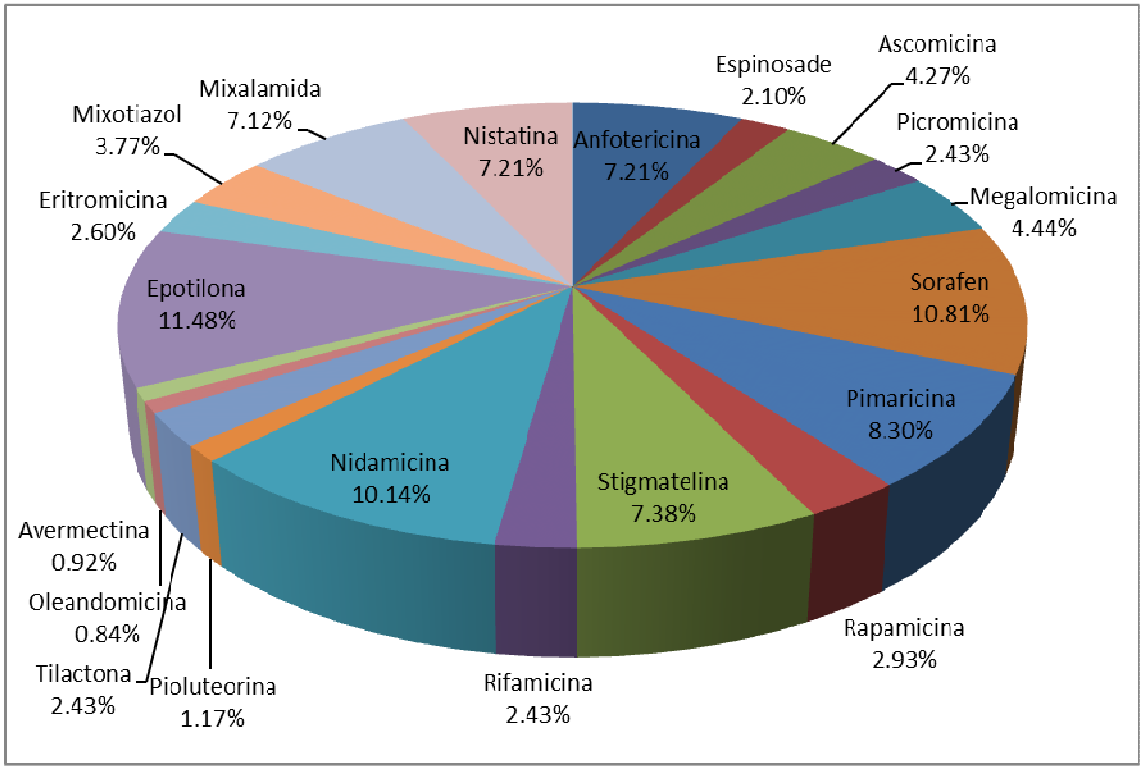


Figura 4.7: Distribuição dos hits obtidos na comparação entre os modelos da região AT de PKSs tipo I modulares (separados por metabólito produzido) contra o banco de proteínas do CAMERA.

A tabela 4.3 mostra a distribuição dos hits por modelo utilizado.

Tabela 4.3: Número de hits obtidos com o HMMER entre os modelos de PKSs tipo I iterativas e o banco de proteínas do CAMERA.

Metabólito produzido pela PKS	Domínio KS	Domínio AT	Domínio ACP
Anfotericina	324	1621	86
Espinosade	570	729	25
Ascomicina	311	636	51
Picromicina	786	588	29
Megalomicina	742	198	53
Sorafen	629	245	129
Pimaricina	304	713	99
Rapamicina	247	152	35
Stigmatelina	1256	1275	88
Rifamicina	363	1307	29
Nidamicina	451	158	121
Pioluteorina	1324	631	14
Tilactona	1091	457	29
Oleandomicina	479	449	10
Avermectina	292	634	11
Epotilona	1505	577	137
Eritromicina	305	174	31
Mixotiazol	1459	384	45
Mixalamida	1877	788	85
Nistatina	376	1036	86
TOTAL	14691	12752	1193

Apenas uma sequência do banco de proteínas do CAMERA apresentou similaridade com os 3 domínios de PKS, a mesma encontra-se na tabela 4.4.

Tabela 4.4: Sequência do banco CAMERA que apresentou similaridade com os modelos dos 3 domínios utilizados.

Identificador	Descrição	Sequência
JCVI_PEP_1 105139030331	/sample_name=GS000a /number_of_sites=2 /site_id_1=JCVI_SITE_GS000_S11 /location_1="Sargasso Station 11" /region_1="Sargasso Sea" /country_1=Bermuda /site_depth_1="5 m" /chlorophyll_density_1="0.17 (0.09+/-0.02) mg/M3" /salinity_1="36.7 ppt" /temperature_1="20.5 C" /water_depth_1=">4200 m" /site_id_2=JCVI_SITE_GS000_S13 /location_2="Sargasso Station 13" /region_2="Sargasso Sea" /country_2=Bermuda /site_depth_2="5 m" /chlorophyll_density_2="0.17 (0.09+/-0.02) mg/M3" /salinity_2="36.6 ppt" /temperature_2="20 C" /water_depth_2=">4200 m"	GGRLDAGTLDRDLGFFEMGMDSLMALD VRTHLENALGIPLSVALLFDHPTVNALAD FLAEQASGTAPDAHVAPAQAQSVPPPQQP RPVAPAIDAREAGTPEPIAIVGMSCRFPGA AHDLDAYWQLLNDGVDAISEVPRERWD VDAYYDPDPEAPGRMYSRFGGFLDDVD QFDPAFFRITPREAAAMDPQQRLLLEVSH EALEHAGIPVDSLKGSRTGVFVGITTNDY ANLQLRNGGGSGIDGYFFTGNPLNTAAG RISYGLGVQGPSMAIDTACSSSLTAIHTAS QNLRSGECDVAIAGGVNLILSPDNSFAVSR TRPLAPDGRCKTFDAAADGFLRNEGCRA AG

Esta sequência foi submetida ao SEARCHPKS, e o resultado pode ser visto na figura 4.8.

Figura 4.8: Regiões de PKSs identificadas pelo programa SEARCHPKS na similar aos modelos dos 3 domínios utilizados na busca com o HMMER contra o CAMERA.

.

4.1.5 – Busca por PKSs tipo I Iterativa no CAMERA:

Os 5 melhores hits entre cada sequência de KS tipo I iterativa e o banco do CAMERA pode ser visto em S3. As sequências foram extraídas do banco com o FASTACMD e incluídos nas análises filogenéticas.

4.1.6 – Busca por PKSs tipo II no banco de proteínas do CAMERA:

Obtivemos um total de 4950 hits entre o modelo de KS tipo II e o banco de proteínas do CAMERA. As sequências dos 10 melhores hits foram extraídas com o FASTACMD e incluídas nas análises filogenéticas. A tabela com os 10 melhores hits pode ser conferida no material suplementar S4.

4.2 – Parâmetros físico-químicos das amostras coletadas:

A tabela 4.5 mostra os parâmetros medidos imediatamente após cada coleta.

Tabela 4.5 – Medição de parâmetros físico-químicos das amostras coletadas. Temperatura, salinidade e pH medidos no momento de cada coleta

	Temperatura	Salinidade	pH
Coleta 1	25,40º C	35,12%	7,82
Coleta 2	23,10º C	36,34%	7,73
Coleta 3	24,20º C	36,52%	7,34
Coleta 4	22,45º C	38,45%	7,65
Coleta 5	21,23º C	35,84%	7,38
Coleta 6	23,01º C	36,39%	7,34
Coleta 7	22,52º C	36,08%.	7,78
Coleta 8	23,20 º C	36,21%.	7,50

4.3 – Obtenção de DNA de alto peso molecular:

Com as amostras obtidas nas coletas 1 a 6, não foi possível obter DNA com qualidade e quantidade suficiente para as análises posteriores.

A partir das 50 membranas utilizadas para filtrar a água da coleta 7, foram feitas duas extrações, uma com um total de 35 membranas (onde foram filtrados 100 litros de água), e outra com 15 membranas (onde foram filtrados 40 litros de água). A figura 4.9 mostra a presença de DNA com grande degradação e diversos pesos moleculares em gel de agarose 1%

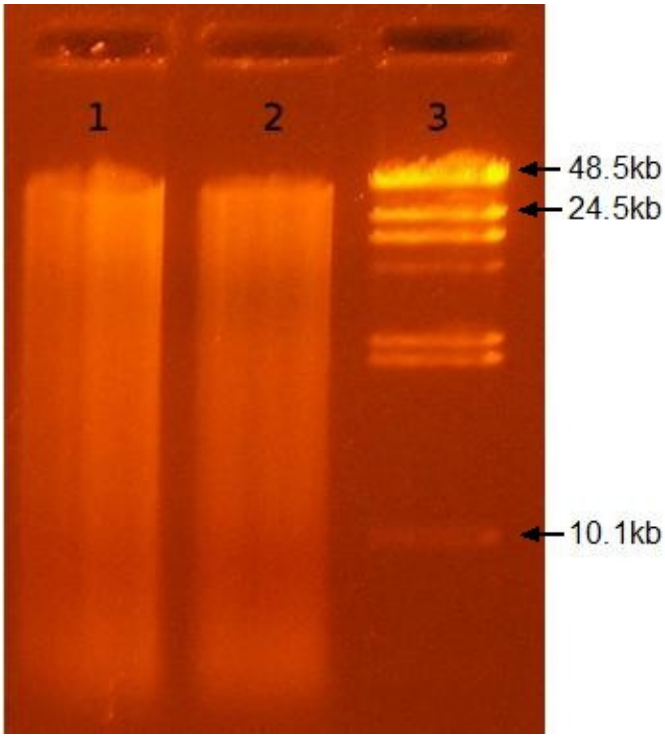


Figura 4.9: Gel de agarose 1% para verificação das extrações de DNA ambiental referente à amostra da coleta 7: 1 – Alíquota com 5 µl de DNA da primeira extração (a partir de 100 litros); 2 – Alíquota com 5 µl de DNA da segunda extração (a partir de 40 litros); 3 - Marcador de peso molecular High Range (500ng) (Fermentas).

Após a extração, todo o DNA ressuspendido em 140 µl de TE foi inserido em um poço grande em gel de agarose 1% de baixo ponto de fusão, e após a purificação da banda excisada, obtivemos um total de 2 µg de DNA com alto peso molecular (entre 25 e 40 mil pares de bases) como pode ser visto na figura 4.10.

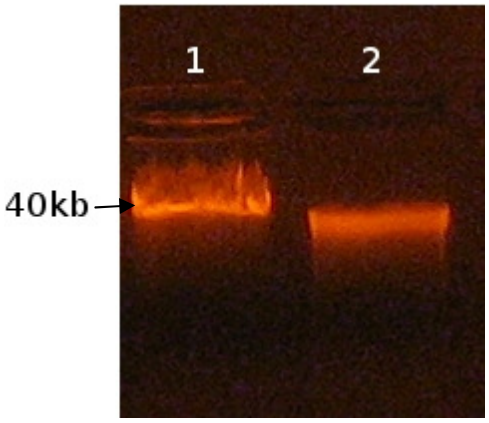


Figura 4.10: Gel de agarose 1% para verificação do DNA ambiental referente à amostra 7 após seleção e purificação de DNA de alto peso molecular: 1 – DNA controle EPICENTRE com 100 ng e 40kb; 2 – Alíquota com 1 µl do DNA ambiental com as pontas reparadas.

Como resultado da primeira extração realizada com as amostras da coleta 8, obteve-se um total de 1,75 µg de DNA, ressuspendido em 50 µl (35 ng/ µl), porém com degradação e fragmentos de tamanhos diversos (figura 4.11).

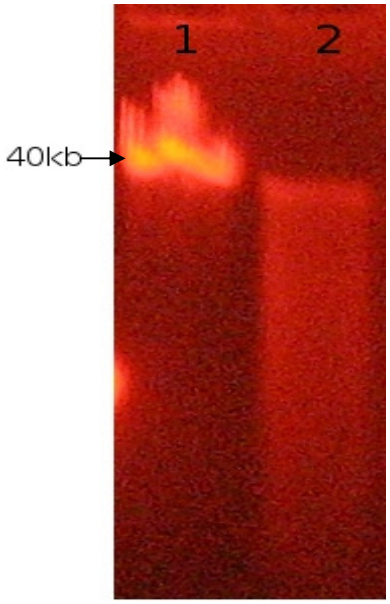


Figura 4.11: Gel de agarose 1% para verificação da primeira extração de DNA ambiental, referente à amostra da coleta 8: 1 - DNA controle do kit EPICENTRE com 100ng e 40kb; 2 - Alíquota com 2 µl da extração de DNA ambiental com o "Metagenomic DNA Isolation Kit for Water" da EPICENTRE.

Na segunda e na terceira extração, realizadas com uma membrana cada, obtivemos um total de 1,25 µg e 5 µg, respectivamente, ambos ressuspendidos em 50 µl de TE (figura 4.12).

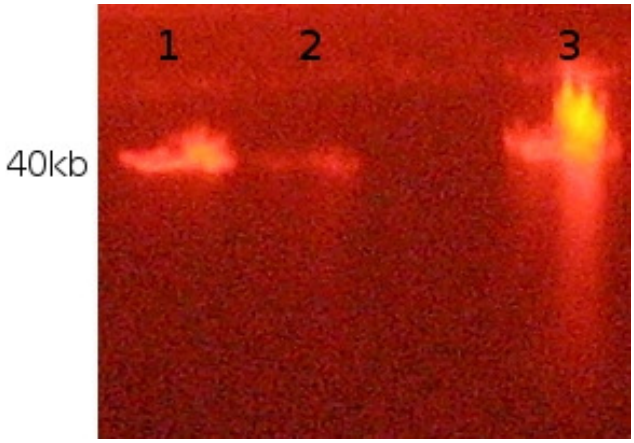


Figura 4.12: Gel de agarose 1% para verificação da segunda e da terceira extração de DNA ambiental referente à amostra da coleta 8: 1 - Alíquota do DNA controle do kit EPICENTRE com 100ng e 40kb; 2 - Alíquota com 2 µl do DNA da segunda extração realizada com o "Metagenomic DNA Isolation Kit for Water" da EPICENTRE; 3 – Alíquota com 2 µl do DNA da terceira extração realizada com o mesmo kit.

Com a amostra da segunda extração, na qual foi obtido DNA de maior peso molecular e pouco degradado, realizamos a reação de reparo das pontas segundo instruções do kit EPICENTRE, seguido de precipitação de DNA e verificação em gel de agarose 1% (figura 4.13). Obteve-se concentração de 55 ng/ µl e foram utilizados 6 µl desta amostra para a ligação ao vetor.

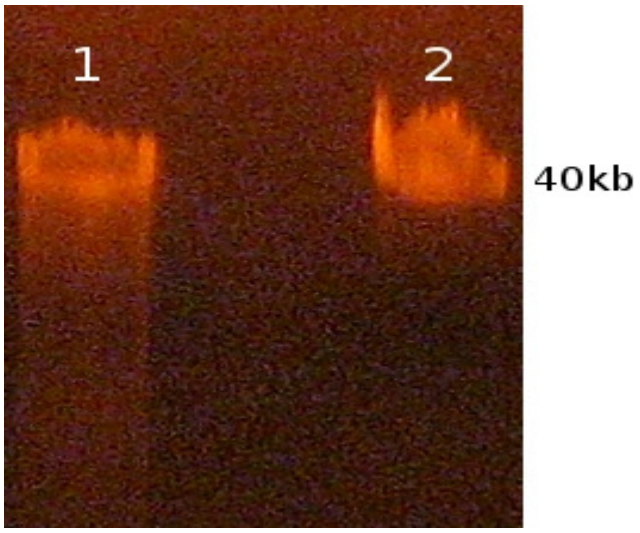


Figura 4.13: Gel de agarose 1% para verificação do DNA ambiental referente à amostra da coleta 8 (segunda extração), após reação de reparo das pontas e precipitação segundo instruções do kit EPICENTRE: 1 - Alíquota com 1 µl do DNA da segunda extração, após reparo das pontas e precipitação; 2 - DNA controle do kit EPICENTRE com 100ng e 40kb.

4.4 – Análises de biodiversidade baseada em rDNA

4.4.1 – Análise de sequências de rDNA 16S (bactérias).

Todas as análises de genes ribossomais foram realizadas com DNA obtido com a coleta 8.

Obtivemos um total de 225 sequências de rDNA bacteriano (16S), das quais 220 foram classificadas como bactérias utilizando o RDP classifier. A tabela 4.6 mostra a classificação filogenética obtida.

Tabela 4.6: Distribuição filogenética das sequências de 16s rDNA classificadas como bactérias através do RDP classifier.

