

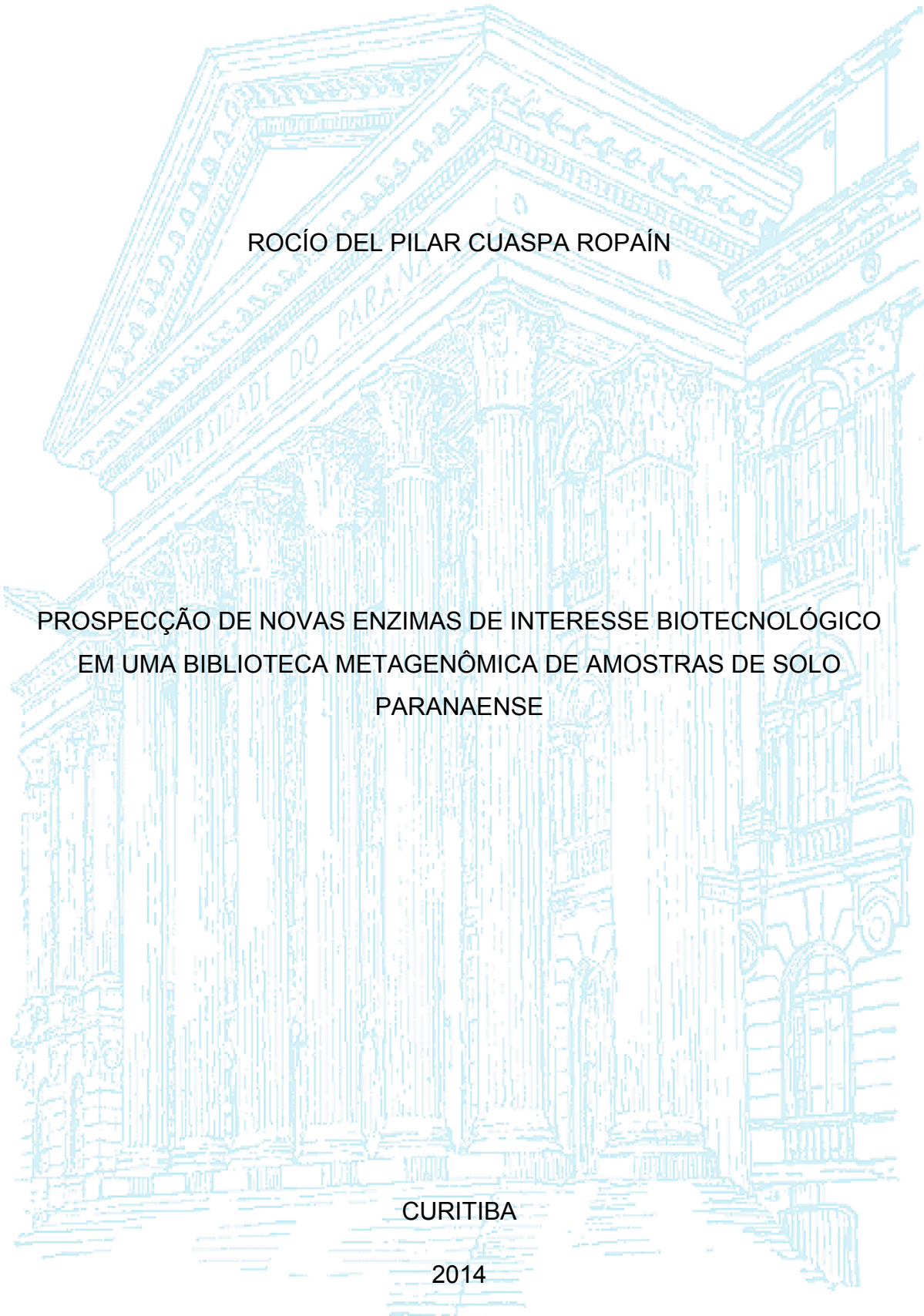
UNIVERSIDADE FEDERAL DO PARANÁ

ROCÍO DEL PILAR CUASPA ROPAÍN

PROSPECÇÃO DE NOVAS ENZIMAS DE INTERESSE BIOTECNOLÓGICO
EM UMA BIBLIOTECA METAGENÔMICA DE AMOSTRAS DE SOLO
PARANAENSE

CURITIBA

2014





ROCÍO DEL PILAR CUASPA ROPAÍN

PROSPECÇÃO DE NOVAS ENZIMAS DE INTERESSE BIOTECNOLÓGICO
EM UMA BIBLIOTECA METAGENÔMICA DE AMOSTRAS DE SOLO
PARANAENSE

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências – Bioquímica, no Curso de Pós- Graduação em Ciências – Bioquímica, Setor de Ciências Biológicas, Universidade Federal do Paraná.

Orientadora: Prof.^a Leda Satie Chubatsu
Co-orientadores: Helisson Faoro,
Marco Aurelio S. de Oliveira

CURITIBA

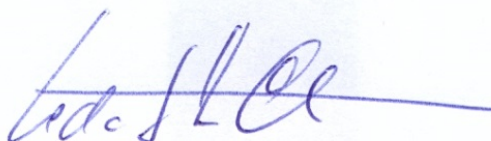
2014

TERMO DE APROVAÇÃO

ROCÍO DEL PILAR CUASPA ROPAÍN

PROSPECÇÃO DE NOVAS ENZIMAS DE INTERESSE BIOTECNOLÓGICO
EM UMA BIBLIOTECA METAGENÔMICA DE AMOSTRAS DE SOLO
PARANAENSE

Dissertação aprovada como requisito parcial à obtenção do grau de Mestre em Ciências – Bioquímica, no Curso de Pós- Graduação em Ciências – Bioquímica, Setor de Ciências Biológicas, Universidade Federal do Paraná, pela seguinte banca examinadora:



Prof^a. Dr^a. Leda Satie Chubatsu
Orientadora – Departamento de Bioquímica e Biologia Molecular, UFPR.



Prof. Dr. Marcelo Müller dos Santos
Departamento de Bioquímica e Biologia Molecular, UFPR.



Profa. Dra. Fabiane Gomes de Moraes Rego
Departamento de Análise Clínica, UFPR

AGRADECIMENTOS

A Deus, que faz tudo possível.

A meu esposo Dany, por seu amor, apoio e força para romper muitas barreiras.

A minha família em Colômbia e outras partes do mundo, mas sempre presente comigo.

Ao povo Brasileiro, especialmente o Paranaense, por sua acolhida aconchegante, amável e alegre.

Aos doutores Leda Satie Chubatsu, Helisson Faoro e Marco Aurelio Schüler de Oliveira pela orientação, co-orientação deste projeto, ajuda na inserção ao mundo científico, colaboração na realização dos experimentos e por todos os ensinamentos.

Aos doutores Marcelo Müller dos Santos e Viviane Paula Martini e o doutorando Robson Alnoch, por seus valiosos aportes nesse trabalho.

A todo o pessoal do Departamento de Bioquímica e Biologia Molecular presente e passado, incluindo professores, servidores, técnicos, colegas e amigos que fazem deste programa uma boa família e um excelente conceito.

À Universidade Federal do Paraná por todo o bem-estar universitário oferecido no meu segundo lar.

À CAPES, CNPq e Fundação Araucária pelo auxílio financeiro.

“A *gente* tem que sonhar, senão as coisas não acontecem”.

Oscar Niemeyer

SUMÁRIO

RESUMO	vii
ABSTRACT.....	viii
RESUMEN.....	ix
LISTA DE FIGURAS.....	x
LISTA DE TABELAS.....	xi
LISTA DE ABREVIATURAS E SIGLAS.....	xii
INTRODUÇÃO E REVISÃO DE LITERATURA.....	13
1. Definição e importância da Metagenômica.....	13
2. Interesse biotecnológico	17
3. Potencial do solo paranaense.....	18
JUSTIFICATIVA.....	20
OBJETIVOS.....	21
MATERIAIS E MÉTODOS.....	22
1. Estratégia de trabalho.....	22
2. Análise bioinformática.....	23
3. Desenho e análise de oligonucleotídeos iniciadores	24
4. Amplificação dos genes de interesse	24
5. Clonagem dos produtos de amplificação.....	25
6. Expressão da enzima de interesse.....	26
7. Purificação de proteína por cromatografia de afinidade	28
8. Espectrometria de massas	29
9. Caracterização funcional da proteína	29
RESULTADOS E DISCUSSÃO.....	30
1. Seleção de sequências gênicas	30
2. Amplificação dos genes de interesse	33
3. Clonagem dos produtos de amplificação.....	35
4. Expressão de proteínas	37
5. Análises bioinformáticas das proteínas C e D.	48
6. Proteína ativadora de lipase.	50
7. Análise da aproximação metagenômica	52
CONCLUSÕES E CONSIDERAÇÕES FINAIS	54
REFERÊNCIAS	56
SUPLEMENTOS.....	60

RESUMO

A metagenômica é o estudo do material genético presente em uma amostra, de forma independente de cultivo e identificação. Esta análise tem fornecido muita informação sobre o potencial genético e da capacidade metabólica e funcional de uma comunidade microbiana. Calcula-se que a grande maioria dos microrganismos ainda permanecem desconhecidos nas condições experimentais atuais. O solo, por sua grande diversidade em microrganismos, oferece um grande potencial em produtos naturais de interesse para indústrias biotecnológicas, incluindo as enzimas como efetivos biocatalisadores. A biblioteca metagenômica utilizada neste projeto foi construída a partir de amostras de solo do Paraná de área natural conservada da Floresta Atlântica e de um solo artificial rico em conteúdo lipídico. Essa biblioteca foi analisada utilizando ferramentas bioinformáticas para a identificação de nove sequências gênicas codificadoras para enzimas com algum interesse biotecnológico. Essas sequências foram amplificadas, três delas foram clonadas em vetores de expressão, e verificada sua expressão em *Escherichia coli*. Uma enzima transcriptase reversa não foi expressa e as enzimas lipase e prolina aminopeptidase foram superexpressas, entretanto apresentaram-se insolúveis apesar de diversas modificações incluindo estirpe bacteriana, temperatura de incubação, concentração de sal, entre outros. Análises realizadas com a lipase indicaram a necessidade de uma proteína foldase (chaperona) para a obtenção de uma enzima ativa. A identificação do fosmídeo da biblioteca metagenômica contendo o gene para a lipase poderá levar à identificação da sequência da foldase correspondente e posteriormente a sua coexpressão, permitindo a análise da atividade desta lipase.

Palavras chave: análise *in silico*, biotecnologia, expressão de proteínas, lipases, metagenoma, microrganismos.

ABSTRACT

Metagenomics is the study of genetic material present in a sample, independently of its culture and identification. This analysis has provided abundant information about genetic potential and metabolic and functional capability of a microbial community. It is estimated that most of microorganisms remain unknown yet under current experimental conditions. Because of its enormous microbial diversity, the soil offers a potential in interesting natural products for biotechnological industries, including enzymes as effective biocatalyzers. The metagenomic library used in this project was constructed from soil samples of a conserved natural area of the Atlantic Forest and from artificial source rich in lipid content, in Parana (Brazil). This library was analyzed using bioinformatic tools for the identification of nine nucleotide sequences codifying for enzymes of biotechnological interest. These sequences were amplified, three of which were cloned into expression vectors, and evaluated its expression in *Escherichia coli*. The reverse transcriptase enzyme was not expressed and lipase and proline aminopeptidase enzymes were overexpressed, however they were found in the insoluble fraction of the protein extract despite several modifications including bacterial strain, incubation temperature, and salt concentration, among others. Analyses carried out with the lipase pointed out for the necessity of a foldase protein (chaperone) in order to obtain an active enzyme. The identification of the fosmid from the metagenomic library containing the lipase gene will make possible the identification of the correspondent foldase and its posterior coexpression, leading to the activity analysis of this lipase.

Key-words: *in silico* analyses, biotechnology, protein expression, lipases, metagenome, microorganisms.

RESUMEN

La metagenómica es el estudio del material genético presente en una muestra, de forma independiente de su cultivo o identificación. Este tipo de estudio ha proporcionado abundante información sobre el potencial genético y la capacidad metabólica y funcional de una comunidad microbiana. Se estima que la gran mayoría de los microorganismos todavía permanecen desconocidos bajo las condiciones experimentales actuales. El suelo, por su gran diversidad en microorganismos, ofrece un gran potencial en productos naturales de interés para industrias biotecnológicas, especialmente en enzimas como efectivos biocatalizadores. La biblioteca metagenómica utilizada en este proyecto fue construida a partir de muestras de suelo de un área natural conservada de la Floresta Atlántica y de un suelo artificial rico en lípidos, en Paraná (Brasil). Esta biblioteca fue analizada utilizando herramientas bioinformáticas para identificación de nueve secuencias génicas que codifican para enzimas de algún interés biotecnológico. Esas secuencias fueron amplificadas, tres de ellas fueron clonadas en vectores de expresión y fue evaluada su expresión en *Escherichia coli*. De ellas, la enzima transcriptasa reversa no fue expresada y las enzimas lipasa y prolina aminopeptidasa fueron superexpresadas, aunque se presentaron como proteínas insolubles a pesar de varias modificaciones como estirpe bacteriana, temperatura de incubación, concentraciones de sal, entre otros. Análisis realizados con la lipasa señalaron la necesidad de una proteína foldasa (chaperona) para la obtención de la enzima activa. La identificación del fosmídeo de la biblioteca metagenómica que contiene el gen para la lipasa podrá conducir a la identificación de la secuencia de la correspondiente foldasa y posteriormente su coexpresión, permitiendo la evaluación de la actividad de esta lipasa.

Palabras clave: análisis *in silico*, biotecnología, expresión de proteínas, lipasas, metagenoma, microorganismos.

LISTA DE FIGURAS

Figura 1. Metagenômica e suas duas estratégias.	15
Figura 2. Eletroforese dos produtos de amplificação das sequências selecionadas.	35
Figura 3. Eletroforeses dos produtos de restrição dos clones de C, D e F.	36
Figura 4. Eletroforese do extrato de <i>E. coli</i> BL21(DE3) expressando as proteínas C e D.	38
Figura 5. Eletroforese do extrato de <i>E. coli</i> Rosetta expressando as proteínas C e D.	39
Figura 6. Placas de avaliação de atividade enzimática de colônias de <i>E. coli</i> Rosetta.	40
Figura 7. Eletroforese das frações solúveis contendo as proteínas C e D expressas em <i>E. coli</i> Rosetta submetidas a diferentes condições de surfactantes durante a lise celular.	42
Figura 8. Eletroforese das frações solúveis contendo as proteínas C e D expressas em <i>E. coli</i> Rosetta submetidas a diferentes condições de Triton X-100 e glicerol durante a lise celular.	43
Figura 9. Eletroforese das frações contendo as proteínas C e D submetidas durante a lise celular a diferentes concentrações de ureia.	44
Figura 10. Eletroforese das frações eluídas durante a purificação da proteína C.	45
Figura 11. Espectrometria Maldi-Tof dos peptídeos trípticos da proteína C.	46
Figura 12. Modelo tridimensional em fitas das proteínas C e D geradas <i>in silico</i> por homologia de proteínas.	49
Figura 13. Filograma das famílias de lipase.	50

LISTA DE TABELAS

Tabela 1. Condições testadas no tampão de lise para solubilização de proteínas.	27
Tabela 2. Composição do SDS-PAGE para análise de proteínas.....	28
Tabela 3. Enzimas de interesse biotecnológico selecionadas neste projeto....	31
Tabela 4. Características das sequências selecionadas para amplificação e clonagem gênica.	32
Tabela 5. Sequências dos pares de <i>primers</i> desenhados para cada produto..	33
Tabela 6. Condições de amplificação.....	34
Tabela 7. Vetores e sítios de restrição utilizados para a clonagem dos fragmentos amplificados.	36
Tabela 8. Identificação da proteína C (lipase) por espectrometria de massas.	47

LISTA DE ABREVIATURAS E SIGLAS

APS = Persulfato de amônio (*Ammonium Persulfate*)

BSA= albumina sérica bovina (*bovine serum albumin*)

DMSO= dimetilsulfóxido

DNA= ácido desoxirribonucleico

DO= densidade óptica

HCCA= ácido α -ciano-4-hidroxicinimico (*α -Cyano-4-hydroxycinnimic acid*)

IPTG= Isopropil β -D-1-tiogalactopiranosídeo

LB= Luria Bertani

MS= espectrometria de massas (*mass spectrometry*)

pb= pares de bases

PCR= Reação em Cadeia da Polimerase

rpm= rotações por minuto

SDS= dodecil sulfato de sódio

SDS-PAGE= Eletroforese em gel de poliacrilamida desnaturante (*sodium doceeryl sulfate-polyacrylamide gel eletroforesis*)

T_m= Temperatura de anelamento (*temperature of melting*)

TRIS= Tris(hidroximetil)-aminometano

UV= ultra violeta

xg= força relativa à gravitação

INTRODUÇÃO E REVISÃO DE LITERATURA

1. Definição e importância da Metagenômica

A metagenômica estuda os genomas totais dos microrganismos presentes em uma amostra. Ela é caracterizada como um conjunto de técnicas e também uma área de pesquisa que utiliza o material genético obtido diretamente do ambiente sem necessidade de cultivo prévio, identificação ou amplificação individual (Committee on Metagenomics: Challenges and Functional Applications, 2007). Igualmente, o termo metagenoma é definido como o DNA microbiano isolado de uma amostra ambiental e que representa o DNA coletivo de todos os microrganismos presentes (Streit e Daniel, 2010).

Desde que o termo metagenoma foi introduzido por Handelsman *et al.* (1998), esta análise tem fornecido muita informação sobre o potencial genético e da capacidade metabólica e funcional de uma comunidade microbiana. Sem diferenciar entre genes expressos e não expressos, a metagenômica tem duas principais funções de importância tanto acadêmica como industrial: a descrição da biodiversidade em termos ecológicos de riqueza e abundância de grupos taxonômicos; e a prospecção biotecnológica de novas enzimas, metabólitos e produtos (Schloss e Handelsman, 2003).

Calcula-se que 99% dos microrganismos em alguns ambientes sejam desconhecidos pela incapacidade de cultivo nas condições de laboratório (Schloss e Handelsman, 2003). Assim, esta nova ferramenta vem sendo utilizada na análise de muitos ambientes com uma rica biodiversidade tais como solos, oceanos, ar, tecidos vegetais e animais, incluindo humanos (Su *et al.*, 2012), como também em ambientes extremos e altamente especializados como águas termais, terrenos hipersalinos ou ambientes constantemente frios (Simon e Daniel, 2011).

Diferentes métodos analíticos independentes de cultivo têm sido desenvolvidos desde que Pace *et al.* (1985) propuseram a clonagem direta de DNA ambiental. O processo consiste na extração do DNA após um processo cuidadoso de purificação e separação dos diversos compostos orgânicos e inorgânicos da amostra. Este material genômico pode ser clonado em vetores adequados e mantido em bactérias

hospedeiras para gerar uma biblioteca metagenômica que pode ser submetida a triagens funcionais de alto rendimento para localizar atividades específicas; ou pode ser submetido à identificação de genes de interesse por homologia de sequências em bancos de dados. Dessa forma, a análise metagenômica para identificação de novos genes pode ser desenvolvida através de duas estratégias: uma baseada na função e outra baseada na sequência. A seleção depende de vários fatores como o tipo de biblioteca construída, o *loci* genético ou atividade funcional de interesse, o tempo e recursos disponíveis para caracterizar a enzima (Kakirde *et al.*, 2010).

No método baseado na função, é pesquisada na biblioteca metagenômica a expressão de um fenótipo particular conferido pelo DNA clonado a uma bactéria hospedeira, como por exemplo, *Escherichia coli* (Figura 1). Esta estratégia envolve uma triagem de alto rendimento permitindo a identificação do clone recombinante que possui o gene de interesse e gera um produto ativo. Esta metodologia dependente da expressão tem sido aplicada na seleção de clones positivos que adquiriram resistência a antibióticos ou a metais pesados (Schmieder e Edwards, 2012). A princípio, esta triagem baseada na atividade requer a expressão concertada dos genes ambientais localizados em um fragmento dado de DNA, independentemente do seu tamanho e estrutura, sendo que então, o reconhecimento de promotores e elementos reguladores pela maquinaria de tradução do hospedeiro de expressão é essencial para a expressão funcional de bibliotecas de genes (Troeschel *et al.*, 2010).

Por sua vez, o método baseado na sequência de nucleotídeos está fundamentado na identificação por comparação de sequências que codificam enzimas de interesse ou de motivos funcionais altamente conservados para o desenho de *primers* para amplificação por PCR e sondagem de genes de bibliotecas de DNA (Ferrer *et al.*, 2009; Schloss e Handelsman, 2003). Alternativamente, em consequência das novas metodologias de sequenciamento de DNA, pode ser realizado o sequenciamento direto do DNA ou metagenoma, com ou sem clonagem prévia. A partir das informações obtidas, a caracterização da proteína envolverá a amplificação da sequência gênica de interesse a fim de permitir sua expressão numa célula hospedeira e sua caracterização (Figura 1).

Esta metodologia baseada na similaridade de sequências, embora necessite de um menor número de etapas experimentais, requer uma alta capacidade tecnológica não apenas de sequenciadores de DNA como programas de bioinformática. Atualmente têm sido desenvolvidas muitas ferramentas na análise de grandes quantidades de informação que facilitam a predição comparação e anotação de genes,

fazendo com que este método seja uma grande possibilidade na funcionalidade biológica de microrganismos (Lee e Lee, 2013).

Suas principais limitações são: 1) baixa cobertura do metagenoma, uma vez que genes de diferentes organismos estão presentes em diferentes concentrações no DNA usado para o sequenciamento, 2) a integração e filtração de sequências gênicas e evidências experimentais para designação funcional de genes desconhecidos, organismos, comunidades e até redes funcionais, e 3) aspectos computacionais de arquivamento, análise e visualização de um vasto número de sequências de DNA liberados nas bases de dados (Vieites *et al.*, 2010).

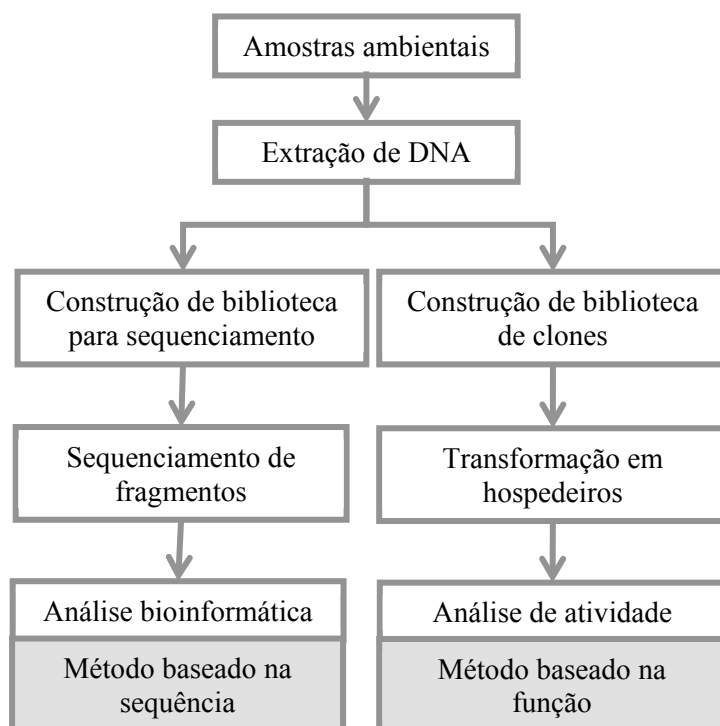


Figura 1. Metagenômica e suas duas estratégias.

Representação esquemática dos tipos de análises para obtenção de novas moléculas. Modificado de Schmieder e Edwards (2012).

A escolha de cada um desses dois métodos apresenta vantagens e desvantagens. O método baseado na função permite a identificação de um clone que expresse enzimas ativas de interesse na aplicação biotecnológica; entretanto, clones podem não ser identificados em decorrência da ausência de expressão no tipo de

hospedeiro sob condições experimentais usadas, gerando relativamente poucos produtos novos (Ekkers *et al.*, 2012).

Por outro lado, no método baseado na sequência, em teoria não existe limite na identificação de genes e a sua expressão pode ser desenhada e ajustada ao produto de interesse, obtendo maior alcance de exploração da diversidade genética, limitada apenas pela capacidade bioinformática, informação existente nas bases de dados (Faoro, 2010). Além disso, avanços na tecnologia de sequenciamento de DNA de segunda geração continuam estendendo as possibilidades da metagenômica. Em comparação à técnica tradicionalmente mais usada de pirosequenciamento (454/Roche), o sequenciamento utilizando leituras curtas (SOLiD e Illumina) é considerado um novo paradigma, já que permite maior rendimento e cobertura com custos muito menores (Teeling e Glockner, 2012).

A rápida e substancial redução de custos no sequenciamento de nova geração tem acelerado o desenvolvimento da metagenômica baseada na sequência e superando a necessidade de clonagem pelo sequenciamento do DNA ambiental diretamente (Thomas *et al.*, 2012; Teeling e Glockner, 2012). Algumas das plataformas de sequenciamento de nova geração incluem o pirosequenciador GS-FLX 454 (Roche), MiSeq, HiSeq, e Genome Analyzer II (Illumina), SOLiD system (Life Technologies/Applied Biosystems), Ion Torrent (Life Technologies), e o PacBio RS II (Pacific Biosciences) (Culligan *et al.*, 2013).

Assim, os avanços nas tecnologias do sequenciamento acompanharam a extensão do uso da metagenômica em maior complexidade. Adicionalmente, a metagenômica oferece altas perspectivas que será possivelmente utilizada de forma tão comum e frequente como qualquer outro método de laboratório (Thomas *et al.*, 2012).

Uchiyama e Miyazaki (2010) desenvolveram um terceiro método de triagem metagenômico denominado expressão de genes induzida pelo substrato, SIGEX (do inglês *substrate-induced gene expression*). Nele, a biblioteca metagenômica é construída utilizando um vetor que contém um gene repórter *downstream* do sítio de inserção para clonagem. Assim, se a expressão do gene é ativada, os clones positivos são identificados pelo sinal derivado do produto do gene repórter. Este método foi desenvolvido para detecção de genes catabólicos, mas não detecta atividade enzimática; ainda assim, em uma triagem de ultra-alto rendimento, que pode detectar até 30.000 clones por segundo.

2. Interesse biotecnológico

Os microrganismos têm um papel fundamental na história do desenvolvimento biotecnológico desde vários milênios atrás. Os primeiros processos envolvidos mediante a utilização de fermentações datam do ano 6000 a.C. com a preparação de bebidas alcoólicas e pão.

Em seu ambiente natural os microrganismos podem conter genes que codificam ou geram vias biodegradativas ou biossintéticas de interesse acadêmico e industrial não previamente identificadas por métodos dependentes de cultivo (Kakirde *et al.*, 2010). Por isso, a diversidade biológica é um importante recurso não somente pelos serviços ambientais que disponibiliza, mas também como potencial recurso no desenvolvimento de oportunidades para bioprospecção (Pylro *et al.*, 2013).

Entre esses produtos naturais de interesse estão incluídas as enzimas como efetivos biocatalisadores, que visam melhorar e gerar processos de produção mais limpos, reduzir energia e matéria prima para produção de biocombustíveis renováveis, degradação de poluentes e geração de novos medicamentos (Kennedy *et al.*, 2008; Pylro *et al.*, 2013).

Como foi afirmado por Lorenz e Eck (2005), diferentes indústrias têm interesse em explorar os recursos dos microrganismos não cultivados por diversas razões: 1) As enzimas são catalisadores ideais e funcionam eficientemente em aplicações específicas; 2) oferecem novidade em sequências gênicas para evitar infringir patentes na aplicação de processos tecnológicos; 3) apresentam uma máxima diversidade como ferramentas para biotransformações, em comparação com a química tradicional sintética; e 4) produzem metabólitos que escapam das condições regulares de laboratório mas que podem ser clonados e expressos em forma heteróloga. A indústria mundial de enzimas é a maior beneficiada para detergentes, aplicações alimentícias, agricultura, processamento têxtil, papel e couro. Assim, a metagenômica fornece uma oportunidade sem precedentes para uma aplicação industrial.

Em estudos atuais da metagenômica encontram-se numerosos produtos com potencial aplicação: glicosil hidrolases, esterases, lipases, porfirinas, moléculas de 'quorum sensing', celulases, xilanases, quitinases, fosfatases, pectinases, amilases, lacases, lactonases, antibióticos, genes de resistência a antibióticos, peptídeos antifúngicos, DNA polimerases, elementos genéticos móveis, probióticos, loci tolerantes a sal, sintetases de policetídeos, enzimas degradadoras de fenol,

degradadoras de dimetilsulfopropionato, proteínas fluorescentes, genes do metabolismo de polihidroxibutirato, etc (Lee e Lee, 2013; Ferrer *et al.*, 2009).

Uma das novas enzimas mais predominantes encontradas a partir de metagenoma de solo é a esterase/lipase, como importante biocatalisador para múltiplas aplicações biotecnológicas, muitas delas comercialmente disponíveis (Böttcher *et al.*, 2010; Lee e Lee, 2013). Suas características mais interessantes são o não requerimento de cofatores, destacada estabilidade em solventes orgânicos, ampla especificidade de substrato, estereoseletividade, e seletividade posicional (Lee *et al.*, 2004).