Filo	%
Firmicutes	**0,9%**
Actinobactéria	**2,3%**
Verrucomicrobia	**0,5%**
Proteobactéria	**47,7%**
Bacteroidetes	**14,1%**
Cianobactéria	**30,5%**
Bactéria não classificada	**4,0%**

A figura 4.14 mostra a classificação das sequências quanto aos gêneros obtidos com o RDP classifier.

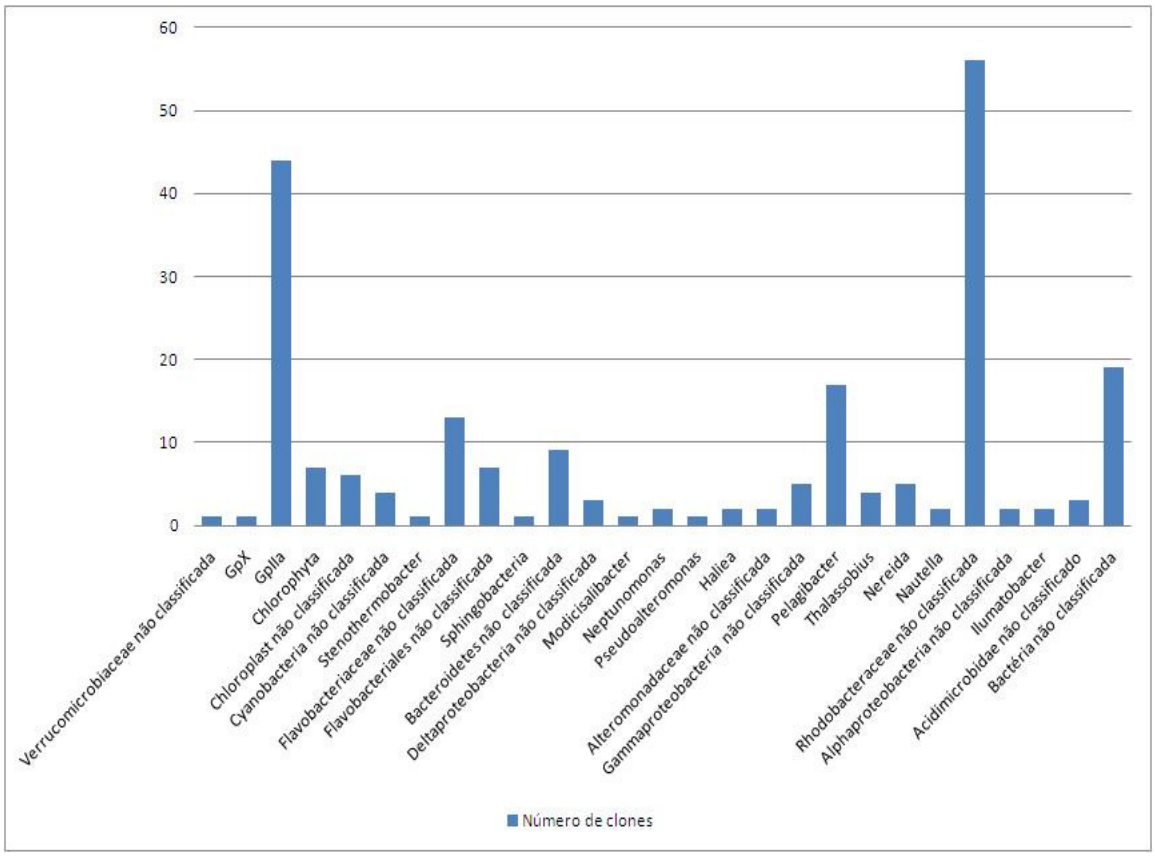
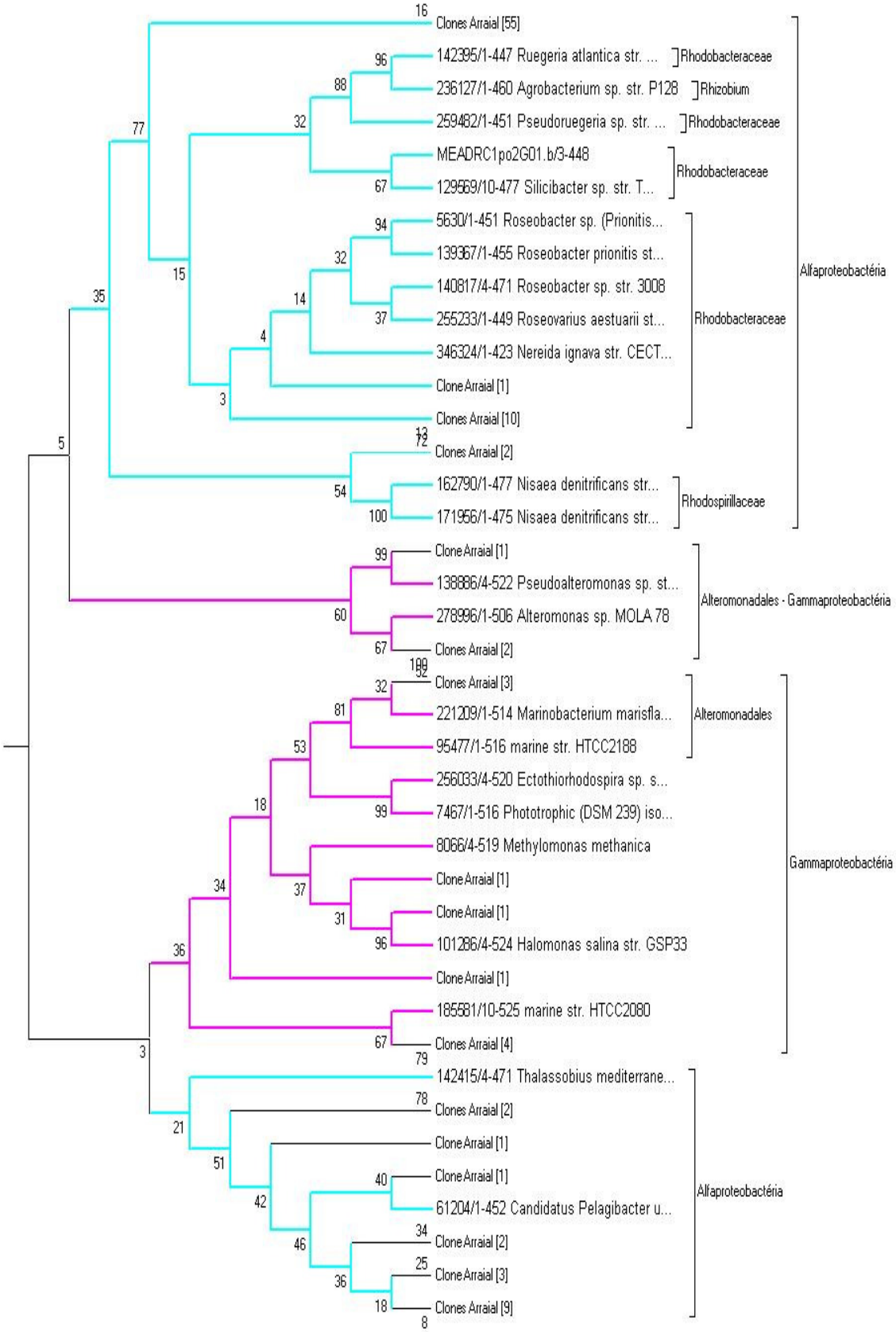
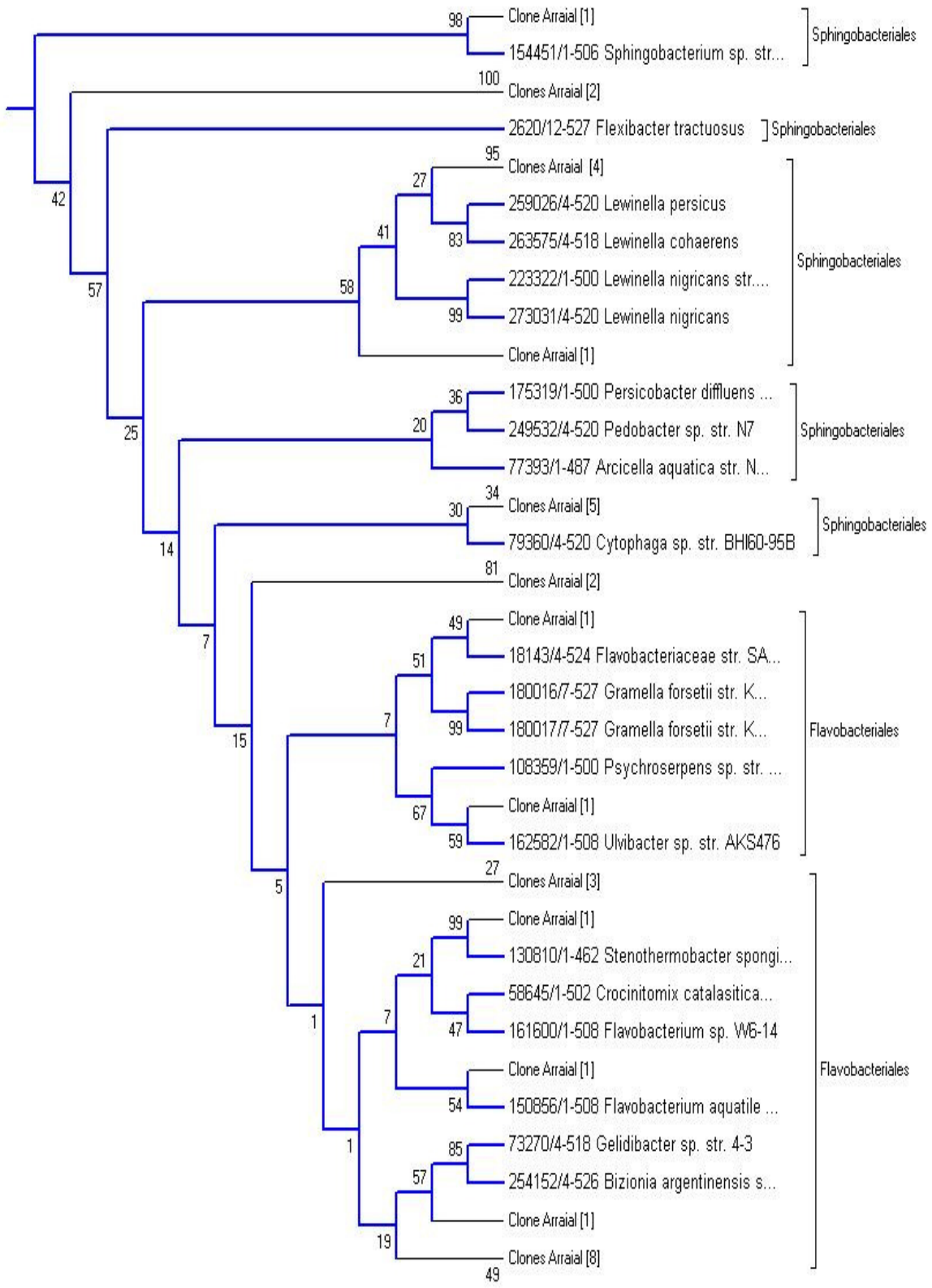


Figura 4.14: Classificação das sequências de rDNA 16S, amplificadas a partir da amostra de DNA ambiental referente à coleta 8, quanto ao gênero, utilizando RDP Classifier.

Já a figura 4.15 mostra um filograma construído com genes vizinhos obtidos no Greengenes. Os clados contendo apenas genes sequenciados neste estudo foram condensados.


Clones Arraial [55]
142395/1-447 Ruegeria atlantica str. ...
Rhodobacteraceae
236127/1-460 Agrobacterium sp. str. P128
Rhizobium
259482/1-451 Pseudoruegeria sp. str. ...
Rhodobacteraceae
MEADRC1po2G01.b/3-448
129569/10-477 Silicibacter sp. str. T...
Rhodobacteraceae
5630/1-451 Roseobacter sp. (Prionitis...
139367/1-455 Roseobacter prionitis st...
140817/4-471 Roseobacter sp. str. 3008
255233/1-449 Roseovarius aestuarii st...
346324/1-423 Nereida ignava str. CECT...
Clone Arraial [1]
Clones Arraial [10]
Rhodobacteraceae
Alfaproteobactéria
Clones Arraial [2]
162790/1-477 Nisaea denitrificans str...
171956/1-475 Nisaea denitrificans str...
Rhodospirillaceae
Clone Arraial [1]
138886/4-522 Pseudoalteromonas sp. st...
278996/1-506 Alteromonas sp. MOLA 78
Clones Arraial [2]
Alteromonadales - Gammaproteobactéria
Clones Arraial [3]
221209/1-514 Marinobacterium marisfla...
95477/1-516 marine str. HTCC2188
Alteromonadales
256033/4-520 Ectothiorhodospira sp. s...
7467/1-516 Phototrophic (DSM 239) iso...
8066/4-519 Methylomonas methanica
Clone Arraial [1]
Clone Arraial [1]
101286/4-524 Halomonas salina str. GSP33
Clone Arraial [1]
185581/10-525 marine str. HTCC2080
Clones Arraial [4]
Gammaproteobactéria
142415/4-471 Thalassobius mediterrane...
Clones Arraial [2]
Clone Arraial [1]
Clone Arraial [1]
61204/1-452 Candidatus Pelagibacter u...
Clone Arraial [2]
Clone Arraial [3]
Clones Arraial [9]
Alfaproteobactéria


Clone Arraial [1]
154451/1-506 Sphingobacterium sp. str...
Sphingobacteriales
Clones Arraial [2]
2620/12-527 Flexibacter tractuosus
Sphingobacteriales
Clones Arraial [4]
259026/4-520 Lewinella persicus
263575/4-518 Lewinella cohaerens
223322/1-500 Lewinella nigricans str....
273031/4-520 Lewinella nigricans
Clone Arraial [1]
Sphingobacteriales
175319/1-500 Persicobacter diffluens ...
249532/4-520 Pedobacter sp. str. N7
77393/1-487 Arcicella aquatica str. N...
Sphingobacteriales
Clones Arraial [5]
79360/4-520 Cytophaga sp. str. BHI60-95B
Sphingobacteriales
Clones Arraial [2]
Clone Arraial [1]
18143/4-524 Flavobacteriaceae str. SA...
180016/7-527 Gramella forsetii str. K...
180017/7-527 Gramella forsetii str. K...
108359/1-500 Psychroserpens sp. str. ...
Clone Arraial [1]
162582/1-508 Ulvibacter sp. str. AKS476
Flavobacteriales
Clones Arraial [3]
Clone Arraial [1]
130810/1-462 Stenothermobacter spongi...
58645/1-502 Crocinitomix catalasitica...
161600/1-508 Flavobacterium sp. W6-14
Clone Arraial [1]
150856/1-508 Flavobacterium aquatile ...
73270/4-518 Gelidibacter sp. str. 4-3
254152/4-526 Bizionia argentinensis s...
Clone Arraial [1]
Clones Arraial [8]
Flavobacteriales