3. Potencial do solo paranaense

O solo é o maior componente da maioria dos ambientes terrestres e é considerado o ecossistema com maior diversidade em comunidades microbianas nativas (Lee e Lee, 2013). Estima-se que um grama de solo pode conter até 10 bilhões de microrganismos e milhares de espécies diferentes incluindo bactérias, arqueas, vírus e microrganismos eucarióticos (Kakirde *et al.*, 2010; Delmont *et al.*, 2011; Teeling e Glockner, 2012). Esses microrganismos são responsáveis pela maior parte dos ciclos biogeoquímicos globais, que incluem os ciclos biogeoquímicos do carbono, nitrogênio, enxofre e fósforo (Su *et al.*, 2012). Sua importância também é dada devido a que as mudanças ambientais por fatores naturais ou pressão antropogênica são seguidas por respostas rápidas no metabolismo e na capacidade reprodutiva dos microrganismos ali presentes, resultando em alterações qualitativas e quantitativas da composição dos habitats (Kisand *et al.*, 2012; Faoro *et al.*, 2010).

Os microrganismos mais abundantes no solo pertencem aos filos *Proteobacteria*, *Acidobacteria*, *Actinobacterium*, *Firmicutes*, e *Verrucomicrobia* (Lee e Lee, 2013) podendo ser encontrados até 385 gêneros diferentes utilizando diversos protocolos (Delmont *et al.*, 2011). Mesmo assim, são os organismos menos estudados (Venter *et al.*, 2004).

O solo e sua ampla biodiversidade oferecem perspectivas promissoras na busca de novas funções de interesse biotecnológico (Van Elsas *et al.*, 2008), e seu enorme potencial pode ser revelado pelo uso da metagenômica como método abrangente.

Na aplicação de métodos independentes de cultivo tem-se realizado diferentes análises de acordo com os tipos de solo, seja com fins bioprospectivos ou de descrição ecológica. Por exemplo, a diferença na diversidade bacteriana dos solos para agricultura e as florestas; os solos de pastagem e sua resposta às mudanças climáticas; a influência da luz e baixos níveis de oxigênio em comunidades microbianas de cavernas; o solo da Antártica como um importante 'hotspot' de ambientes extremos; análises filogenéticas de solos associados à contaminação com metais e hidrocarbonetos, entre outros (Su *et al.*, 2012).

É estimado que o Brasil possui 20% da diversidade mundial em macro-organismos biológicos, como um dos 17 países catalogados como megadiversos em plantas e animais, mas apesar da sua importância, a diversidade microbiana ainda é considerada como pouca conhecida (Pylro *et al.*, 2013).

Na região sul do Brasil, no estado do Paraná, encontra-se o sistema montanhoso da Serra do Mar, que abriga a Floresta Atlântica, um dos 25 pontos mais biodiversos do planeta e que conta com proteção da UNESCO desde 1992 (Unesco, 2009). Esta área conservada tem sido caracterizada em sua diversidade de fauna e flora, e particularmente seu solo indicou altos níveis de diversidade influenciados pela altitude, fatores físico-químicos, temperatura, disponibilidade de oxigênio e baixa presença humana (Faoro *et al.*, 2010), constituindo-se um potencial sítio na prospecção metagenômica.

No entanto, a utilização de solo rico em conteúdo lipídico como fonte de DNA microbiológico também é uma das estratégias aplicadas na tentativa de melhorar as possibilidades de obter enzimas envolvidas na degradação de compostos lipídicos. Embora estas enzimas possam ser obtidas de fontes animais e vegetais, as fontes microbianas possuem características úteis tais como: baixo custo de produção, alto rendimento, diversidade em atividade catalítica, e mais ampla especificidade de substrato (Glogauer *et al.*, 2011).

JUSTIFICATIVA

Estudos anteriores realizados no Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná permitiram a construção de bibliotecas metagenômicas a partir de amostras com diferentes tipos de solo, que levaram a identificação de novas enzimas de interesse biotecnológico (Couto *et al.*, 2010; Glogauer *et al.*, 2011; Faoro *et al.*, 2011).

Com a aquisição de um sequenciador automático de DNA de segunda geração (sistema SOLiD™ 4, da *Life Technologies*) foi realizado o sequenciamento da biblioteca metagenômica SM construída a partir de amostras de solo Paranaense selecionadas. Este sequenciamento amplia o potencial de exploração dessas bibliotecas e permite a identificação de novos genes que codificam para proteínas de interesse biotecnológico.

OBJETIVOS

1. Objetivo geral

Realizar prospecção e caracterizar a atividade de novas enzimas de interesse biotecnológico em uma biblioteca metagenômica construída a partir de amostras de solo paranaense.

2. Objetivos específicos

- A partir do sequenciamento de DNA, realizar análise bioinformática de uma biblioteca metagenômica para enzimas de interesse industrial.
- Amplificar e clonar os genes de interesse identificados.
- Superexpressar, purificar e descrever a atividade da enzima identificada.

MATERIAIS E MÉTODOS

1. Estratégia de trabalho

A biblioteca metagenômica utilizada foi construída a partir de amostras de solo da Floresta Atlântica Paranaense e de solo contaminado com gordura animal da estação de tratamento de efluentes de uma indústria de processamento e embalagem de carnes e laticínios localizada em Carambeí, Paraná (Faoro *et al.*, 2011; Glogauer *et al.*, 2011). Estas bibliotecas já foram testadas na prospecção, isolamento e caracterização de outros produtos biotecnológicos (lipases, quitinases, amilases e microbicidas).

Na construção da biblioteca metagenômica, as amostras de solo foram submetidas a processo de purificação de DNA, subclonagem em fasmídeos e transformação em células *E. coli* estirpe EPI300. Foram pré-selecionados os clones que mostraram alguma característica de interesse (halo de atividade ou coloração) pelo método baseado na função. Utilizou-se triagem com meios contendo tributirina ou tricaprilina, altas concentrações de cloreto de sódio, presença de cálcio e goma arábica. De um total aproximado de 2.500 clones com atividade lipolítica ou produção de composto colorido, o material fosmidial de 192 clones foi purificado individualmente, e posteriormente misturado e sequenciado utilizando a plataforma de sequenciamento SOLiD4 (Life Technologies). Os dados obtidos neste sequenciamento foram agrupados em sequências contíguas (*contigs*) e deram origem a uma biblioteca *in silico* denominada SM (do inglês *Soil Metagenome*). Esta etapa foi realizada pelo Dr. Helisson Faoro, no Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná.

Neste projeto essa biblioteca foi alvo de uma prospecção, analisando utilizando bioinformáticas para a identificação de sequências gênicas codificadoras para enzimas com algum interesse biotecnológico. Essas sequências foram amplificadas e clonadas em vetores adequados e a expressão de cada enzima foi verificada em cepa bacteriana. Duas enzimas foram selecionadas para serem caracterizadas funcionalmente.

2. Análise bioinformática

O sequenciamento da biblioteca metagenômica SM resultou num total de $2,7 \times 10^6$ nucleotídeos agrupados em 18265 *contigs*. Estas sequências foram analisadas para a identificação de genes codificadores para enzimas com algum interesse biotecnológico.

Esta prospecção *in silico* consistiu nos seguintes passos: a busca por similaridade de sequências e conservação de domínios utilizando a plataforma *CLC WorkBench* (QIAGEN®), o programa *MEGAN* (Huson *et al.*, 2011), assim como a análise manual utilizando a ferramenta *BLAST* (NCBI). O *CLC Genomic WorkBench*, como plataforma de análise de dados para DNA, RNA e proteínas, permitiu fazer o alinhamento massivo dos *contigs* sob certos parâmetros e manipular a informação numa interface gráfica. Os *contigs* obtidos foram comparados com o banco de dados não redundante (*nr*, do inglês *non-redundant*) para identificação de potenciais alvos. O programa *MEGAN* fez a análise dos *contigs* a partir dos resultados da comparação com o banco *nr*. Utilizando a informação da base de dados *KEGG* (Kanehisa *et al.*, 2004) foi possível criar árvores de grupos funcionais que facilitaram a identificação de enzimas envolvidas numa determinada via metabólica.

Após a primeira seleção de sequências foi feita avaliação de fases de leitura abertas (ORF, do inglês *Open Reading Frames*) utilizando o programa *FramePlot 2.3.2* (Ishikawa e Hotta, 1999) baseado na busca de códons de início e parada de síntese proteica. Também foi feita análise de domínios conservados de proteína utilizando *Conserved Domains* (NCBI) e *InterPro Scan* (EMBL-EBI), visando prever a função biológica a partir dos domínios funcionais.

Já os parâmetros utilizados para a seleção dos genes de interesse foram os seguintes: sequências maiores do que 500 pb que aumentam a probabilidade de identificar uma sequência codificante completa; identidade em torno de 50% com o primeiro *hit* da busca em *BlastX*, que permitiu identificar sequências tanto conservadas quanto novas; cobertura da sequência alvo próxima a 100% em relação a uma enzima provavelmente completa; organismo alvo que fosse diferente de *E. coli*, que foi o organismo utilizado para construção da biblioteca metagenômica; conteúdo de citosinas e guaninas (C+G) na sua sequência nucleotídica entre 50 e 70%, ausência de domínios proteicos transmembrana, pela dificuldade de solubilização durante a purificação e exclusão daquelas com domínios de função desconhecida (ou DUFs), e aquelas contendo peptídeo sinal para facilitar as condições de purificação.

A seleção definitiva das sequências foi baseada na revisão de literatura para pesquisa de enzimas com atividade de interesse biotecnológico e por sua aplicabilidade principalmente nas indústrias têxtil e farmacêutica.

Foi feita uma posterior avaliação *in silico* dos parâmetros proteicos destas enzimas, tais como: características físico-químicas utilizando o programa *ProtParam* (Gasteiger *et al.*, 2005); análise de utilização de códons de baixa frequência em *E.coli* (<http://people.mbi.ucla.edu/sumchan/caltor.html>) e em outros organismos (Fuhrmann *et al.*, 2004); análise hidrofobicidade utilizando o programa *Peptide Property Calculator* (https://www.genscript.com/ssl-bin/site2/peptide_calculation.cgi); modelagem tridimensional *in silico* por homologia de proteínas em *Swiss-Model* (Arnold *et al.*, 2006); análise filogenética das famílias de lipases estabelecidas por Arpigny e Jaeger (1999) utilizando o programa *ClustalW* (Larkin *et al.*, 2007) pelo método *neighbor-joining*, utilizando as ferramentas da *EMBL-EBI* (Goujon *et al.*, 2010); análise de vizinhança de genes procarióticos e predição de associações de ação interativa pelo *String* (Franceschini *et al.*, 2013); alinhamento múltiplo de sequências por *T-coffee* (Di Tommaso *et al.*, 2011; Notredame *et al.*, 2000) e busca de sítios proteicos funcionais utilizando as ferramentas *Prosite* (Sigrist *et al.*, 2013) e *Pfam* (Punta *et al.*, 2012).

3. Desenho e análise de oligonucleotídeos iniciadores

Para amplificação das sequências gênicas selecionadas da biblioteca metagenômica, foram desenhados oligonucleotídeos iniciadores (*primers*) com os seguintes critérios: clonagem nos vetores de expressão pET-28a ou pET-29a (Novagen), inserção dos sítios de restrição adequados para clonagem e expressão da proteína contendo um polipeptídeo de 6 histidinas (cauda His-tag) na região N-terminal ou C-terminal da proteína.

As sequências dos *primers* foram analisadas utilizando o Programa *Oligo Analyzer* (IDT) para determinação de comprimento adequado, conteúdo de C+G em torno a 50%, temperatura de anelamento próxima entre o par de *primers*, e menor possibilidade de formação de grampo e dimerização entre os pares de *primers*.

4. Amplificação dos genes de interesse

As sequências escolhidas foram amplificadas no termociclador *Master Cycler Gradient* (Eppendorf) e foram padronizadas as condições de amplificação por PCR,

tais como temperaturas de anelamento dos *primers* com a fita molde, composição da solução tampão (KCl ou $(\text{NH}_4)_2\text{SO}_4$) e presença de aditivos (DMSO ou Betaina). Como molde para a amplificação foi utilizada a mistura de DNA dos 192 clones fosmidiais da biblioteca metagenômica (SM). Foram utilizadas *Taq* Polimerase de produção local ou *Phusion High-Fidelity* (Thermo Scientific), enzima de alta fidelidade e processividade.

Os produtos das amplificações foram analisados por eletroforese em gel de agarose 1% corados com brometo de etídio e visualizados em transiluminador UV acoplado a um sistema de fotodocumentação (UVP).

Quando necessário, foi feita purificação do produto de PCR ou de restrição a partir do gel utilizando *NucleoSpin Gel and PCR clean-up* (Macherey-Nagel).

5. Clonagem dos produtos de amplificação

Os produtos de amplificação foram clonados em vetores adequados segundo procedimento descrito (Sambrook *et al.*, 1989). Duas estratégias foram utilizadas: inserção dos produtos de PCR diretamente no vetor pTZ57R utilizando sistema de clonagem A/T (Thermo Scientific) ou inserção nos vetores pET-28a ou pET-29a (Novagen) utilizando T4 DNA ligase (Thermo Scientific) após restrição com endonucleases de restrição adequadas. No primeiro caso os plasmídeos construídos foram posteriormente digeridos com enzimas de restrição adequadas para recuperação dos fragmentos de interesse e posterior subclonagem nos vetores pET-28 ou pET-29.

Bactérias hospedeiras de *E. coli* estirpe TOP10 (Invitrogen) foram preparadas para quimio-competência. Assim, um pre-inóculo da célula crescido durante a noite a 37°C em 3 mL de meio Luria Bertani (LB) (composto por Cloreto de Sódio (NaCl) 10 g/L, extrato de levedura 5 g/L e caseína hidrolisada 10 g/L). Foi inoculado conservando a relação de 1 mL do pré-inóculo em cada 100 mL de meio LB e foi crescido até atingir uma DO_{600} aproximada entre 0,4 e 0,6. Após crescimento as células foram lavadas 3 vezes em solução de CaCl_2 100 mM e finalmente ressuspensas em solução de Cloreto de Cálcio 85 mM contendo glicerol 15% v/v, aliquotadas em tubos de polipropileno e estocadas a -45°C. A transformação com os plasmídeos construídos foi feita por choque térmico utilizando solução de CaCl_2 conforme descrito por Sambrook *et al.* (1989). Desta forma, a suspensão contendo células quimio-competentes foi pré-incubada em gelo com o plasmídeo por 30 minutos, submetida a choque térmico de 42 °C por 1 minuto e colocada de novo em gelo, seguido por

incubação em meio LB por 30 minutos. Após esse período de recuperação, alíquotas de 150 μ L foram plaqueadas em meio LB-ágar (meio LB contendo ágar 15 g/L) contendo o antibiótico de seleção (canamicina 100 μ g/mL (Km^{100}) ou ampicilina 250 μ g/mL (Amp^{250})). Após incubação durante a noite a 37°C as colônias resistentes ao antibiótico foram analisadas.

A purificação de plasmídeos foi feita pelo método de lise alcalina (Sambrook *et al.*, 1989) a partir de colônia resistente ao antibiótico de seleção. Após crescimento durante a noite em meio LB na presença do antibiótico de seleção, 3 mL de cultura foram centrifugados a 14100 xg por 30 segundos e as células ressuspendidas em solução GET (glucose 50 mM, EDTA pH 8,0 10 mM e Tris-HCl pH 8,0 25 mM). As células foram lisadas com solução contendo SDS 1% e NaOH 0,2 M, e posteriormente o DNA genômico e proteínas foram precipitados com adição de solução KAcf (acetato de potássio 3M pH 5,2). A mistura foi centrifugada a 14100 xg por 5 minutos e o sobrenadante transferido para um novo tubo. A solução foi extraída com uma mistura de Fenol:Clorofórmio:Álcool isoamílico (25:24:1) e a fase aquosa foi recuperada após centrifugação a 14100 xg por 5 minutos. O DNA plasmidial foi precipitado com a adição de 0,6 volumes de Isopropanol puro, e recuperado por centrifugação a 14100 xg por 5 minutos. O precipitado foi lavado com 1 volume de etanol 70% e, após secagem em estufa a 37°C por 30 minutos, o DNA foi dissolvido em 30 μ L água ultrapura.

A confirmação de clonagem dos fragmentos de DNA foi realizada por análise de restrição e sequenciamento de DNA. A digestão com enzimas de restrição foi realizada conforme indicação do fornecedor da enzima. O sequenciamento de DNA foi feito baseado na interrupção da síntese por didesoxirribonucleotídeos utilizando o sistema *Big Dye Terminator* (Applied Biosystems) e sequenciador automático *Genetic Analyzer ABI3500* (Applied Biosystems). Os *primers* utilizados foram específicos para cada sequência. As sequências foram alinhadas utilizando o programa *ClustalW* e visualizadas no *BioEdit* (Hall, 1999).

6. Expressão da enzima de interesse

Para a expressão das enzimas de interesse em *E. coli* foram testadas as estirpes BL21 (DE3) e Rosetta (DE3)pLYSs (Novagen). Estas estirpes são capazes de expressar a RNA polimerase de fago T7, permitindo a expressão de proteínas

dependentes deste promotor. A estirpe Rosetta possibilita ainda a expressão de genes contendo códons de baixa utilização em *E. coli*.

Foi feita uma avaliação da atividade enzimática em meio sólido. A atividade lipase foi realizada em placa ágar com tributirina 1% para extrato celular bruto e com LA com tributirina 1% para célula inteira. A atividade protease em placa ágar com leite desnatado 2% para extrato celular bruto e LA com leite desnatado 2% para célula inteira. Foram incubados a 37°C, na presença ou ausência de IPTG 0,5 mM. A presença da atividade esperada foi visualizada através da formação de um halo translúcido ao redor da colônia correspondendo à hidrólise do composto.

Para a expressão da proteína de interesse a cultura bacteriana foi crescida em meio LB até atingir uma DO_{600} aproximada de 0,5 e foi induzida pela adição de 0,5 mM de IPTG ao cultivo bacteriano. Foram testados os seguintes tempos e temperaturas de indução: 2h/37°C, 3h/30°C, 6h/18°C e 20h/16°C, sob agitação constante a 250 rpm.

Após cultivo, as células foram recuperadas por centrifugação a 4700 xg por 10 minutos e posteriormente ressuspensas em tampão de lise contendo Tris-HCl 100mM pH 7 com diferentes condições testadas para obtenção de maior quantidade de proteína solúvel (Tabela 1). As células foram lisadas por sonicação utilizando pulsos de 10 em 10 segundos num total de 3 minutos. O extrato celular foi separado entre as frações solúvel e insolúvel por centrifugação a 13.000 xg por 5 minutos.

Tabela 1. Condições testadas no tampão de lise para solubilização de proteínas.

Parâmetro	Condições
Força iônica	NaCl, KCl (50 mM, 100 mM, 500 mM)
pH	6,8; 8,0; 8,8
Detergentes	Tween 20, Triton X-100, N-Lauroil Sarcosina (0,01%; 0,1%; 0,5%; 1%)
Glicerol	1%, 5%, 10%
DTT	1 mM
Ureia	2, 4, 5, 6, 8 M
	CellLytic™ IB (Sigma)

As proteínas do extrato celular bruto e das frações solúveis e insolúveis foram quantificadas pelo método de Bradford (Bradford, 1976) utilizando como padrão de proteína albumina de soro bovino (BSA) e analisadas de acordo com sua massa

molecular por eletroforese em condição desnaturante em gel de poli(acrilamida) (SDS-PAGE) 12%, (Laemmli, 1970). Os géis foram preparados com os reagentes descritos na Tabela 2. Para análise, 5 µg de proteína em tampão de amostra (Tris-HCl pH 6,8 62,5 mM, glicerol 10%, SDS 2%, β-mercaptoetanol 5% e azul de bromofenol 0,01%) após aquecimento a 100 °C por 3 minutos. A eletroforese foi realizada a 180 V durante 60 minutos em tampão de corrida (Tris base 25 mM, glicina 250 mM e SDS 1% p/v), coradas em solução de água destilada/metanol/ácido acético 5:5:1 com Coomassie-Blue R-250 (Sigma) 1% p/v durante a noite e descoradas por 1 hora em solução descorante (água destilada/metanol/ácido acético 5:5:1). A foto do gel foi capturada em sistema de fotodocumentação (UVP) com luz visível.

Tabela 2. Composição do SDS-PAGE para análise de proteínas.

Componente	Gel de separação 12%	Gel de empilhamento 4%
Água ultrapura	2,17 mL	1,58 mL
Tris-HCl pH 8,8 1,5M	1,25 mL	-
Tris-HCl pH 6,8 0,5M	-	650 µL
SDS 10%	50 µL	25 µL
Acrilamida/bis-acrilamida 19:1, 40% (Sigma)	1,5 mL	243,5 µL
Persulfato de amônio 10%	25 µL	12,5 µL
TEMED	2,5 µL	5 µL

7. Purificação de proteína por cromatografia de afinidade

Após preparação do extrato proteico, a enzima expressa como uma proteína de fusão com cauda His-tag amino- ou carboxi-terminal, foi purificada por cromatografia de afinidade em coluna *HiTrap* Chelating (GE Healthcare) carregada com níquel. Nessa coluna, a resina de Sepharose carrega-se com solução saturante de níquel e as histidinas da cadeia polipeptídica são ligadas nela até que outro composto deslocador como o imidazol, desfaça a ligação. A enzima foi eluída utilizando um gradiente crescente de imidazol até 1 mol/L em sistema automatizado de cromatografia líquida rápida de proteína (FPLC) *Äkta* (GE Healthcare).

As proteínas das frações eluídas foram quantificadas com reativo de Bradford (Sigma) utilizando um padrão de proteína de BSA e analisadas de acordo com sua massa molecular por eletroforese em gel 12% de poliacrilamida (SDS-PAGE) com as condições apresentadas no item 6. As frações contendo a maior quantidade de proteína e menor contaminação foram misturadas. Essas frações foram dialisadas em membrana de tamanho de poro de 12 kDa contra 2 L tampão de diálise (Tris-HCl pH 8,0 50 mM, KCl 100 mM, com e sem Triton X-100 0,5%), para excluir das amostras a maior quantidade de imidazol e das condições utilizadas para solubilização de proteínas (Tabela 1).

8. Espectrometria de massas

Confirmação da identidade da enzima purificada foi feita através de análise em espectrômetro de massa *MALDI-ToF/Tof Autoflex II* (Bruker Daltonics), pelo método “*Peptide Mass Fingerprinting*” (PMF) e MS/MS. Os espectros foram obtidos utilizando o software *FlexControl 2.0* (Bruker Daltonics) e analisados no programa *Flex Analysis 2.0* (Bruker Daltonics).

Para essa análise, amostras da proteína purificada foram tomadas a partir de corte na banda no gel de poliacrilamida, descoradas, desidratadas, submetidas a hidrólise de ligações Arginina-Lisina pela enzima tripsina, secagem e mistura com a matriz de ionização HCCA, segundo modificações do protocolo de digestão triptica de Shevchenko *et al.* (1996).

9. Caracterização funcional da proteína

A atividade enzimática de extrato celular e da proteína purificada foi avaliada em substratos artificiais. Para a lipase foi utilizado o *p*-nitrofenol-butilato 1 mM em tampão Tris-HCl pH 7,5 50 mM, que ao ser hidrolisado passa de incolor a cor amarela que pode ser detectado com $\lambda=410$ nm em espectrofotômetro. Para a atividade de protease foi utilizado colágeno marcado *Azo coll* (Sigma) em tampão fosfato de potássio monobásico pH 7,0 100 mM, que ao ser digerido enzimaticamente pode ser determinado por absorbância em 520 nm segundo protocolo de Chavira Jr *et al.* (1984).

RESULTADOS E DISCUSSÃO

1. Seleção de sequências gênicas

A partir dos 18265 *contigs* da biblioteca metagenômica SM foi feita uma seleção para aqueles com comprimento superior a 500 pb visando a identificação de sequências gênicas completas. Aproximadamente 1535 *contigs* foram selecionados nessa primeira etapa. Em função das diferentes rotas de processamento via KEGG envolvidas no metabolismo geral (Kanehisa e Goto, 2000), foram selecionados 2100 *contigs*. Após análise de fases de leitura aberta considerando a sequência completa, organismo identificado diferente de *E. coli* e conteúdo de citosinas e guaninas em torno a 50%, foram selecionados 700 *contigs*. Foi feita uma análise para domínios proteicos e foram selecionados 44 *contigs* com função biológica conhecida. Baseado em dados da literatura e análise de bioinformática, 9 sequências gênicas foram selecionadas (Tabela 3). A partir deste ponto as sequências e seus subprodutos serão nomeados com as letras da Tabela 4. As sequências de nucleotídeos e seus produtos estão disponíveis no Suplemento 1. Informações referentes aos domínios proteicos estão disponíveis no Suplemento 2 e os alinhamentos com o primeiro *hit* do *BlastX* estão no Suplemento 3.

Tabela 3. Enzimas de interesse biotecnológico selecionadas neste projeto.

Enzima	Reação envolvida	Importância	Referência
Glicosil hidrolase	Desramificação do glicogênio: hidrólise de ligações α -1,6-glicosídicas de cadeias externas fosforiladas do glicogênio.	Análise de polissacarídeos, produção de etanol como biocombustível e degradação de glicogênio.	(Song <i>et al.</i> , 2010)
Triptofanase	Conversão reversível de triptofano a indol, piruvato e amônia.	Produção industrial de L-triptofano para nutrição, alimentos e medicamentos.	(Kawasaki <i>et al.</i> , 1993)
Triacilglicerol lipase	Hidrólise de ligações éster carboxílicas para liberação de ácidos graxos e glicerol.	Amplas aplicações industriais: alimentos, detergentes, química e bioquímica.	(Gupta <i>et al.</i> , 2004)
Peptidase prolina específica	Catálise da liberação de um resíduo prolina N-terminal de um peptídeo pela clivagem do grupo amida na presença de Magnésio.	Degradação de compostos organofosforados; em alimentos no amadurecimento de queijos; aplicações biomédicas.	(Theriot <i>et al.</i> , 2009)
Metaloprotease	Hidrólise de ligações peptídicas com atividade dependente de Zinco em matriz extracelular, gelatina, queratina.	Processamento de compostos proteicos, e <i>design</i> de novas drogas inibitórias.	(Marchler-Bauer <i>et al.</i> , 2013)
Transcriptase reversa	Síntese de cDNA a partir de RNA.	Indústria farmacêutica e de pesquisa científica, estudos celulares e outros.	(Belfort <i>et al.</i> , 2011)
Deidratase	Biossíntese de poliidroxicanoato, na geração de fontes de carbono e energia pela degradação de ácidos graxos.	Produção de termoplásticos degradáveis, ou 'plástico verde', substituindo o plástico sintético.	(Wang <i>et al.</i> , 2010)
Amilase	Degradação do amido e outros substratos polissacarídeos pela hidrólise das ligações alfa-1,4 glicosídicas.	Produção de biocombustíveis, proteínas, açúcares e químicos.	(Sun <i>et al.</i> , 2010)

Tabela 4. Características das sequências selecionadas para amplificação e clonagem gênica.

As sequências foram renomeadas com letra única. Para cada atividade são apresentados o tamanho do gene, porcentagem de citosinas e guaninas e o *contig* correspondente da biblioteca metagenômica SM, assim como o organismo, identidade, e cobertura correspondentes ao primeiro *hit* do *BlastX* (NCBI).

Nome	Enzima (produto)	Tamanho (pb)	Conteúdo C+G (%)	Contig fonte	Organismo	Identidade (%)	Cobertura (%)
A.	glicosil hidrolase	2172	63,3	1139 ou AbFP2*	<i>Azospirillum brasilense</i> Sp245	76	88
B.	triptofanase	1419	54,9	623	<i>Singulisphaera acidiphila</i> DSM	58	74
C.	triacilglicerol lipase	914	56,8	335	<i>Rhodofera ferrireducens</i> T118	77	88
D.	peptidase prolina especifica	921	54,9	35	<i>Alicyclobacillus acidocaldarius</i> DSM 446	56	74
E.	metaloprotease	697	54,5	221	<i>Schlesneria paludicola</i>	76	95
F.	transcriptase reversa	1424	57,9	2122	<i>Syntrophobacter fumaroxidans</i> MPOB	67	80
G.	deidratase	502	55,7	2026	<i>Tistrella mobilis</i> KA081020-065	73	93
S.	amilase sem peptídeo sinal	1872	57,0	Mafil PB12**	<i>Candidatus Koribacter versatilis</i> Ellin345	48	98
W.	amilase com peptídeo sinal	1801	57,0	Mafil PB12**	<i>Candidatus Koribacter versatilis</i> Ellin345	48	98

*AbFP2 corresponde ao genoma de *Azospirillum brasilense* FP2, utilizado como fonte de DNA por sua alta similaridade com *A. brasilense* Sp245, encontrado na biblioteca metagenômica.

**Mafil P PB12 corresponde à biblioteca metagenômica da Floresta Atlântica Paranaense.

2. Amplificação dos genes de interesse

Foram desenhados pares de oligonucleotídeos iniciadores (“*primers*”) para cada sequência segundo os parâmetros descritos em materiais e métodos. As sequências dos *primers* desenhados e sua temperatura de anelamento estão descritas na Tabela 5.

Tabela 5. Sequências dos pares de *primers* desenhados para cada produto.

Estão sublinhados os diferentes sítios de restrição inseridos e a temperatura de anelamento (*T_m*).

Sequências reconhecidas pelas enzimas de restrição: *Nde*I: CATATG, *Eco*RI: GAATTC, *Xho*I CTCGAG, *Nhe*I: GCTAGC, *Nco*I: CCATGG, *Bam*HI: GGATCC.

Produto	Primers	Sequência	<i>T_m</i> (°C)
A	F-c01139	GCTCATATGGGATGCGAAGCCATG	64
	R-c01139	CAGGAATTCGACAGTTGGGTCGGTTC	
B	F-c623	GGCATATGAGATTTCCATCTGAGCCGTTC	56
	R-c623	CAGAATTCGATGAGGCAGAGCGCTGGAAG	
C	F-c335	GTACATATGCTGGCCGCCAGCGCCAC	67
	R-c335	GACTCGAGCAGACCCATGCCCTGCAG	
D	F-c35	GTAGCTAGCATGAACACGGTCAAACGG	53
	R-c35	CTACTCGAGTCAAACGAGCTTTCCCGCCTC	
E	F-c221	GTCCATGGAAGAGCGGCCGCTGACTG	48
	R-c221	CACTCGAGCAGTTTCTTATACGGCAAG	
F	F-c2122	GTAGCTAGCATGACGGCAAGAGACGCAGA	54
	R-c2122	GTCTCGAGTCATGAAGTACGCTGGACTTTG	
G	F-c2026	GTCATATGATTTTAATTAGTAGTCGGGT	51
	R-c2026	CTCTCGAGTCTTGCCTTCCTGCGCATC	
S	LSC29amy5s	ACTCATATGGCACAGCCGGCCGCTGGA	61
	LSC29amy3	TCGGATCCTGCACCCGATAGACCGTCA	
W	LSC9amy5w	GCTCATATGACAATCGCCGGTCCAATT	61
	LSC29amy3	TCGGATCCTGCACCCGATAGACCGTCA	

As condições determinadas e padronizadas para amplificação das sequências de interesse estão listadas na Tabela 6, sendo a fonte de DNA uma solução do material fosmidial misturado da biblioteca metagenômica SM. As sequências B, C, D, E, F, S e W foram amplificadas utilizando uma *Taq* DNA Polimerase de produção local e sem necessidade de aditivos. Por outro lado, a sequência A (glicosil hidrolase) não

foi amplificada apesar da utilização de uma DNA Polimerase termoestável de alta fidelidade e processividade, maior tempo de extensão e gradiente de temperatura de anelamento. Esta sequência apresentou maior similaridade com um gene de *Azospirillum brasilense* estirpe Sp245 (Tabela 4, Suplemento 3), assim, foi decidido utilizar o DNA genômico de *A. brasilense* estirpe FP2 como fonte para a amplificação do gene A. Esta escolha é decorrente da longa experiência deste laboratório com *A. brasilense* FP2 (Pedrosa e Yates, 1984). Para a amplificação do gene foram testadas diferentes temperaturas para anelamento dos *primers* e a presença dos aditivos Betaína e DMSO, sendo que foi conseguida uma amplificação com DMSO 5% (Tabela 6).

A sequência G foi amplificada com DNA polimerase termoestável de alta fidelidade e processividade, descrita em materiais e métodos, com a temperatura de anelamento calculada segundo recomendações do fabricante.

Tabela 6. Condições de amplificação.

Reagentes utilizados para amplificar as diferentes sequências.

Reagente	Sequências B, C, D, E, F, S e W	Sequências A e G
DNA Polimerase	Produção local	Phusion High-Fidelity (Thermo Scientific)
Tampão	10x (NH ₄) ₂ SO ₄ (Fermentas): 1x	5x Phusion HF: 1x
MgCl ₂ 25 mM	1 mM	-
dNTP 10 mM	0,3 mM	0,2 mM
Primer F 10 µM	1 µM	0,5 µM
Primer R 10 µM	1 µM	0,5 µM
DNA	200-300 µg	200-300 µg
Água ultrapura	Suficiente para 25 µL	Suficiente para 20 µL
Aditivos	-	DMSO 5%

Todas as sequências gênicas foram amplificadas com sucesso de acordo com o tamanho esperado (Figura 2). Considerando os diversos produtos de amplificação referente à sequência A (glicosil hidrolase), a banda com o tamanho esperado (2172 pb) foi purificada a partir do gel utilizando *NucleoSpin Gel and PCR clean-up* (Macherey-Nagel) e posteriormente clonada.

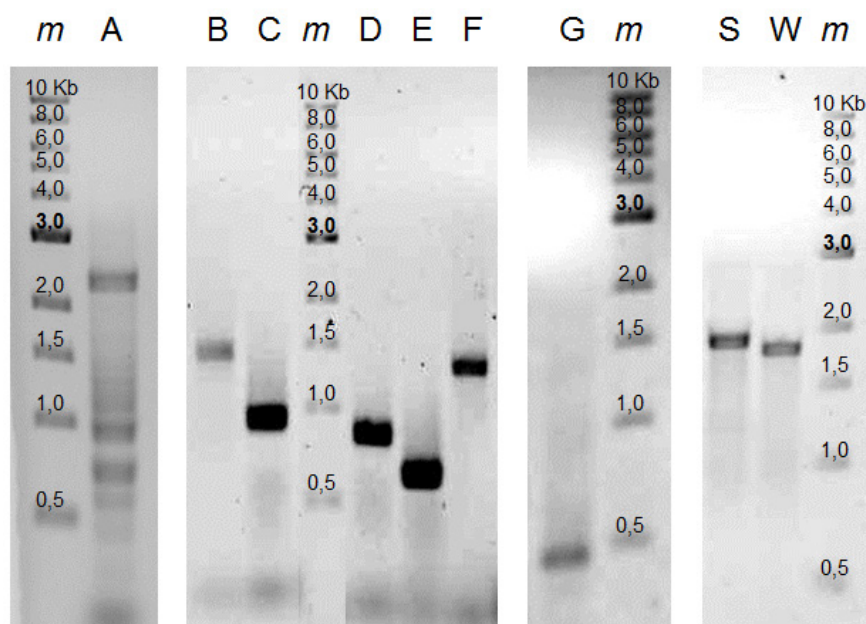


Figura 2. Eletroforese dos produtos de amplificação das sequências selecionadas.

Gel de agarose 1% em tampão TAE 1X. *m* indica os marcadores de massa molecular em kb (*1 kb ladder* da NEB). As demais linhas correspondem a cada um dos produtos gênicos segundo denominação descrita na Tabela 4. O DNA foi corado em solução de brometo de etídio (0,05%) e a imagem obtida em sistema UVP. Os tamanhos esperados em pb para cada produto incluindo as inserções dos sítios de restrição são: A: 2172, B: 1419, C: 914, D: 921, E: 697, F: 1424, G: 502, S: 1788, W: 1872.

3. Clonagem dos produtos de amplificação

Os produtos de amplificação foram digeridos com as endonucleases de restrição indicadas na Tabela 7 e posteriormente ligados no vetor correspondente pET-28a ou pET-29a (mapas mostrados no Suplemento 4) digerido com as mesmas enzimas. Esses produtos de ligação foram transformados em células hospedeiras como descrito na seção materiais e métodos.

Todos os fragmentos amplificados foram utilizados para clonagem nos vetores de expressão, entretanto, somente os produtos C, D e F foram confirmados. Esses vetores foram nomeados pET-C, pET-D e pET-F, respectivamente. A Figura 3 mostra o perfil de restrição desses clones. A confirmação de clonagem desses fragmentos também foi realizada por sequenciamento de DNA (dados não mostrados).

Tabela 7. Vetores e sítios de restrição utilizados para a clonagem dos fragmentos amplificados.

Os mapas e sítios de restrição dos vetores pET-28a e pET-29a estão mostrados no Suplemento 4.

Sequência	Enzima	Vetor	Enzimas de restrição
A	glicosil hidrolase	pET-29a	NdeI/EcoRI
B	triptofanase	pET-29a	NdeI/EcoRI
C	triacilglicerol lipase	pET-29a	NdeI/XhoI
D	peptidase prolina especifica	pET-28a	NheI/XhoI
E	metaloprotease	pET-28a	NcoI/XhoI
F	transcriptase reversa	pET-28a	NheI/XhoI
G	deidratase	pET-29a	NdeI/XhoI
S	amilase sem peptídeo sinal	pET-29a	NdeI/BamHI
W	amilase com peptídeo sinal	pET-29a	NdeI/BamHI

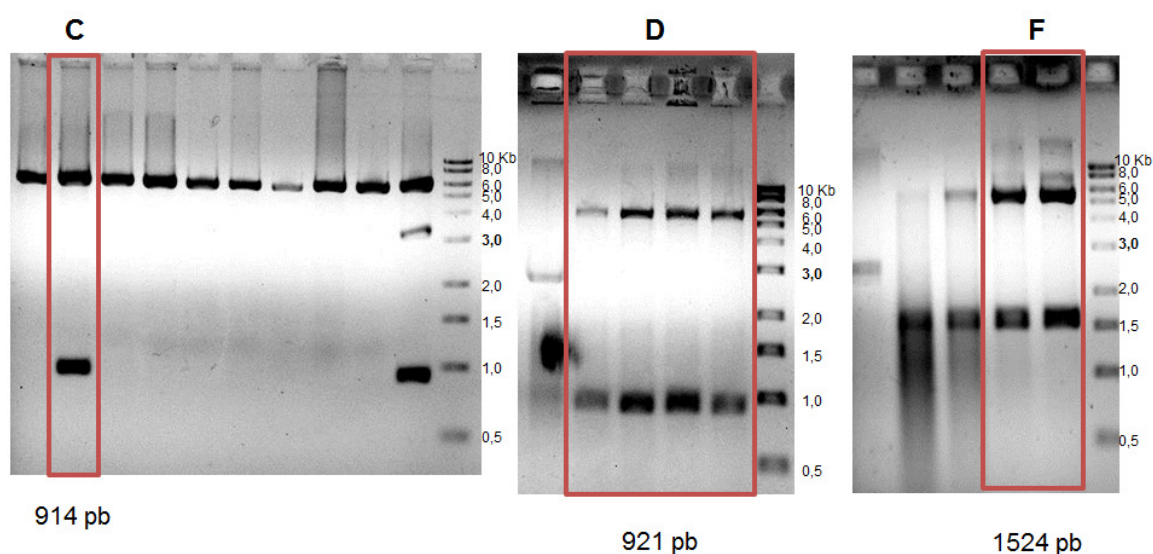


Figura 3. Eletroforeses dos produtos de restrição dos clones de C, D e F.

Gel de agarose 1% em tampão TAE 1X. São apresentados todos os clones avaliados para cada plasmídeo construído e indicados em quadro os clones corretos que foram selecionados. Na direita de cada gel está o marcador de pares de bases (NEB). Os clones foram digeridos com suas enzimas respectivas listadas na Tabela 7 para o qual se esperava a liberação do fragmento correspondente ao vetor pET-29a ou pET-28a com 5370 pb e o fragmento de tamanho da amplificação, indicado embaixo de cada gel. O clone F foi digerido com as enzimas *Xba*I e *Xho*I e o produto esperado corresponde a 100 pb de adição de acordo com o sítio de policlonação do vetor pET-28a (Suplemento 4).

Contudo, as sequências A, B, E, G, S e W foram abandonadas depois de múltiplas tentativas para resolver os inconvenientes próprios da clonagem encontrados na restrição, ligação, transformação e manutenção dos clones. Para alguns deles inclusive foi tentada a alternativa de ligar diretamente o produto de amplificação no vetor pTZ57R conforme o sistema de pontas coesivas A-T quando amplificado com a *Taq* DNA polimerase de produção local. Particularmente para as sequências B, E e G, foi encontrado que sua dificuldade foi decorrente da introdução do sítio de restrição perto das extremidades dos *primers*. Segundo recomendações dos fabricantes, as endonucleases envolvidas na digestão dessas sequências requerem pelo menos 3 bases nucleotídicas entre o sítio de restrição composto por 5 bases nucleotídicas e o extremo 3' ou 5'. Como observado na Tabela 5, nos *primers* F-c623, F-c221, F-c2026, existem somente 2 nucleotídeos de espaçamento, o que diminui até 100% em reações duplas sua chance para ser reconhecido pelas endonucleases de restrição.

4. Expressão de proteínas

Os clones confirmados pET-C, pET-D e pET-F, foram transformados na estirpe de *E. coli* BL21(DE3). Foi avaliada a superexpressão das proteínas de interesse por eletroforese em gel de poliacrilamida desnaturante (*SDS-PAGE*) do extrato bruto após lise celular das bactérias hospedeiras e separação das frações solúvel e insolúvel. A proteína C (lipase) e a proteína D (prolina aminopeptidase) foram superexpressas e permaneceram na fração insolúvel (Figura 4), mas a proteína F (transcriptase reversa) não foi superexpressa (não mostrado).

Após confirmação da expressão das proteínas C (lipase) e D (prolina aminopeptidase) e devido à insolubilidade das mesmas, foi testada a solubilização mediante uso de tampão de sonicação contendo NaCl ou KCl em alta e baixa concentração (500 e 50 mM) em diferentes pH (6,8; 8,0 e 8,8). As proteínas continuaram na fração insolúvel do extrato celular em todas as condições (dados não apresentados).

Apesar da maior parte da proteína expressa apresentar-se na fração insolúvel do extrato celular, foi avaliada a atividade enzimática lipase da proteína C em LA-tributirina 1% para bactéria e ágar-tributirina 1% para extrato celular bacteriano; e para atividade protease da proteína D em LA-leite 1% para bactéria e ágar-leite 1% para extrato celular, segundo indicado em materiais e métodos. Para nenhum dos testes foi encontrado halo de clareamento de atividade hidrolítica (dados não mostrados).

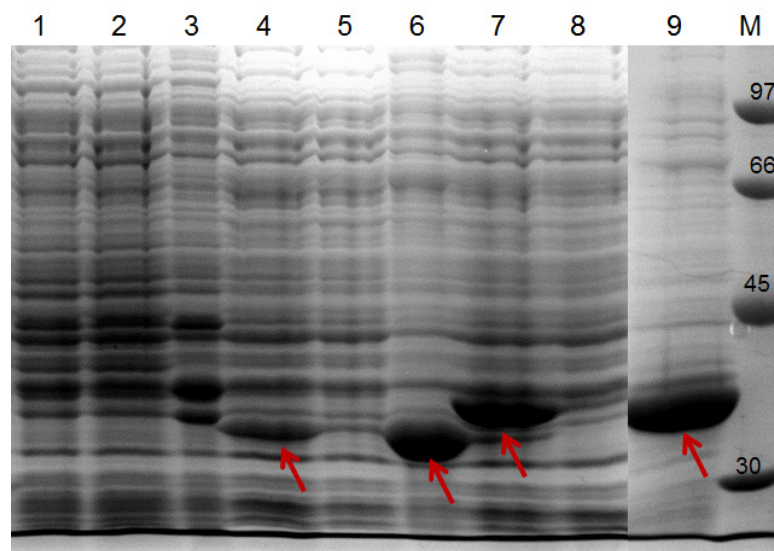


Figura 4. Eletroforese do extrato de *E. coli* BL21(DE3) expressando as proteínas C e D.

Gel de SDS poliacrilamida 12% (SDS-PAGE). Com setas estão indicadas as bandas correspondentes às proteínas C (32 kDa) e D (37 kDa). M indica o marcador de massa molecular em kDa (GE Healthcare). As proteínas foram coradas com *Coomassie blue*. **1:** extrato celular bruto *E. coli* BL21(DE3) controle sem transformação, **2:** controle fração solúvel, **3:** controle fração insolúvel, **4:** proteína C extrato celular bruto, **5:** proteína C fração solúvel, **6:** proteína C fração insolúvel. **7:** proteína D extrato celular bruto, **8:** proteína D fração solúvel, **9:** proteína D fração insolúvel.

Apesar da presença de códons raros na sequência de C (lipase), D (prolina aminopeptidase) e F (transcriptase reversa), as duas primeiras foram superexpressas apesar de apresentarem-se na fração insolúvel do extrato celular. Em vista disso, foi resolvido utilizar a estirpe *E. coli* Rosetta (DE3)pLysS a fim de verificar aumento de solubilidade (para C e D) e expressão (para F).

Após transformação na estirpe Rosetta, encontrou-se uma alta concentração das proteínas C (lipase), D (prolina aminopeptidase), como observado na Figura 5, no entanto, ao serem separadas as frações, estas apresentaram-se insolúveis.

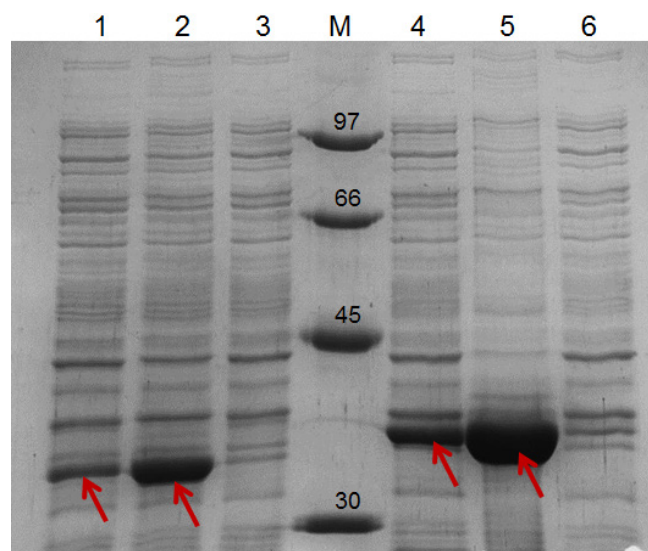


Figura 5. Eletroforese do extrato de *E.coli* Rosetta expressando as proteínas C e D.

Gel de SDS poliacrilamida 12% (SDS-PAGE). As proteínas foram coradas com *Coomassie blue*. Com setas estão indicadas as bandas correspondentes às proteínas C (32 kDa) e D (37 kDa). M indica o marcador de massa molecular em kDa (GE). **1:** proteína C extrato celular bruto, **2:** proteína C fração insolúvel, **3:** proteína C fração solúvel, **4:** proteína D extrato celular bruto, **5:** proteína D fração insolúvel, **6:** proteína D fração solúvel.

As bactérias hospedeiras estirpe Rosetta contendo o vetor pET-C (da lipase) plaqueadas em meio ágar com tributirina não apresentaram atividade, enquanto em meio LB-ágar com tributirina e com IPTG formaram halo de atividade hidrolítica depois de duas semanas de incubação (Figura 6a). Já com o vetor pET-D (da prolina aminopeptidase) a atividade foi evidenciada no extrato celular bruto com ágar leite desnatado 2% após uma semana de incubação (Figura 6b).

Os extratos celulares brutos também foram avaliados em placa de ágar contendo tributirina 1% ou leite desnatado 1% e 0,1%, na presença ou ausência de CaCl_2 0,1 mM como cofator de algumas enzimas hidrolíticas. Em todos os casos não houve a formação de halo frente ao controle positivo (Pronase 20 mg/mL) (dados não mostrados).

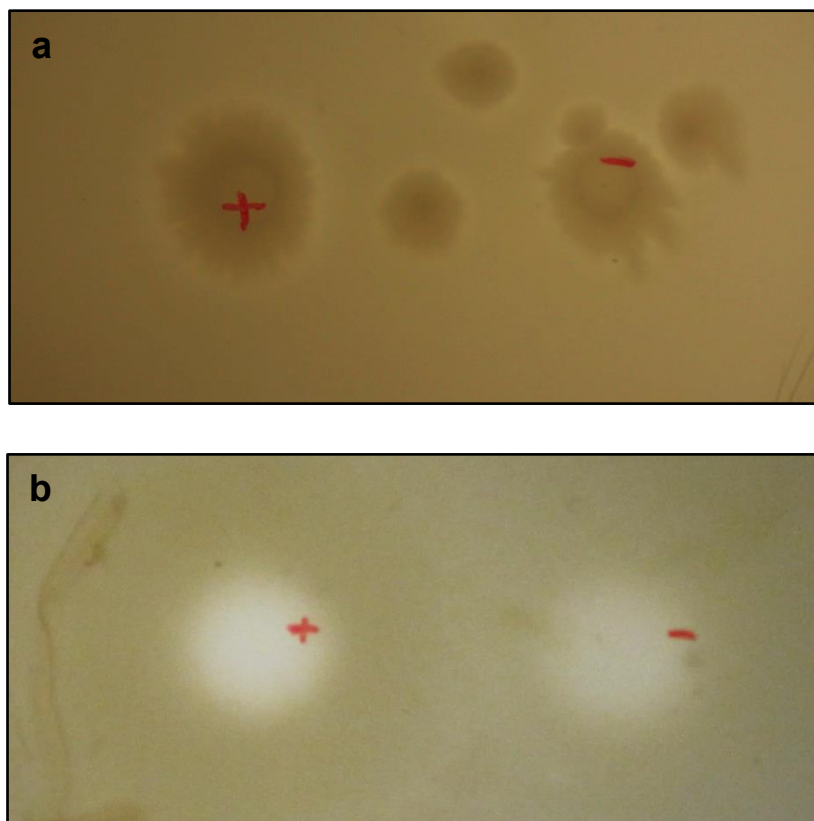


Figura 6. Placas de avaliação de atividade enzimática de colônias de *E.coli* Rosetta.

Observa-se halo de clareamento de meio em volta da colônia como indicativo de atividade enzimática extracelular do hospedeiro contendo o plasmídeo com inserto (+) e sem plasmídeo (-). **a:** Colônia contendo o vetor pET-C em meio LB-ágar com tributirina 1% e IPTG 0,5 mM após 2 semanas de incubação. **b:** Colônia contendo o vetor pET-D em meio ágar-leite desnatado 2% após 1 semana de incubação.

Apesar da transformação na estirpe Rosetta, a proteína F (transcriptase reversa) não foi superexpressa provavelmente devido ao elevado número de códons de baixa frequência em *E. coli* (Suplemento 5). Por análise *in silico*, sua expressão efetiva poderia ser conseguida em organismos vegetais tais como *Arabidopsis thaliana*, *Avena sativa*, *Anabaena variabilis*, *Oryza sativa*, *Zea mays*, *Musa acuminata*, *Nicotiana tabacum* ou *Pisum sativum*. Esta análise pode sugerir que esta sequência pode ser originária de genoma vegetal ou, por sua função transcriptase reversa, de retrovírus associados a tecidos vegetais, mesmo apresentando alta similaridade com *Syntrophobacter fumaroxidans* MPOB (Tabela 4), organismo cujo habitat são sedimentos de água doce e biorreatores anaeróbios (Plugge *et al.*, 2012).

Até agora, *E. coli* tem sido o organismo mais utilizado como hospedeiro de expressão na maioria dos estudos metagenômicos. Essa preferência ocorre por vários atributos: 1) possui uma alta eficiência de transformação, 2) permite diversidade no reconhecimento de sinais de expressão exógenos, 3) carece de genes de modificação de restrição e recombinação homóloga, e 4) versatilidade nos sinais de tradução de mRNA; não obstante essas vantagens, *E. coli*, como qualquer outro hospedeiro, é incapaz de expressar todo DNA exógeno devido a diferenças na maquinaria de transcrição, tradução e pós-tradução do organismo originário (Culligan *et al.*, 2013).

Algumas estratégias propostas por Culligan *et al.* (2013) para melhorar a expressão heteróloga são a utilização de hospedeiros alternativos ou duplos, a modificação de vetores e a otimização de códons; propostas por Uchiyama e Miyazaki (2009) são a engenharia de ribossomos e outros fatores envolvidos na transcrição e tradução, assim como tomar vantagem da biologia sintética como uma aproximação mais radical.

Como hospedeiros diferentes a *E. coli* têm sido utilizadas bactérias do solo pertencentes aos gêneros *Sphingomonas*, *Burkholderia*, *Bacillus*, *Acidobacterium* e *Verrucomicrobium*, e como alternativas incluem *Streptomyces*, *Rhizobium* e *Pseudomonas* spp. e mutantes de *Lysobacter enzymogenes* e *Pseudomonas fluorescens* (Van Elsas *et al.*, 2008).

As análises bioinformáticas de hidrofobicidade indicaram que as sequências de aminoácidos das proteínas C (lipase) e D (prolina aminopeptidase) apresentavam muitas regiões hidrofóbicas (Suplemento 6), podendo explicar a insolubilidade das proteínas expressas, apesar disso foram tentadas várias alternativas para aumentar a solubilização. Nas diferentes condições de tempo e temperatura de indução o resultado continuou semelhante, então foi estabelecida a temperatura de 30 °C e incubação por 3 horas para continuar com os próximos testes. Nem mudanças na força iônica ou pH do tampão de sonicação levaram a um aumento de solubilidade, tendo sido estabelecido o tampão contendo 50 mM Tris-HCl pH 6,8 e KCl 100 mM, condições que funcionam para uma ampla gama de proteínas em ambiente fisiológico. Deste modo, foi testada a adição de diversos componentes surfactantes que permitiram observar uma pequena diferença, como observado na Figura 7, com os detergentes Triton X-100 1% e N-Lauroil sarcosina em concentrações de 0,01%, 0,1%, e 1%, na presença de glicerol 10%.

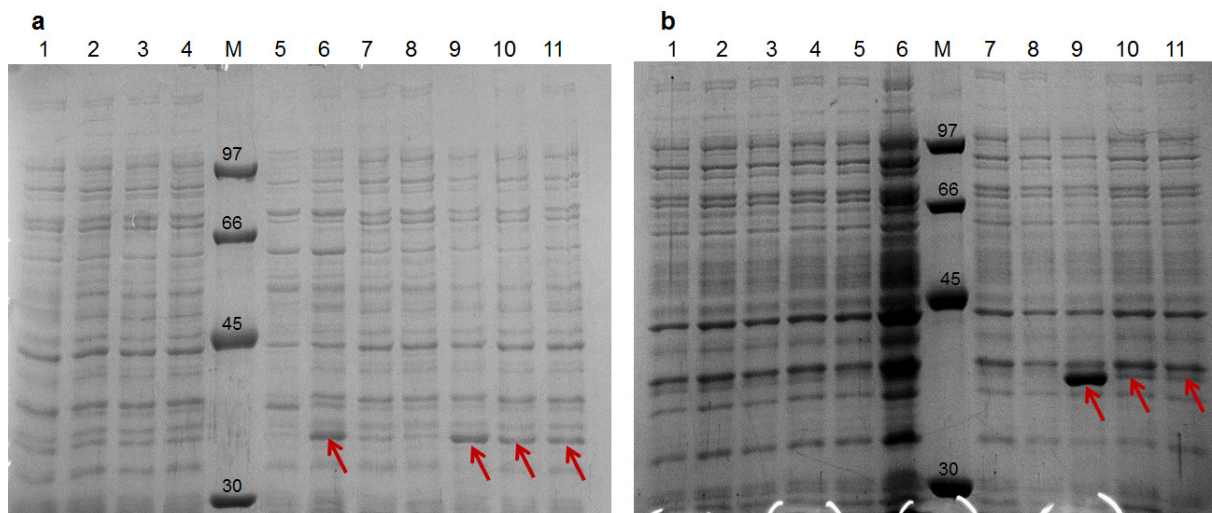


Figura 7. Eletroforese das frações solúveis contendo as proteínas C e D expressas em *E. coli* Rosetta submetidas a diferentes condições de surfactantes durante a lise celular.

Gel de SDS poliacrilamida 12% (SDS-PAGE). As proteínas foram coradas com *Coomassie blue*. M indica o marcador de massa molecular em kDa (GE Healthcare). **Painel a** corresponde às frações solúveis da proteína C (32 kDa) e **painel b** às frações da D (37 kDa). As setas indicam as proteínas de interesse em maior concentração na fração solúvel do extrato celular: **Linhas 1 a 3**, presença de Tween 20; **linhas 4 a 6**, presença de Triton X-100, e **linhas 7 a 9**, N-lauroil-sarcosina, todas nas concentrações de 0,01%, 0,1% e 1%, respectivamente. **Linha 10** DTT 1mM + glicerol 10%, e **linha 11**: Imidazol + glicerol 10%.

Considerando que as proteínas foram obtidas em maior quantidade na fração solúvel na presença de Triton X-100 e glicerol, foram testados esses dois componentes diferencialmente em diferentes concentrações, e foi encontrado que os dois juntos exerciam diferença (Figura 8), sendo selecionado Triton X-100 0,5% e glicerol 5% para testar atividades enzimáticas em substratos artificiais como descrito na seção 9 dos materiais e métodos. Essas atividades tanto do extrato bruto, quanto da fração solúvel das duas proteínas foram nulas (resultados não mostrados).

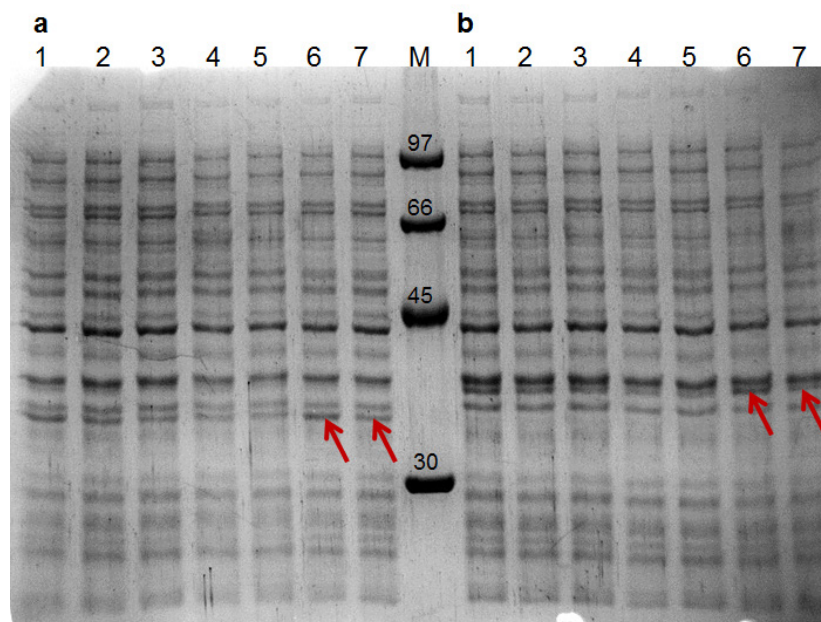


Figura 8. Eletroforese das frações solúveis contendo as proteínas C e D expressas em *E. coli* Rosetta submetidas a diferentes condições de Triton X-100 e glicerol durante a lise celular.