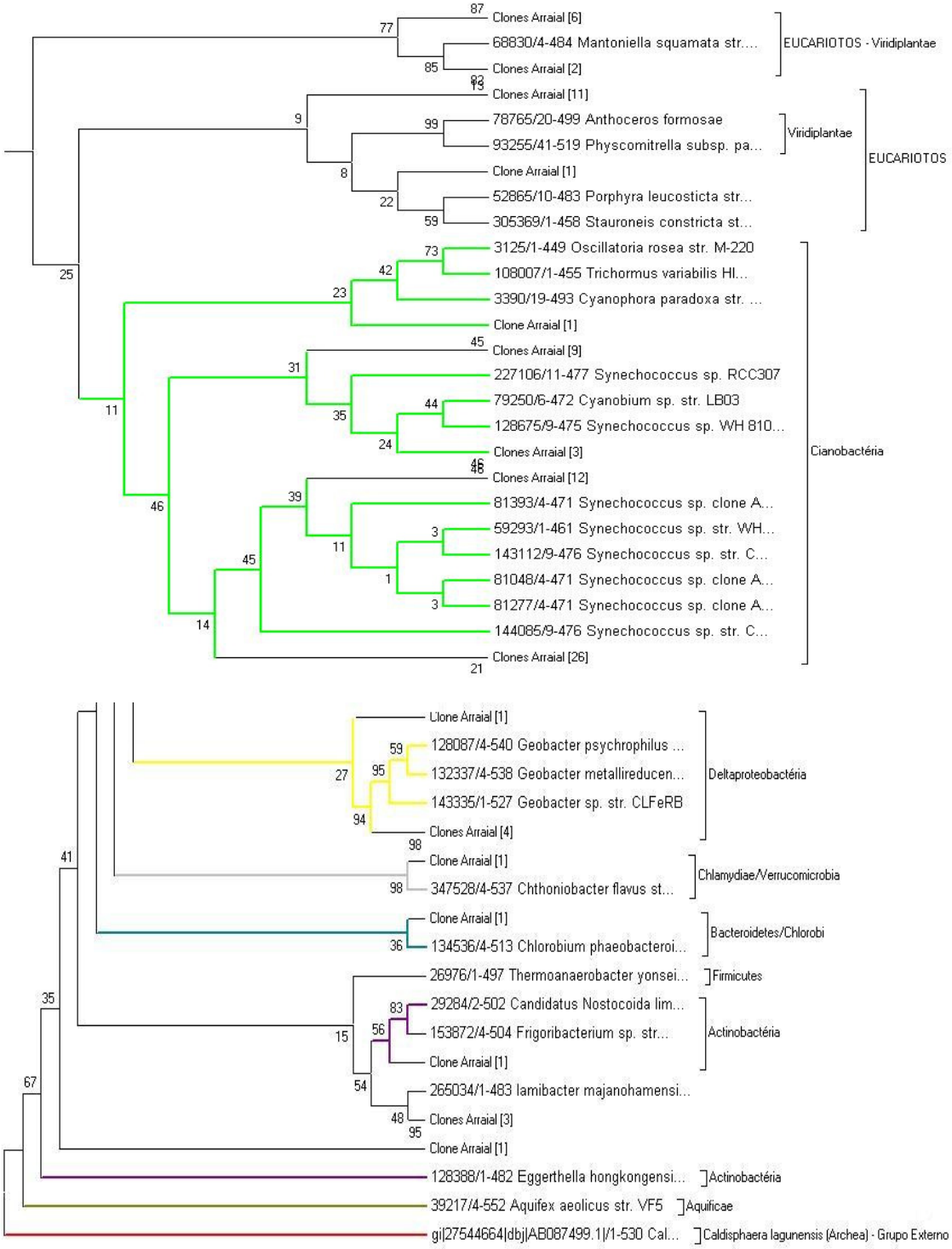
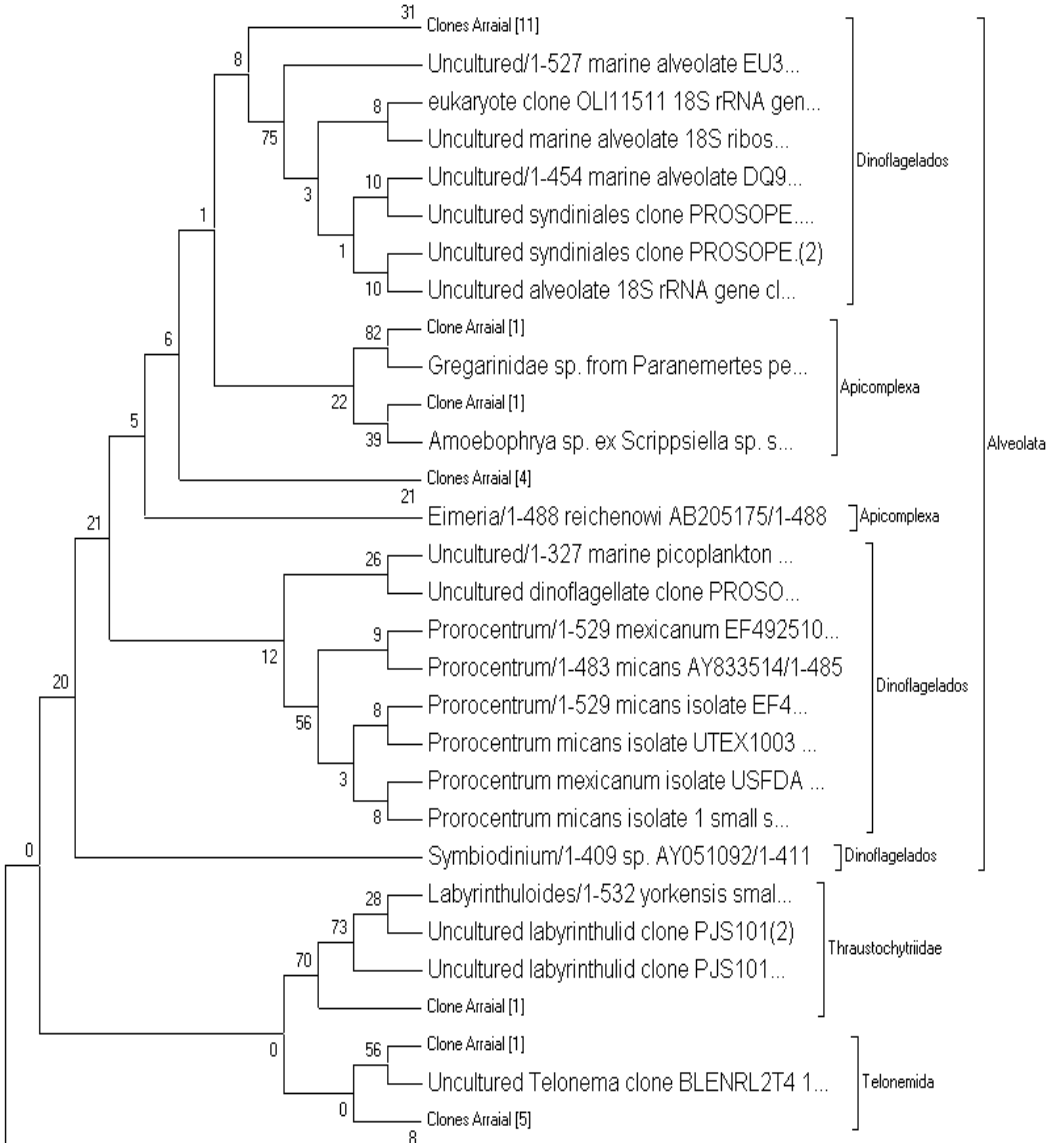
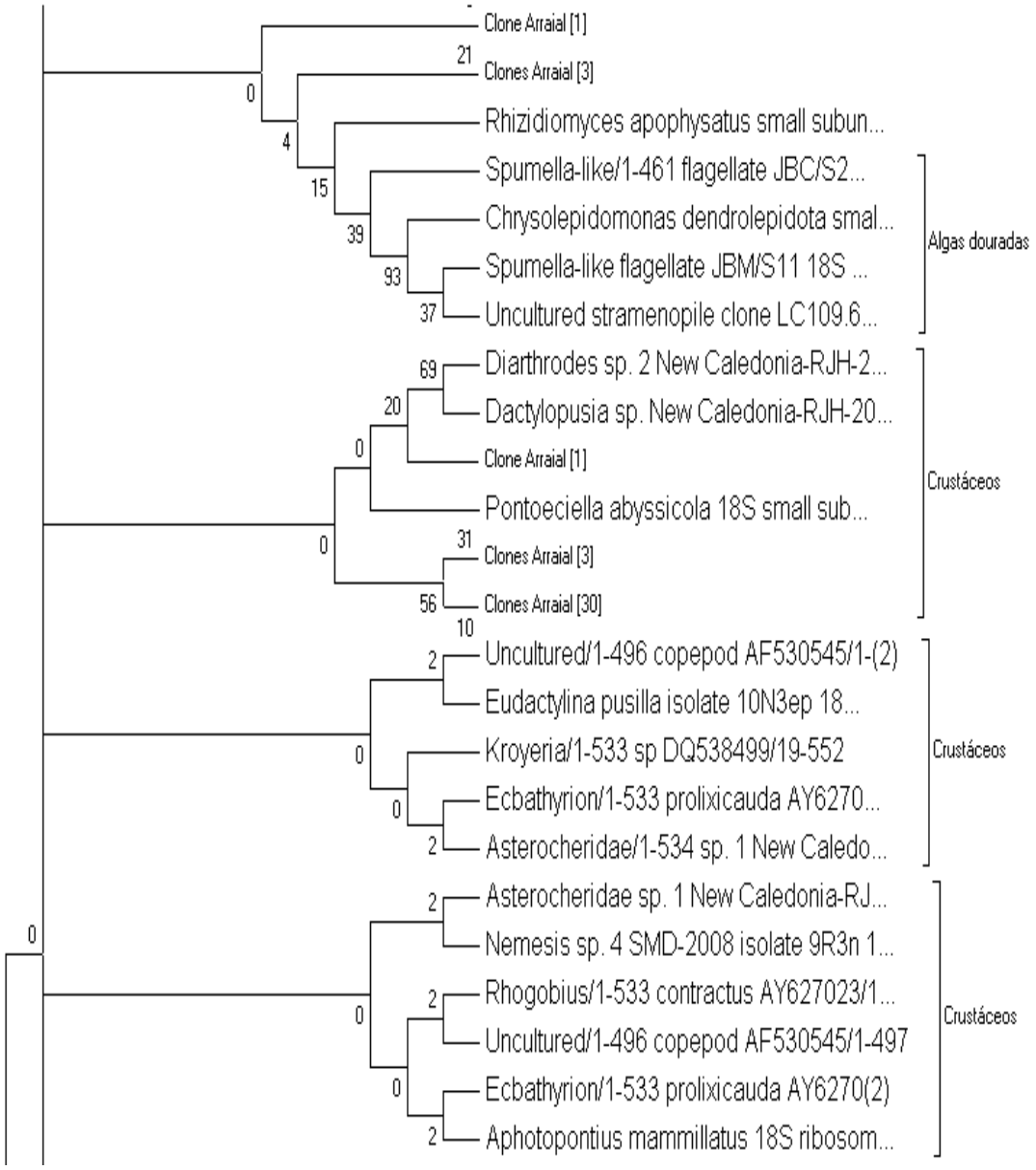


Figura 4.15: Árvore filogenética com ramos condensados, construída com o programa MEGA 4.0 utilizando método de agrupamento com vizinhos e análise de *bootstrap* com valor 100, exibindo agrupamentos dos genes de 16S ambientais obtidos em Arraial do Cabo com vizinhos obtidos no Greengenes. Ramos em azul claro exibem clados com sequências de Alfaproteobactérias, em rosa os clados com Gamaproteobactérias, em azul escuro as Bacteriodetes, em verde as Cianobactérias, em amarelo as Deltaproteobactérias, em cinza as sequências de Chlamydiae, em roxo as Actinibactérias e em vermelho o grupo externo (archeas). Os clados que possuem apenas as sequências obtidas neste estudo foram condensados e os valores entre chaves se referem ao número de sequências por clado.

4.4.2 – Análise de genes de 18S (eucariotos):

Foram obtidas 103 sequências com qualidade (PHRED>15) e tamanho adequado (cerca de 400 pares de base) para as análises. As mesmas foram submetidas ao BLAST (blastn) contra o banco NR do NCBI. Uma árvore foi construída com vizinhos obtidos no SILVA e com os 3 melhores hits obtidos com o BLAST contra o NR (Figura 4.16).




Clone Arraial [1]
Clones Arraial [3]
Rhizidiomyces apophysatus small subun...
Spumella-like/1-461 flagellate JBC/S2...
Chrysolepidomonas dendrolepidota smal...
Spumella-like flagellate JBM/S11 18S ...
Uncultured stramenopile clone LC109.6...
Algas douradas
Diarthrodes sp. 2 New Caledonia-RJH-2...
Dactylopusia sp. New Caledonia-RJH-20...
Clone Arraial [1]
Pontoeciella abyssicola 18S small sub...
Clones Arraial [3]
Clones Arraial [30]
Crustáceos
Uncultured/1-496 copepod AF530545/1-(2)
Eudactylina pusilla isolate 10N3ep 18...
Kroyeria/1-533 sp DQ538499/19-552
Ecbathyrion/1-533 prolixicauda AY6270...
Asterocheridae/1-534 sp. 1 New Caledo...
Crustáceos
Asterocheridae sp. 1 New Caledonia-RJ...
Nemesis sp. 4 SMD-2008 isolate 9R3n 1...
Rhogobius/1-533 contractus AY627023/1...
Uncultured/1-496 copepod AF530545/1-497
Ecbathyrion/1-533 prolixicauda AY6270(2)
Aphotopontius mammillatus 18S ribosom...
Crustáceos

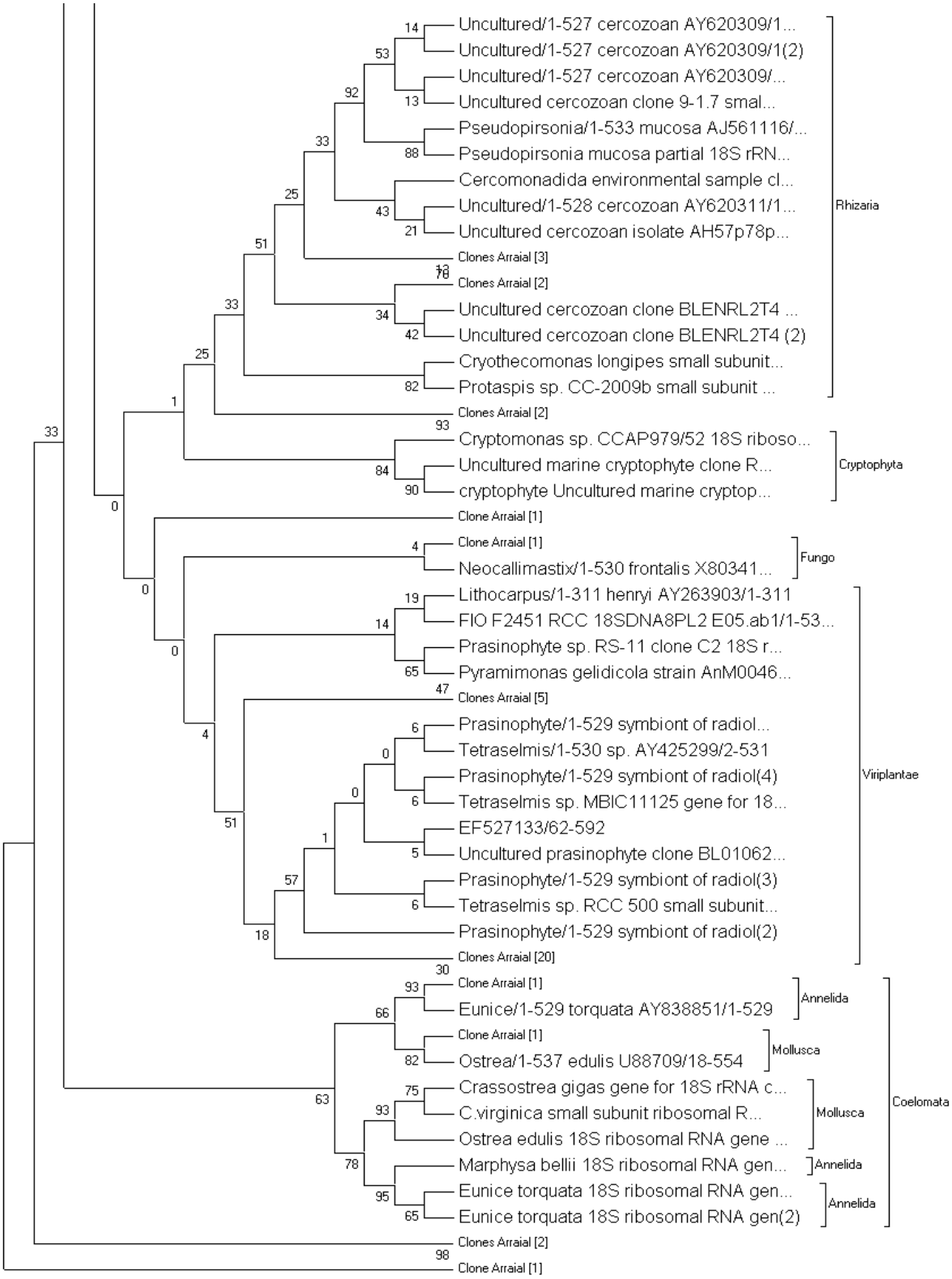


Figura 4.16: Árvore filogenética com ramos condensados, construída com o programa MEGA 4.0 utilizando método de agrupamento com vizinhos e análise de *bootstrap* com valor 100, exibindo o agrupamento das sequências de rDNA 18S com vizinhos do banco de dados SILVA e obtidos com a busca utilizando o programa BLAST (blastn) contra o banco de sequências nucleotídicas não redundante do NCBI (NR). Os clados que possuem apenas as sequências obtidas neste estudo foram condensados e os valores entre chaves se referem ao número de sequências por clado.

4.5 – Amplificação e sequenciamento de regiões KS de PKSs ambientais

Com o DNA extraído a partir da amostra da coleta 7, obtivemos amplificação e o produto de PCR apresentou o tamanho esperado (700pb) para o par de iniciadores utilizado (figura 4.17). Já com o DNA obtido na coleta 8, não houve amplificação. Em ambos os casos, o controle positivo (DNA genômico de *Streptomyces*) apresentou amplificação de DNA com tamanho esperado.