Gel de SDS poliacrilamida 12% (SDS-PAGE). As proteínas foram coradas com *Coomassie blue*. M indica o marcador de massa molecular em kDa (GE Healthcare). **Painel a** corresponde às frações da proteína C (32 kDa) e **painel b** às frações da D (37 kDa). Com setas estão indicadas as seleções das condições. **1:** Glicerol 10%, **2:** Glicerol 5%, **3:** Glicerol 1%, **4:** Triton X-100 1%, **5:** Triton X-100 0,5%, **6:** Glicerol 10% + Triton X-100 1%, e **7:** Glicerol 5% + Triton X-100 0,5%.

Em vista desses resultados, foi decidido solubilizar completamente a proteína e depois purificá-la. Essa solubilização foi realizada através de um gradiente de ureia no tampão de lise celular. Na Figura 9a é possível observar que a proteína C (lipase) foi solubilizada a partir de 4 M de ureia, enquanto que a proteína D (prolina aminopeptidase) foi solúvel a partir de 8 M (Figura 9b). A proteína solubilizada foi purificada por cromatografia de afinidade por ligação a metal. Frações obtidas foram analisadas por SDS-PAGE e são apresentadas na Figura 10. É possível observar que a proteína C foi purificada eficientemente e que há pouca presença de contaminantes. As frações contendo maior quantidade de proteína foram agrupadas e dialisadas para diminuir a concentração de ureia na amostra. Durante o processo de diálise, a proteína formou agregados evidenciando um processo de precipitação que também impediu quantificar apropriadamente a proteína purificada e cuja atividade enzimática foi nula com substratos artificiais como descrito na seção 9 dos materiais e métodos.

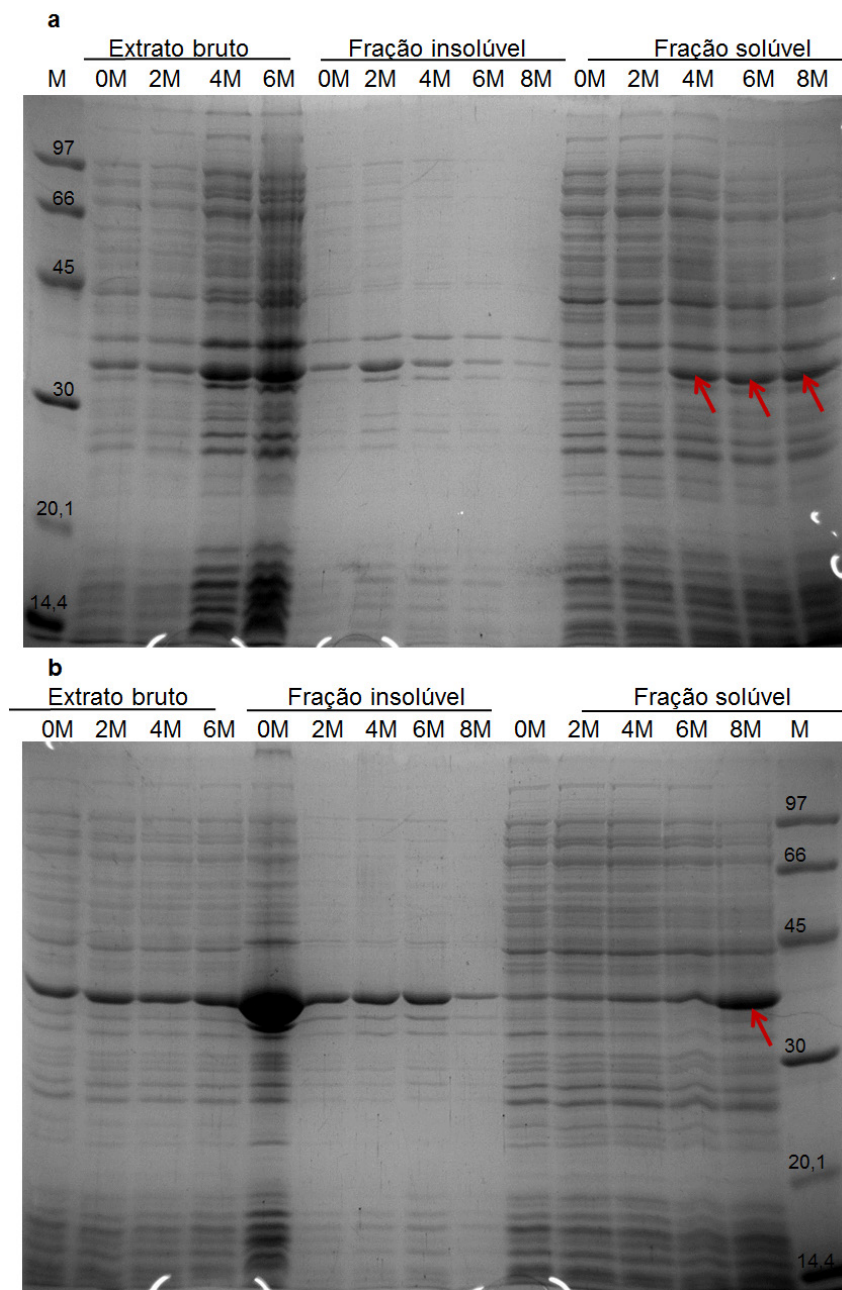


Figura 9. Eletroforese das frações contendo as proteínas C e D submetidas durante a lise celular a diferentes concentrações de ureia.

Gel de SDS poliacrilamida 12% (SDS-PAGE). As proteínas foram coradas com *Coomassie blue*. M indica o marcador de massa molecular em kDa. Mostram-se as concentrações de ureia em Molaridade. Com setas estão indicadas as proteínas em melhores condições de solubilidade. **Painel a** corresponde às frações da proteína C (32 kDa) e **painel b** às frações da D (37 kDa).

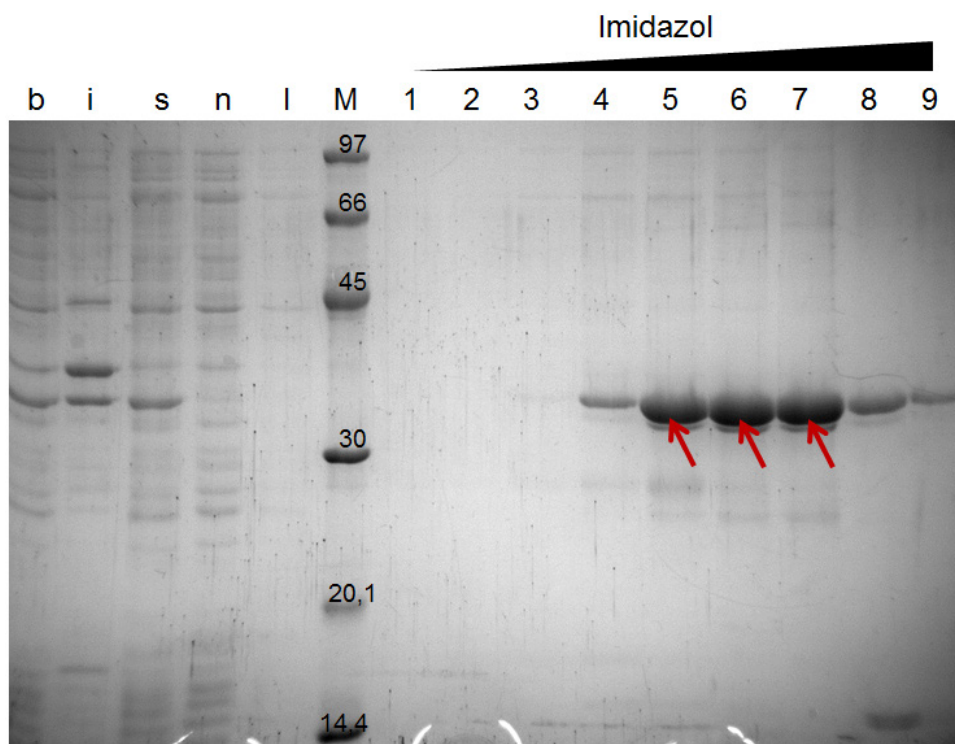


Figura 10. Eletroforese das frações eluídas durante a purificação da proteína C.

Gel de SDS poliacrilamida 12% (SDS-PAGE). As proteínas foram coradas com *Coomassie blue*. M indica o marcador de massa molecular em kDa (GE Healthcare). As setas indicam as frações coletadas contendo a proteína C (31,83 kDa). **b**: extrato celular bruto, **i**: fração insolúvel, **s**: fração solúvel, **n**: fração de não ligação na coluna, **l**: fração da lavagem da coluna. **1-9**, primeiras 9 frações eluídas com o gradiente de imidazol.

O Reagente CellLytic™ IB (Sigma) foi utilizado como uma alternativa na solubilização de agregados de proteínas ou corpos de inclusão substituindo o tampão de lise, segundo instruções do fabricante. A proteína foi purificada de forma semelhante à solubilização com ureia, entretanto, após diálise a proteína também precipitou e não apresentou atividade enzimática.

Para confirmação da identidade da proteína C (lipase) foi feita uma análise por espectrometria de massa tipo MALDI-ToF, uma lista de massas moleculares dos peptídeos gerados experimentalmente de uma amostra (Figura 11) foi comparada com massas moleculares de peptídeos deduzidos *in silico* a partir da sequência nucleotídica extraída da biblioteca metagenômica SM. Por comparação dos picos de massas manualmente foram identificados 9 dos 21 picos gerados (Tabela 8), confirmando que se tratava da mesma molécula peptídica.

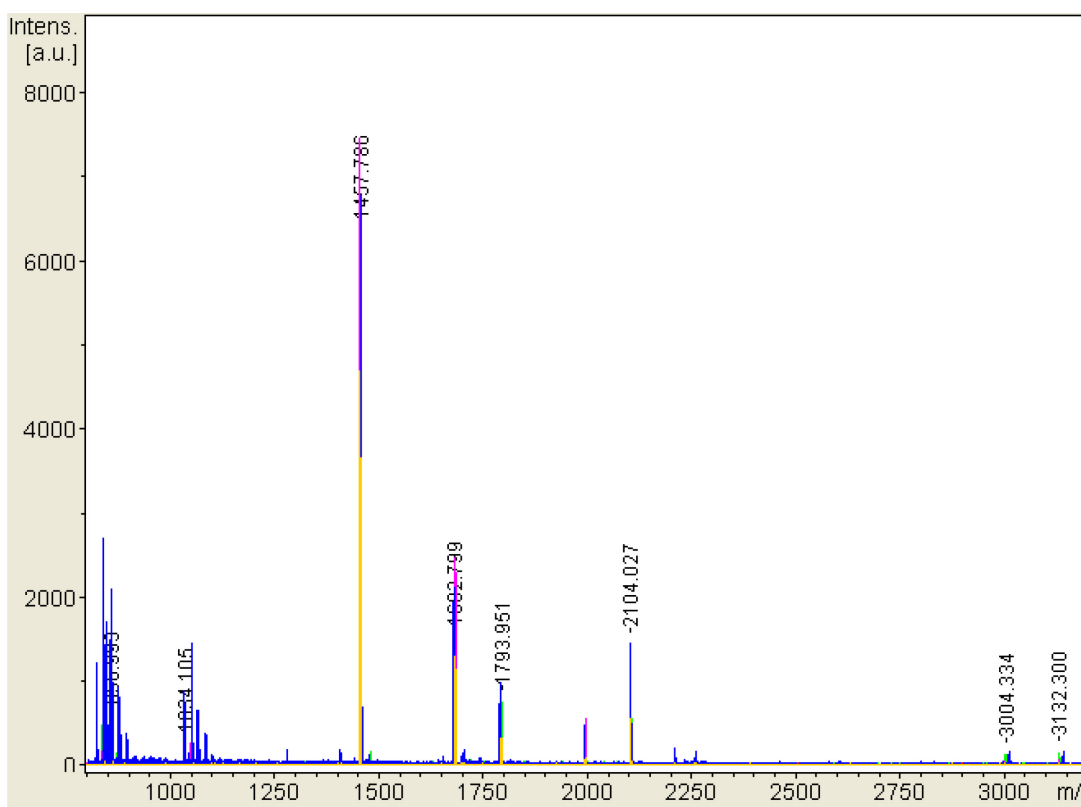


Figura 11. Espectrometria MALDI-ToF dos peptídeos tripticos da proteína C.

Estão indicadas as massas que foram encontradas na análise *in silico*.

Tabela 8. Identificação da proteína C (lipase) por espectrometria de massa.

São apresentados os valores das massas peptídicas tripticas determinadas *in silico* e determinados por Maldi-tof com erro de 0,20 Da.

Range	Sequências peptídicas	Massa <i>in silico</i>	Massa por espectrometria
1-22	MLAASATFTAAQTQAAGYTQTR	2259.095	-
52-71	SGGATVYTPQVSAANSTEV	1993.970	1993.977
52-80	SGGATVYTPQVSAANSTEV	3003.551	3004.334
52-81	SGGATVYTPQVSAANSTEV	3131.646	3132.300
72-80	GEQLLLEVK	1027.591	-
72-81	GEQLLLEVKK	1155.686	-
72-90	GEQLLLEVKKIVAVTGKPK	2049.256	-
81-90	KIVAVTGKPK	1039.675	1034.105
81-104	KIVAVTGKPKVNLIGHSHGGPTIR	2478.455	-
82-90	IVAVTGKPK	911.580	-
82-104	IVAVTGKPKVNLIGHSHGGPTIR	2350.360	-
91-104	VNLIGHSHGGPTIR	1456.790	1456.569
105-125	YVASVRPDLVASATSVAGV	2103.132	2103.552
126-144	GSAVADILLGIAPPGSLSR	1793.005	1792.578
126-156	GSAVADILLGIAPPGSLSR	2962.670	-
145-156	DVITTIATGLGK	1187.676	-
157-185	LLSLLSGSSTLPQNSLAAQSL	2772.487	-
243-256	NDGLVSSCSSLGK	1402.651	1408.548
243-259	NDGLVSSCSSLGKVIR	1770.905	-
290-308	QHANRLQGMGLLEHHHHH	2288.099	-
295-308	LQGMGLLEHHHHH	1681.801	1681.343

5. Análises bioinformáticas das proteínas C e D.

Na ausência de atividade específica nas proteínas C (lipase) e D (prolina aminopeptidase) foram analisadas e identificadas nas sequências de aminoácidos, características da família das α/β hidrolases que incluem as lipases e peptidases. Seus modelos tridimensionais *in silico* confirmaram a presença de domínios de 6 e 7 folhas β para as proteínas C e D, respectivamente (Figura 12a e 12b), sugerindo que estas proteínas possuem pelo menos uma estrutura terciária conservada. Também nessas sequências foi identificada a presença do domínio pentapeptídico conservado G-X-S-X-G contendo o resíduo ligante de metal no sítio catalítico (Suplemento 7). Não obstante, esta análise não é suficiente para analisar seu sítio catalítico nem concluir acerca de sua atividade enzimática, em vista disso, prossegue-se com as análises da proteína C.

A análise filogenética determinou que a proteína C agrupa-se à família das lipases verdadeiras, relacionada com as subfamílias I.1 e I.2. e mais relacionada com *Vibrio cholerae* (Figura 13). Segundo Arpigny e Jaeger (1999) este grupo de lipases apresenta três características fundamentais para sua atividade enzimática: em sua grande maioria precisam de uma proteína ativadora de lipases, nomeada foldase, cuja sequência codificadora encontra-se usualmente no mesmo operon da sequência da lipase; apresenta o sítio de ligação para um íon divalente como cálcio e requer sua presença como cofator catalítico; e contém dois resíduos de cisteína, importantes para a estabilização do sítio ativo da enzima.

Na análise de sequência de aminoácidos da proteína C o resíduo serina 97 foi identificado como sítio de ligação para o íon cálcio (Suplemento 7a e 7b), bem como dos dois resíduos de cisteína nas posições 198 e 249 que poderiam formar ponte dissulfeto (não mostrado).

Deste modo, para avaliar a atividade, seria necessário a coexpressão com uma proteína ativadora foldase como presenciado em vários organismos bacterianos, predominantemente do filo proteobacteria (Suplemento 8).

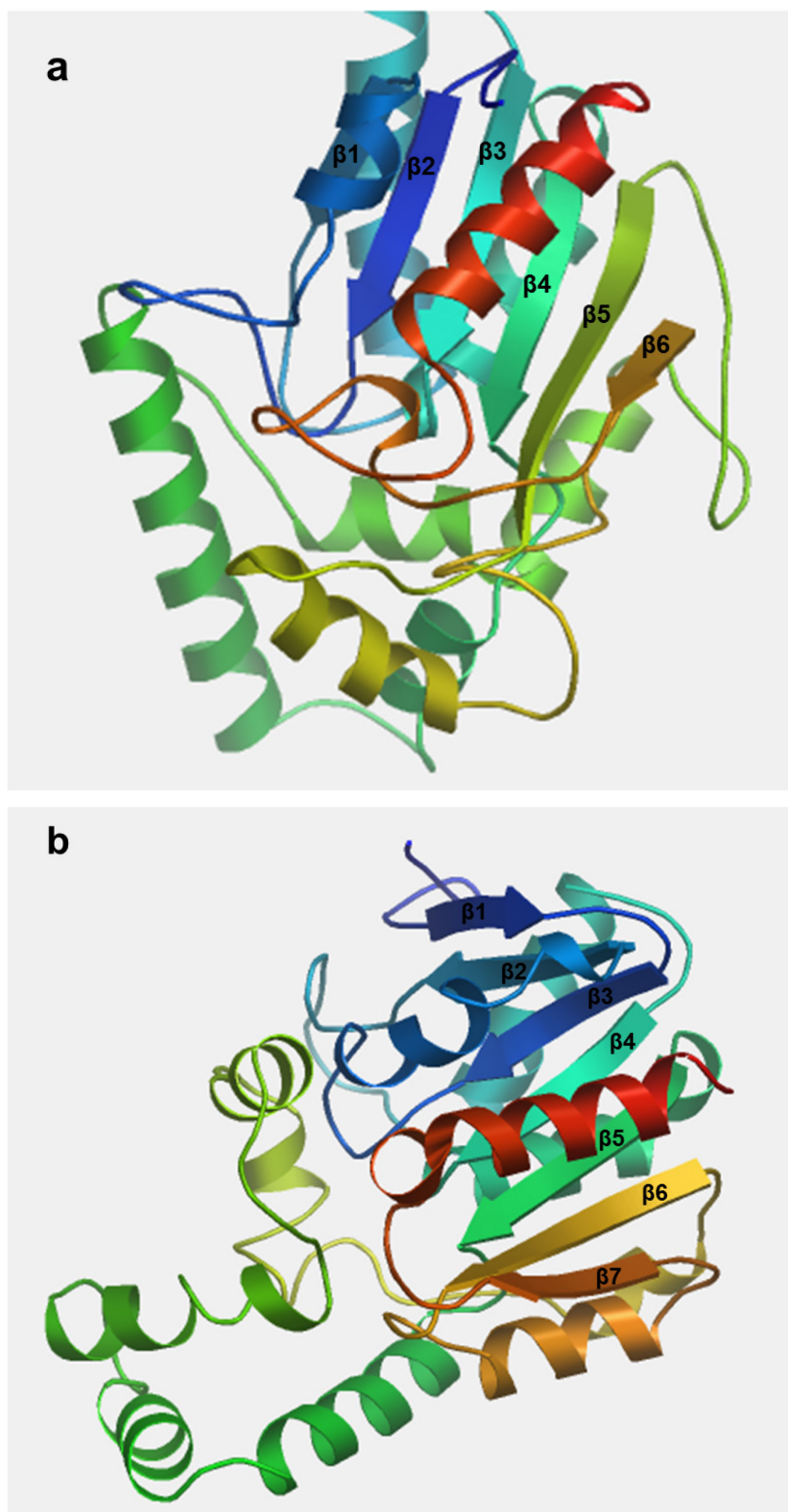


Figura 12. Modelo tridimensional em fitas das proteínas C e D geradas *in silico* por homologia de proteínas.

Estão indicadas as folhas β características da família de proteínas α/β hidrolases. Gerado em *Swiss-Model* (Arnold *et al.*, 2006). **Painel a:** Proteína C, e **painel b:** Proteína D.

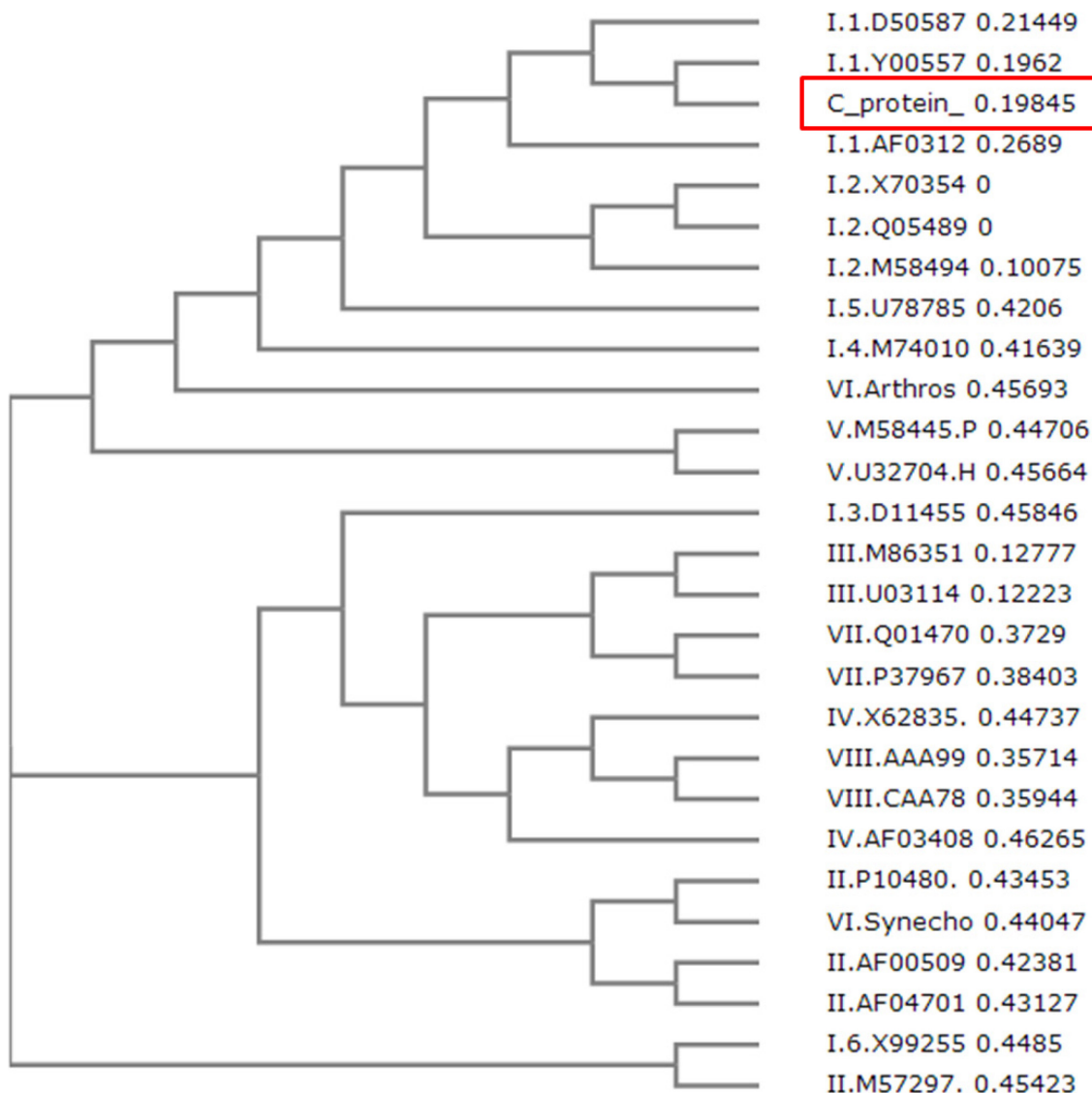


Figura 13. Filograma das famílias de lipase.

Realizado em *ClustalW* pelo método de *Neighbor-joining*. Indica-se para cada grupo a família (I-VIII), número de acesso no NCBI, organismo e as distâncias. Baseado em Arpigny e Jaeger (1999).

6. Proteína ativadora de lipase.

Tem sido demonstrado que as foldases ativam somente sua respectiva lipase e que poderiam ser substituídas com eficiência limitada entre espécies proximamente relacionadas (Rosenau *et al.*, 2004). O mecanismo molecular de funcionamento das chaperonas não é entendido completamente, mas sabe-se que ajudam as lipases a

superar a barreira energética na via produtiva de dobramento da proteína (Rosenau *et al.*, 2004). Alternativamente, tem se visto em condições *in vitro* que a presença de glicerol 40% no tampão de dobramento estabiliza a conformação nativa ao estimular interações hidrofóbicas dentro de proteínas não enoveladas. No entanto, o glicerol não pode ser generalizado como substituinte das foldases (Rosenau *et al.*, 2004).

Foram feitas várias tentativas para encontrar uma atividade enzimática lipase, tendo como requisito a coexpressão da proteína C com uma proteína foldase:

Sua respectiva foldase poderia encontrar-se no mesmo *contig* de origem da sequência gênica de C. No entanto, o *contig* 335 tem um comprimento de 991 pb, e a sequência extraída para codificar essa proteína tem um comprimento de 897 pb, portanto, esse *contig* não podia conter outra sequência codificante, com comprimento estimado de 960 pb.

Em vista da indisponibilidade da sequência da foldase referente a proteína C, foram desenhados *primers* para amplificar a sequência foldase com base no *hit* de mais alta similaridade do *BLAST* para a sequência nucleotídica C. Esta sequência pertence a *Rhodofera ferrireducens* T118 e apresenta o domínio conservado 'lipase chaperona' (foldase). Foram testadas todas as condições de PCR incluindo presença de aditivos e DNA polimerases termoestáveis de alta fidelidade, mas o amplificado não foi obtido.

Foi feita uma coexpressão da proteína C com o plasmídeo pLip XL-1B (Müller-Santos *et al.*, 2006), que contém o operon *liphp* constituído por um gene de lipase e um gene de foldase, provenientes da bactéria *Burkholderia cepacia*, que apresentou uma relação filogenética próxima. Esse vetor foi gentilmente disponibilizado pelo Dr. Marcelo Müller dos Santos e induzido segundo suas instruções. A coexpressão em bactéria não produziu atividade lipase em meio sólido com tributirina em relação com a expressão dos vetores independentes.

A coexpressão do vetor pET-C com o vetor *lifC6G9* (Martini *et al.*, 2011), contendo uma foldase para coexpressão com sua respectiva lipase de origem metagenômica, não mostrou uma atividade lipase diferente à previamente obtida. Esse vetor foi gentilmente cedido pela Dra. Viviane Martini e Robson Alnoch.

Durante a realização destes ensaios, a biblioteca metagenômica *SM* foi analisada a fim de identificar o fosmídeo contendo a sequência codificadora para a proteína C. A identificação do clone foi realizada por amplificação do gene para a proteína C a partir dos *primers* utilizados neste trabalho. O fosmídeo F-10 será

sequenciado e a sequência codificadora para a foldase deverá ser determinada permitindo sua amplificação e posterior coexpressão com a proteína C.

7. Análise da aproximação metagenômica

Neste projeto foram utilizadas as duas estratégias metagenômicas: baseada na função e baseada na sequência. A biblioteca utilizada consistiu em 192 clones positivos para alguma atividade identificada pela estratégia baseada na função e cujo DNA fosmidial depois foi sequenciado ($2,7 \times 10^6$ pb) e analisado para a identificação de 44 genes de interesse dos quais foram selecionadas somente 9 sequências para sua caracterização. Desta forma, segundo Ferrer *et al.* (2009), o descobrimento de uma enzima em uma biblioteca de expressão, seguida da identificação do mesmo gene na biblioteca de insertos e sequenciamento do fragmento identificado, constitui um meio poderoso de maximizar o processo de descobrimento e identificar aqueles novos organismos que estão produzindo essas enzimas.

No entanto, a identificação de um gene de interesse através de expressão metagenômica funcional é tradicionalmente baixa e pode variar amplamente de 1 positivo por cada 2,7 megabases de DNA até 1 por 3979.5 megabases (Culligan *et al.*, 2013). Isto pode ser influenciado por diversos fatores: a fonte de DNA metagenômico, o tamanho do gene de interesse, sua abundância dentro da biblioteca metagenômica, o sistema de vetor e hospedeiro selecionado, a triagem utilizada e a capacidade do hospedeiro para expressar o gene (Uchiyama e Miyazaki, 2009).

Já além do desafio da seleção desses 192 clones iniciais a partir de 3300 clones procedentes de outras bibliotecas metagenômicas, ainda durante o processo de análise de sequências foi vista a limitação de que uma grande parte desse material metagenômico (aproximadamente 13265 dos 18265 *contigs*) foi descartado devido à insuficiência de tamanho (inferior a 500 pb), limitação da metagenômica baseada na sequência também assinalada por outros autores (Van Elsas *et al.*, 2008).

75% dos clones sequenciados da biblioteca metagenômica SM utilizado neste projeto foram identificados a partir de atividade lipase. Esse pré-tratamento constitui uma das estratégias na superação de limitações da metagenômica baseada na função (Van Elsas *et al.*, 2008; Schloss e Handelsman, 2003), e apesar disso, só foi identificado por homologia em bases de dados um gene sob essa denominação ou relacionados.

Venter *et al.* (2004) afirmam que o sequenciamento massivo de DNA de uma amostra ambiental é o método mais prático para examinar seu conteúdo genômico, em comparação com os estudos convencionais baseados em PCR. Não obstante, neste projeto foi encontrado que, em concordância com Simon e Daniel (2011), para tomar uma vantagem completa dessa grande quantidade de informação são necessárias melhores ferramentas de análise integrada e bases de dados compreensivas.

Adicionalmente, na pesquisa por novas enzimas e bioatividades a partir de metagenomas tão ricos como os de solos, são requeridos esforços multidisciplinares que combinem a bioinformática, a química analítica e tecnologia de alto rendimento, microbiologistas, enzimologia, bioengenharia e usuários finais, entre outros (Lee e Lee, 2013; Ferrer *et al.*, 2009). O esforço na superação das limitações da metagenômica deverá ser focado na geração de novos hospedeiros e sistemas de expressão, a correta anotação funcional de genes e um valioso aporte da comunidade científica no processo global de intercâmbio, comparação e avaliação crítica dos resultados metagenômicos (Culligan *et al.*, 2013; Thomas *et al.*, 2012).

Com o vasto volume de proteínas não caracterizadas existentes em bases de dados públicas e os amplos recursos genéticos ambientais não explorados, a busca por genes para novas enzimas de interesse deve ser melhor dirigida, como afirmam Uchiyama e Miyazaki (2009) não simplesmente fazendo um *Blast*, mas uma predição funcional guiada pela estrutura.

Recentemente a metagenômica têm sido complementada com outras ferramentas: as análises expressão gênica e produção de proteínas das comunidades microbianas gerando a metatranscriptômica e a metaproteômica, que são aproximações com o potencial de permitir entender a dinâmica funcional *in situ* e os processos evolutivos dos consórcios microbianos (Simon e Daniel, 2011). Além disso, tem surgido o termo “metagenômica sintética” como uma derivação do estudo metagenômico que envolve a busca de sequências de interesse mediante mineração em bases de dados ou conjuntos de dados metagenômicos, seguido de síntese química dos genes selecionados (Culligan *et al.*, 2013). Esta aproximação teve sucesso em 94% dos clones na obtenção de enzimas metil haleto transferases importantes na indústria química e de combustíveis (Bayer *et al.*, 2009), e sua potencialidade poderia ser estendida para outras áreas.

CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste trabalho foram analisadas sequências da biblioteca metagenômica SM para a identificação de sequências codificadoras para enzimas de interesse biotecnológico. Nove sequências foram escolhidas para amplificação. Das sequências amplificadas, três foram clonadas em vetores pET28a ou pET29a. As sequências codificadoras para uma lipase e para uma prolina aminopeptidase foram superexpressas em *E. coli*, entretanto, ambas apresentaram-se insolúveis, apesar de diversas modificações incluindo estirpe bacteriana, temperatura de incubação, concentração de sal, entre outros. Análises realizadas com a lipase indicaram a necessidade de uma proteína foldase (chaperona) para a obtenção de uma enzima ativa. Ensaio com a coexpressão de chaperonas relacionadas a duas outras lipases não resultaram em atividade. A identificação do fosmídeo da biblioteca metagenômica SM contendo o gene para a lipase poderá levar à identificação da sequência da foldase correspondente e posteriormente sua coexpressão, permitindo a análise da atividade desta lipase.

Para uma avaliação completa das sequências propostas neste projeto teriam de ser feitas várias modificações: na sequência A (glicosil hidrolase), buscar sua amplificação a partir da biblioteca metagenômica SM e não do genoma de *A. brasilense* FP2, já que apesar de sua similaridade os *primers* possivelmente são específicas para sua fonte original. Nas sequências B (triptofanase), E (metaloprotease), e G (deidratase), deveriam ser redesenhados os *primers* de amplificação tendo em conta a inserção do número de bases suficientes para uma digestão eficiente pelas enzimas de restrição. Já para a expressão da proteína F (transcriptase reversa) deveria ser feita uma síntese da sequência com modificação de seus códons raros por códons de uso frequente em *E. coli* ou pode ser utilizado outro organismo hospedeiro de expressão.

Para posteriores trabalhos baseados na aproximação metagenômica para busca de proteínas de interesse das indústrias biotecnológicas, a prospecção pode ser dirigida para enzimas mais específicas dentro do amplo espectro das reações envolvidas no metabolismo ou outras vias. Adicionalmente, deve ser buscado maior suporte das ferramentas bioinformáticas existentes assim como deve ser fortalecida sua expansão em conteúdo curado e sua criação para novas aplicações.

No Brasil, tendo recentes avanços importantes biotecnológicos na metagenômica, o sequenciamento de nova geração, a bioinformática e os novos procedimentos de extração de ácidos nucleicos, junto com as análises convencionais, existe uma grande oportunidade de estudar a enorme biodiversidade do microbioma para ser transformado em riqueza biotecnológica. Essa busca pode ser facilitada através de iniciativas de colaboração entre comunidades científicas nacionais e internacionais como o recentemente constituído “*Brazilian Microbiome Project*” (BMP) (Pylro *et al.*, 2013).

REFERÊNCIAS

- Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22, 2: 195-201.
- Arpigny JL, Jaeger KE (1999) Bacterial lipolytic enzymes: classification and properties. *Biochemical Journal* 343, 1: 177-183.
- Bayer TS, Widmaier DM, Temme K, Mirsky EA, Santi DV, Voigt CA (2009) Synthesis of Methyl Halides from Biomass Using Engineered Microbes. *Journal of the American Chemical Society* 131, 18: 6508-6515.
- Belfort M, Curcio MJ, Lue NF (2011) Telomerase and retrotransposons: reverse transcriptases that shaped genomes. *Proc Natl Acad Sci U S A* 108, 51: 20304-20310.
- Böttcher D, Schmidt M, Bornscheuer U. cap. 11. Screens for Active and Stereoselective Hydrolytic Enzymes. In: Streit, W. R. e Daniel, R. (Ed.). (2010) *Metagenomics*, v.668, p.169-176. *Methods in Molecular Biology*. Humana Press: ISBN 978-1-60761-822-5.
- Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72, 1-2: 248-254.
- Chavira Jr R, Burnett TJ, Hageman JH (1984) Assaying proteinases with azocoll. *Anal Biochem* 136, 2: 446-450.
- Committee on Metagenomics: Challenges and Functional Applications NRC. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington: The National Academies Press, ISBN 9780309106764. Disponível em: < http://www.nap.edu/openbook.php?record_id=11902 >.
- Couto GH, Glogauer A, Faoro H, Chubatsu LS, Souza EM, Pedrosa FO (2010) Isolation of a novel lipase from a metagenomic library derived from mangrove sediment from the south Brazilian coast. *Genet Mol Res* 9, 1: 514-523.
- Culligan EP, Sleator RD, Marchesi JR, Hill C (2013) Metagenomics and novel gene discovery: Promise and potential for novel therapeutics. *Virulence* 5, 3: 0-1.
- Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* 77, 4: 1315-1324.
- Di Tommaso P, Moretti S, Xenarios I, Orobitz M, Montanyola A, Chang J-M, Taly J-F, Notredame C (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res* 39, suppl 2: W13-W17.
- Ekkers DM, Cretoiu MS, Kielak AM, Elsas JD (2012) The great screen anomaly--a new frontier in product discovery through functional metagenomics. *Appl Microbiol Biotechnol* 93, 3: 1005-1020.
- Faoro H (2010) Thesis: Prospecção metagenômica de biocatalisadores da microbiota de solos da Floresta Atlântica Doutorado em Ciências-Bioquímica. Pós-graduação em Ciências - Bioquímica, Federal University of Parana, Curitiba. 214 p.
- Faoro H, Alves AC, Souza EM, Rigo LU, Cruz LM, Al-Janabi SM, Monteiro RA, Baura VA, Pedrosa FO (2010) Influence of soil characteristics on the diversity of bacteria in the Southern Brazilian Atlantic Forest. *Appl Environ Microbiol* 76, 14: 4744-4749.
- Faoro H, Glogauer A, Souza EM, Rigo LU, Cruz LM, Monteiro RA, Pedrosa FO (2011) Identification of a new lipase family in the Brazilian Atlantic Forest soil metagenome. *Environmental Microbiology Reports* 3, 6: 750-755.

- Ferrer M, Beloqui A, Timmis KN, Golyshin PN (2009) Metagenomics for Mining New Genetic Resources of Microbial Communities. *J Mol Microbiol Biotechnol* 16, 1-2: 109-123.
- Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41, D1: D808-D815.
- Fuhrmann M, Hausherr A, Ferbitz L, Schödl T, Heitzer M, Hegemann P (2004) Monitoring dynamic expression of nuclear genes in *Chlamydomonas reinhardtii* by using a synthetic luciferase reporter gene. *Plant Mol Biol* 55, 6: 869-881.
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, J. M. (Ed.). (2005) *The Proteomics Protocols Handbook*, p. 571-607. Humana Press: ISBN 978-1-58829-343-5.
- Glogauer A, Martini VP, Faoro H, Couto GH, Muller-Santos M, Monteiro RA, Mitchell DA, De Souza EM, Pedrosa FO, Krieger N (2011) Identification and characterization of a new true lipase isolated through metagenomic approach. *Microb Cell Fact* 10: 54.
- Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R (2010) A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res* 38, suppl 2: W695-W699.
- Gupta R, Gupta N, Rathi P (2004) Bacterial lipases: an overview of production, purification and biochemical properties. *Appl Microbiol Biotechnol* 64, 6: 763-781.
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: 95-98.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5, 10: R245-249.
- Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21, 9: 1552-1560.
- Ishikawa J, Hotta K (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol Lett* 174, 2: 251-253.
- Kakirde KS, Parsley LC, Liles MR (2010) Size Does Matter: Application-driven Approaches for Soil Metagenomics. *Soil Biol Biochem* 42, 11: 1911-1923.
- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28, 1: 27-30.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32, suppl 1: D277-D280.
- Kawasaki K, Yokota A, Oita S, Kobayashi C, Yoshikawa S, Kawamoto S, Takao S, Tomita F (1993) Cloning and characterization of a tryptophanase gene from *Enterobacter aerogenes* SM-18. *J Gen Microbiol* 139, 12: 3275-3281.
- Kennedy J, Marchesi JR, Dobson AD (2008) Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb Cell Fact* 7: 27.
- Kisand V, Valente A, Lahm A, Tanet G, Lettieri T (2012) Phylogenetic and functional metagenomic profiling for assessing microbial biodiversity in environmental monitoring. *PLoS One* 7, 8: e43630.
- Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 21: 2947-2948.

- Lee MH, Lee SW (2013) Bioprospecting Potential of the Soil Metagenome: Novel Enzymes and Bioactivities. *Genomics Inform* 11, 3: 114-120.
- Lee S-W, Won K, Lim H, Kim J-C, Choi G, Cho K (2004) Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl Microbiol Biotechnol* 65, 6: 720-726.
- Lorenz P, Eck J (2005) Metagenomics and industrial applications. *Nat Rev Microbiol* 3, 6: 510-516.
- Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* 41, D1: D348-D352.
- Martini VP, Glogauer A, Iulek J, Souza EMD, Müller-Santos M, Pedrosa FDO, Krieger N (2011) Isolamento e Caracterização de uma Nova Lipase (LipC6G9) com aplicações em biocatálise. 34a Reunião Anual da Sociedade Brasileira de Química. Florianópolis, SC: Sociedade Brasileira de Química (SBQ)
- Müller-Santos M, De Souza EM, Pedrosa FDO, Baratti JC, Mitchell DA, Krieger N (2006) Determination of lipase activity using image analysis. *Anal Biochem* 351, 2: 305-307.
- Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302, 1: 205-217.
- Pace NR, Stahl DA, Lane DJ, Olsen G (1985) Analyzing natural microbial populations by rRNA sequences. *ASM News* 51: 4-12.
- Pedrosa FO, Yates MG (1984) Regulation of nitrogen fixation (*nif*) genes of *Azospirillum brasilense* by *nifA* and *ntr* (*gln*) type gene products. *FEMS Microbiol Lett* 23, 1: 95-101.
- Plugge CM, Henstra AM, Worm P, Swarts DC, Paultisch-Fuchs AH, Scholten JC, Lykidis A, Lapidus AL, Goltsman E, Kim E *et al.* (2012) Complete genome sequence of *Syntrophobacter fumaroxidans* strain (MPOBT). *Standards in Genomic Sciences* 7, 1.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* 40, D1: D290-D301.
- Pylro V, Roesch L, Ortega J, Amaral A, Tótola M, Hirsch P, Rosado A, Góes-Neto A, Da Costa Da Silva A, Rosa C *et al.* (2013) Brazilian Microbiome Project: Revealing the Unexplored Microbial Diversity—Challenges and Prospects. *Microb Ecol*: 1-5.
- Rosenau F, Tommassen J, Jaeger K-E (2004) Lipase-specific foldases. *ChemBioChem* 5: 152-161.
- Sambrook J, Fritsch E, Maniatis T. (1989). *Molecular Cloning: A Laboratory Manual*. 2nd edition. New York: Cold Spring Harbor Laboratory, ISBN 0879695773.
- Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology* 14, 3: 303-310.
- Schmieder R, Edwards R (2012) Insights into antibiotic resistance through metagenomic approaches. *Future Microbiol* 7, 1: 73-89.
- Shevchenko A, Wilm M, Vorm O, Mann M (1996) Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. *Analytical Chemistry* 68, 5: 850-858.
- Sigrist CJA, De Castro E, Cerutti L, Cuče BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. *Nucleic Acids Res* 41, D1: D344-D347.
- Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. *Appl Environ Microbiol* 77, 4: 1153-1161.
- Song H-N, Jung T-Y, Park J-T, Park B-C, Myung PK, Boos W, Woo E-J, Park K-H (2010) Structural rationale for the short branched substrate specificity of the glycogen debranching enzyme GlgX. *Proteins: Structure, Function, and Bioinformatics* 78, 8: 1847-1855.

- Streit W, Daniel R, Eds. (2010) *Metagenomics: Methods and Protocols*. Methods in Molecular Biology: Humana Press, v.668, 341 p.
- Su C, Lei L, Duan Y, Zhang KQ, Yang J (2012) Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Appl Microbiol Biotechnol* 93, 3: 993-1003.
- Sun H, Zhao P, Ge X, Xia Y, Hao Z, Liu J, Peng M (2010) Recent advances in microbial raw starch degrading enzymes. *Appl Biochem Biotechnol* 160, 4: 988-1003.
- Teeling H, Glockner FO (2012) Current opportunities and challenges in microbial metagenome analysis-a bioinformatic perspective. *Briefings in Bioinformatics* 13, 6: 728-742.
- Theriot CM, Tove SR, Grunden AM. Chapter 3 Biotechnological Applications of Recombinant Microbial Prolidases. In: Allen I. Laskin, S. S. e Geoffrey, M. G. (Ed.). (2009) *Advances in Applied Microbiology*, v.Volume 68, p.99-132. Academic Press: ISBN 0065-2164.
- Thomas T, Gilbert J, Meyer F (2012) *Metagenomics - a guide from sampling to data analysis*. *Microb Inform Exp* 2, 1: 3.
- Troeschel S, Drepper T, Leggewie C, Streit W, Jaeger K-E. cap. 8. Novel Tools for the Functional Expression of Metagenomic DNA. In: Streit, W. R. e Daniel, R. (Ed.). (2010) *Metagenomics*, v.668, p.117-139. Methods in Molecular Biology. Humana Press: ISBN 978-1-60761-822-5.
- Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current Opinion in Biotechnology* 20, 6: 616-622.
- Uchiyama T, Miyazaki K. cap. 10. Substrate-Induced Gene Expression Screening: A Method for High-Throughput Screening of Metagenome Libraries. In: Streit, W. R. e Daniel, R. (Ed.). (2010) *Metagenomics*, v.668, p.153-168. Methods in Molecular Biology. Humana Press: ISBN 978-1-60761-822-5.
- Unesco. List of biosphere reserves which are wholly or partially World Heritage sites. 2009. Disponível em: < <http://www.unesco.org/new/en/natural-sciences/environment/ecological-sciences/biosphere-reserves/world-network-wnbr/wnbr/> >. Acesso em: 19/02/2014.
- Van Elsas JD, Costa R, Jansson J, Sjolting S, Bailey M, Nalin R, Vogel TM, Van Overbeek L (2008) The metagenomics of disease-suppressive soils - experiences from the METACONTROL project. *Trends Biotechnol* 26, 11: 591-601.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 5667: 66-74.
- Vieites J, Guazzaroni M-E, Beloqui A, Golyshin P, Ferrer M. cap. 1. Molecular Methods to Study Complex Microbial Communities. In: Streit, W. R. e Daniel, R. (Ed.). (2010) *Metagenomics*, v.668, p.1-37. Methods in Molecular Biology. Humana Press: ISBN 978-1-60761-822-5.
- Wang H, Guo J, Pang H, Zhang X (2010) Identification and characterization of a new monoamine oxidase type C-like dehydratase from *Phytophthora capsici* involved in polyhydroxyalkanoates biosynthesis. *Biotechnol Lett* 32, 11: 1719-1723.

SUPLEMENTOS

Suplemento 1. Sequências de nucleotídeos selecionadas e suas traduções.

Suplemento 2. Análise bioinformática de domínios proteicos das 9 sequências selecionadas.

Suplemento 3. Alinhamentos com a sequência de amino ácido de maior similaridade encontrada no GenBank utilizando o programa *BlastX*.

Suplemento 4. Mapas dos vetores pET-28a e pET-29a (Novagen).

Suplemento 5. Utilização de códons raros nas sequências C, D e F em *E.coli*.

Suplemento 6. Análise de hidrofobicidade das sequências peptídicas de C (lipase) e D (prolina aminopeptidase).

Suplemento 7. Análise de domínios conservados das proteínas C (lipase) e D (prolina aminopeptidase).

Suplemento 8. Análise de identificação de proteínas chaperonas de lipase.

Suplemento 1. Sequências de nucleotídeos selecionadas e suas traduções.

As sequências encontram-se no formato FASTA. **a.** Sequências de nucleotídeos, **b.** sequências de aminoácidos.

a. Sequências de nucleotídeos

```
>contig01139 glycosyl hydrolase FP2 pET-29b NdeI/EcoRI PolyHisC-ter (A)
gctcatATGGGATGCGAAGCCATGCCCGGTGCCGTAAGGGACGGATCGGCAATCCAGATGG
ACGCGGATACGGGGCGGACCGGCACACCCAGTCCCCCGCTCGGGGAGCTGCGGGCGCC
GCTGGGAGCGATCCCCACCGGCCAGGGGACGAACTTCAGCGTGTTTTCGAAGCACGCGA
CCGGAATCGAGCTGCTGCTTTTCGACCGCGCCGAGAATGCGGGGCCTGCGCGCGTGATC
CACCTCGATCCCTCCACCCATCGCACCTATCATTACTGGCACGTGTTTCTTCCCGGCGTGA
CGGCCGGCCAAATCTATGGGTATCGCGCCGAGGGGCCATGGGACCCCGCCAACGGCCTG
CGTTTCGATCGGGACAAGCTTCTCCTTGATCCCTATGGCCGCGCGGTGGTCGTTCCGGAC
CGCTACAGCCGCGACGACATCCGCAAACCCGGTGACGATTGCGGCGGCGCCATGAAAAG
CGTGGTTGTGGATCCGGGGTCTACGATTGGGAAGGCGACGCACCGTTGCGGCGGTCGT
CCGCGCAGACCATCGTGTATGAGATGCATGTGCGCGGATTCACACGCCATCCCAGTCCG
GCGTCGGCGGGAAGACCCGCGGCACGTTTCCCGGGTGATCGAGAAGATTCCGTACCTG
CAGAAATTGGGCGTCACCGCGGTGGAAGTCTGCGCGGTGTTCCAATTCGATGCGCAGGAT
TGCCCGCCGGGAAAGGTCAACTACTGGGGATACGCGCCGGTCTCGTTCTTTGCGCCGCAT
GCGGCGTACAGTTCGCGTTCCGATCCGCTCGGTCCGCTGGACGAGTTCGCGGACATGGT
CAAGGCGCTGCACCGCGGCGGCATCGAGGTATCCTCGACGTGTTTTCAACCACACGG
CGGAAGGCGACCACAACGGGCGGACCCCTGTGTTCCGAGGGTTGGACAATCCCACCTACT
ATCTTCTCGAGGACGACCGATCCCGCTACGCCAACTACACGGGCACCGGGAACACCTTGA
ACGCCAATCACCTGTGTCGTCGCGCATGATCGTGGACAGCCTTCGATACTGGGTGCGAGA
CGATGCATGTGGACGGCTTCCGCTTCGATCTGGCCTCGGTTCTGTCCCGCGACACATCGG
GGCATTCCGATCCCAGAACGCGCCGATCCTGTGGACATCGAAACGGAACCGGCGCTTGGC
GGTACGAAGCTGATCGCCGAGGCTTGGGATGCGGCGGGCCTTTATCAGGTGCGGAGTTT
CGTCGGCGACAGCTGGAAGGAATGGAACGGACGGTTCGTCGACGACGTGCGTGCGTTTT
TCCGCGGTGAACCGGATCGGTGACCCAGATCGCGGACCGCATCCTGGGCAGCCCGGAG
ATCTACGGGCATGAGGAGCGGGAGGCTGAACAAAGCGTCAACTTCGTGACCTGCCATGAC
GGTTTTACGCTCAACGACGTGGTGTCTACGATCGCAAGCACAACGAGGCGAACGGCGAG
GACAACCGCGATGGCGCGGACGACAACCGCAGTTGGAAGTGCGGCGTGAAGGCCCGAG
CGACGATCCGGCAATCGAGCGGCTGAGAAGCCGCCAAGTGAAGAACCTCCTGACCGTGA
CGATGCTATCGCTCGGCATCCCGATGATCACGATGGGCGACGAGGCGGCGCACCCAG
TCCGGCAACAACAACGCCTACTGCCAGGACAACGAGACCAGTTGGTTTGATTGGACGCTG
GTCGAGACGCACGCGGACGTGCACCGGTTTCGTCACTGTGCTCAACACACGCCGAAGTCT
GCGGGACAGGATGTACGAGCGGGTGCCTTTGCGGCAGCTGCTCCGGAAGGCGAAGATAA
CCTGGCACGGCGTGACCCCGAGCAACCGGATTGGAGCCGCGATTCCCACAGCATCGCC
GTCCAAGCGAAGGTGGAGCAGGGTCCGGTTCGCATTTACCTGATCCTGAACGCCTACTGG
AAACCGCTTTCGTTGACCTTCCGCCAGCAAATGACGGTCTGTGTTGGCCCATGGCGCCGA
TGGATCGACACCTCGCTCGACCCCCATGTGATATCGTCGAGTGGAACTTGGCGCCGACG
CTGAGTGAGTTGTCCTATCTGGTGGAAAGCGCGGTCCGTCGTCGTGCTGATCCATGACGCC
GGCGAACCGACCAACTGTGCaattccTG
```

```
>Contig 623 – triptofanase INV-COM pET-29b NdeI/EcoRI PolyHisC-ter (B)
ggcatATGAGATTTCCATCTGAGCCGTTCAAATCAAAGTGGTGGAAACCCATTTCGTGCGACAA
CGCGCGAGGAGCGGGACCGTCTGCTGCGCGAGGCGGCTACAACCTGTTTCATGTGCCG
GCAGAGAGTGTGTACGTCGATCTGTTGACCGACAGCGGCACATCCGCCATGAGCGACAAC
CAATGGGCGGGCCTGATGCTCGGTGACGAGTCCCTATGCCGGCAGCAAGAATATTACCAG
TTTCAAGAGGTGCTACGGTCCATCTTCGGTTACAAGCACGTCATTCCACACACCGAGGGG
CGCATGGCGGAAAATCTCCTCTTACCACCATCGTGAAGCCAGGCATGTGCGTCCCGAAC
AATATTCATTTTGATACTACGCGTGCAAACGTTGAGCATCAAGGAGCACAGGCCCTGGATA
TGGTGGTCAAAGAGGCTTACGACCCGCACTGCGAGTTGCCATTCAAGGGAAATATGGACT
TGGTTCGTTTGGAGGGGACGATCAATCGAGTCCGGCGGGACCACATTCCGTTGGTCATGT
TGACGATACCAACAACAGCGGAGGCGGGCAACCGGTCTCGATGGACAATATCCGTGCGA
CGCGCGTACTGCTCAACCGTTACGACATTCTCTTTTTTCGACGCCTGCCGGTTTGGCGGA
AAATTGTTTTTTCATTAAGGAGCGCGAGCCGGGCTATGACGGGGTATCGATTCTCGATATC
```


GCGCAGGAACCTCTCAGCTATGGCGATGGCTGCACGATGTCCGCCAAAAAGGATGGCCTC
 GTGAATATCGGAGGATTTCTCAGCCTGAAGAATGATCAGTGGGCACAAGACATTACCAACA
 TGCTCATCTGGTGAAGGCTTTTCGACCTACGGAGGATTAGCAGGCCGTGATCTCGAAG
 CGATGGCCAGAGGGTTGCGCGAAGTGTGGACGAAGACTACCTGAGCTTCCGGATAGGG
 CAGGTCCGCTATCTAGGAGAGTTGCTGGATCAAGCCGGGGTGCCATTCTCAAACCGATC
 GGGGGGCACGCCGTCTATCTCAACGCGAAAGAATTTCTGCCTCACATCCAGCAGGCCCAA
 TTTCCCGCCAAAGCGTTGGTGGCCGCCCTTTATCGGGAGTATGGTATCAGGGGGGTGGAG
 ATCGGCACCGTAATGTTTGAAAGACAGATTCCGCGACCGGTGCGAACGATCTATCCAGAG
 CTCGAGATGGTTCGGCTGGCCATCCCCGTGCGGTGTATACGAACATGCAGATCACCTAT
 GTCGCCGAATCAATCATCGAGCTGTATCAGCGGCGCGAGATGATTCATGGACTCGCCTTG
 ACCTATGAGGCCTCCGTCTGCGGCACTTACGCGCCGATTTACGGAACCTCATGACCAG
 TCCCTCTTCCAGCGCTCTGCCTCAAtcgaattctg

>contig_335 Triacylglycerol lipase INV-COM pET-29b NdeI/XhoI PolyHisC-ter (C)
 gtacatagCTGGCCGCGCAGCGCCACTTTTACTGCCGCCAGACTCAGGCTGCAGGTTACACC
 CAGACCCGTTATCCGATTGACTGGTACACGGTTTATTCGGCTTCGATAACATCGGGCCGG
 TGGAATATTTCTACGGCATCCCGTCTGCGCTGCGTTCGGCGGCGCAACGGTGTACACGC
 CGCAAGTGTGCGCAGCCAACAGCACCGAAGTGC GCGGTGAACAGCTCTTACTGGAAGTAA
 AAAAAATCGTCGCGGTACCCGGTAAACCGAAAGTGAATCTGATCGGTACAGCCACGGCG
 GGCCTACCATCCGTTATGTGGCTTCGGTACGTCCGGATCTGGTTCGCTCGGCAACTTCAG
 TGGCAGGCGTTAAACAAAGTTCTGCGGTGGCGGATATTCTGCTGGGCATCGCGCCGCCG
 GGCAGTTTGTACGCGATGTGATTACGACGATTGCTACCGGCCTCGGTAAATTGCTGTCTG
 CTGCTGTCTGGCAGTTCCACATTGCCACAGAATTCACTGGCGGCCGCGCAATCCTTGTCAA
 CGGCGGGTTTCGGCAAATTCATGCTGCGCATCCGGCCGGTTTACCGAGCACAGCTTGCG
 GCGAGGGCGCGTATCAGGTCAATGGTGTTCCTATTTCTCATGGAGTGGCGCATCGAATT
 ACACCAATGTGCTGGATATTCTGGATCCGGCGCTGGCAGTAACCGGTCTGGCTTTCGGCG
 GCGGAAAAATGATGGTCTGGTGTCTTCTGTTCCAGCCATCTGGGTAAGGTCAATTCGCGA
 TGATTATGCGATGAACCATGCGGATGAAATCAACCAGAGCGTGGGTATTGTGAATCTGTTT
 GAAGTGAATCCGGTGTGCGGTGTTCCGCCAGCACGCCAACC GCCTGCAGGGCATGGGTCT
 Gctcgagtc

>Contig 35 proline-specific peptidase pET-28a NheI/XhoI PolyHisN-ter (D)
 gctagcATGAACACGGTCAAAACGGGCGGCGTTTCAGATGGTGTCCATAGACGGTTCGGTTTCA
 AGTCTGGACGAAACGAATAGGCGCCGGTCCGCGCACGATGCTGACTTGCATGGCGGTC
 CCGCTCCACTCACGAATACTTCAATGCTTTCGAGGATTTTCTGCCGCCAATGGGATTCA
 GCTCATTTACTATGATCAGCTCGGATCGGGCACTCCGATCAGCCCGATAATCCGGCCCTT
 TGGGTGGTTCGAGCGCTTTCGTGACGAGGTAGAACAGGTCAGAGCGGCGCTCGGCCTCAC
 CGGATTCTACTTATATGGCCACTCCTGGGGCGGTATGCTCGCAATCGAATACGCGTTGAAA
 TACCAAAGCCATCTCAAGGGTCTGATCATTTCGAACATGACTGCCAGCATAGCGTCGTATG
 TACTTACGTTAATGAGCTGCGGCGCCAACACTACCCGCCGAGTCGCAGCGCATTCTGGAAC
 GATACGAAGCGACCGGGGAATACACAGCACCGAATATGAGAAAGTCATGTTCCGAGAAA
 TTTATTCCAGGCACTTATGCCGCCTTGTCCATGGCCGGAACCTTTGATCCGGATGATCCG
 GCATATGAACCAGAAGGTCTACAACAAGATGCAGGGTCCCAATGAATTCGTGGTCACCGG
 AACCTTCAAAGACTGGGATCGATGGGACGATATCAAAAATATAAACGTCCCCACTCTGTTAT
 CAGTGGGCCGTTTCGATACGATGAGCGTGGCCGACGTCGAAAGGATGGGCACCCTGATTC
 CCAATGCGCGAGTATCGATTTGCGAGACGGGCAGCCATTGTTCAATGTACGACGATCAGG
 AGCGCTATTTTGAAGATCTGGTTCGCTTCATCAAAGATGTGAGGCGGGAAAGCTCGTTTG
 Actcgagtag