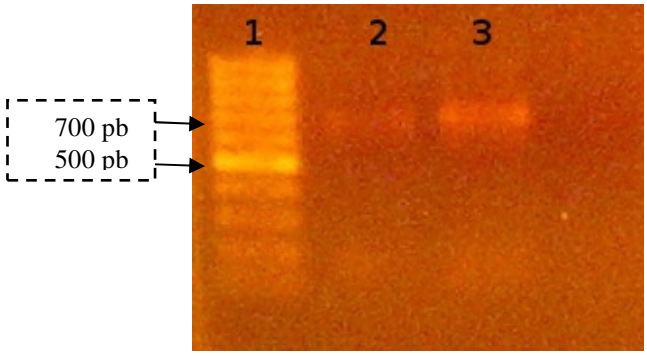


Figura 4.17: Gel de agarose 1% para verificação da reação de PCR realizada para amplificação da regiao KS de PKS tipo I a partir da amostra de DNA ambiental da coleta 7: 1 – Marcador de peso molecular "100bp" (500ng) (Fermentas); 2 – Alíquota da reação de PCR realizada com DNA ambiental da coleta 7, exibindo amplificação de DNA com 700 pares de base; 3 – Alíquota da reação de PCR realizada com o controle positivo (DNA de actinomiceto produtor de PKS tipo I) exibindo amplificação de DNA com 700 pares de base.

A partir da clonagem do produto de PCR amplificado com o DNA da coleta 7, obtivemos um total de 96 colônias brancas . Estas foram submetidas ao sequenciamento, e foram obtidas 7 sequências com qualidade (PHRED>15) (tabela 4.7). Destas, 5 possuem similaridade com regiões KS de PKSs tipo I. A tabela 4.8 mostra os 3 melhores hits obtidos com o blastp entre as 5 sequências e o banco NR do NCBI.

Tabela 4.7: Sequências de DNA obtidas a partir do sequenciamento do produto de PCR realizado para amplificação de região KS com o DNA da amostra 7.

Clone	Sequência
MEADPKSKS7H11	CCTTTAGGGGCCATGGGCCTCGACGAATGGAACATCCCCAGCACCTTTACGGCCA TCGCTCAAAGCGGCCTTAATCACAGCACGTTGCGCTGCGCCGTTGGGCGCCAAA ATCCCAAAAGTCCTTCCACCATGATTAACAGCGGATCCGCGAATTACCGCAAAAA TCCGATCGTTGTCACGGATGGCATCAAACATTCGTTTTAACAAAACACCTCCGCA CCCCTCACCTCGCACATAGCCGTGGGCGGACTCGTCAAAGGGCCAACATGCGCC TGTGGGAAAGAGCATTCTGGCTTTGGAAAGTGAGATGAAGGTATCCGCCAACAA ACACACGTTTACACCGCCGGCGAGCGCTTTTTCTCACTCGCGTCTTCGCAAACTC TGAATCGCATCATGAACCGCAACCAGTGCTGAACTGCATGCAGGGTCAATCGACT TCGATGGACCTGTAAAATTGAAAACATAGGAAAGTCGGTTTGCACAGATCGAGA GGGCTCCGCCCGCCCCA
MEADPKSKS7H10	CCTTTAGGGGCCATGGGCCTCGACGTATTGAACATCCTCAGCACCCTTACGTCCA TCGCTCAAAGCGGCCTTAATCACAGCACGTTGCGCTGCGCCGTTGGGCGCCAAA ATCCCAAAAGTCCTTCCACCATGATTAACAGCGGATCCGCAAATTACCGCAAAAA TCCGATCGTTGTCACGGATGGCATCAAACATTCGTTTTAACAAAACACCTCCGCA CCCCTCACCTCGCACATAGCCGTTGGCGGACTCGTCAAAGGGCCAACATGCGCC TGTGGAAGAGAGCATTCTGGCTTTGGAAAGTGAGATGAAGGTATCCGCCAACAA ACACACGTTTACACCGCCAGCGAGCGTTTTTTCACACTCGCGTCTTCGCAAACTC TGAATCGCATCATGAACCGCACCCAGGGCTGAACTGCATGCAGTGTCAATCGACT TCAAGGGACCTGTAAAATTGAAAACATAGGAAAGTCGGTTTGCACAGATCGAGA GGGCTCCCCCCGTCCCACA ATGCGCATCC AAGTGATAAA TTGA
MEADPKSKS7F11	CGCGGCCGCCGGCAGGCCGACCAAAGGGGAGAGCCCCCAACGCGGGGGAGGC AAAGCCCGAGAAAACAAAAGGGCCACCAAAAAAGCGGGGCGAAACCAGGGGC AAAGCGGGCCCCGGGGGGAAAAGGAAACCCGCCCACAACCCCCCA
MEADPKSKS7E06	AATTGCTTTGGATCCCCACCACCGGGTGTTGCTGACGTGTTCTTCCCCCGCGATG GAAAACGCAAATATTGATCCGGTCTCGTTGTCCCAAAGCCCGACGGGCGTGTTCG TTGGGGCAGGTCAAACGGATTACTCCCGTGTGATGAGCCATTTTGATAACTCATTT TATCACTTGGATGCCCATTGGGGGACGGGCGGACCCCTCTCGATCTGTGCAAACC GACTCTCCTATGTCTTCAATTTTACAGGTCCATCAAATTCAATTGACACTGCATGC ATTTCACCACTGGTTGCGGTTCATGATGCAATTCAAATTTTGCAAAAACCCGAGT GTGACGCTGCCCTCGCTGGCGGTGTAAACGTGTGTCTGTCGGCGGATACCTTCTT CTCACTTTCCAAAGCCAGAATGCTCTCTCCCACAGGCGCATGTCGGCCCTTTGAT GAGTCCGCCAACGGCTATGTGCGAGGTGAGGGGTGCGGAGTTGTTCTGTTAAAA CGAAGGTCTGATGCCATCCGTGACACCGATCGGATTT
MEADPKSKS7E10	TCGTTTCCGCAGCAGCGGCTGCCATTAAAAACATCATGGGAGGCCCTAAAAGATG CCGGCATTCCACCCTCTAGTTTATTCGAGTCAAATACCGGGGTTTTCACCGCCATC TTAAATCATGACTACTCTGATTTGATGTTAATGAAAGGTCTAAAAAAATATGCAAA CCCATACTCGGCATTAAGTTATTGGGGTTGTATTGCCGCAGGTAAAATATCGTATTT TTTAGGCTTAAATGGCCCAAGCTTGGCTGTTGATACCGGGGGCTCAGCGTCTATCA TCAGCGTACACAAAGCGGGCAAAAGCTTACGTAACCAAAAAAGTGATATCGCAC TTGCAGGAGGGTGTCACGTTGTGTTTTCCCCAAAGCGGGTCATGAATTATTGCCG

	CGTTGGTGTGTTATCCCCTAGCGGCATGTGTAAAACGTTTTCTGATGATGCCGATG GGTTCGCCAAAGGAGAAGGTTGCGGCATACTGGTTTTAAAACGCTTAAGTGATGC ATTGAGTCATGGTGATAGAATTTATGCCGGGGTAAAAGGCACAGCGATAAACCAC GATGGTGCCAGTGCAGGATTAACTGTGCCAAGGGGGCCTGCACAAGAAAAAGTG ATCAAGGCCGCATTACATCATGCCGGATTAAAGGCGACCGATATCGATTACGTCAA AGC
MEADPKSKS7H05	AATTCTGTCAGTGCCGTGGCCCTCGACGGTTAAGGCGCGGGCGCTGCTTTCAAA CTCTCTCAGCATGTTGTAGGTGCGCTGGAACTGCTGCGCGTTGGGACCGCGCTGG CTGCGGATTTGCGGTACCAGGGTGTTGGCTATAACGTGGGTGAGTACATCCTCGA CCGTCTCGGGACCGCCGCGGTTACGCATCCTCAGCCGCGGCAGTCCGGCGGCTA TTGCCGCGCGG
MEADPKSKS7F02	GCAGCACCGGCTGCCATTAAAAACATCAGGGGAGGCCCTTAAAAATGCCGGCTT TCCACCCTCTAGTTTATTCAAGTCAAATACCGGGGTTTTCACCGCCATCTTAAATC ATGACTACTCTGATTTGATGTTAATGAAAGGTCTAAAAAAATATGCAAACCCATAC TCGGCATTAATTTATTGGGGTTGTATTGCCGCAGGTAAAATATCGTATTTTTTAGGC TTAAATGGCCCAAGCTTGGCTGTTGATACCGGGGGCTCAGCGTCTATCATCAGCG TACACAAAGCGTGCAAAAGCTTACGTAACCAAAAAAGTGATATCGCACTTGCAG GAGGGTGTCACGTTTTGTTTTCCCCAAAGCGGGTCATGAATTATTGCCGCGTTGG TGTGTTATCCCCTAGCGGCATGTGTAAAACGTTTTCTGATGATGCCAATGGGTTCG CCAAAGGAAAAGGTTGCGGCATACTGTTTTTAAAACGCTTAAGTGATGCATTGAG TCATGGTGATAAAATTTATGCCGGGGAAAAAGGCACACCAATAAACCACGATGGT GCCAGTGCAGGATAAACGGTGCCAA

Tabela 4.8: Três melhores hits, obtidos com o programa BLAST (blastn), entre cada sequência obtida na amplificação de regioões KS a partir da amostra da coleta 7 e o banco de sequências não redundantes (NR) do NCBI.

Clone	Hit	Score	E-value
ks7E06.b_2	ref|NP_870253.1| polyketide synthase [Rhodopirellula baltica ...	280	1,00E-73
	ef|YP_002378014.1| Erythronolide synthase., Oleoyl-(acyl-car...	263	2,00E-68
	gb|ACC99565.1| type I polyketide synthase [uncultured bacterium]	262	2,00E-68
ks7E10.b_1	gb|AAW84213.1| modular polyketide synthase [uncultured bacter...	236	1,00E-60
	gb|AAS98783.1| JamL [Lyngbya majuscula]	233	1,00E-59
	gb|AAT70105.1| CurJ [Lyngbya majuscula]	229	3,00E-58
ks7H10_4	ref|NP_870253.1| polyketide synthase [Rhodopirellula baltica ...	207	7,00E-52
	ref|ZP_05029386.1| Beta-ketoacyl synthase, N-terminal domain ...	196	2,00E-48
	dbj|BAF68998.1| polyketide synthase [Microcystis aeruginosa]	195	3,00E-48
ks7H11.b_4	ref|NP_870253.1| polyketide synthase [Rhodopirellula baltica ...	252	2,00E-65
	gb|ACC99565.1| type I polyketide synthase [uncultured bacterium]	228	4,00E-58
	dbj|BAF68998.1| polyketide synthase [Microcystis aeruginosa]	227	8,00E-58
ks7F02.b_2	gb|AAS98783.1| JamL [Lyngbya majuscula]	192	3,00E-47
	gb|AAT70105.1| CurJ [Lyngbya majuscula]	189	2,00E-46
	ref|YP_001865644.1| beta-ketoacyl synthase [Nostoc punctiform...	186	2,00E-45

4.6 - Construção da biblioteca metagenômica

A partir das amostras da coleta 7, obteve-se um total de 501 clones com insertos de aproximadamente 40 mil pares de bases (40kb), distribuídos em 5 placas de 96 poços e 21 tubos de criopreservação.

Já com o material da coleta 8, obteve-se um total de 3500 clones, estocados em placas de 96 poços. Ambas as bibliotecas foram criopreservadas em solução de glicerol e estocadas em -80º C.

4.7 – Extração de fosmídeos dos clones da biblioteca

Com o método de lise alcalina modificado, obtivemos para cada extração a partir de 6 ml de cultura de cada clone um total de 50 µl com aproximadamente 300ng/µl de DNA. Quando foi utilizado um volume total de 48 ml em 8 tubos, e o DNA total dos 8 tubos foram ressuspendidos em 50 µl de água mili-Q, obtivemos uma concentração de 5,5 µg/ µl.

4.8 – Triagem das bibliotecas em busca de PKSs

Com a metodologia utilizada neste estudo, não foi possível localizar em quais clones existem PKSs nas bibliotecas construídas, pois os iniciadores exibiram amplificação inespecífica, para todos os "pools" e clones testados.

4.9 – Análises filogenéticas das regiões KS ambientais

Foram construídas 5 árvores filogenéticas. Em todos os casos o modelo evolutivo sugerido pelo MODELGENERATOR foi o de WAG (Whelan and Goldman, 2001).

A primeira árvore (figura 4.18) mostra a evolução das PKSs tipo I (iterativas e modulares), tendo como raiz as fabB e fabF. As PKSs tipo I iterativas de bactéria aparecem primeiro, seguidas pelas iterativas de fungo e pelas modulares.

A figura 4.19 mostra a árvore filogenética com as sequências da região KS do

PKSDB e do IterDB recortadas apenas na região alinhada com as KS sequenciadas neste estudo. Podemos observar que das 5 sequências de KS obtidas com a amostra de Arraial do Cabo, 3 estão mais próximas de PKSs iterativas e 2 de PKSs modulares.

Com as sequências de região KS obtidas na triagem do banco ambiental do NCBI e do CAMERA, obtivemos as árvores exibidas nas figuras 4.20 e 4.21 respectivamente, nas quais podemos observar diversas sequências agrupadas com PKSs tipo I iterativas e modulares.

Com as sequências do CAMERA e do NCBI que apresentaram similaridade ao modelo de região KS de PKSs tipo II , foi construída uma árvore, utilizando sequências de fabH bacterianas como raiz. A figura 4.22 mostra a filogenia destas sequências perante as sequências curadas obtidas no NCBI.

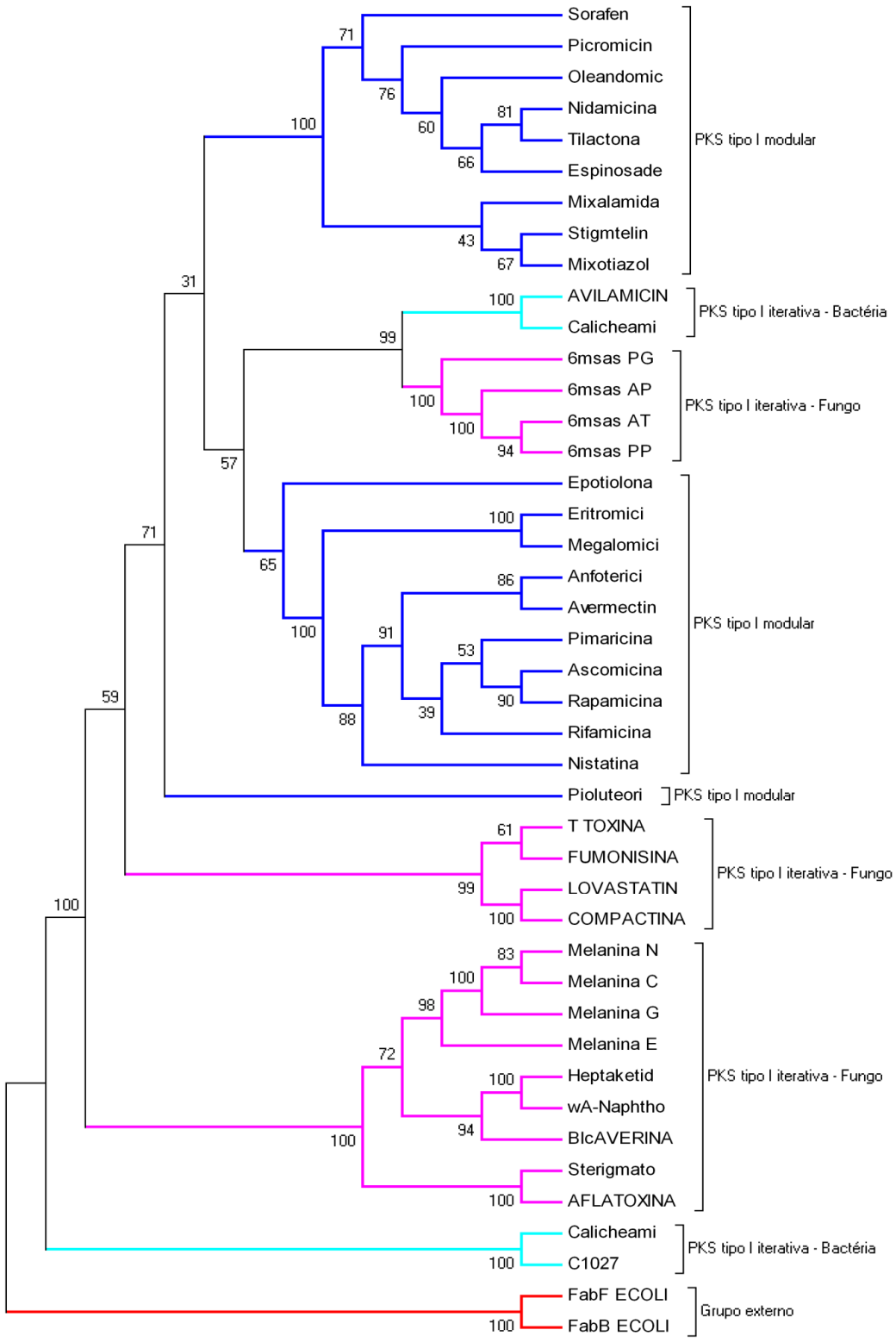


Figura 4.18: Árvore filogenética de regiões KS, construída com o programa PHYLM, utilizando modelo evolutivo de WAG e análise de *bootstrap* com valor 100, exibindo a evolução das PKSs tipo I iterativas e modulares em fungos e bactérias obtidas nos bancos curados PKSDB e IterDB. Em azul escuro, os clados com PKSs tipo I modulares, em azul claro os clados com PKSs tipo I iterativas de bactérias, em rosa as PKSs tipo I iterativas de fungos e em vermelho o grupo externo (domínios de ligação a ácido graxo B e F de *E. Coli*).