>contig_221 metalloprotease pET28b NcoI/XhoI PolyHisC-ter (E)
 gtcctagGAAGAGCGGCGCTGACTGCCGAAGAACAGCGAAGCAAAGAATTTGCGGCTACGA
 TCCTGCGCTTACGGAAAAAGTTTGGGACGAACAGTTCCAAAAATCGGCAAGAAATATTC
 CAAGCCGCACATGGTCTCTTCTCCGAGCAAGTCGATACCGGCTGCGGCTCGGCGCCGT
 CGGCCGTGGGTCCCTTCTATTGCCCGCGGACAAAACCTGTCTATCTCGACCCGACGTTTTT
 CGATGAGCTGCAAAACAAATTGGGCGGTTCCAAGGGCGAATTCTCGCAGGCCTACGTCAT
 TGCTCACGAAGTTGGTTCATCACGTGCAAAACCTGCTGGGATATAGCAGAATCGCGGATGA
 AAATCAGCAGAGCGCACCTAGTAAAGCGAAAGCGAATGAGTGGTCCGTCGCGCTGGAGCT
 GCAAGCCGACTACCTCGCCGGCTGCTGGGCGCACCCAGGCAAAAAGGAATTTTAT
 CGAGCCGGGCGACATCGAGTCCGCGCTCAAGACGGCTAATGCGATCGGGGACACCCTG
 TGCAAAAGCGCGCTACCGGATTCACCTCGCCGGAGAAATACACTCATGGCACATCGGCTC

AGCGTCTCAAATGGTTCCGCGCCGGCTTCGAGACCGGCGACTTGAAAAAGATGAAGGAGC
TTTTTGACTTGCCGTATAAGAACTGctcgagTGA

>contig_2122 Reverse transcriptase INV-COM pET-28a NheI/XhoI PolyHisN-ter (F)
gtagctagcATGACGGCAAGAGACGCGAGAAAGTATGGCTGAGAAGCCCCGAATCCCCTCGGG
AGGTAGCGGTTCGAAATCGCGAGATAACGAGGCGGGTTCGTCAGGGTCAACGGGCAAGGG
GAGAAGACTCTAGCCCGGAGTGCAGCAGTTGATGGAAGCAGTGGTCGAGCGAGGGAAAC
ATGCAGACCGCGCTCCAGCGAGTATGAGCAACAGGGGAGCAGCCGGAGCCGACGGGAT
GACGGTTGATGAACTGAAGCCGCACTTGAGGGAGGAGTGAAGCAGGATCAAAGGAGAAC
TGCTGGCAGGGGAATACCAACCGGAGCCAGTGTCTGAAGGTAGAGATACCGAAGGCAGAG
GGAAAGGGGTGCGAAAGCTTGGCATCCCGACGGTGGTGGACCGGCTGATCCAGCAGGC
GTTGCATCAAGTACTAAGCCCGATCTTTGAGCCAGGATTTTCGGAATCGAGCTATGGCTTT
CGGCCAGGCCGGGACGACAGGATGCGGTGCGGCAAGCACGGGCATATGTGGGTGAAG
GGCGGCGGTGGTTCGTAGATATCGACTTGAGAAATTCCTTTGACCGAGTTAATCACGACA
AAATGATGTCGCGGCTAGCGAGGCGGATCAAGGATAAGCGGATACTGCGGATGATCCGAA
GGTACCTACAGGCTGGAATGATGGAAGGCGGGCTGGTGACACAGAGGAGAGAGGGGACG
CCGACAGGGCGGGCCGCTATCGCCGCTGTTGTCGAACATTCTGCTGGATGAGCTGGACAA
GGAAGTGGAGAGACGAGGGCACAAGTTCTGCCGGTACGCCGACGATTGCAATGTGTATGT
GCGGAGTGAAGTGCAGGGGAGCGAGTGAAGGAGTGCATCACAAGGTTTCTTGAAGGC
GACTGCGGCTGAAGGTAAACGAGGAGAAGAGCGCAGTAGAGCGGCCGTGGAAGAGGAAG
TTTCTGGGCTACACAATGACGTGGCACCTGGAACCGCAATAAAGGTAGCGGAGAAGTCCG
GTGAAACGACTAAAGGTGAAGCTGCGGGAGATTCTGCGGCAGGGCCGAGGACGCAACAT
TGGGAGACTAATCGAGGAAGAACTAACACCGCTGCTGAGAGGCTGGATGAACTATTTCCG
GCTGGCGGAGGTGAAAGGAATCTTCGAGGAGTTAGATAGCTGGATACGGCGGAAGTTGAG
GTGTGTAATCTGGCGGCAATGGAAGCGCACCTGGGCGCGGGTAAAGGGACTGATGAAGC
GTGGTTTGGAGAGGGATCGGGCGCTGAAATCAGCGACTAATGGGCGAGGGCCATGGTGG
AACGCTGGGGCCTCGCACATGCACGAGGCCTTCCCAAGACATACTTTGATCGCTGCGGT
TTGGTGTGCTGCTAGATCAACGGCTCAAAGTCCAGCGTACTTCATGActcgagac

> contig_2026 deidratase pET-29b NdeI/XhoI PolyHisC-ter (G)
gtcatatgATTTTAATTAGTAGTCGGGTTGGGAGTACAGAAATGAGCGGACTTTATTTTCGAGGA
GTTCAAAGTCGGGCAGCTGTTCCATCATGCGATCACTCGCACGGTCACCGAGACCGACAA
CCTCCTTTTACCACGCTGACTCACAATCCCGAGCCCTGCACCTCGATGTCGAGTTTCGTG
AAACAAACCGAGTTCGGTCAGCGCCTTGTCAATAGCATTTCACCCTGGGCTTATGATCG
GGTTTTCGGTTCGTTGACACCACCTTGGGCGAGACCGTCCCAATCTGGGGATGAACGAT
GTTTCGATTCGCCAACCGGTCTTCATCGGCGATACGTTGCGAGCACAAAGCACGGTGCTG
GAAATGCGTGAAAGCAAGTACGACCTGATGCAGGCATCGTAGTGTTCGAGCACCGATGC
CTGAACCAGCGCGACGAGGAAGTTCGGTACTGCAAGCGGTCCGCTTGTGCGCAGGAA
GGCAAGActcgagagTAG

>amilase S pET29b NdeI/BamHI PolyHisC-ter (S)
actcatatgGCACAGCCGGCCGCTGGACCGAGGGTCTCCAAGGTTGAACCGCCGAATTGGTG
GATGGACTTTGCACCGACGGTCATGTTTCTGCTCTACGGTGAGAACCTTGGGGGAGCCAA
TGTTTCAGTTGATTATCCGAACGCGGTGGTTGCCAAAGTTCTACCTCAACCCGATGGGAAG
CACCTCTTTGTCTGGTTGAGTTTTTATCCAGGAACCCGCCGGGAGATGTCGTGATTCAG
TGAAGACCGCTCCGGCGAGACAAGCGTGCCTGTGCCATTGTTGACGCGCTCGCCCCAG
GAAGGACGGTTTCAAGGAATCACGCGCGACGATGTCATCTACCTAATCATGCCCGACCGC
TTTGCGGACGGCGATCCCGCAAACAACATGCCGCCAGGTGCTGCTCCAGGTACTTACGAC
CGGAGCGGCGCTAAGACCTATCACGGTGGCGACCTGAAGGGAATCCAAGAGCACCTGCC
GTACTTGAAGGACCTTGGAGTGACGGCGCTTTGGCTCACGCCGTTGTACGACAATGACAA
CTCAACTTCGGATTACCACGGCTACGGAGCAGTGGACGAATACGCCGTTGAAGATCACTT
CGGCACGATGAAGTCTTATCAAGACTTGGTAGCCGCTGCTCATCAGGTCCGGCCTGAAGGT
GATGCTGGACATGGTTCCCAACCATGTTGGCCCCAAGCATCCCTGGGCGACGTCACAGCC
CGCTCCGATTGGCTGCACGGAACGACCGAGCATCATCTCGACACAGACTATTACTATGC
GCCTGTCACTGATCCCCACGCAGTCAAGGCAAATTATGTGAGCGCGCTCGAAGGCTGGTT
CGCTGACGTCTTACCAGACCTGGCGCAGGAGAATCCGCTGGTGGCGCTATACCTGCTCCA
GAATGCCAGTGGTGGACAGAGAGTGGCGGGGTCGACGGGATTCCGTTGACACATTTCC
CTACGTGCCCGGAGCTTTTGGCAGTACTACCACCGGGATTGTTCTCCCATTTTCCAAAC
TTCTTACGGTCCGGAGATATACTAAGTACACCGGATGACGTCGATTGGGCTGCG
GGGACAGCCGGTTCGACGGCATCGACACCCACCTGACAACGCCGTTTCGATTTCCCAATG

AACGCCGCGATTTCGCGAAGTCGTTGCTCATGGAGCTTCGGCGAAGAAGATTGTGGACGTG
 CTTCCGCCAGGACCGGCTGTATCCGCATCCTGAGCTGCTGGTCACCTTCATCGGCAATCAC
 GACATGAAGCGCTTCTGACCGATGCCAACGGCTCACAGGAAAAGCTGAAGCTGGCATT
 TCGTTGCTGGCCACACTGCGCGGAATTCCTCAGCTTTATTACGGCGATGAAATCGGCATGA
 CCGGCGGCGACGACCCAGACAATCGCCATGATTTTCCCGGTGGATTTCCCGGCGACCAG
 CACAATGCCTTACACAAACTGGCAGAACGCCGGATGAGCAGGAAATCTTCGCCCACGTA
 CAAACTGTATTGAAGCTTCGACAGGAACACCCTGCCTTGCGCCGAGGCGCGCAAAGCAT
 ATAGCGGTTGGGGACAAGTACTATGCATTCACGCGCGAGGGCGATGGCGAGCGCTTACTG
 ATCGTGTGAACAACGGTGATGCCGAGAACATCACCATTGATCTCAGCGACACCTCCATCG
 CCGATGCGAAAACGATCACCCCGCTCTTCTCCGCCTCCCCGGCACAACCTCCAGGGCAGTT
 TATTGCGCCTGCAACTTGACACACAACAGTCTGACGGTCTATCGGGTGCAGGATCCGA

>amilase W pET29b NdeI/BamHI PolyHisC-ter (W)

gctcatATGACAATCGCCGGTCCAATTCATCATTGTGCCCGTCTTCATCGTTGTGCTCTCGTGT
 TCCCGCTGCTTCTCTCGTTGACTGCAGCTGCACAGCCGGCCGCTGGACCGAGGGTCTCCA
 AGGTTGAACCGCCGAATTGGTGGATGGACTTTGCACCGACGGTCATGTTTCTGCTCTACG
 GTGAGAACCTTGGGGGAGCCAATGTTTCAGTTGATTATCCGAACGCGGTGGTTGCCAAAG
 TTCTACCTCAACCCGATGGGAAGCACCTCTTTGTCTGGTTGAGTTTTTATCCAGGAACCCG
 CCCGGGAGATGTCGTGATTCACGTGAAGACGCCGTCCGGCGAGACAAGCGTGCCTGTGC
 CATTGTTGCAGCGCTCGCCCCAGGAAGGACGGTTTCAAGGAATCACGCGCGACGATGTCA
 TCTACCTAATCATGCCCGACCGCTTTGCGGACGGCGATCCCGCAAACAACATGCCGCCAG
 GTGCTGCTCCAGGTACTTACGACCGGAGCGGCGCTAAGACCTATCACGGTGGCGACCTGA
 AGGGAATCCAAGAGCACCTGCCGTACTTGAAGGACCTTGGAGTGACGGCGCTTTGGCTCA
 CGCCGTTGTACGACAATGACAACCTCAACTTCGGATTACCACGGCTACGGAGCAGTGGACG
 AATACGCCGTTGAAGATCACTTCGGCACGATGAAGTCCTATCAAGACTTGGTAGCCGCTGC
 TCATCAGGTCGGCCTGAAGGTGATGCTGGACATGGTTCCCAACCATGTTGGCCCCAAGCA
 TCCCTGGGCGACGTCACAGCCCGCTCCGGATTGGCTGCACGGAACGACCGAGCATCATC
 TCGACACAGACTATTACTATGCGCCTGTACTGATCCCCACGCAGTCAAGGCAAATTATGT
 GAGCGCGCTCGAAGGCTGGTTCGCTGACGTCTACCAGACCTGGCGCAGGAGAATCCGC
 TGGTGGCGCTATACCTGCTCCAGAATGCCGAGTGGTGGACAGAGAGTGGCGGGGTGAC
 GGCTTCCGATTGACACATTTCCCTACGTGCCGCGGAGCTTTTGGCAGTACTACCACGCG
 GGATTGTTCTCCATTTTCCAAACTTCTTACGGTCCGGCGAGATATACTCAACTCAGACCCAA
 CGGTGACGTCGTATTGGGCTGGCGGGCAGACCGGGTTCGACGGCATCGACACCCACCTG
 ACAACGCCGTTGATTTCCCAATGAACGCCGCGATTTCGCGAAGTCGTTGCTCATGGAGCTT
 CGGCGAAGAAGATTGTGGACGTGCTTCGCCAGGACCGGCTGTATCCGCATCCTGAGCTGC
 TGGTCACCTTCATCGGCAATCACGACATGAAGCGCTTCTGACCGATGCCAACGGCTCAC
 AGGAAAAGCTGAAGCTGGCATTTCGTTGCTGGCCACACTGCGCGGAATTCCTCAGCTTTA
 TTACGGCGATGAAATCGGCATGACCGGCGGGCGACGACCCAGACAATCGCCATGATTTTCC
 CGGTGGATTTCCCGGCGACCAGCACAATGCCTTACACAAACTGGCAGAACGCCGGATGA
 GCAGGAAATCTTCGCCACGTACAAACTGTATTGAAGCTTCGACAGGAACACCCTGCCTTG
 CGCCGAGGCGCGCAAAGCATATAGCGGTTGGGGACAAGTACTATGCATTCACGCGCGA
 GGGCGATGGCGAGCGCTTACTGATCGTGTGAACAACGGTGTGCCGAGAACATCACCAT
 TGATCTCAGCGACACCTCCATCGCCGATGCGAAAACGATCACCCCGCTCTTCTCCGCCTC
 CCCGGCACAACCTCCAGGGCAGTTTATTGCGCCTGCAACTTGACACACAACAGTCTGACGGT
 CTATCGGGTGCAGGATCCGA

b. Sequências de aminoácidos

>A_ glicosil-hidrolase

MGCEAMPGAVRDGSAIQMDADTGRGTGTPSPARGAAAPLGAIPTGQGTNFSVFSKHATGIELL
 LFDRAENAGPARVIHLDPSTHRTYHYWHVFLPGVTAGQIYGRAEGPWDPANGLRFDRDKLLL
 DPYGRAVVVDPDRYSRDDIRKPGDDCGGAMKSVVDPGSDWEGDAPLRRSSAQTIYEMHV
 RGFTRHPSSGVGGKTRGTFAGLIEKIPYLQKLGVTAVELLPVFQFDAQDCPPGKVNWGYAPV
 SFFAPHAAYSSRSDDLGPLDEFDRDMVKALHRGGIEVILDVFNHTAEGDHNGPTLCFRGLDNP
 TYYLLEDDRSRYANYTGTGNTLNANHPVVRMIVDSLRYWVETMHVDGFRFDLASVLSRDT
 GHPIPNAPILWDIETEPALAGTKLIAEAWDAAGLYQVGSFVGDSWKEWNGRFRDDVRAFFRGE
 PGSVTQIADRILGSPEIYGHEEREAEQSVNFVTCHDGFTLNDVVSYDRKHNEANGEDNRDGAD
 DNRSWNCGVEGSPDDPAIERLRSRQVKNLLVTMLSLGIPMITMGDEARRTQSGNNNAYCQD
 NETSWFDWTLVETHADVHRFVTLNTRRSLRDRMYERVPLRQLLRKAKITWHGVTPEQPDWS
 RDSHSIAVEAKVEQGRLRIYLILNAYWKPLRFDLPPANDGRVGPWRRWIDTSLDPPCDIVEWNL
 APTLSELSYLVEARSVVLIHDAGEPTQLSNS

>B_ triptofanase

MRFPSEPFKIKVVEPIRRTTREERDRLREAGYNLFHVPAESVYVDLLTDSGTSAMSDNQWAG
 LMLGDESYAGSKNYHFEEVRSIFGYKHVIPTHQGRMAENLLFTTIVKPGMCPVNNIHFDTR
 ANVEHQGAQALDMVVKEAYDPHCELPFKGNMDLVRLEGTINRVGRDHIPLVMLTITNNSGGGQ
 PVSMDNIRATRVLNRYDIPLFFDACRFAENCFKIKEREKPGYDGVSIQELFSYGDGCTMSA
 KKDGLVNIGGFLSLKNDQWAQDITNMLILVEGFSTYGGLAGRDLEAMARGLREVLEDEDYLSFRI
 GQVRYLGELLDQAGVPILKPIGGHAVYLNKEFLPHIQQAQFPAQALVAALYREYGIQVEIGTV
 MFGKTDSATGRTIYPELEMVRLAIPRRVYTNMQUYVAESIIELYQRREMIHGLALTYEASVLRHF
 TARFTELHDQSLFQRSASSNS

>C_ lipase_ His-Tagged

MLAASATFTAQTQAAGYTQTRYPIVLVHGLFGFDNIGPVEYFYGIPSALRSGGATVYTPQVSA
 ANSTEVRGEQLLLEVKKIVAVTGKPKVNLIGHSHCGPTIRYVASVRPDLVASATSVAGVNGGSA
 VADILLGIAPPGSLSRDVTITATGLGKLLSLLSGSSTLPQNSLAAAQSLSTAGSAKFNAHPAGL
 PSTACGEGAYQVNGVAYFSWSGASNYTNVLDILDPALAVTGLAFGGAKNDGLVSSCSSHLGK
 VIRDDYAMNHADINQSVGIVNLFVNPVSVFRQHANRLQGMGLLEHHHHHH

Obs. É indicado o pentapeptídeo conservado característico da família das α/β hidrolases.

>D_ prolina-peptidase_ His-Tagged

MGSSHHHHHHSSGLVPRGSHMASMNTVKTGGVQMVSIDGRFQVWTKRIGAGPPTMLTLHGG
 PGSTHEYFEFCDFLPPNGIQLIYYDQLGSGNSDQPDNPALWVVERFRDEVEQVRAALGLTGF
 YLYGHSWGGMLAIEYALKYQSHLGLIISNMTASIASYVTVNELRRQLPAESQRILERYEATGE
 YTAPEYKVMFGEIYSRHLRCLAPWPEPLIRMIRHMNQKVYNKMQGPNEFVVTGTFKDWDWRW
 DDIKNINVPNTLLSVGRFDTMSVADVERMGTLIPNARVVICETGSHCSMYDDQERYFEDLVRFIK
 DVEAGKLV

Obs. É indicado o pentapeptídeo conservado característico da família das α/β hidrolases.

>E_ metaloprotease

MEERPLTAAEQRSKEFAATILRFTEKVVWDEQFQKIGKKYSKPHMVLVSEQVDTGCGSAPSAVG
 PFYCPADKTVYLDPTFFDELQNKLGSGKGEFSQAYVIAHEVGHVQNLGYSRIADENQQSAP
 SKAKANESVRLQLADYLAGVWAHGGKKEFNIEPGDIESAIKTANAIGDDRLQKRATGFTSP
 EKYTHGTSARLKWFRAGFETGDLKMKELFDLPYKLL-

>F_ transcriptase-reversa_ His-tagged

MGSSHHHHHHSSGLVPRGSHMMTARDAESMAEKPESSHSGGSGRKS RDNEAGASRV TARGE
 DSSPECEQLMEAVVERGNMQTALQRVMSNRGAAGADGMTVDELKPHLREEWKRIKGELLAG
 EYQPEPVLKVEIPKAEGKGVKRLGIPTVVDRLIQALHQLSPIFEPGFSESSYGFPRGRSAQD
 AVRQARAYVGEGRRWVVDIDLEKFFDRVNHDKMMSRLARRIKDKRILRMIRRYLQAGMMEGG
 LVTQRREGTPQGGPLSPLLSNILLDELDELKELERRGHKFCRYADDCNVYVRSRSAGERVKESIT
 RFLERRLRKLVNEEKSAVERPWKRKFLGYTMTWHLEPRIKVAENSVKRLKVKLREILRQGRGR
 NIGRLIEEELTPLLRGWMNYFRLAEVKGIFEELDSWIRRLRCVIWRQWKRTWARVKGLMKRG
 LERDRALKSATNGRGPWWNAGASHMHEAFPKEYFDRCLVSLLDQRLKVQRTS-

>G_deidratase

MILISSRVGSTEMSGLYFEEFKVGLFHHAITRTVTETDNLLFTTLTHNPQPLHLDVEFVKQTEF
GQRLVNSIFTLGLMIGVSVGDTTLTGTTVANLGMNDVRFANPVFIGDTLRAQSTVLEMRESKSRP
DAGIVVFEHRCLNQRDEEVGYCKRSALMRRKARLES

>S_amilase-sem-peptideo-sinal

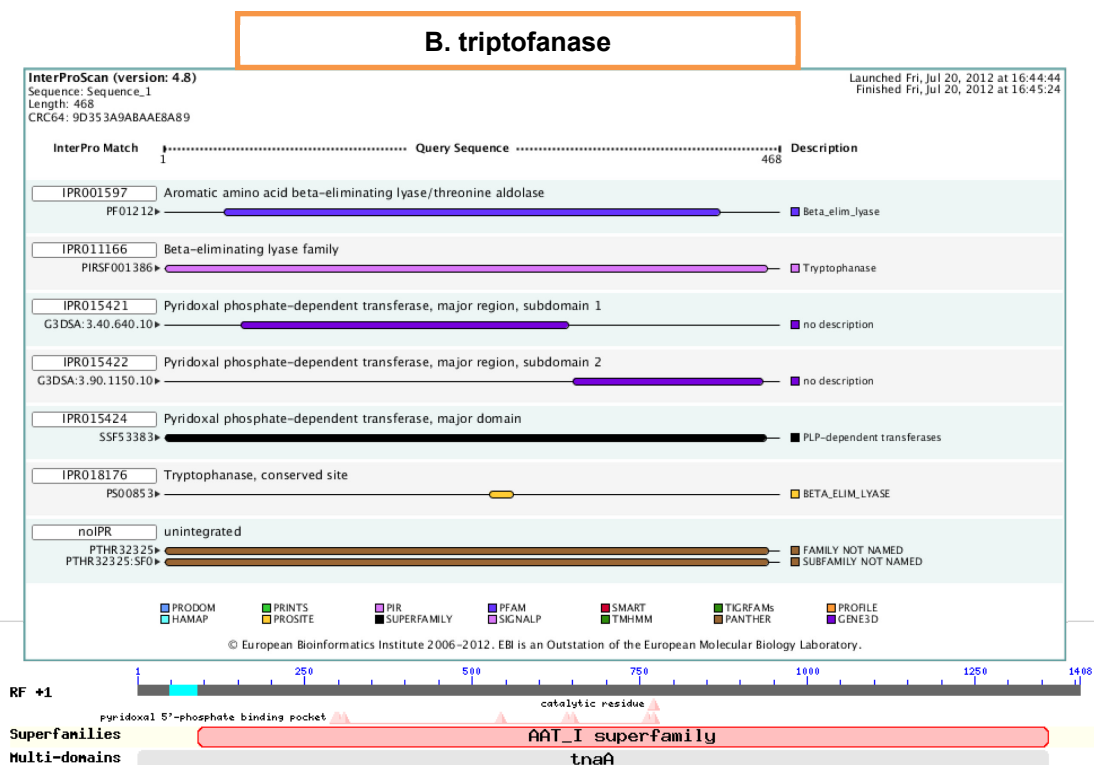
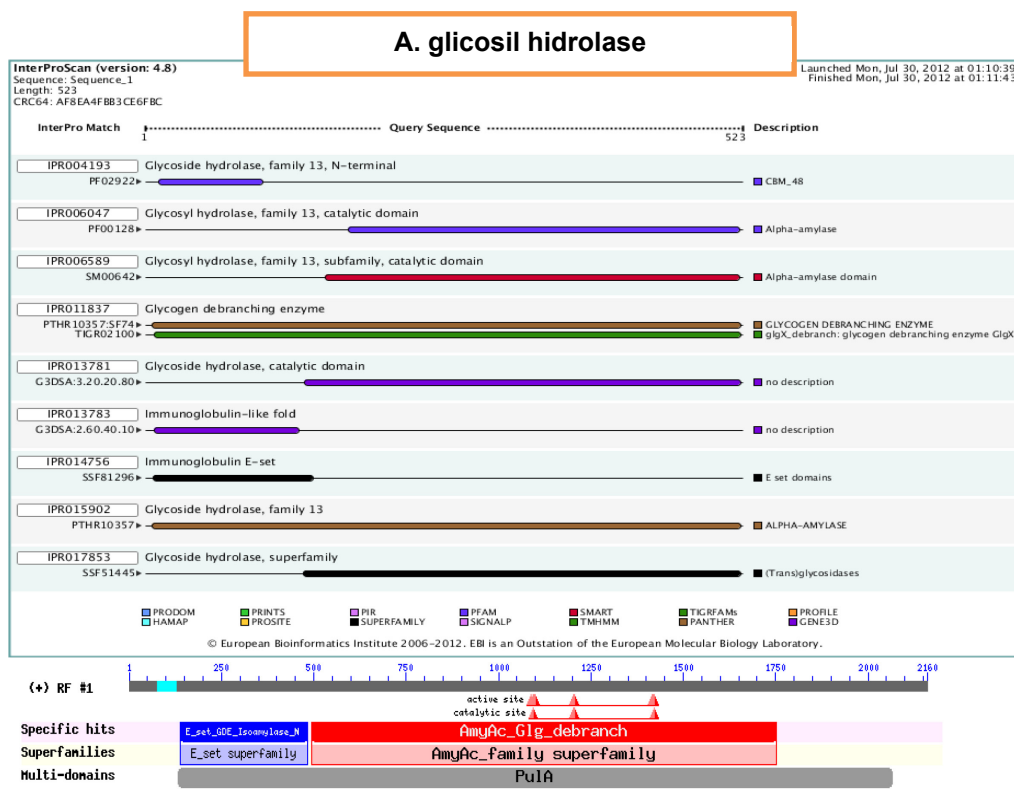
MAQPAAGPRVSKVEPPNWWMDFAPTVMFLLYGENLGGANVSVDYPNAVVAKVLPQPDGKHL
FVWLSFYPGTRPGDVVIHVKTSPGETSVPVPLLQRSPQEGRFQGITRDDVIYLIMPDRFADGDP
ANNMPPGAAPGTYDRSGAKTYHGGDLKGIQEHLPLYKDLGVTALWLTPLYDNDNSTSDYHGY
GAVDEYAVEDHFVTMKSQDLVAAAHQVGLKVMLDMVFNHVGPKHPWATSQPAPDWLHGTT
EHLDTDYYPVTPHAVKANYVSALEGWFADVLPDLAQENPLVALYLLQNAEWWTESGGV
DGFRIIDFPYVPRSFQYHAGLFSHFNFVTVGEIYNSDPTVTSYWAGGQTGFIDGIDHLTT
PFDFPMNAIREVVAHGASAKKIVDLRQDRLYPHPELLVTFIGNHDMKRFLTDANGSQEKLKL
AFSLLATLRGIPQLYYGDEIGMTGGDDPDNRHDFPGGFGDQHNAFTQTGRTPDEQEIFAHVQ
TVLKLREQHPALRRGAQKHIAVGDKYYAFTREGDGERLLIVLNNNGDAENITIDLSDTSIADAKTIT
PLFSASPAQLQGSLRLQLAHNSLTVYRVQDP

>W_amilase-com-peptideo-sinal

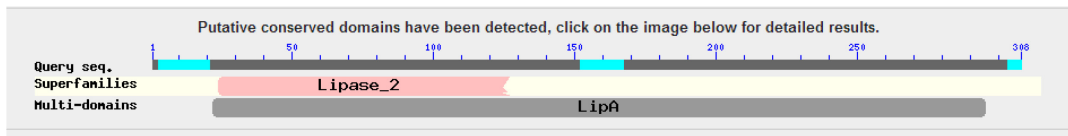
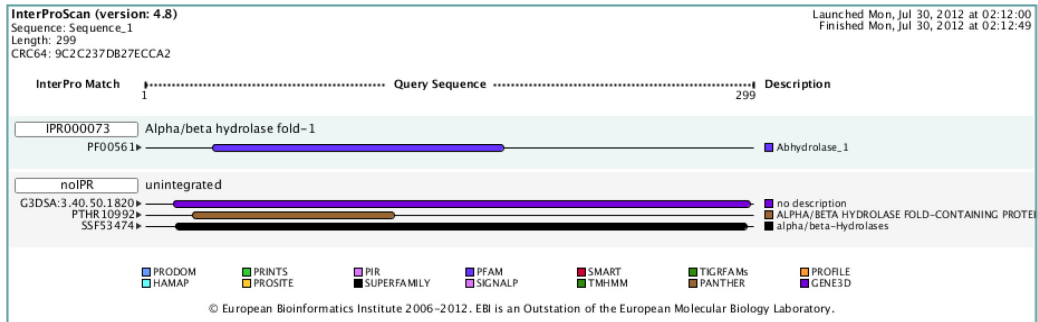
MTIAGPIHHCARLHRCALVFLLLLSLTAAQPAAGPRVSKVEPPNWWMDFAPTVMFLLYGENL
GGANVSVDYPNAVVAKVLPQPDGKHLFVWLSFYPGTRPGDVVIHVKTSPGETSVPVPLLQRSP
QEGRFQGITRDDVIYLIMPDRFADGDPANNMPPGAAPGTYDRSGAKTYHGGDLKGIQEHLPLY
KDLGVTALWLTPLYDNDNSTSDYHGYGAVDEYAVEDHFVTMKSQDLVAAAHQVGLKVMLD
MVPNHVGPCKHPWATSQPAPDWLHGTT EHLDTDYYPVTPHAVKANYVSALEGWFADVL
PDLAQENPLVALYLLQNAEWWTESGGVDGFRIDTFPYVPRSFQYHAGLFSHFNFVTVGEI
YNSDPTVTSYWAGGQTGFIDGIDHLTTPDFPMNAIREVVAHGASAKKIVDLRQDRLYPH
ELLVTFIGNHDMKRFLTDANGSQEKLKLAFSLLATLRGIPQLYYGDEIGMTGGDDPDNRHDFPG
GFGDQHNAFTQTGRTPDEQEIFAHVQTVLKLREQHPALRRGAQKHIAVGDKYYAFTREGDG
ERLLIVLNNNGDAENITIDLSDTSIADAKTITPLFSASPAQLQGSLRLQLAHNSLTVYRVQDP

Suplemento 2. Análise bioinformática de domínios proteicos das 9 sequências selecionadas.

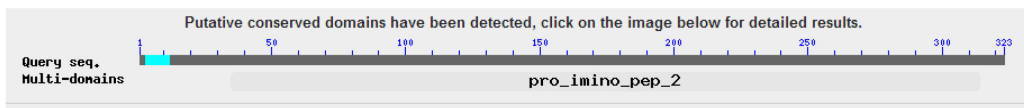
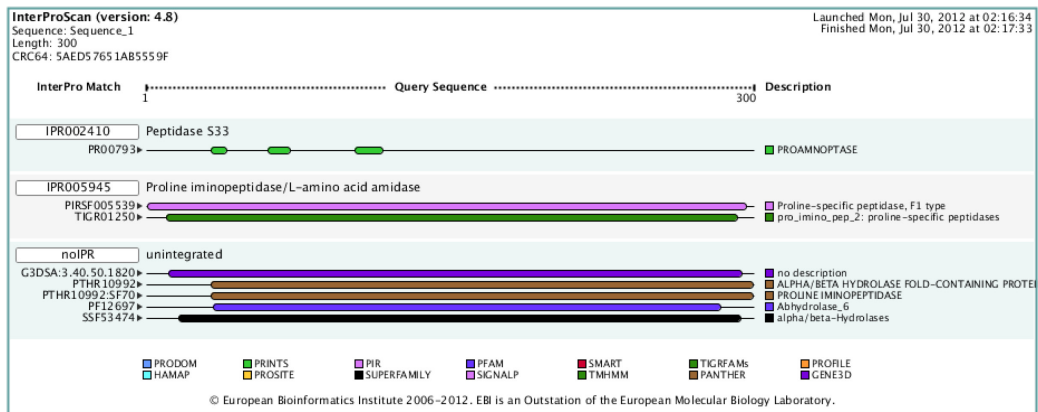
Domínios funcionais conservados de proteína utilizando *InterPro Scan* (EMBL-EBI) e *Conserved Domains* (NCBI).



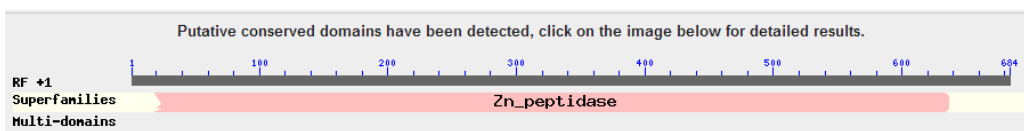
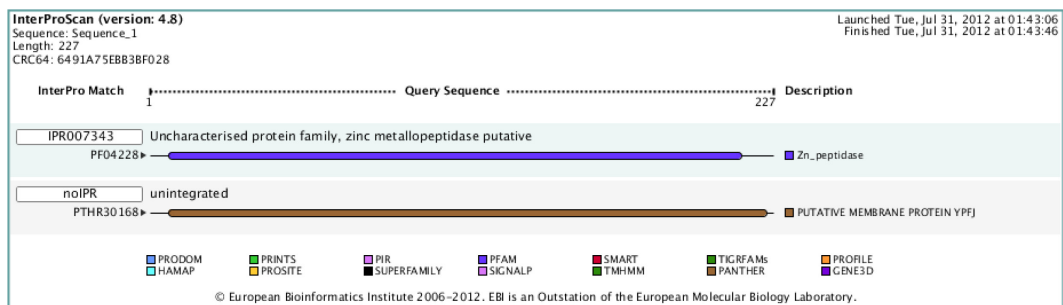
C. triacilglicerol lipase



D. neptidase prolina especifica



E. metaloprotease



F. transcriptase reversa

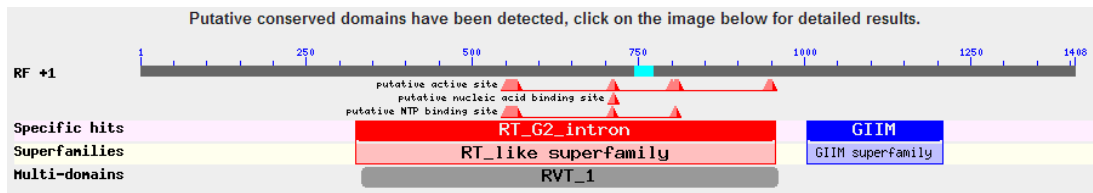
InterProScan (version: 4.8)
 Sequence: Sequence_1
 Length: 468
 CRC64: 2B019F3F62A4DB8C

Launched Tue, Jul 31, 2012 at 02:26:09
 Finished Tue, Jul 31, 2012 at 02:27:26

InterPro Match	Query Sequence	Description
IPRO00477	Reverse transcriptase	
PF00078		RVT_1
P550878		RT_POL
IPRO13597	Group II intron, maturase-specific	
PF08388		GIIM
noIPR	unintegrated	
PTHR19446		REVERSE TRANSCRIPTASES
PTHR19446:SF107		PREDICTED: SIMILAR TO RNA-DIRECTED DNA POLYM
SSF56672		DNA/RNA polymerases

■ PRODOM ■ PRINTS ■ PIR ■ PFAM ■ SMART ■ TIGRFAMS ■ PROFILE
■ HAMAP ■ PROSITE ■ SUPERFAMILY ■ SIGNALP ■ TMHMM ■ PANTHER ■ GENE3D

© European Bioinformatics Institute 2006–2012. EBI is an Outstation of the European Molecular Biology Laboratory.



G. deidratase

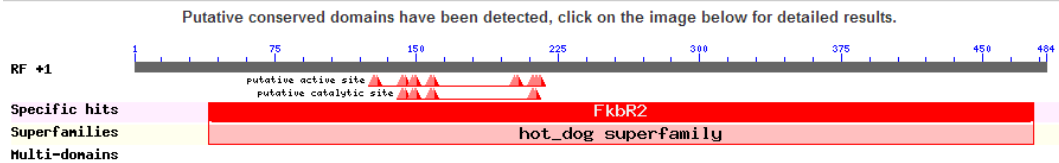
InterProScan (version: 4.8)
 Sequence: Sequence_1
 Length: 150
 CRC64: 99B17C162A5C922F

Launched Thu, Jul 26, 2012 at 21:44:12
 Finished Thu, Jul 26, 2012 at 21:45:28

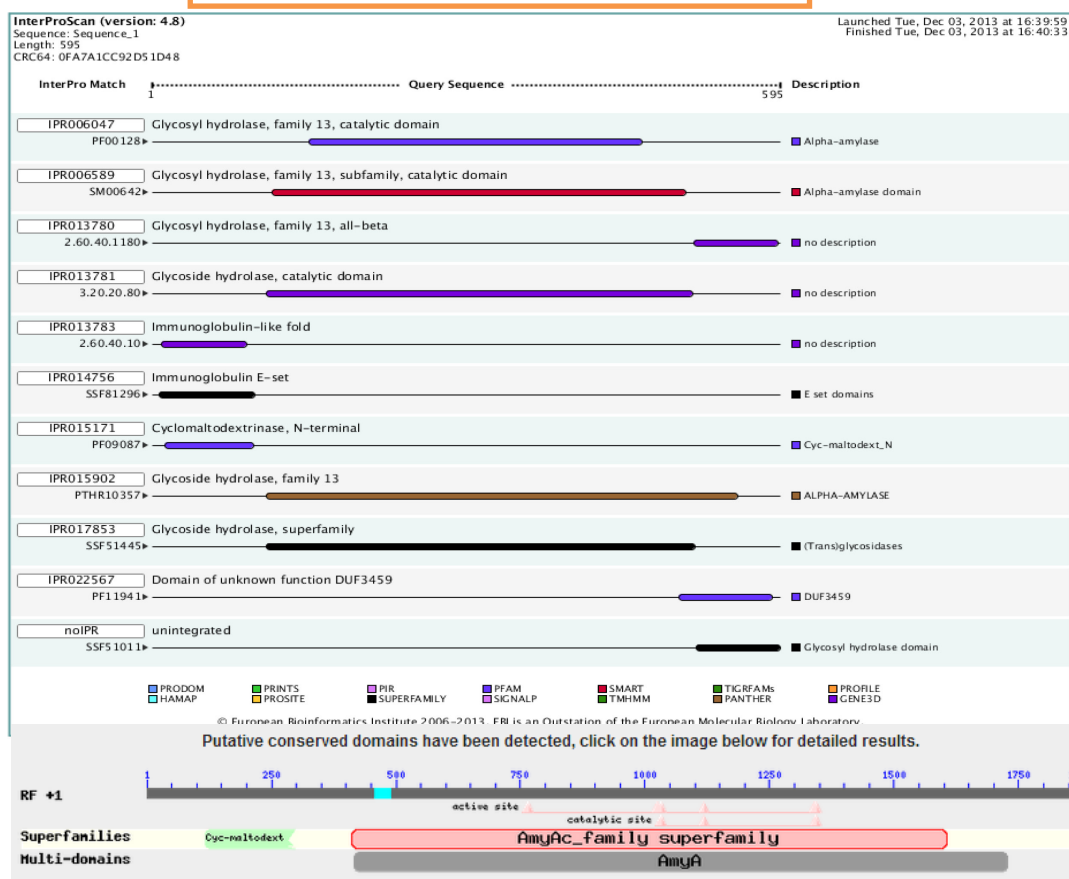
InterPro Match	Query Sequence	Description
IPRO02539	MaoC-like dehydratase	
PF01575		MaoC_dehydratas
noIPR	unintegrated	
G3DSA:3.10.129.10		no description
PTHR13078		FAMILY NOT NAMED
PTHR13078:SF18		POTENTIAL UNCHARACTERIZED PROTEIN
SSF54637		Thioesterase/thiol ester dehydratase-isomerase

■ PRODOM ■ PRINTS ■ PIR ■ PFAM ■ SMART ■ TIGRFAMS ■ PROFILE
■ HAMAP ■ PROSITE ■ SUPERFAMILY ■ SIGNALP ■ TMHMM ■ PANTHER ■ GENE3D

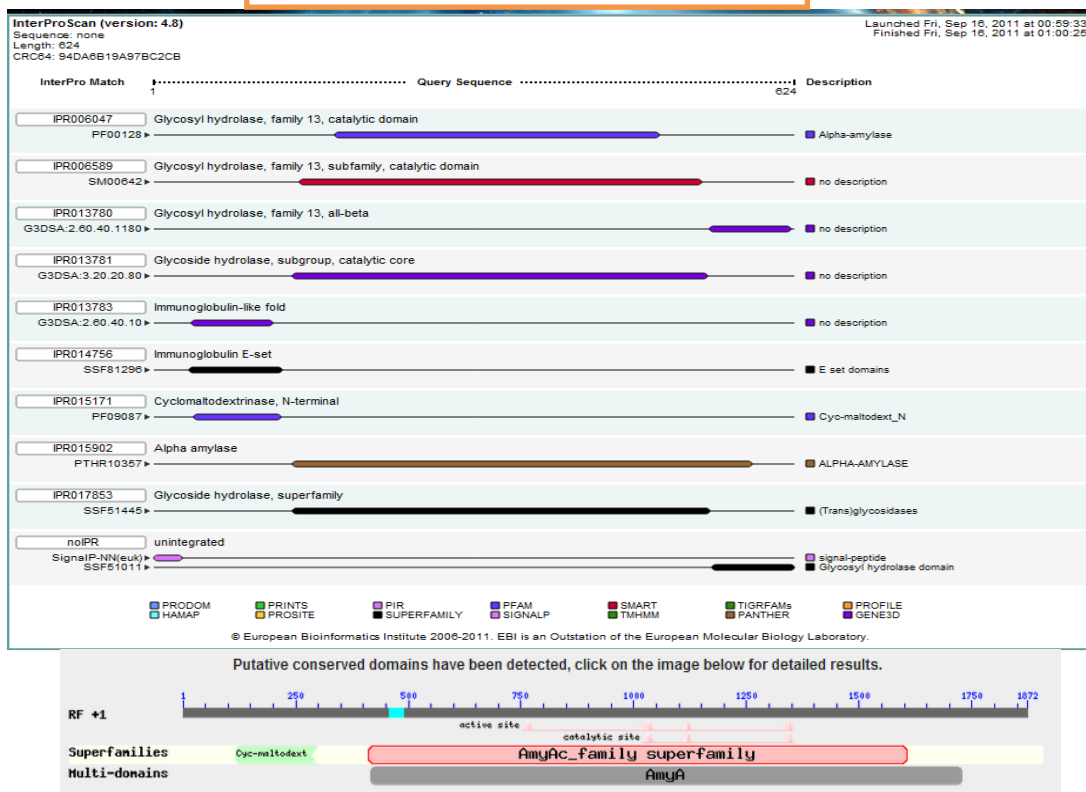
© European Bioinformatics Institute 2006–2012. EBI is an Outstation of the European Molecular Biology Laboratory.



S. amilase sem peptídeo sinal



W. amilase com peptídeo sinal



Suplemento 3. Alinhamentos com a sequência de amino ácido de maior similaridade encontrada no GenBank utilizando o programa *BlastX*.

Para cada sequência encontrada é apresentada o nome da enzima, o organismo ao qual pertence, informações do alinhamento e as sequências submetida (Query) e encontrada (Sbjct).

glycosyl hydrolase (glycogen debranching enzyme) [Azospirillum brasilense Sp245]

Sequence ID: [ref|YP_005032618.1](#) Length: 719 Number of Matches: 1

[▶ See 2 more title\(s\)](#)

Range 1: 33 to 547 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
844 bits(2180)	0.0	Compositional matrix adjust.	391/515(76%)	451/515(87%)	0/515(0%)	+2
Query 29		KGESAPLGATPNKLGVNFSLFSRHASGVQLLFFDREDADQPTRVIRLDPFTNRITYYWHV			208	
Sbjct 33		+G +APLGA P G NFS+FS+HA+G++LL FDR + P RVI LDP T+RTY+YWHV			92	
Query 209		RVGAPVQPGQIYGYRVEGPFDPGRGLRFDPAKVLDDTYGRGVIVPRKYSRETAHSTEDNAA			388	
Sbjct 93		F+PGV GQIYGYR EGP+DP GLRFD K+LLD YGR V VP +YSR+ D+			152	
Query 389		FLPGVITAGQIYGYRAEGPWPDPANGLRFRDRKLLLDYPYGRAVVVPPDRYSRDDIRKPGDDCG			568	
Sbjct 153		TAMKSVVIDPNVYDWECDTPLRRSCSRITILYEMHVRGFTRHPSSGVPEEKRGTYAGMIEK			212	
Query 569		AMKSVV+DP YDWECD PLRRS ++TI+YEMHVRGFTRHPSSGV + RGT+AG+IEK			748	
Sbjct 213		GAMKSVVVDPGSYDWECDAPLRRSSAQTIVYEMHVRGFTRHPSSGVGGKTRGT FAGLIEK			272	
Query 749		IPYLQRLGVTAVELLPVFQFDVQDCPPGLVNYWGYAPVVSFFSPHQAYSSRQDGSQPADEF			928	
Sbjct 273		IPYLQRLGVTAVELLPVFQFDVQDCPPG VNYWGYAPVVSFF+PH AYSSR D GP DEF			332	
Query 929		RDLVKALHRAGIEVILDVVFVNHTAEGDHSGPTLSFKGVDNSTYYILERNRAEYANYSGTG			1108	
Sbjct 333		RD+VKALHR GIEVILDVVFVNHTAEGDH+GPTL F+G+DN TYY+LE +R+ YANY+GTG			392	
Query 1109		RDMVKALHRGGIEVILDVVFVNHTAEGDHNGPTLCFRGLDNPTYLLEDDRSRYANYTGTG			1288	
Sbjct 393		NILNANHPIVRRMILDSVRYWVGMHVDGFRFDLASILARDSSGHPLANPPVLWDIESDP			452	
Query 1289		LAGTK+IAEAWDAAG+YQVGSF+GDSW+EWNGRFRDDVR+FFRG+ GSV QIADRI+GS			1468	
Sbjct 453		ALAGTKLIAEAWDAAGLYQVGSFVGDVSWKEWNGRFRDDVRAFFRGEPSVTQIADRILGS			512	
Query 1469		PEIYGHKEREAEQSVNFVTC HDGFTLNDLVSYNQKHNEENGEENRDGANDNQSWNCGVEG			1573	
Sbjct 513		PEIYGH+EREAEQSVNFVTC HDGFTLND+VSY++KHNE NGE NRDGA+DN+SWNCGVEG			547	
Query 1469		PTDDPAIEKLRNRQVKNFLVTMLSLGMPMILMGD				
Sbjct 513		P+DDPAIE+LR+RQVKN LVTMLSLG+PMI MGD				
Query 1469		PSDDPAIERLRSRQVKNLLVTMLSLGIPMITMGD				

tryptophanase [Singulisphaera acidiphila DSM 18658]

Sequence ID: [ref|YP_007201759.1](#) Length: 458 Number of Matches: 1[▶ See 2 more title\(s\)](#)

Range 1: 1 to 457 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
545 bits(1405)	0.0	Compositional matrix adjust.	264/457(58%)	338/457(73%)	0/457(0%)	+1
Query 1	MRFPEPFKIKVVEPIrrttreerdrllreAGYNLFHVPAESVYVDLLIDSGISAMSDNQ					180
Sbjct 1	M+ EPFKIK+VEP++ TT E+R++ LR+A +NLF +PAE V +DLLIDSGISAMS +Q MKTIIIEPFKIKMVEPLKLTITVEQREQALRKAHFNLFQIPAEVDIIDLLIDSGISAMSADQ					60
Query 181	WAGLMLGDESAYAGSKNYHFEEVRSIFGYKHVIPHQGRMAENLLFTTIIVKPGMCPVNN					360
Sbjct 61	WAG++ GDESYAG++++HFE V+R + G H++PTHQGR +E +LF PG +P+N WAGMIRGDESAYAGARSWFHFENVLRDLTGMHILPTHQGRASERILFELTGGPGKVIPSN					120
Query 361	IHFDTTRANVEHQGAQALDMVVEKAYDPHCELPFKGNMDLVRLEGTINRVGRDHIPVLM					540
Sbjct 121	HFDTTTRAN+EH GA+A+D+V+ E DP PFKGN+D+ +LE I +G + IPL M NHFDTTTRANIEHSGARAVDLVIAEGTDPNRNHPFKGNIDVAKLESLEEIGSERIPLCMA					180
Query 541	TITNNSGGGQPVSMNIRATRVLLNRYDIPLFFDACRFAENCFEIKEREPEGYDGVSI					720
Sbjct 181	T+TNNNSGGGQPV+S N+RA R + +R+ IPLF DACRFAEN I++REPG G S I TVTNNNSGGGQPVSLANLRAVREVCHRHGIPLFLDACRFAENAALIQQREPGQSGRSARAI					240
Query 721	AQELFSYDGGCTMSAKKDGLVNIIGGFLSLKNDQWADITNMLLILVEGFSTYGGLAGRDLE					900
Sbjct 241	A+E+F DG T+SAKKDGLVNIIG L ++ND A ++LIL EGF TYGGLAGRDLE AREMFDLADGATISAKKDGLVNIIGVLLMRNDALALRANDLILTEGFVITYGGLAGRDLE					300
Query 901	AMARGLREVLEDEYLSFRIGQVRYLGELLDQAGVPILKPIGGHAVYLNAAKFLPHIQQAQ					1080
Sbjct 301	AMA+G EVLDEDEYL +R+ V YLGE L AG+PI++P GGHA+Y++A F HI AMAQQFVEVLEDEYLYRRLRSVAYLGEHLLAAGIPIVEPPGGHAIYIDAASFCTHIPPRH					360
Query 1081	FPAQALVAALYREYRIGRVEIGTVMFGKTDSATGRTIYPELEMVRLAIPRRVYTNMQITY					1260
Sbjct 361	FP QALV ALYR GIRGVEIG+VMF D TG T++P +E+VRLA+PRRVYT I Y FPGQALVCALYRHAGIRGVEIGSVMFHSVDPDTGETVHPPMELVRLALPRRVYTQSHIDY					420
Query 1261	VAESIIELYQRREMIHGLALTYEASVLRHFTARFTEL 1371					
Sbjct 421	E++IEL +R+ I GL + LRHFTA+FTE+ TVEAMIELAAQRDSIRGLRIVESPPTLRHFTAKFTEI 457					

triacylglycerol lipase [Rhodofera ferrireducens T118]

Sequence ID: [ref|YP_524598.1](#) Length: 305 Number of Matches: 1[▶ See 2 more title\(s\)](#)

Range 1: 27 to 305 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
425 bits(1092)	3e-146	Compositional matrix adjust.	215/279(77%)	245/279(87%)	0/279(0%)	+1
Query 64	RYPIVLVHGLFGFDNIGPVEYFYGIPSAIRSGGATVYTPQVSAANSTEVRRGEQLLLEVKK					243
Sbjct 27	R+PIVLVHGLFGFDNIGP++Y+YGIPSAIR+ GA VY QVSAANSTEVRRGEQLL+VK+ RFPVIVLVHGLFGFDNIGPLDYWYGIPSALRADGAQVYVYTPQVSAANSTEVRRGEQLLVQVKQ					86
Query 244	IVAVTGKPKVNLIGHSHGGPTIRYVASVRPDLVASATSVAGVNGKSAVADILLGIAPPGS					423
Sbjct 87	I+A TG KVNLIHSHGGPTIRYVASVRPDLVAS TSV GVNKGSVAD+LLG+APPGS ILAAATGASKVNLIGHSHGGPTIRYVASVRPDLVASVTSVGGVNGKSAVADVLLGVAPPGS					146
Query 424	LSRDVITIIAtglgkllsllsgsstlPQNSLAAAQSLSTAGSAKFNAHPAGLPSTACGE					603
Sbjct 147	LS V+ +I GLG ++S LSG S L Q+SLAAAQSLSTAGS KFN AHP GLP+TACGE LSNSVLISITNGLGSIISFLSGGSGLSQDSLAAAQSLSTAGSLKFNLAHPEGLPTTACGE					206
Query 604	GAYQVNGVAYFWSWGSANYNVLDILDPALAVTGLAFGGAKNDGLVSSCSSHLGKIVIRDD					783
Sbjct 207	GAY V GVAYFWSGA YTNV D+LDPALA+T LAF GAKNDGLV+SCSS LG+VIRDD GAYAVRGVAYFWSWGAQPYTNVFDVLDPALALTSLAFNGAKNDGLVSSCSSRLGRVIRDD					266
Query 784	YAMNHADEINQSVGIVNLFVNPVSVFRQHANRLQGMGL 900					
Sbjct 267	YA+NH DE+NQ+VG+VNLFE NPV+++RQ ANRL+ +GL YALNHLDEVNQTVGLVNLFEINPVTLYRQANRLKNLGL 305					

proline-specific peptidase [*Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* DSM 446]Sequence ID: [ref|YP_003185561.1](#) Length: 295 Number of Matches: 1[▶ See 2 more title\(s\)](#)Range 1: 5 to 295 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
361 bits(926)	3e-121	Compositional matrix adjust.	161/291(55%)	213/291(73%)	0/291(0%)	+1
Query 25	VQMVSIDGRFQVWIKRIGAGPPTMLTLHGPGSTHEYFECFEDFLPPNGIQLIYYDQLGS					204
Sbjct 5	+++++ VWT+R+G P ML LHGGPG++HEYFE FE +L GI+L +YDQLGS					64
Query 205	GNSDQPDNPALWVVERFRDEVEQVRAALGLTGFYLYGHSWGGMLAIEYALKYQSHLKGLI					384
Sbjct 65	SDQPD+P+LW ++RFR EV++VR A+GL FYL G SWGGMLA+EYAL + LKGL+					124
Query 385	ISNMTASIASYVTVYNELRRQLPAESQRILERYEATGEYTAPEYKVMFGEIYSRHLCL					564
Sbjct 125	ISNMTASI SYV Y+ LR+QLP E Q L+ +E G+Y + Y++++ +Y +HLCRL					184
Query 565	APWPEPLIRMIRHMNQKVYNKMQGPNEFVVVTGFKDWRDWDIKNINVPILLSVGRFDTM					744
Sbjct 185	PWP+ ++R HMNQ+VYN MQGPNEFVVVG FKDWRW + ++VPTL+ R+DTM					244
Query 745	SVADVERMGTLIPNARVSICETGSHCSMYDDQERYFEDLVRFIKDVEAGKL 897					
Sbjct 245	AD+E M IP AR +IC GSH SM+DD + YF+ +V F++DVEAG+					295

metalloprotease [*Schlesneria paludicola*]Sequence ID: [ref|WP_010585896.1](#) Length: 289 Number of Matches: 1Range 1: 69 to 284 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
350 bits(899)	2e-118	Compositional matrix adjust.	165/217(76%)	190/217(87%)	1/217(0%)	+1
Query 16	TAEEQRSKEFAATILRFTEKQVWDEQFQKIGKKYSKPHMVLFSQVDTGCGSAPSAVGPFY					195
Sbjct 69	+ EE RS++FAATIL +TE VW E F+K G++Y P MVLFS QV+I CG APSAVGPFY					128
Query 196	CPADKTVYLDPTFFDELQNKLGSKGEFSQAYVIAHEVGHVQNLGYSRIADENQQSAP					375
Sbjct 129	CPAD+TVYLDPTFF EL ++LGGG EFSQAYVI HEVGHVQNLGYSRI DE +Q++					187
Query 376	SKAKANESVRLLEQADYLAGVWAHGGKKEFNFIIEPGDIESAIKTANAIGDRLQKQATG					555
Sbjct 188	SK +AN WSVRLELQADYLAGVWAH+G+++F FIE GDIESAI++ANAIGDRLQK+ATG					247
Query 556	FTSPEKYTHGISAQRLLKWFRAFETGDLKMKELFDL 666					
Sbjct 248	FTSPEKYTHGISAQR+KWER GFETGDL K+KELF+L					284

RNA-directed DNA polymerase [Syntrophobacter fumaroxidans MPOB]

Sequence ID: [ref|YP_847308.1|](#) Length: 467 Number of Matches: 1[▶ See 2 more title\(s\)](#)Range 1: 29 to 467 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
598 bits(1543)	0.0	Compositional matrix adjust.	303/440(69%)	360/440(81%)	1/440(0%)	+1
Query 85		AGASRVTARGETDSSPECEQLMEAVVERGNMQTALQRVMSNRGAAGADGMTVDELKPHLRE			264	
		A AS + AR +DS QLMEAVVER NM AL++V +N+G+AG DG++VD L+ LRE				
Sbjct 29		ACASSLAARRDSDRQRTMQLMEAVVERENMFGALRQVEANKGSAGVDGVSVDALRACLRE			88	
Query 265		EWKRIKIGELLAGEYQPEPVLKVEIPKAEGKGVKRLGIPTVVDRLIQQALHQVLSPIFEPG			444	
		W RIK ELL G YQP+PV KVEIPK GKG+R+LGIPTV+DRLIQQAL+QV+ PIF+P				
Sbjct 89		HWPRIKEELLEGRYQPQVVRKVEIPKGGKGMRLGIPTVMDRLIQQALNQVMQPIFDPD			148	
Query 445		FSESSYGFRPGRSAQDAVRQARAYVGEGRWVVDIDLEKFFDRVNHDKMMSRLARRIKDK			624	
		FSESSYGFRPGRSA AV +AR Y RRWVD+DLEKFFDRVNH +M+RLAR+I D+				
Sbjct 149		FSESSYGFRPGRSAHQAVLRAREYAATDRRWVDMDEKFFDRVNHDIIMARLARKIADR			208	
Query 625		RILRMIRRYLQAGMMEGGLVTQRREGTPQGGPLSPLLSNILLdeldekeleRRGHKFCRYA			804	
		R+L++IRRYLQAG M GG+V+ R EGTQGGPLSPLLSNILLD+LDKELE+RGH FCRYA				
Sbjct 209		RVLQLIRRYLQAGSMVGVVSPRTEGTPQGGPLSPLLSNILLDDLDKELEQRGHAFCRYA			268	
Query 805		DDCNVYVRSRAGSERVKESITRFLERRLRKLVNEEKSAVERPWRKFLGYTMTWHLEPRI			984	
		DDCN+YV+SR AG+RV ES+TRFL RL+LKN +KSAV RPW RKFLGY+MT+H PR+				
Sbjct 269		DDCNIVKSRRAQRVLESLTRFLANRLKLVNVDKSAVARPWRKFLGYSMTFHKRPRL			328	
Query 985		KVAENSVKRLKVKLREILRQGRGRNIGRLIEEELTPLLRGWNYFRLAEVKGIFEELDSW			1164	
		+VA V R+K KLRE R GRGRNI R+I EELTP+LRGW+NYFRL+EVKG FEELD W				
Sbjct 329		RVAPAVVDRMKAKLREQFRMGRGRNIRRVI-EELTPVLRGWNYFRLSEVKGNFEELEW			387	
Query 1165		IRRKLRVIVRQWKRTIWARVKGMLKRGLEDRALKSATNNGRGPWWNAGASHMHEAFPKTY			1344	
		IRRK RC+IWRQWKRTI+ R K +MK GL +RA +SA N RGPWWN+GASHM++ FPK +				
Sbjct 388		IRRKFCIIWRQWKRTIYTRAKNMMKCGLGEERAWRSKRNQRPWWNSGASHMNQCFPKRF			447	
Query 1345		FDRCGLVSLLDQRLKVVQRTS 1404				
		F+R GLVSL Q ++Q TS				
Sbjct 448		FERLGLVSLLSQLRRLQCTS 467				

dehydratase [Tistrella mobilis KA081020-065]