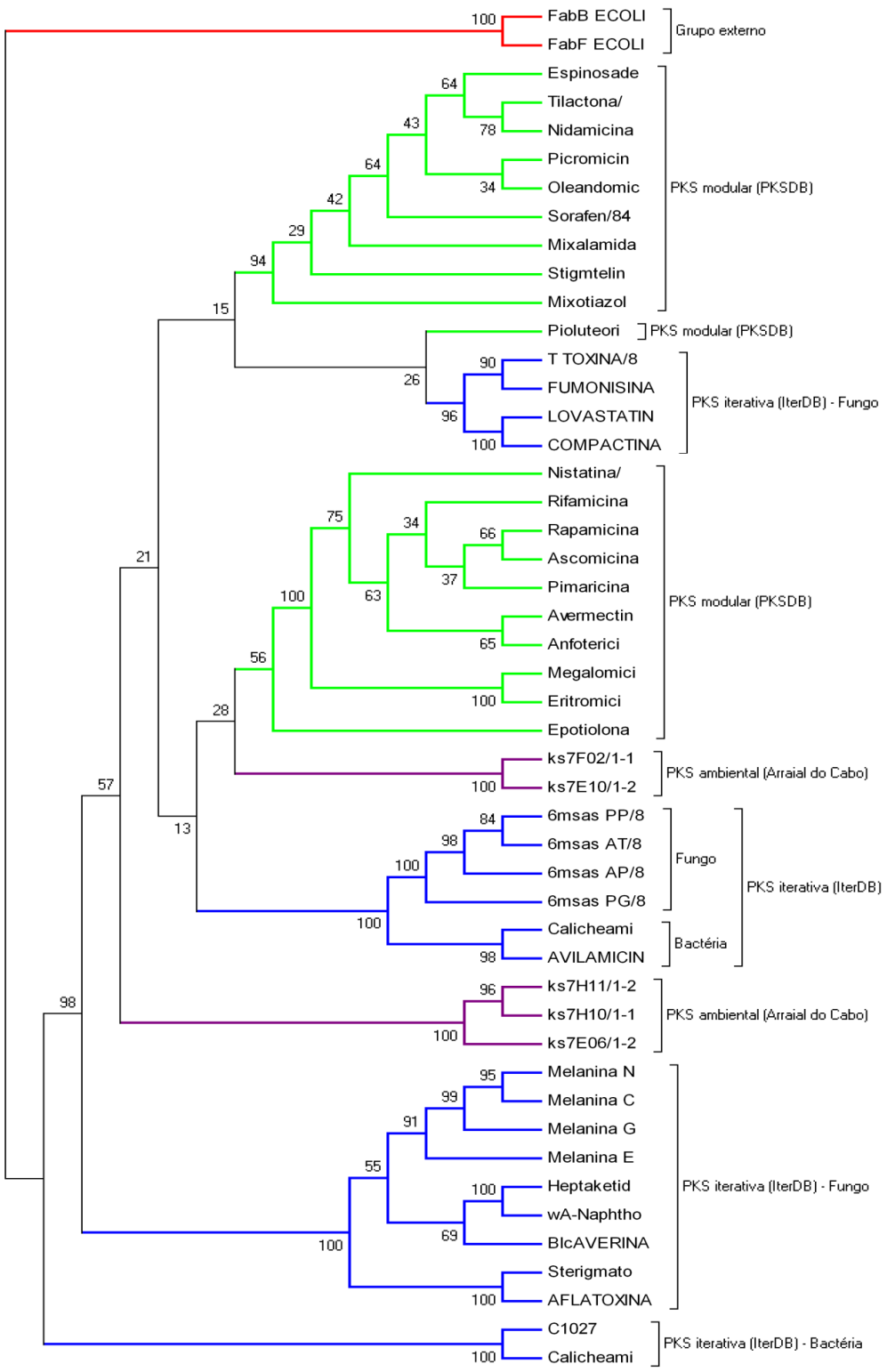


Figura 4.19: Árvore filogenética, construída com o programa PHYLM, utilizando modelo evolutivo de WAG e análise de *bootstrap* com valor 100, com os domínios KS sequenciados a partir da amplificação da amostra 7, situando-os entre os domínos KS de PKSs tipo I iterativas e modulares dos bancos curados PKSDB e IterDB. Em azul os clados com sequências de PKSs iterativas, em roxo as sequências obtidas com amplificação a partir de DNA da coleta 7, em verde as PKSs modulares e em vermelho o grupo externo.

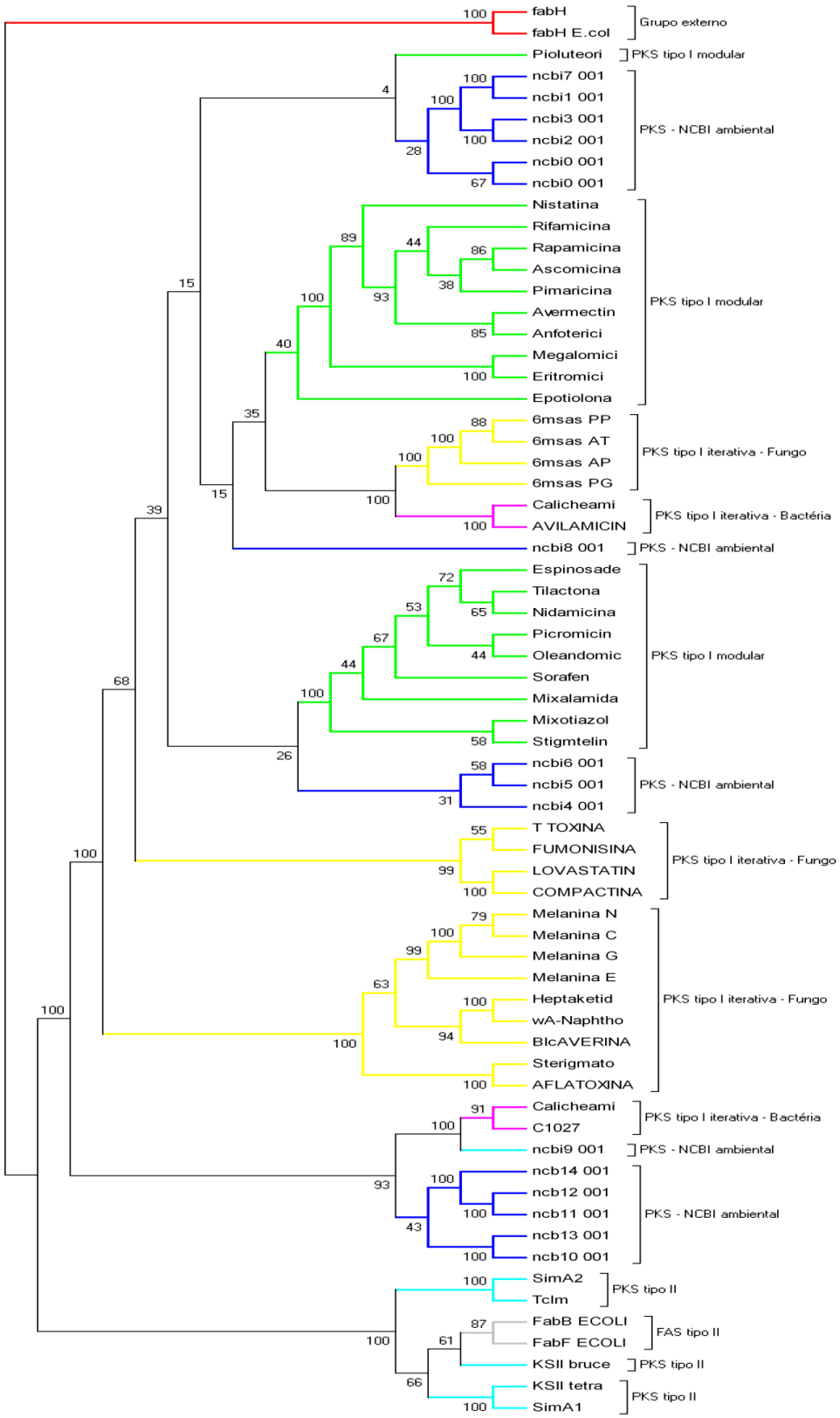


Figura 4.20: Árvore filogenética, construída com o programa PHYLM, utilizando modelo evolutivo de WAG e análise de *bootstrap* com valor 100, com os domínios KS tipo I extraídos do banco ambiental do NCBI (melhores hits obtidos com HMMER), situados entre as regiões KS de FAS tipo II, PKSs tipo I e II obtidos em bancos curados PKSDB e IterDB e também no Genbank (PKS tipoII).

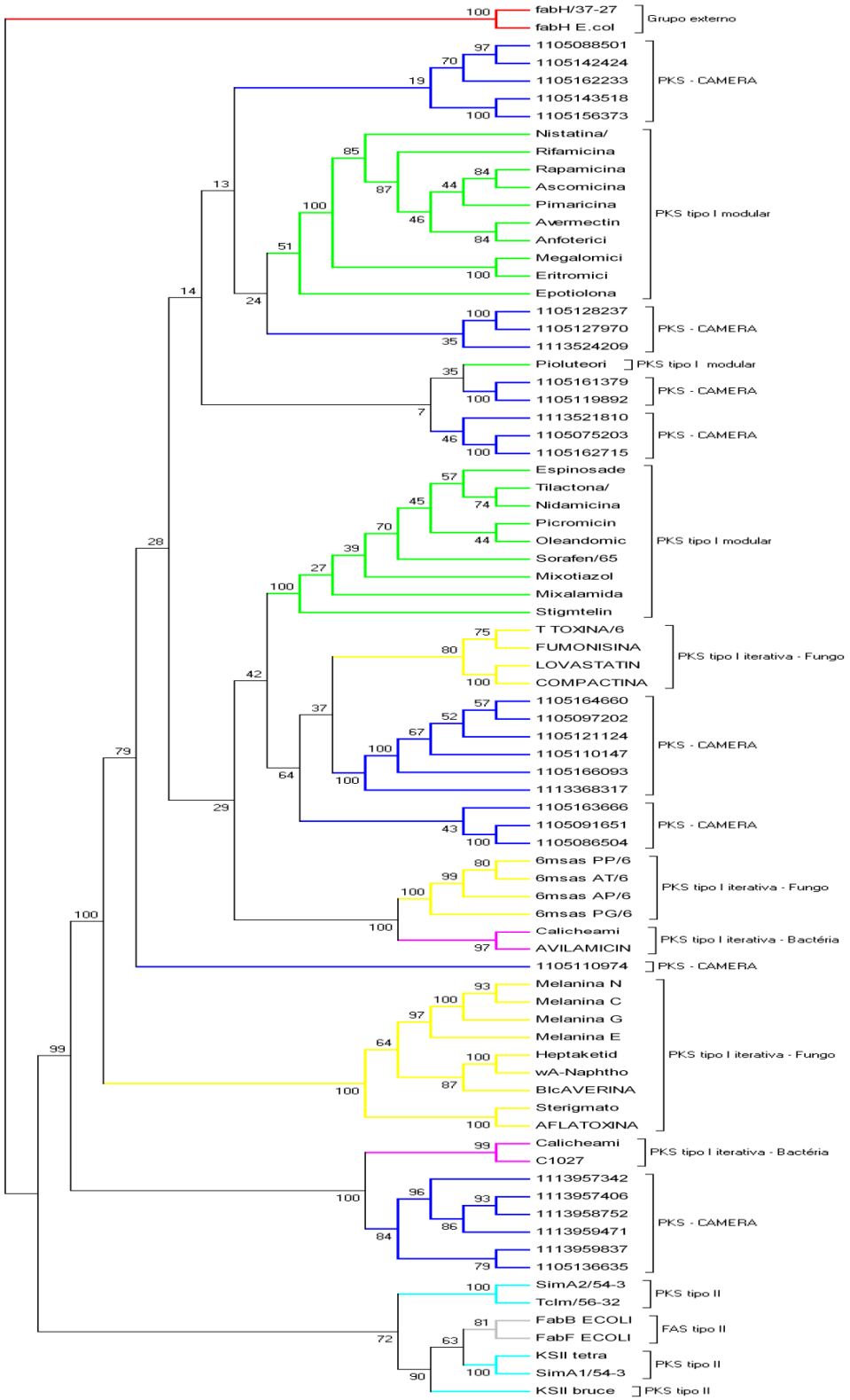


Figura 4.21: Árvore filogenética, construída com o programa PHYLM, utilizando modelo evolutivo de WAG e análise de *bootstrap* com valor 100, mostrando as regiões KS tipo I extraídas do banco CAMERA (melhores hits obtidos com HMMER, situadas entre as FAS tipo II, PKSs tipo I e II obtidas em bancos curados PKSDB e IterDB e também no Genbank (PKS tipo II).

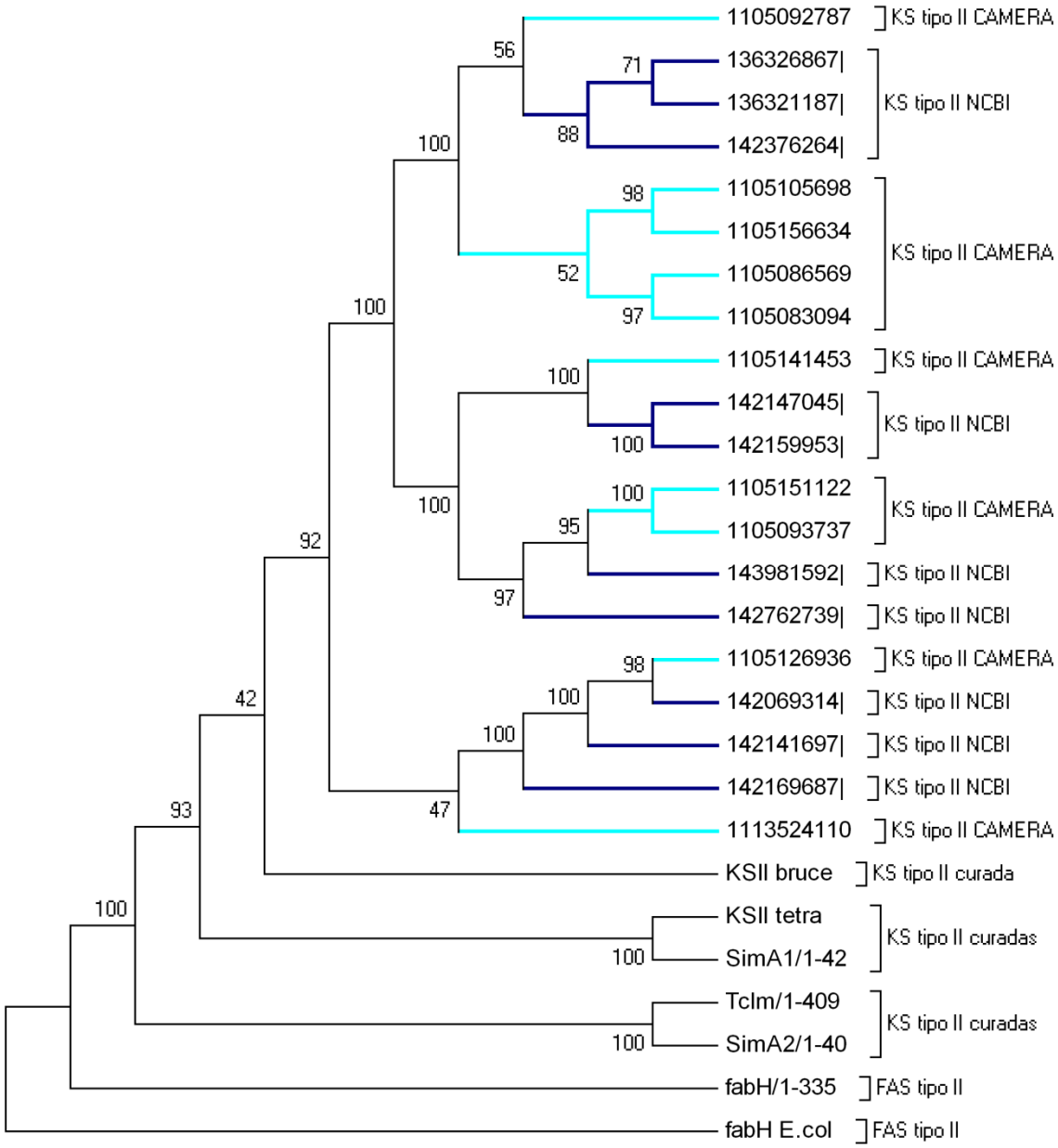


Figura 4.22: Árvore filogenética, construída com o programa PHYLM, utilizando modelo evolutivo de WAG e análise de *bootstrap* com valor 100, situando as regiões KS tipo II ambientais extraídas do banco ambiental do NCBI e do CAMERA, perante as obtidas no Genbank, utilizadas na construção do modelo *hmm* usado na triagem in silico das mesmas, utilizando domínios de ligação a ácido graxo tipo II (fabH) de *E. coli* e *M. bovis* como sequências externas. Em azul claro os clados com sequências do CAMERA e em azul escuro as sequências do banco ambiental do NCBI.

5 - Discussão:

A escolha das enzimas da família PKS nesse estudo foi motivada pela grande importância da mesma para a produção de diversos compostos com atividade de interesse na indústria farmacêutica.

Foerstner e colaboradores realizaram um estudo em 2008 descrevendo uma metodologia de triagem computacional em busca de PKSs modulares em bancos ambientais utilizando modelos *hmm*. Porém eles concluíram em seu trabalho que não é possível apenas com metodologias baseadas em similaridade classificar as PKSs quanto ao tipo nem ao menos separá-las de ácido graxo sintases, sendo necessário para tal realizar análises filogenéticas.

Para comprovar a existência de PKSs em diversos ambientes (principalmente marinhos), foram realizadas triagens computacionais não só em busca de PKSs tipo I modulares (com sequências do PKSDB), mas também buscamos as tipo I iterativas (com sequências do IterDB) e as tipo II (com sequências da região KS tipo II baixadas do NCBI) nos bancos ambientais do NCBI e do CAMERA. Utilizamos o programa HMMER (também utilizado no estudo de Foersnter) para triar as PKSs modulares e as do tipo II, pois o mesmo é muito utilizado em buscas por homologias e se mostra mais sensível na busca por homólogos distantes, por utilizar modelos ocultos de markov (Eddy et al. 2008). Para a triagem em busca de PKSs tipo I iterativas, utilizamos o BLAST (blastp) pois por sua natureza iterativa, existe apenas uma sequência de cada domínio de PKS produtora de cada metabólito, impossibilitando a construção de modelos *hmm* por metabólito, da forma que fizemos para as modulares. Comparando os resultados das triagens, observou-se sobreposição dos melhores hits entre as PKSs tipo I iterativas e modulares (diversas sequências ambientais apresentam hits com diversos modelos e sequências de PKSs tipo I das duas classes), comprovando não ser possível a separação por métodos estritamente baseados em similaridade. Porém, comparando os 10 melhores hits do modelo de KS tipo II com os melhores hits de cada modelo de KS tipo I, observou-se que os construídos com KS tipo II trazem sequências diferentes das tipo I, não havendo sobreposição entre os tipos. Apesar disso, é muito difícil separar sequências de PKSs tipo II das de FAS tipo II com esse método, exatamente como previsto por Foersnter e colaboradores, sendo necessário realizar inferências filogenéticas para tal.

Os bancos PKSDB e IterDB dividem as sequências de PKSs tipo I de acordo com o metabólito produzido e não só pelo organismo produtor. Realizamos buscas com cada modelo específico para os metabólitos produzidos. Os resultados mostram a presença de sequências com alta similaridade a todos os modelos utilizados, tanto no banco do NCBI quanto no banco de proteínas do CAMERA. Estes bancos contam, quase na sua totalidade, com sequências obtidas a partir de ambientes marinhos diversos, ao redor do mundo. Todavia, muitas sequências mostram alta similaridade com mais de um modelo (em alguns casos, com todos os modelos), tornando difícil qualquer inferência quanto ao metabólito produzido pelas sequências ambientais apenas baseando-se em similaridade. Porém, o grande número de hits comprova a hipótese da existência de uma grande diversidade de PKSs nestes ambientes e motiva a busca no litoral brasileiro, conhecido pela sua vasta biodiversidade.

Após a verificação da existência de uma grande variedade de PKSs em ambientes marinhos, o local (Arraial do Cabo) para as coletas foi escolhido, motivado pelo fato de que nosso grupo já realizava um amplo estudo de biodiversidade nessa área, através do sequenciamento de genes ribossomais (rDNA), comprovando a existência de uma vasta diversidade de organismos potencialmente produtores de enzimas de interesse biotecnológico.

O primeiro grande desafio a ser vencido foi a obtenção de DNA ambiental com qualidade, em quantidade suficiente e pureza adequada para as reações enzimáticas subsequentes. Deve-se levar em conta que o método de extração utilizado precisa lisar o maior número de células, abrangendo a maior parte possível da biodiversidade do ambiente estudado (Schmeisser et al. 2007). Os métodos químicos nem sempre conseguem lisar alguns tipos de células, porém são os métodos mais utilizados quando se deseja preservar da melhor forma possível a integridade do DNA, mantendo-o em um tamanho adequado para a clonagem em grandes bibliotecas em vetores como fosmídeos ou BAC. No entanto, os métodos físicos geralmente conseguem lisar até mesmo as células mais resistentes, sendo mais adequados para estudos de biodiversidade baseados em genes ribossomais. Um estudo realizado por Luna e colaboradores em 2006, avaliou três métodos de extração diferentes, realizando seqüenciamento de genes de rDNA para avaliar biodiversidade. Comparando os resultados, foi possível perceber diferenças significativas entre os ribotipos presentes no DNA extraído dos mesmos locais pelos diferentes métodos, comprovando a influência do método e concluindo que estudos

mais completos deveriam utilizar mais de um método combinado.

No presente estudo foram realizadas duas coletas e para cada uma delas foi usado um diferente método de extração. Para a coleta 7, o método utilizado foi o de fenol:clorofórmio modificado e para a coleta 8 foi utilizado um kit comercial (EPICENTRE) que envolve a lise química dos microorganismos e consegue obter DNA com alto peso molecular (~40kb), apto à clonagem direta em fosmídeos ou cosmídeos. O método de fenol:clorofórmio demanda um volume muito maior de água (cerca de 140 litros para se obter um total de 2µg de DNA de alto peso molecular), pois a maior parte do DNA recuperado com este método apresenta tamanho inferior ao mínimo necessário para a construção da biblioteca (25kb) sem a formação de quimeras entre os trechos de DNA ambiental (figura 4.9). Já com a extração realizada com o kit comercial EPICENTRE, com um volume muito menor (2 litros), foi possível obter um total de 1250 ng de DNA de alto peso molecular e com pureza adequada (figura 4.12).

Porém, para a escolha da metodologia de extração do DNA e a escolha do vetor para a construção de uma biblioteca não deve se levar em conta apenas estes fatores. Os alvos moleculares a serem triados na biblioteca devem influenciar a decisão, pois alguns genes podem estar presentes apenas em organismos de difícil lise, e a utilização de um método estritamente químico pode levar à não obtenção do DNA de interesse. A utilização de vetores grandes nem sempre é necessária e alguns estudos ainda são conduzidos utilizando vetores menores como plasmídeos, como por exemplo, o estudo de Gabor e colaboradores (2004), no qual o DNA extraído foi clonado em plasmídeos e a biblioteca foi triada em busca de novas amilases. Entretanto, a utilização de vetores maiores, como cosmídeos, fosmídeos e BAC, é vantajosa pois torna possível a triagem em busca de enzimas grandes ou até óperons inteiros.

As PKSs são em geral grandes e por isso os trabalhos que visam à localização das mesmas em geral envolvem a criação de grandes bibliotecas de cosmídeos ou fosmídeos, como por exemplo, o trabalho realizado por Jiao e colaboradores em 2007, no qual foram localizados fragmentos de genes PKS em uma biblioteca de fosmídeos construída a partir de sedimento do mar do leste da China.

Pequenas alíquotas do DNA extraído nas coletas 7 e 8 foram primeiramente submetidas a reações de PCR para amplificação de regiões KS com a utilização de iniciadores degenerados, na tentativa de comprovar a existência de PKSs nas amostras. O uso destes iniciadores permite a localização de sequências novas, enquanto o uso de

iniciadores específicos apenas leva a redescoberta de sequências já conhecidas. Porém, o uso de iniciadores degenerados pode levar a muitos falsos positivos (amplificação inespecífica) e até mesmo a falsos negativos.

Apenas um par de iniciadores (degKS2F.i e degKS5R.i) obteve sucesso na amplificação de regiões KS, porém apenas com DNA extraído pelo método de fenol-clorofórmio (amostra 7), que envolve uma etapa anterior de vigorosa agitação das membranas de filtragem e consequente facilitação da lise dos microorganismos (figura 4.17). Diversos fatores podem estar envolvidos neste resultado, como por exemplo, o fato de que as coletas foram realizadas em épocas distintas e pode ser que a microbiota local tenha uma grande variação durante o ano. Outra hipótese é a influência da metodologia de extração, que pode levar à lise diferenciada de organismos e também leva a diferentes concentrações e purezas de DNA.

Das sequências obtidas a partir da amplificação com o par de iniciadores degKS2F.i e degKS5R.i, apenas 7 possuíam qualidade suficiente para análises posteriores. Destas, 5 apresentaram similaridade com a região KS das enzimas PKSs (tabela 4.8), como o esperado, enquanto as outras duas apresentaram similaridade apenas com sequências hipotéticas. A maior parte dos melhores hits obtidos nas análises com o BLAST mostra similaridade com sequências de PKSs de Cianobactérias. Futuramente com a utilização de iniciadores específicos para estas sequências, a biblioteca construída com o mesmo DNA utilizado na reação de PCR, será triada em busca de clones que possuam esta sequência, e os que possuírem serão subclonados e sequenciados, para que se possível, obtenha-se a PKS inteira (ou a maior parte). Porém, este número provavelmente subestima a diversidade de PKSs no ambiente, pois só foi realizada uma única reação de PCR com um volume de 25 µl para a clonagem, e a concentração de DNA utilizada na ligação ao vetor foi pequena, fazendo com que a eficiência fosse extremamente baixa. Infelizmente a maioria dos trabalhos para busca de PKSs diretamente de DNA ambiental é realizada a partir de solo ou sedimento, tornando difícil a comparação entre os resultados obtidos com este estudo. Um exemplo de trabalho realizado com solo de floresta é o publicado por Pang e colaboradores em 2008, no qual foram amplificadas e sequenciadas 38 regiões KS tipo I. As análises filogenéticas realizadas no mesmo estudo mostraram que foi possível obter 14 regiões KS tipo I diferentes e 9 do tipo II.

Para tentar melhor compreender os motivos do insucesso da amplificação de PKSs na amostra da coleta 8, foi realizado um mapeamento preliminar da biodiversidade nesta amostra, a partir do sequenciamento de rDNA. Os resultados mostram a presença de filos bacterianos conhecidamente produtores de PKSs, como, por exemplo, um grande número de sequências classificadas como cianobactérias e proteobactérias (tabela 4.6). Isto levanta a hipótese de que provavelmente existiriam sequências de PKSs na amostra, em baixa abundância, porém a técnica de PCR utilizada não teria sido eficaz na amplificação das mesmas. Os motivos para a ineficácia da amplificação podem ser diversos, como por exemplo, a utilização de iniciadores não capazes de anelar nas PKSs específicas daqueles organismos. Porém, as sequências de PKS obtidas com a amostra da coleta 7 mostraram alta similaridade às PKSs de cianobactérias como *Nostoc puntiforme* e *Lyngbya majuscula* (tabela 4.8). Isto mostra que, se o par de iniciadores foi capaz de amplificar sequências deste filo bacteriano na amostra 7, deveria também amplificar na amostra 8, tornando pouco provável a hipótese da inespecifidade do par de iniciadores.

Infelizmente, por escassez de DNA da amostra da coleta 7, não foi possível realizar estudos de biodiversidade, para comparação com a coleta 8 e realizar inferências mais precisas sobre os fatores que influenciaram a diferença de resultado na amplificação de PKSs entre as coletas.

Algumas sequências de 16S agruparam na árvore com genes de eucariotos (figura 4.15), porém a classificação realizada com o RDP classifier mostra que estas sequências provavelmente pertencem a cloroplastos (figura 4.14) explicando a proximidade com eucariotos, em sua maioria, fotossintetizantes.

As análises de rDNA de eucariotos mostram escassez de fungos (o maior grupo de eucariotos produtores de PKSs de interesse para a indústria) na amostra da coleta 8. Porém, a presença de dinoflagelados novamente faz com que não seja possível concluir que o motivo do insucesso na amplificação de PKSs com esta amostra seja pela ausência de microorganismos produtores, pois cerca de 25 espécies de dinoflagelados produzem aproximadamente 45 PKSs diferentes (Rein & Barrone, 1999). Dois gêneros de dinoflagelados produtores de PKS foram indicados como vizinhos dos clones de Arraial pelas análises do BLAST e pelo alinhamento com vizinhos do SILVA: *Prorocentrum*, conhecidamente produtor de ácido octadaico e *Symbiodinium*, que Snyder e colaboradores (2003) demonstraram possuir genes produtores de PKS

similares aos de *Bacillus subtilis*. Porém as sequências destes gêneros não foram incluídas nos clados que possuíam sequências de clones de Arraial, ficando localizadas em clados vizinhos, sustentados por baixos valores de *bootstrap*.

Há também na amostra uma grande diversidade de crustáceos, algas, moluscos e anelídeos, fato já esperado para o ambiente marítimo estudado.

Apesar do insucesso na amplificação de regiões KS na amostra 8, foram construídas duas bibliotecas de fosmídeos, uma para amostra 7 e outra para amostra 8. A pureza do DNA obtido com o método de fenol:clorofórmio (amostra 7) não se mostrou satisfatória para a clonagem em fosmídeos, havendo forte inibição das reações enzimáticas necessárias, resultando em uma biblioteca com um número muito menor de clones do que o esperado para a quantidade de DNA utilizado na reação de ligação ao vetor. Espera-se de 10^3 a 10^6 clones para 250ng de DNA com alto peso molecular utilizado em uma reação de ligação, e foram obtidos apenas 501 clones. Além disso, houve uma grande perda de DNA nos processos de seleção de DNA de alto peso e purificação. Já a clonagem de 250 ng da amostra de DNA da coleta 8 resultou em uma biblioteca com 3500 clones.

O mesmo par de iniciadores de PCR utilizado na amplificação de regiões KS diretamente do DNA ambiental foi usado na tentativa de triar a biblioteca construída com a amostra 7, em busca de PKSs. Porém, a metodologia se mostrou ineficaz, pois as reações exibiram falsos positivos, provavelmente amplificação inespecífica de sequências do vetor ou da célula hospedeira (*E. coli*).

Posteriormente, para comprovar a hipótese de que através de filogenia é possível realizar a classificação das PKSs ambientais, foram realizadas diversas análises filogenéticas. Kodama e colaboradores em 2005 realizaram um amplo estudo sobre a evolução de PKSs em bactérias. Eles utilizaram métodos Bayesianos, além de máxima verossimilhança, máxima parcimônia e agrupamento com vizinhos para a construção de árvores filogenéticas dos domínios KS e AT, na tentativa de entender não apenas a evolução das PKSs de tipo I e II, como também explicar a ligação evolutiva entre as mesmas e as FAS. Os resultados mostram as sequências de fabH (FAS tipo II) de arqueias como ancestral comum a todos os outros tipos de FAS e PKSs, sendo as PKSs tipo II bacterianas as primeiras a surgirem, antes mesmo das fasF bacterianas e mitocondriais. Posteriormente surgem as FAS tipo I bacterianas e de fungos, as PKSs tipo I iterativas de bactérias, e depois as PKSs tipo I modulares, provavelmente

originadas por duplicação gênica a partir das iterativas. Por último aparecem as iterativas de fungos e algumas de bactérias, seguidas pelas FAS tipo I presentes em animais.

Para comparar os nossos resultados com os de outros estudos, construímos uma árvore apenas utilizando apenas sequências de bancos curados (sem sequências ambientais), pelo método de máxima verossimilhança utilizando o modelo evolutivo de WAG, com análise de *boostrap* com valor 100 e obtivemos resultados similares aos de Kodama, exceto pelo fato de que em nossas árvores, algumas PKSs iterativas de fungos aparecem antes das modulares (logo após separação entre as iterativas de bactérias e as demais), enquanto outras depois, mostrando uma intercalação entre as classes (figura 4.18). Isto sugere uma coevolução entre iterativas e modulares após o surgimento das modulares em bactérias.

Outras duas árvores foram construídas com o mesmo método, porém incluindo as sequências ambientais (do NCBI e do CAMERA) (figura 4.20 e 4.21 respectivamente) que mostraram alta similaridade com as PKSs dos bancos curados. Através destas árvores foi possível situar as sequências ambientais e inferir classificação das mesmas. No caso das PKSs tipo I, obtivemos a separação entre as iterativas e modulares, porém das 10 sequências que apresentaram similaridade com o modelo de KS tipo II, algumas agruparam com FAS tipo II e outras com PKSs tipo II, ficando bastante clara a separação das mesmas perante as tipo I, mas mostrando que em alguns casos os modelos de tipo II falham na distinção entre FAS e PKS.

As análises filogenéticas de regiões KS sequenciadas em nosso estudo mostram que, apesar do resultado das análises com o BLAST indicarem similaridade com sequências modulares de PKSs bacterianas, 3 sequências geradas neste estudo são mais próximas das PKSs iterativas de fungos e duas mais próximas de PKSs modulares. Porém, este par de iniciadores foi utilizado por Schirmer e colaboradores em 2005 apenas para triar bibliotecas metagenômicas construídas com microbiota de esponja marinha (*Discodermia dissoluta*) em busca de PKSs modulares, e seus resultados (inclusive filogenéticos) mostram uma ampla diversidade das mesmas em diversos filos de bactérias e não são observadas sequências iterativas de fungos. Uma explicação possível para a diferença entre os resultados do estudo de Schirmer e o resultado deste estudo pode estar na natureza da amostra utilizada, visto que em nosso estudo foi utilizada água do mar, ao invés de microbiota de esponja marinha.

6 – Conclusões

- Ambientes marinhos possuem ampla diversidade de sequências com alta similaridade a PKSs tipo I e II.

- Métodos tradicionais de extração de DNA utilizando fenol e clorofórmio se mostram ineficientes para a construção de bibliotecas metagenômicas a partir de amostras de água da zona pelágica marinha.

- É possível que na biblioteca construída com o DNA da coleta 7 existam PKSs, pois regiões KS da mesma estão presentes no DNA ambiental utilizado para a construção da mesma. Já a construída com a amostra da coleta 8 não apresentou regiões KS de PKSs tipo I, porém foi possível amplificar genes de rDNA de uma grande diversidade de cianobactérias, proteobactérias e dinoflagelados, o que torna possível a existência de PKSs na mesma.

- É necessário utilizar um grupo maior de iniciadores para triar em busca de PKSs em amostras de DNA ambiental, pois a diversidade desta família de enzimas é muito grande para ser contemplada com um pequeno grupo de iniciadores.

- As regiões KS sequenciadas neste estudo estão filogeneticamente relacionadas a KSs tipo I modulares e iterativas.

- Através de construção de árvores filogenéticas foi possível inferir classificação de PKSs quanto ao tipo e também quanto a classes (iterativas ou modulares).

7 – Referências bibliográficas

Amann R, Ludwig W. Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. FEMS Microbiol Rev. 2000; 24(5):555-65.

Amann R, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews 1995; 59(1):143-169.

Ansari MZ, Yadav G, Gokhale RS, Mohanty D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Research 2004; 32(Web Server Issue):W405-W413.

Azam F. Oceanography: Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 1998; 280:694-696.

Cane DE, Walsh CT, Khosla C. Harnessing the biosynthetic code: combinations, permutations, and mutations. Science 1998; 282(5386):63-8.

Castoe TA, Stephens T, Noonan BP, Calestani C. A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. Gene 2007; 392(1-2):47-58.

Chelius MK, Triplett EW. The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. Microbial Ecology 2001; 41(3):252-263.

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) [banco de dados na Internet]. Disponível em: http://camera.calit2.net.

Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G, Martinez A, Kolvek SJ, Hopke J, Osburne MS, August PR, Nalin R, Guérineau M, Jeannin P, Simonet P, Pernodet JL. Recombinant Environmental Libraries Provide Access to MicrobialDiversity for Drug Discovery from Natural Products. Applied and Enviromental Microbiology 2003; 69(1):49-55.

Couto ECG, Da Silveira FL, Rocha GRA. Marine Biodiversity in Brazil: The current status. Gayana 2003; 67(2): 327-340.

Cowan D, Meyer Q, Stafford W, Muyanga S, Cameron R, Wittwer P. Metagenomic gene discovery: past, present and future. Trends Biotechnology 2005;23(6):321-319.

Derakshani M, Lukow T, Liesack W. Novel bacterial lineages at the (sub)division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. Applied and Enviromental Microbiology 2001; 67(2):623-31.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi

D, Hu P,. Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Enviromental Microbiology 2006; 72(7):5069-5072.

Eddy SR. Profile hidden Markov models. Bioinformatics. 1998; 14(9):755-763.
Frandsen RJN. Polyketide synthases [*homepage* na Internet]. Disponível em: http://www.rasmusfrandsen.dk/ny_side_8.htm.

Gokhale RS, Sankaranarayanan R, Mohanty D. Versatility of polyketide synthases in generating metabolic diversity. Current Opinion in Structural Biology 2007; 17(6):736–743.

Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 2002; 30(14):3059-3066.

Kennedy J, Marchesi JR, Dobson AD. Marine metagenomics: Strategies for the discovery of novel enzymes with biotechnological applications from marine ecosystems. Microbial Cell Factories 2008; 21(**7**):27.

Lal R, Kumari R, Kaur H, Khanna R, Dhingra N, Tuteja D. Regulation and manipulation of the gene clusters encoding type-I PKSs. Trends in Biotechnology 2000; 18(6):264-274.

Luna GM, Dell'Anno A, Danovaro R. DNA extraction procedure: a critical issue for bacterial
diversity assessment in marine sediments. Enviromental Microbiology 2006;8(2):308-320.

Martinez A, Kolvek SJ, Yip CL, Hopke J, Brown KA, MacNeil IA, Osburne MS. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. Applied and Enviromental Microbiology 2004; 70(4):2452-63.

National Center for Biotechnology Information (NCBI) [banco de dados na Internet]. Disponível em: http://www.ncbi.nlm.nih.gov.

Pang MF, Tan GYA, Abdullah N, Lee CW, Ng CC. Phylogenetic Analysis of Type I and Type II Polyketide Synthase from Tropical Forest Soil. Biotechnology 2008; 7(4): 660-668.

Pastre R, Marinho AMR, Rodrigues-Filho E, Souza AQL, Pereira JO. Diversidade de policetídeos produzidos por espécies de *Penicillium* isoladas de *Melia azedarach* E *Murraya paniculata.* Quim. Nova 2007; 30(8):1867-1871.

PKSDB [Internet]. Disponível em: http://linux1.nii.res.in/~pksdb/polyketide.html).

[banco de dados na Internet]. Disponível em: http://linux1.nii.res.in/~pksdb/DBASE/pageALL4.html.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 2007; 35(21):7188-7196.

Rein KS, Borrone J. Polyketides from dinoflagellates: origins, pharmacology and biosynthesis. Comparative Biochemistry and Physiology – Part B: Biochemistry and Molecular Biology 1999; 124(2):117-131.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PloS Biol. 2007; 5(3):e77.

Schirmer A, Gadkari R, Reeves,CD, Ibrahim F, DeLong EF, Hutchinson1 CR. Metagenomic Analysis Reveals Diverse Polyketide Synthase Gene Clusters in Microorganisms Associated with the Marine Sponge *Discodermia dissoluta.* Applied and Enviromental Microbiology 2005; 71(8):48440-4849.

Schmeisser C, Steele H, Streit WR. Metagenomiscs, biotechnology with non-culturable microbes. Applied Microbiology and Biotechnology 2007; 75(5):955-962.

Shen B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Current Opinion in Chemical Biology 2003;7(2):285-95.

Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, Bali V, Batra N. Metagenomics: Concept, methodology, ecological inference and recent advances. Biotechnology J. 2009; 4(4): 480–494.

Snyder RV, Gibbs PDL, Palacios A, Abiy L, Dickey R, Lopez JV, Rein KS. Polyketide Synthase Genes from Marine Dinoflagellates. Marine Biotechnology 2003; 5, 1-12.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004; 304(5667):66-74.

Uchiyama T, Watanabe K. Substrate-induced gene expression (SIGEX) screening of metagenome libraries. Nature Protocols 2008; 3, 1202 – 1212.

Watanabe A. and Ebizuka Y. Unprecedented Mechanism for Chain Length Determination in Fungal Aromatic Polyketide Synthases. Chemistry and Biology 2004; 11(8):1101-1106.

Wawrik B, Kerkhof L,Zylstra GJ, Kukor JJ. Identification of Unique Type II Polyketide Synthase Genes in Soil. Applied and Environmental Microbiology 2005; 71(5): 2232-2238.

Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 2001; 18(5):691-9.

Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proceedings of the National. Academy of Sciences U S A 1998; 95(12):6578-6583.

Yadav G, Gokhale RS, Mohanty D. SEARCHPKS: A program for detection and analysis of polyketide synthase domains. Nucleic Acids Research 2003; 31(13):3654-3658.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biology 2007; 5(3):e16

8 - Material Suplementar

S1 – Tabela com os 3 melhores hits entre as sequências KS tipo I iterativa e o banco ambiental do NCBI.

Query	Descricao query	GI	Descricao hit	Score	Evalue
AFLAT_001_KS_001.seq	Expect = 7e-71 length= 431(374- 805)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	815	2e-85,
AFLAT_001_KS_001.seq	Expect = 7e-71 length= 431(374- 805)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	757	1e-78,
AFLAT_001_KS_001.seq	Expect = 7e-71 length= 431(374- 805)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	697	7e-72,
AFLAT_001_KS_001.seq	Expect = 7e-71 length= 431(374- 805)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	690	5e-71,
AFLAT_001_KS_001.seq	Expect = 7e-71 length= 431(374- 805)	gb|EBK82670.1|	hypothetical protein GOS_8668700 [marine metagenome]	677	2e-69,
AVILA_001_KS_001.seq	Expect = e-114 length= 435(25- 460)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	1003	1e-107,
AVILA_001_KS_001.seq	Expect = e-114 length= 435(25- 460)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	965	1e-103,
AVILA_001_KS_001.seq	Expect = e-114 length= 435(25- 460)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	918	2e-97,
AVILA_001_KS_001.seq	Expect = e-114 length= 435(25- 460)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	907	4e-96,
AVILA_001_KS_001.seq	Expect = e-114 length= 435(25- 460)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	892	2e-94,
BIKAV_001_KS_001.seq	Expect = 3e-72 length= 430(346- 776)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	730	1e-75,
BIKAV_001_KS_001.seq	Expect = 3e-72 length= 430(346- 776)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	729	2e-75,
BIKAV_001_KS_001.seq	Expect = 3e-72 length= 430(346- 776)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	698	7e-72,
BIKAV_001_KS_001.seq	Expect = 3e-72 length= 430(346- 776)	gb|ECZ67020.1|	hypothetical protein GOS_2142050 [marine metagenome]	676	2e-69,
BIKAV_001_KS_001.seq	Expect = 3e-72 length= 430(346- 776)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	670	1e-68,
C1027_001_KS_001.seq	Expect = 4e-54 length= 455(4- 459)	gb|EBG62110.1|	hypothetical protein GOS_9402449 [marine metagenome]	703	2e-72,
C1027_001_KS_001.seq	Expect = 4e-54 length= 455(4- 459)	gb|EDE78607.1|	hypothetical protein GOS_1070273 [marine metagenome]	538	2e-53,
C1027_001_KS_001.seq	Expect = 4e-54 length= 455(4- 459)	gb|EDF15758.1|	hypothetical protein GOS_1004776 [marine metagenome]	531	2e-52,
C1027_001_KS_001.seq	Expect = 4e-54 length= 455(4- 459)	gb|EDD32563.1|	hypothetical protein GOS_1321998 [marine metagenome]	513	2e-50,
C1027_001_KS_001.seq	Expect = 4e-54 length= 455(4- 459)	gb|ECU98303.1|	hypothetical protein GOS_2983389 [marine metagenome]	485	4e-47,
CALEN_001_KS_001.seq	Expect = 4e-56 length= 458(4- 462)	gb|EBG62110.1|	hypothetical protein GOS_9402449 [marine metagenome]	694	2e-71,
CALEN_001_KS_001.seq	Expect = 4e-56 length= 458(4- 462)	gb|EDE78607.1|	hypothetical protein GOS_1070273 [marine metagenome]	610	1e-61,

CALEN_001_KS_001.seq	Expect = 4e-56 length= 458(4- 462)	gb|EDF15758.1|	hypothetical protein GOS_1004776 [marine metagenome]	582	2e-58,
CALEN_001_KS_001.seq	Expect = 4e-56 length= 458(4- 462)	gb|EDD32563.1|	hypothetical protein GOS_1321998 [marine metagenome]	565	2e-56,
CALEN_001_KS_001.seq	Expect = 4e-56 length= 458(4- 462)	gb|EDH54350.1|	hypothetical protein GOS_586729 [marine metagenome]	557	1e-55,
CALOR_001_KS_001.seq	Expect = e-112 length= 423(13- 436)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	979	1e-104,
CALOR_001_KS_001.seq	Expect = e-112 length= 423(13- 436)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	975	1e-104,
CALOR_001_KS_001.seq	Expect = e-112 length= 423(13- 436)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	975	1e-104,
CALOR_001_KS_001.seq	Expect = e-112 length= 423(13- 436)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	961	1e-102,
CALOR_001_KS_001.seq	Expect = e-112 length= 423(13- 436)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	885	1e-93,
COMPA_001_KS_001.seq	Expect = 6e-84 length= 438(9- 447)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	907	3e-96,
COMPA_001_KS_001.seq	Expect = 6e-84 length= 438(9- 447)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	868	1e-91,
COMPA_001_KS_001.seq	Expect = 6e-84 length= 438(9- 447)	gb|ECV65921.1|	hypothetical protein GOS_2857438 [marine metagenome]	905	6e-96,
COMPA_001_KS_001.seq	Expect = 6e-84 length= 438(9- 447)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	843	1e-88,
COMPA_001_KS_001.seq	Expect = 6e-84 length= 438(9- 447)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	839	3e-88,
FUMON_001_KS_001.seq	Expect = 3e-75 length= 418(31- 449)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	832	2e-87,
FUMON_001_KS_001.seq	Expect = 3e-75 length= 418(31- 449)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	814	2e-85,
FUMON_001_KS_001.seq	Expect = 3e-75 length= 418(31- 449)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	814	2e-85,
FUMON_001_KS_001.seq	Expect = 3e-75 length= 418(31- 449)	gb|ECV65921.1|	hypothetical protein GOS_2857438 [marine metagenome]	800	9e-84,
FUMON_001_KS_001.seq	Expect = 3e-75 length= 418(31- 449)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	783	7e-82,
LOVAS_001_KS_001.seq	Expect = 2e-82 length= 438(9- 447)	gb|ECV65921.1|	hypothetical protein GOS_2857438 [marine metagenome]	906	5e-96,
LOVAS_001_KS_001.seq	Expect = 2e-82 length= 438(9- 447)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	868	1e-91,
LOVAS_001_KS_001.seq	Expect = 2e-82 length= 438(9- 447)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	839	3e-88,
LOVAS_001_KS_001.seq	Expect = 2e-82 length= 438(9- 447)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	837	5e-88,
LOVAS_001_KS_001.seq	Expect = 2e-82 length= 438(9- 447)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	833	1e-87,
MS_AP_001_KS_001.seq	Expect = 9e-96 length= 420(15- 435)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	911	3e-96,
MS_AP_001_KS_001.seq	Expect = 9e-96 length= 420(15- 435)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	893	3e-94,
MS_AP_001_KS_001.seq	Expect = 9e-96 length= 420(15- 435)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	886	2e-93,
MS_AP_001_KS_001.seq	Expect = 9e-96 length= 420(15- 435)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	846	9e-89,

MS_AP_001_KS_001.seq	Expect = 9e-96 length= 420(15- 435)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	841	4e-88,
MS_PG_001_KS_001.seq	Expect = 4e-93 length= 426(45- 471)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	897	5e-95,
MS_PG_001_KS_001.seq	Expect = 4e-93 length= 426(45- 471)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	894	1e-94,
MS_PG_001_KS_001.seq	Expect = 4e-93 length= 426(45- 471)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	844	7e-89,
MS_PG_001_KS_001.seq	Expect = 4e-93 length= 426(45- 471)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	847	3e-89,
MS_PG_001_KS_001.seq	Expect = 4e-93 length= 426(45- 471)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	802	5e-84,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECZ67020.1|	hypothetical protein GOS_2142050 [marine metagenome]	801	6e-84,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECU92432.1|	hypothetical protein GOS_2992767 [marine metagenome]	783	7e-82,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	754	2e-78,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	745	2e-77,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	744	2e-77,
MS_PP_001_KS_001.seq	Expect = 2e-94 length= 422(35- 457)	gb|ECV65921.1|	hypothetical protein GOS_2857438 [marine metagenome]	722	1e-74,
WA_NA_001_KS_001.seq	Expect = 2e-72 length= 430(379- 809)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	775	6e-81,
WA_NA_001_KS_001.seq	Expect = 2e-72 length= 430(379- 809)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	729	2e-75,
WA_NA_001_KS_001.seq	Expect = 2e-72 length= 430(379- 809)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	755	2e-78,
WA_NA_001_KS_001.seq	Expect = 2e-72 length= 430(379- 809)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	737	2e-76,
WA_NA_001_KS_001.seq	Expect = 2e-72 length= 430(379- 809)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	712	1e-73,
STERG_001_KS_001.seq	Expect = 3e-69 length= 431(383- 814)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	808	1e-84,
STERG_001_KS_001.seq	Expect = 3e-69 length= 431(383- 814)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	736	3e-76,
STERG_001_KS_001.seq	Expect = 3e-69 length= 431(383- 814)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	711	2e-73,
STERG_001_KS_001.seq	Expect = 3e-69 length= 431(383- 814)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	686	1e-70,
STERG_001_KS_001.seq	Expect = 3e-69 length= 431(383- 814)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	686	1e-70,
AFMEL_001_KS_001.seq	Expect = 1e-69 length= 430(378- 808)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	758	7e-79,
AFMEL_001_KS_001.seq	Expect = 1e-69 length= 430(378- 808)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	705	1e-72,
AFMEL_001_KS_001.seq	Expect = 1e-69 length= 430(378- 808)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	730	1e-75,
AFMEL_001_KS_001.seq	Expect = 1e-69 length= 430(378- 808)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	710	3e-73,
AFMEL_001_KS_001.seq	Expect = 1e-69 length= 430(378- 808)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	690	6e-71,

THNCL_001_KS_001.seq	Expect = 3e-71 length= 431(384- 815)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	712	1e-73,
THNCL_001_KS_001.seq	Expect = 3e-71 length= 431(384- 815)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	703	1e-72,
THNCL_001_KS_001.seq	Expect = 3e-71 length= 431(384- 815)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	701	3e-72,
THNCL_001_KS_001.seq	Expect = 3e-71 length= 431(384- 815)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	696	1e-71,
THNCL_001_KS_001.seq	Expect = 3e-71 length= 431(384- 815)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	670	9e-69,
THNED_001_KS_001.seq	Expect = 3e-63 length= 431(369- 800)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	681	5e-70,
THNED_001_KS_001.seq	Expect = 3e-63 length= 431(369- 800)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	651	2e-66,
THNED_001_KS_001.seq	Expect = 3e-63 length= 431(369- 800)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	630	4e-64,
THNED_001_KS_001.seq	Expect = 3e-63 length= 431(369- 800)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	620	7e-63,
THNED_001_KS_001.seq	Expect = 3e-63 length= 431(369- 800)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	607	2e-61,
THNGL_001_KS_001.seq	Expect = 2e-68 length= 431(380- 811)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	665	4e-68,
THNGL_001_KS_001.seq	Expect = 2e-68 length= 431(380- 811)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	613	4e-62,
THNGL_001_KS_001.seq	Expect = 2e-68 length= 431(380- 811)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	649	3e-66,
THNGL_001_KS_001.seq	Expect = 2e-68 length= 431(380- 811)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	631	4e-64,
THNGL_001_KS_001.seq	Expect = 2e-68 length= 431(380- 811)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	620	7e-63,
THNND_001_KS_001.seq	Expect = 5e-70 length= 431(385- 816)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	747	1e-77,
THNND_001_KS_001.seq	Expect = 5e-70 length= 431(385- 816)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	713	1e-73,
THNND_001_KS_001.seq	Expect = 5e-70 length= 431(385- 816)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	740	9e-77,
THNND_001_KS_001.seq	Expect = 5e-70 length= 431(385- 816)	gb|EDH19292.1|	hypothetical protein GOS_649688 [marine metagenome]	717	4e-74,
THNND_001_KS_001.seq	Expect = 5e-70 length= 431(385- 816)	gb|ECV73088.1|	hypothetical protein GOS_2845443 [marine metagenome]	705	1e-72,
T_TOX_001_KS_001.seq	Expect = 2e-84 length= 423(13- 436)	gb|ECW24235.1|	hypothetical protein GOS_2755572 [marine metagenome]	872	4e-92,
T_TOX_001_KS_001.seq	Expect = 2e-84 length= 423(13- 436)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	835	7e-88,
T_TOX_001_KS_001.seq	Expect = 2e-84 length= 423(13- 436)	gb|ECU92436.1|	hypothetical protein GOS_2992782 [marine metagenome]	793	5e-83,
T_TOX_001_KS_001.seq	Expect = 2e-84 length= 423(13- 436)	gb|ECV65921.1|	hypothetical protein GOS_2857438 [marine metagenome]	791	1e-82,
T_TOX_001_KS_001.seq	Expect = 2e-84 length= 423(13- 436)	gb|EBP75987.1|	hypothetical protein GOS_7859092 [marine metagenome]	766	8e-80,

S2 – Tabela com os 10 melhores hits entre o modelo de KS tipo II e o banco ambiental do NCBI.

GI	Descricao hit	Score	Evalue
gi|138342234|gb|EBZ08133.1|	hypothetical protein GOS_	267.8	1.4e-74
gi|135010976|gb|EBE91475.1|	hypothetical protein GOS_	266.3	4.1e-74
gi|135229787|gb|EBG30572.1|	hypothetical protein GOS_	261.0	1.7e-72
gi|137343641|gb|EBT56761.1|	hypothetical protein GOS_	259.2	5.6e-72
gi|142100043|gb|ECV58982.1|	hypothetical protein GOS_	258.9	7.1e-72
gi|141603579|gb|ECS53448.1|	hypothetical protein GOS_	258.6	8.4e-72
gi|141055520|gb|ECP32792.1|	hypothetical protein GOS_	257.6	1.8e-71
gi|142131736|gb|ECV82225.1|	hypothetical protein GOS_	256.0	5.2e-71
gi|140316139|gb|ECK93209.1|	hypothetical protein GOS_	254.9	1.1e-70
gi|142635149|gb|ECZ48413.1|	hypothetical protein GOS_	253.2	3.7e-70

S3 – Tabela com os 3 melhores hits de cada sequência KS tipo I iterativa e o banco do CAMERA

Query	GI	Score	Evalue
AFLAT_001_KS_001.seq	JCVI_PEP_1105143518691	672	7e-69,
AFLAT_001_KS_001.seq	JCVI_PEP_1105142424805	626	2e-63,
AFLAT_001_KS_001.seq	JCVI_PEP_1105156373951	614	4e-62,
AFLAT_001_KS_001.seq	JCVI_PEP_1105162715115	607	3e-61,
AFLAT_001_KS_001.seq	JCVI_PEP_1105127970587	598	2e-60,
AVILA_001_KS_001.seq	JCVI_PEP_1105119892957	787	3e-82,
AVILA_001_KS_001.seq	JCVI_PEP_1105086504081	779	2e-81,
AVILA_001_KS_001.seq	JCVI_PEP_1105110974887	777	4e-81,
AVILA_001_KS_001.seq	JCVI_PEP_1105128237707	773	1e-80,
AVILA_001_KS_001.seq	JCVI_PEP_1105127970587	769	4e-80,
BIKAV_001_KS_001.seq	JCVI_PEP_1105143518691	645	8e-66,
BIKAV_001_KS_001.seq	JCVI_PEP_1105142424805	619	9e-63,
BIKAV_001_KS_001.seq	JCVI_PEP_1105162233823	603	7e-61,
BIKAV_001_KS_001.seq	JCVI_PEP_1105156373951	600	2e-60,
BIKAV_001_KS_001.seq	JCVI_PEP_1105162715115	599	2e-60,
C1027_001_KS_001.seq	JCVI_PEP_1113958752148	729	2e-75,
C1027_001_KS_001.seq	JCVI_PEP_1113959471766	716	6e-74,
C1027_001_KS_001.seq	JCVI_PEP_1113957406222	714	1e-73,
C1027_001_KS_001.seq	JCVI_PEP_1105136635202	712	2e-73,
C1027_001_KS_001.seq	JCVI_PEP_1113959837648	692	3e-71,
CALEN_001_KS_001.seq	JCVI_PEP_1105136635202	702	2e-72,
CALEN_001_KS_001.seq	JCVI_PEP_1113959837648	655	6e-67,

CALEN_001_KS_001.seq	JCVI_PEP_1113958752148	652	2e-66,
CALEN_001_KS_001.seq	JCVI_PEP_1113957406222	646	7e-66,
CALEN_001_KS_001.seq	JCVI_PEP_1113957342758	622	4e-63,
CALOR_001_KS_001.seq	JCVI_PEP_1105119892957	859	1e-90,
CALOR_001_KS_001.seq	JCVI_PEP_1105086504081	853	7e-90,
CALOR_001_KS_001.seq	JCVI_PEP_1105143518691	839	3e-88,
CALOR_001_KS_001.seq	JCVI_PEP_1105161379171	807	1e-84,
CALOR_001_KS_001.seq	JCVI_PEP_1105127970587	805	3e-84,
COMPA_001_KS_001.seq	JCVI_PEP_1105119892957	737	2e-76,
COMPA_001_KS_001.seq	JCVI_PEP_1105097202953	737	2e-76,
COMPA_001_KS_001.seq	JCVI_PEP_1105110147351	730	1e-75,
COMPA_001_KS_001.seq	JCVI_PEP_1105121124937	716	6e-74,
COMPA_001_KS_001.seq	JCVI_PEP_1105166093349	714	9e-74,
FUMON_001_KS_001.seq	JCVI_PEP_1105119892957	711	2e-73,
FUMON_001_KS_001.seq	JCVI_PEP_1105097202953	667	3e-68,
FUMON_001_KS_001.seq	JCVI_PEP_1105091651557	665	5e-68,
FUMON_001_KS_001.seq	JCVI_PEP_1105086504081	653	1e-66,
FUMON_001_KS_001.seq	JCVI_PEP_1113368317834	645	9e-66,
LOVAS_001_KS_001.seq	JCVI_PEP_1105097202953	746	2e-77,
LOVAS_001_KS_001.seq	JCVI_PEP_1105121124937	740	1e-76,
LOVAS_001_KS_001.seq	JCVI_PEP_1105110147351	739	1e-76,
LOVAS_001_KS_001.seq	JCVI_PEP_1105119892957	739	1e-76,
LOVAS_001_KS_001.seq	JCVI_PEP_1105164660723	715	8e-74,
MS_AP_001_KS_001.seq	JCVI_PEP_1105119892957	798	2e-83,
MS_AP_001_KS_001.seq	JCVI_PEP_1105119892957	797	2e-83,
MS_AP_001_KS_001.seq	JCVI_PEP_1105161379171	753	3e-78,
MS_AP_001_KS_001.seq	JCVI_PEP_1105161379171	745	2e-77,
MS_AP_001_KS_001.seq	JCVI_PEP_1105088501523	746	2e-77,
MS_PG_001_KS_001.seq	JCVI_PEP_1105119892957	794	5e-83,
MS_PG_001_KS_001.seq	JCVI_PEP_1105161379171	763	2e-79,
MS_PG_001_KS_001.seq	JCVI_PEP_1105143518691	759	6e-79,
MS_PG_001_KS_001.seq	JCVI_PEP_1105088501523	758	8e-79,
MS_PG_001_KS_001.seq	JCVI_PEP_1105142424805	736	3e-76,
MS_PP_001_KS_001.seq	JCVI_PEP_1105119892957	717	4e-74,
MS_PP_001_KS_001.seq	JCVI_PEP_1105086504081	706	8e-73,
MS_PP_001_KS_001.seq	JCVI_PEP_1105162233823	704	1e-72,
MS_PP_001_KS_001.seq	JCVI_PEP_1105161379171	698	7e-72,
MS_PP_001_KS_001.seq	JCVI_PEP_1105088501523	676	2e-69,
WA_NA_001_KS_001.seq	JCVI_PEP_1105143518691	633	2e-64,
WA_NA_001_KS_001.seq	JCVI_PEP_1113521810274	627	1e-63,
WA_NA_001_KS_001.seq	JCVI_PEP_1113524209040	626	1e-63,
WA_NA_001_KS_001.seq	JCVI_PEP_1105162715115	625	2e-63,
WA_NA_001_KS_001.seq	JCVI_PEP_1105110974887	615	3e-62,
STERG_001_KS_001.seq	JCVI_PEP_1105143518691	646	6e-66,

STERG_001_KS_001.seq	JCVI_PEP_1105142424805	630	6e-64,
STERG_001_KS_001.seq	JCVI_PEP_1105162715115	618	1e-62,
STERG_001_KS_001.seq	JCVI_PEP_1105127970587	611	7e-62,
STERG_001_KS_001.seq	JCVI_PEP_1113521810274	605	4e-61,
AFMEL_001_KS_001.seq	JCVI_PEP_1105143518691	615	3e-62,
AFMEL_001_KS_001.seq	JCVI_PEP_1105142424805	613	5e-62,
AFMEL_001_KS_001.seq	JCVI_PEP_1105162715115	601	1e-60,
AFMEL_001_KS_001.seq	JCVI_PEP_1113521810274	593	9e-60,
AFMEL_001_KS_001.seq	JCVI_PEP_1105110974887	592	1e-59,
THNCL_001_KS_001.seq	JCVI_PEP_1105127970587	635	1e-64,
THNCL_001_KS_001.seq	JCVI_PEP_1105143518691	626	2e-63,
THNCL_001_KS_001.seq	JCVI_PEP_1105142424805	604	6e-61,
THNCL_001_KS_001.seq	JCVI_PEP_1113521810274	597	3e-60,
THNCL_001_KS_001.seq	JCVI_PEP_1105162715115	593	9e-60,
THNED_001_KS_001.seq	JCVI_PEP_1105143518691	579	4e-58,
THNED_001_KS_001.seq	JCVI_PEP_1105142424805	569	6e-57,
THNED_001_KS_001.seq	JCVI_PEP_1113521810274	558	1e-55,
THNED_001_KS_001.seq	JCVI_PEP_1105156373951	549	1e-54,
THNED_001_KS_001.seq	JCVI_PEP_1105162715115	533	1e-52,
THNGL_001_KS_001.seq	JCVI_PEP_1105143518691	563	3e-56,
THNGL_001_KS_001.seq	JCVI_PEP_1105142424805	562	4e-56,
THNGL_001_KS_001.seq	JCVI_PEP_1105162715115	537	3e-53,
THNGL_001_KS_001.seq	JCVI_PEP_1105127970587	536	4e-53,
THNGL_001_KS_001.seq	JCVI_PEP_1113524209040	513	2e-50,
THNND_001_KS_001.seq	JCVI_PEP_1105143518691	636	1e-64,
THNND_001_KS_001.seq	JCVI_PEP_1105127970587	605	4e-61,
THNND_001_KS_001.seq	JCVI_PEP_1105142424805	602	1e-60,
THNND_001_KS_001.seq	JCVI_PEP_1105156373951	595	6e-60,
THNND_001_KS_001.seq	JCVI_PEP_1105162715115	587	5e-59,
T_TOX_001_KS_001.seq	JCVI_PEP_1105163666183	765	1e-79,
T_TOX_001_KS_001.seq	JCVI_PEP_1105143518691	728	2e-75,
T_TOX_001_KS_001.seq	JCVI_PEP_1105086504081	724	6e-75,
T_TOX_001_KS_001.seq	JCVI_PEP_1105119892957	711	2e-73,
T_TOX_001_KS_001.seq	JCVI_PEP_1105121124937	708	4e-73,

S4 – Tabela com os 10 melhores hits entre o modelo de KS tipo II e o banco do CAMERA

GI	Descricao hit	Score	Evalue
JCVI_PEP_1105158479575	/read_id=JCVI_READ_10929634771	267.8	1e-73
JCVI_PEP_1105118441005	/read_id=JCVI_READ_10923437547	266.3	2.9e-73
JCVI_PEP_1113974862398	/orf_id=JCVI_ORF_1113974862397	262.9	3.2e-72
JCVI_PEP_1113974858724	/orf_id=JCVI_ORF_1113974858723	262.9	3.2e-72
JCVI_PEP_1105088527057	/read_id=JCVI_READ_10955210332	261.0	1.2e-71
JCVI_PEP_1105110582005	/read_id=JCVI_READ_10911423638	259.7	2.9e-71
JCVI_PEP_1105132971563	/read_id=JCVI_READ_297914 /beg	259.2	4e-71
JCVI_PEP_1112698896770	/orf_id=JCVI_ORF_1112698896769	259.1	4.2e-71
JCVI_PEP_1112708689790	/orf_id=JCVI_ORF_1112708689789	259.1	4.2e-71
JCVI_PEP_1105161569947	/read_id=JCVI_READ_10955218739	258.9	5.1e-71