Sequence ID: [ref|YP_006373147.1|](#) Length: 156 Number of Matches: 1[▶ See 2 more title\(s\)](#)Range 1: 1 to 149 [GenPept](#) [Graphics](#)

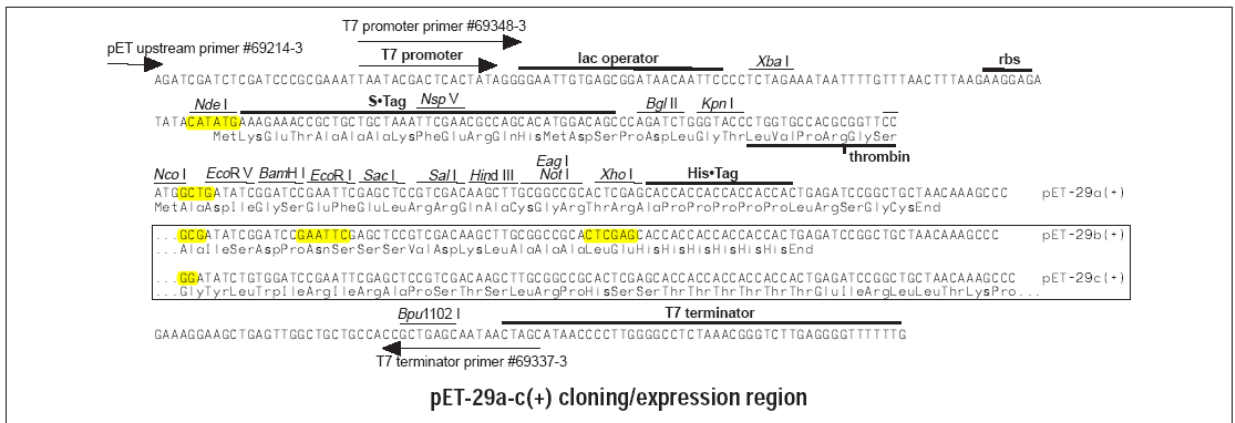
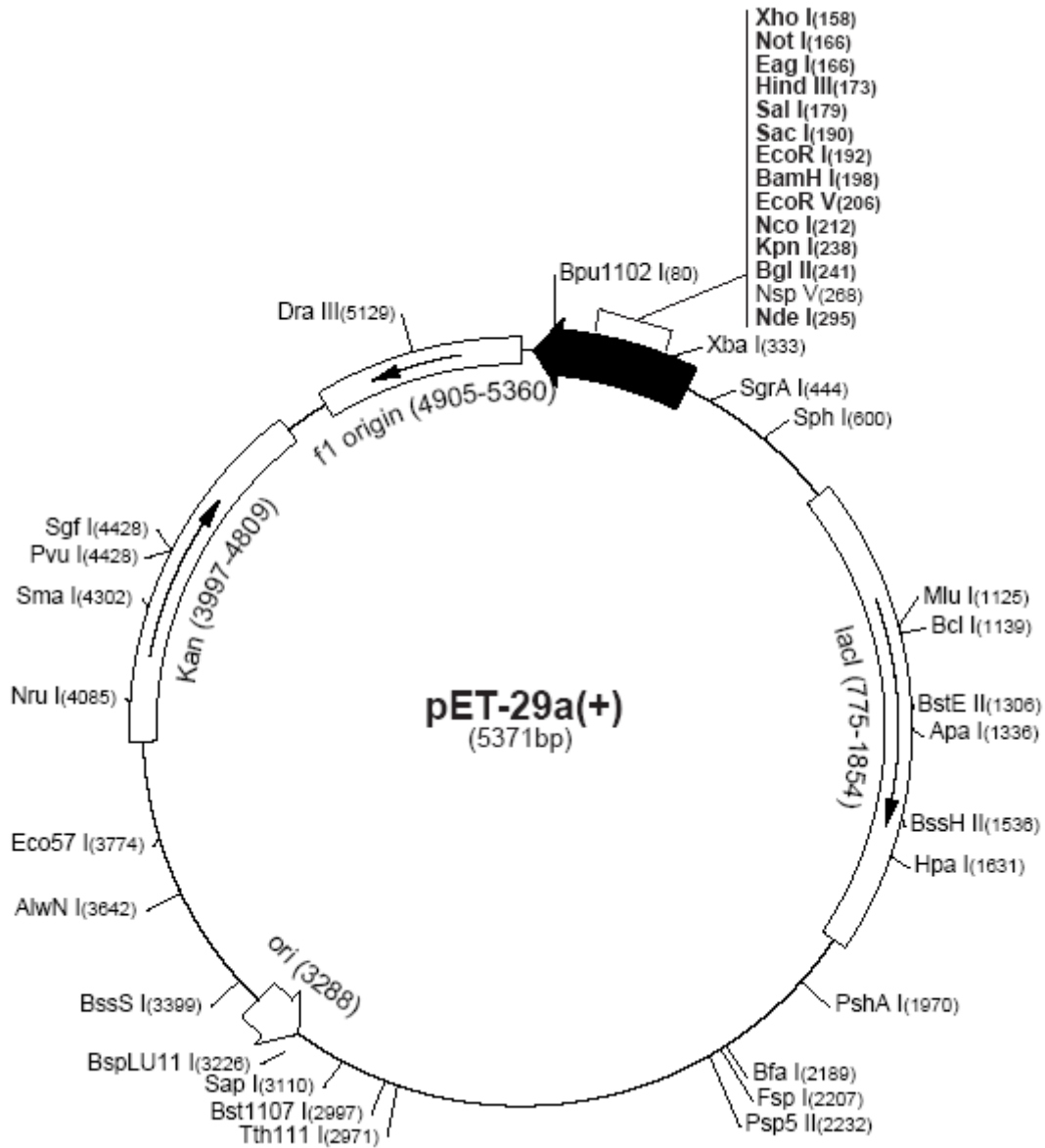
▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
232 bits(592)	6e-75	Compositional matrix adjust.	109/149(73%)	125/149(83%)	0/149(0%)	+1
Query 34		MSGLYFEFVKVQQLFHHAITRIVTETDNLFTLTHNPQLHLDDVEFVKQTEFGQRLVNS			213	
		MSGL+FEF VQ F HA+TRIVTETDNL FTLTHNPQLHLDD E F K TEF G+ LVNS				
Sbjct 1		MSGLFFFEFTVGRQFEHAVTRIVTETDNLITLTHNPQLHLDAEFCKGTEFGRIIVNS			60	
Query 214		IFTLGLMIGVSVGDTTLGTTIVANLGMNDVRFANPVFIGDILRAQSTVLEMRESKSRPDAG			393	
		IFT GLM+GVSVGDTTLGTTIVANLGM V+F PV +GDTLRA + V+E+R+S+SRPDAG				
Sbjct 61		IFTFGLMLGVSVGDTTLGTTIVANLGMGVKFFPKPVHVGDTLRAATEVVELRDSRSPDAG			120	
Query 394		IVVFEHRCLNQRDEEVGYCKRSALMRRKA 480				
		+ VFEH+C NQR E V C+R ALM+R +				
Sbjct 121		VAVFEHKCFNQRGEMVAICRRFALMKRAS 149				

alpha amylase [Candidatus Koribacter versatilis Ellin345]

Sequence ID: [ref|YP_590578.1](#) Length: 610 Number of Matches: 1[▶ See 2 more title\(s\)](#)

Range 1: 1 to 609		GenPept	Graphics	▼ Next Match	▲ Previous Match	
Score	Expect	Method	Identities	Positives	Gaps	Frame
576 bits(1484)	0.0	Compositional matrix adjust.	295/615(48%)	392/615(63%)	7/615(1%)	+1
Query	28	CARLHRCALVFPLLLSLTAAQAAGPRVSKVEPPNWWMDFAPTVMFLLYGENLGGANVS				207
Sbjct	1	MCRLSR--LVAYLLFSCALFAQ---APKISKVDPNPNWWANYPHSPMLLLTGENLANAKVS				55
Query	208	VDYPNAVVAKVLQPDPGKHLFVWLSFYPGTRPGDVVIHVKTIPSGETSVPVPLLQRSPQEG				387
Sbjct	56	ANYPHLKITKSESSADGRYVVFVYLDEQKDLKPGTAHFSVQTAGGNTALDFVFDKRPSELE				115
Query	388	RFQGITRDDVIYLIMPDRFADGDPannmppgaapgTYDRSGAKTYHGGDLKGIQEHLPYL				567
Sbjct	116	RAQGLNASDTIYLIMPDRFADGDPNNP A YDR+ YHGGDLKG+ +HL YL				174
Query	568	KDLGVTAALWLTPLYDNDNSTSDYHGYGAVDEYAVEDHFGTMKSYQDLVAAAHQVGLKVML				747
Sbjct	175	HDLGVSTVWLT PWWKNDGNSADYHGYHVTDFYGIEDHFGNMKDLQQMVSAAHGKGMKVML				234
Query	748	DMVPHVGPKHPWATSQPAPDWLHGTTTEHLLDTDYYYAPVTDPHAVKANYVSALEGWFAD				927
Sbjct	235	DYVVNHTGPFHPWAEHPPTPTWLHGTPAKHPQPKYNFWPLVDPHGTQADRTPVLEGWFVD				294
Query	928	VLPDLAQENPLVALYLLQNAEWWTESGGVDGFRIDTFPYVPRSFQYHAGLFSHFPPNF				1107
Sbjct	295	RLPDLNVDDPKLTEYLIDNGLWMMETASLDGYRLDTPYSSREFWSKWHKALFEVYPRTF				354
Query	1108	TVGEIYNSDPTVTSYWAGGQTGFIDGIDHTLTPFDPFMNAAREVVVAHGASAKKIVDVL				1287
Sbjct	355	TIGEVSDGDPAVVSFFQGGKREYDGDIDSGVTTVDFDPTMYAIRDVLIRQQPASKLQEVLE				414
Query	1288	QDRLYPHPELLVTFIGNHDMKRELT DANGSQEKLKLAFLSLATLRGIPQLYYGDEIGMTG				1467
Sbjct	415	HDALYPNPAVLVPPFIGNHDKPRFMGEKATVPELNAAAASLLTLRGIPQLYAGDEIAMPG				474
Query	1468	GDDPDNRHDFPGGFGDQHNAFTQTGRTPDEQEIFAHVQTVLKLQEHAPALRRGAQKHIA				1647
Sbjct	475	GEDPDNRDRFPGGFAGDPQNAFTASGRTPQQEAF AHLQKLLQLRQKQKALQSGEQTDLE				534
Query	1648	VGDKYYAFTREGDGERLLIVLNNG-DAENITIDLSDTSIADAKTITPLFSASPAQLQGS				1824
Sbjct	535	SSEKGFAYYRVSGDDRVLIVLNSGSDAQTI AIPKVQTPANATSFTALDSAATAQTSGDS				594
Query	1825	LRLQLAHNSLTVYRV 1869				
Sbjct	595	VIANVPGMTVAIFQV 609				



Suplemento 5. Utilização de códons raros nas sequências C, D e F em *E.coli*.

Análise feita no programa *Rare Codon Caltor* (<http://people.mbi.ucla.edu/sumchan/caltor.html>).

C lipase

ATG CTG GCC GCC AGC GCC ACT TTT ACT GCC GCC CAG ACT CAG GCT GCA GGT TAC ACC CAG ACC CGT TAT CCG ATT GTA CTG GTA CAC GGT TTA TTC GGC TTC GAT AAC ATC **CGG** CCG GTG GAA TAT TTC TAC GGC ATC CCG TCT GCG CTG CGT TCC GGC GGC GCA **ACG** GTG TAC **ACG** CCG CAA GTG TCG GCA GCC AAC AGC ACC GAA GTG CCG GGT GAA CAG CTC TTA CTG GAA GTA AAA AAA ATC GTC GCG GTC ACC GGT AAA CCG AAA GTG AAT CTG ATC GGT CAC AGC CAC GGC **GGG** CCT ACC ATC CGT TAT GTG GCT TCG GTA CGT CCG GAT CTG GTT GCG TCG GCA ACT TCA GTG GCA GGC GTT AAC AAA GGT TCT GCG GTG GCG GAT ATT CTG CTG GGC ATC GCG CCG CCG GGC AGT TTG TCA CCG GAT GTG ATT **ACG** **ACG** ATT GCT ACC GGC CTC GGT AAA TTG CTG TCG CTG CTG TCT GGC AGT TCC ACA TTG CCA CAG AAT TCA CTG GCG GCC GCG CAA TCC TTG TCA **ACG** GCG GGT TCG GCA AAA TTC AAT GCT GCG CAT CCG GCC GGT TTA CCG AGC ACA GCT TGC GGC GAG GGC GCG TAT CAG GTC AAT GGT GTT GCC TAT TTC TCA TGG AGT GGC GCA TCG AAT TAC ACC AAT GTG CTG GAT ATT CTG GAT CCG GCG CTG GCA GTA ACC GGT CTG GCT TTC GGC GGC GCG AAA AAT GAT GGT CTG GTG TCT TCC TGT TCC AGC CAT CTG GGT AAG GTC ATT CCG GAT GAT TAT CCG ATG AAC CAT CCG GAT GAA ATC AAC CAG AGC GTG GGT ATT GTG AAT CTG TTT GAA GTG AAT CCG GTG TCG GTG TTC CCG CAG CAC GCC AAC CCG ATG GGT CTG CTC GAG

The Number of Bases in the above Sequence = 906

The Number of Codons in the above Sequence = 302

Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	0
	CGG	0
	AGG	0
	AGA	0
Glycine	GGA	0
	GGG	2
Isoleucine	AUA	0
Leucine	CUA	0
Proline	CCG	0
Threonine	ACG	5

Repeated and/or Consecutive Rare Codons

ACG **ACG** = 1

D proline aminopeptidase

ATG AAC **ACG** GTC AAA **ACG** GGC GGC GTT CAG ATG GTG TCC **AUA** GAC GGT **CGG** TTT CAA GTC TGG **ACG** AAA **CGA** **AUA** GGC GCC GGT CCG CCG **ACG** ATG CTG ACT CTG CAT GGC GGT **CGG** GGC TCC ACT CAC GAA TAC TTC GAA TGC TTC GAG GAT TTT CTG CCG **CGG** AAT **GGG** ATT CAG CTC ATT TAC TAT GAT CAG CTC **GGA** TCG GGC AAC TCC GAT CAG **CGG** GAT AAT CCG GCC CTT TGG GTG GTC GAG CCG TTT CGT GAC GAG GTA GAA CAG GTC **AGA** GCG GCG CTC GGC CTC ACC **GGA** TTC TAC TTA TAT GGC CAC TCC TGG GGC GGT ATG CTC GCA ATC GAA TAC CCG TTG AAA TAC CAA AGC CAT CTC AAG GGT CTG ATC ATT TCG AAC ATG ACT GCC AGC **AUA** GCG TCG TAT GTT ACT TAC GTT AAT GAG CTG **CGG** CCG CAA **CUA** **CGG** GCC GAG TCG CAG CCG ATT CTG GAA **CGA** TAC GAA GCG ACC **GGG** GAA TAC ACA GCA CCA GAA TAT GAG AAA GTC ATG TTC **GGA** GAA ATT TAT TCC **AGG** CAC TTA TGC CCG CTT GCT CCA TGG CCG GAA CCT TTG ATC **CGG** ATG ATC **CGG** CAT ATG AAC CAG AAG GTC TAC AAC AAG ATG CAG GGT **CGG** AAT GAA TTC GTG GTC ACC **GGA** ACC TTC AAA GAC TGG GAT **CGA** TGG GAG GAT ATC AAA AAT **AUA** AAC GTC **CGG** ACT CTG TTA TCA GTG GGC CGT TTC GAT **ACG** ATG AGC GTG GCC GAC GTC GAA **AGG** ATG GGC ACC CTG ATT **CGG** AAT CCG **CGA** GTA TCG ATT TGC GAG **ACG** GGC AGC CAT TGT TCA ATG TAC GAC GAT CAG GAG CCG TAT TTT GAA GAT CTG GTT CCG TTC ATC AAA GAT GTC GAG GCG **GGA** AAG CTC GTT TGA

The Number of Bases in the above Sequence = 903

The Number of Codons in the above Sequence = 301

Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	4
	CGG	4
	AGG	2
	AGA	1
Glycine	GGA	5
	GGG	2
Isoleucine	AUA	4
Leucine	CUA	1
Proline	CCG	7
Threonine	ACG	6

Repeated and/or Consecutive Rare Codons

CGA **AUA** = 1

F transcriptase reversa

ATG **ACC** GCA **AGA** GAC GCA GAA AGT ATG GCT GAG AAG **CCG** GAA TCC CAC TCG **GGA** GGT AGC GGT **CGG** AAA TCG **CGA** GAT
AAC GAG GCG GGT GCG TCA **AGG** GTC **ACC** GCA **AGG** **GGA** GAA GAC TCT AGC CCG GAG TGC GAG CAG TCG TCG TCG TCG TCG TCG TCG
GTC GAG **CGA** **GGG** AAC ATG CAG ACC GCG CTC CAG **CGA** GTG ATG AGC AAC **AGG** **GGA** GCA GCC **GGA** GCC GAC **GGG** ATG **ACC**
GTT GAT GAA CTG AAG CCG CAC TTG **AGG** GAG GAG TGG AAG **CGG** ATC AAA **GGA** GAA CTG CTG GCA **GGG** GAA TAC CAA CCG
GAG CCA GTG CTG AAG GTA GAG **AUA** CCG AAG GCA GAG **GGG** AAA **GGG** GTG **CGA** AAG CTT GGC ATC CCG **ACC** GTG GTG GAC
CGG CTG ATC CAG CAG GCG TTG CAT CAA GTA **CUA** AGC CCG ATC TTT GAG CCA **GGA** TTT TCG GAA TCG AGC TAT GGC TTT
CGG CCA GGC **GGG** AGC GCA CAG GAT GCG GTG **CGG** CAA GCA **CGG** GCA TAT GTG GGT GAA **GGG** **CGG** **CGG** TGG GTC GTA
GAT ATC GAC TTG GAG AAA TTC TTT GAC **CGA** GTT AAT CAC GAC AAA ATG ATG TCG **GGG** **CUA** GCG **AGG** **CGG** ATC AAG GAT
AAG **CGG** **AUA** CTG **CGG** ATG ATC **CGA** **AGG** TAC **CUA** CAG GCT **GGA** ATG ATG GAA GGC **GGG** CTG GTG ACA CAG **AGG** **AGA** GAG
GGG **ACC** CCG CAG GGC **GGG** CCG **CUA** TCG CCG CTG TTG TCG AAC ATT CTG CTG GAT GAG CTG GAC AAG GAA CTG GAG **AGA**
CGA **GGG** CAC AAG TTC TGC **CGG** TAC GCC GAC GAT TGC AAT GTG TAT GTG **CGG** AGT **CGA** AGT GCC **GGG** GAG **CGA** GTG AAG
GAG TCG ATC ACA **AGG** TTT CTT GAA **AGG** **CGA** CTG **CGG** CTG AAG GTA AAC GAG GAG AAG AGC GCA GTA GAG **CGG** CCG TGG
AAG **AGG** AAG TTT CTG GGC TAC ACA ATG **ACC** TGG CAC CTG GAA CCG **CGA** **AUA** AAG GTA GCG GAG AAC TCG GTG AAA **CGA**
CUA AAG GTG AAG CTG **CGG** GAG ATT CTG **CGG** CAG GGC **CGA** **GGA** CCG AAC ATT **GGG** **AGA** **CUA** ATC GAG GAA GAA **CUA** ACA
CCG CTG CTG **AGA** GGC TGG ATG AAC TAT TTC **CGG** CTG GCG GAG GTG AAA **GGA** ATC TTC GAG GAG TTA GAT AGC TGG **AUA**
CGG **CGG** AAG TTG **AGG** TGT GTA ATC TGG **CGG** CAA TGG AAG CCG ACC TGG GCG **CGG** GTA AAG **GGA** CTG ATG AAG CGT GGT
TTG GAG **AGG** GAT **CGG** GCG CTG AAA TCA GCG ACT AAT **GGG** **CGA** **GGG** CCA TGG TGG AAC GCT **GGG** GCC TCG CAC ATG CAC
GAG GCC TTT **CCG** AAG ACA TAC TTT GAT GCG TGC GGT TTG GTG TCG CTG **CUA** GAT CAA **CGG** CTC AAA GTC CAG CGT ACT
TCA TGA

The Number of Bases in the above Sequence = 1407

The Number of Codons in the above Sequence = 469

Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	14
	CGG	26
	AGG	12
	AGA	5
Glycine	GGA	10
	GGG	15
Isoleucine	AUA	4
Leucine	CUA	8
Proline	CCG	2
Threonine	ACC	6

Repeated and/or Consecutive Rare Codons

AGG **GGA** = 2

CGA **GGG** = 1

GGG **CGG** **CGG** = 1

CGG **CUA** = 1

AGG **CGG** = 1

CGG **AUA** = 1

CGA **AGG** = 1

AGG **AGA** = 1

GGG **ACC** = 1

AGA **CGA** **GGG** = 1

AGG **CGA** = 1

CGA **AUA** = 1

CGA **CUA** = 1

CGA **GGA** = 1

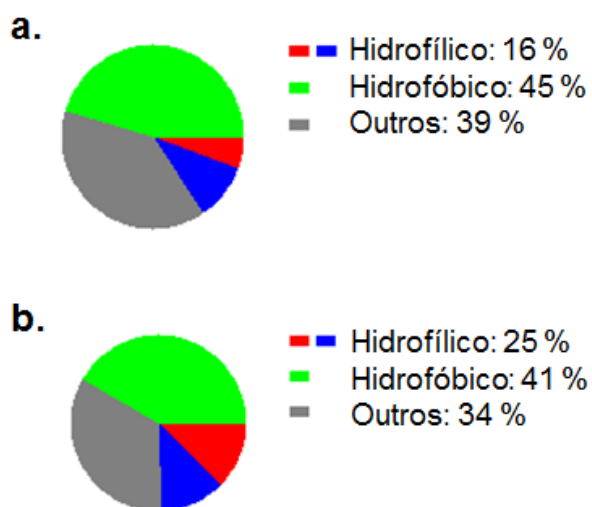
GGG **AGA** **CUA** = 1

AUA **CGG** **CGG** = 1

GGG **CGA** **GGG** = 1

Suplemento 6. Análise de hidrofobicidade das sequências peptídicas de C (lipase) e D (prolina aminopeptidase).

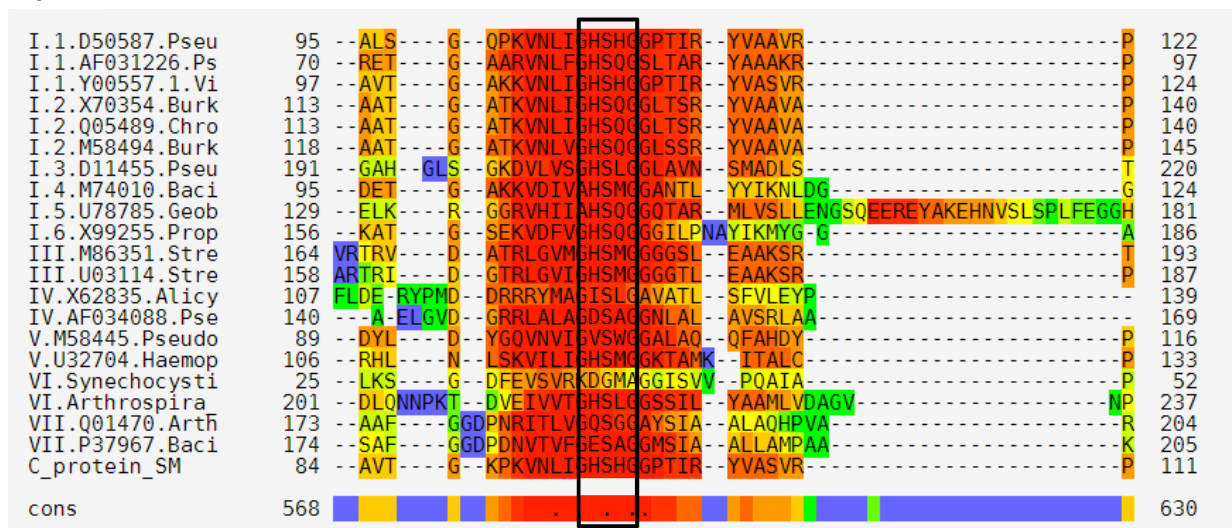
Composição dos peptídeos segundo os tipos de aminoácidos. Análise feita utilizando o programa *Peptide Property Calculator* (https://www.genscript.com/ssl-bin/site2/peptide_calculation.cgi). Foram analisadas a sequências peptídicas incluindo a cauda de poli-histidinas. **a.** sequências peptídicas de C (lipase), **b.** de D (prolina aminopeptidase).



Suplemento 7. Análise de domínios conservados das proteínas C (lipase) e D (prolina aminopeptidase).

Painel a. Alinhamento múltiplo realizado em *T-cofee* (Di Tommaso *et al.*, 2011; Notredame *et al.*, 2000) da sequência da proteína C (lipase) com as sequências representativas das famílias de lipases de acordo com Arpigny e Jaeger (1999). Está mostrado no quadro o pentapeptídeo conservado G-X-S-X-G. **Painel b:** Análise realizado em *Prosite* (Sigrist *et al.*, 2013) na sequência peptídica C assinalando o aminoácido S (serina) da posição 97, como parte do sítio ativo que pode ligar a um íon metálico, sendo este aminoácido parte da tríada catalítica presente em todas as famílias de lipases. **Painel c:** Alinhamento múltiplo realizado em *T-cofee* (Di Tommaso *et al.*, 2011) da sequência da proteína D (prolina aminopeptidase) com sequências representativas de peptidases envolvendo prolina no sítio de clivagem. Está mostrado no quadro o pentapeptídeo conservado G-X-S-X-G. **Painel d:** Análise de família de proteínas pelo *Pfam* (Punta *et al.*, 2012) da sequência da proteína D, identificando a família de α/β hidrolases.

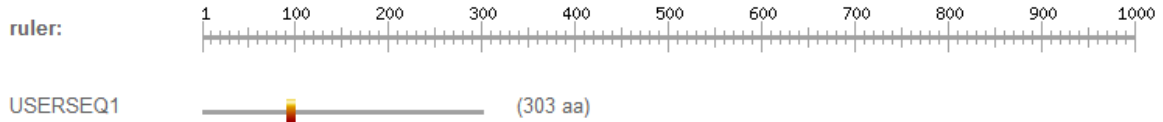
a.



b.

hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by PS00120 LIPASE_SER Lipases, serine active site :



91 - 100: VNLIGHSHGG

Predicted feature:

ACT_SITE 97 Charge relay system (By similarity) [condition: none]

c.

BAK03239.1-Hord	159	DLVADIEKLRQHL	-----DI	PEWQVFGGSWG	STLALAYSQTHPKVTGIVLRGIFLL	RK	212
ACF78312.1-Zea	138	DLVADIEKLREHL	-----GI	PEWQVFGGSWG	STLALAYSQEHDPKVTGLVLRGIFLL	RK	191
EEF40239.1-Ric	163	DLISDIEKLREYL	-----QI	PEWQVFGGSWG	STLALAYSQAHPNKVTGLVLRGIFLL	RK	216
CBI25074.3-Viti	156	DLVNDIEKLREHL	-----EI	PEWQVFGGSWG	STLALAYSQSHDPKVTGMVLRGIFLL	RK	209
NP_179037.2-Ara	149	DLVNDIEKLREHL	-----KI	PEWLVFGGSWG	STLALAYSQSHDPKVTGLVLRGIFLL	RK	202
ACS88344.1-Phan	84	DLVKDIEKIREHL	-----EV	EKWVFGGSWG	STLSLAYAQSYPERVKSLVLRGIFLL	RK	137
NP_641261.1-Xan	87	DLVADIERLRTHL	-----GV	DRWQVFGGSWG	STLALAYAQTHPQQVTELVRGIFLL	RR	140
BAA23336.1-Serr	90	HLVADIERLREMA	-----GV	EOWLFGGSWG	STLALAYAQTHPERVSEMVLRGIFLL	RK	143
CAC40647.3-Aspe	136	NIVRDCEAVRRCL	MTDYPEDK	RKWSITGQSF	GFCAVTYLSIPEGLAEAFICGGLPPL	VD	195
AF439997.1-Rasa	134	NIVRDCEAVRKCL	TADYPEEN	QKWSVLGQSF	GFCAVTYLSKFPEGLREVFTTGGLPPL	VN	194
NP_191713.1-Ara	189	NIVKDAEFIRVRL	VP-----KA	DPWTILGQSF	GFCALTYLSFAPEGLKQVLLITGGIPPI	GKA	246
XP_002511924.1R	187	NIVNDAEFIRVRL	VP-----DA	EPWTILGQSY	GFCAVTYLSFAPHGLKQVLLITGGIPPI	ISNG	244
AEV23270.2-Trit	159	DLVADIEKLRQHL	-----DI	PEWQVFGGSWG	STLALAYSQTHPKVTGIVLRGIFLL	RK	212
AAW89640.1-Neis	84	DLVADIEKVREML	-----GI	GKWLFGGSWG	STLSLAYAQTHPERVKGVLVLRGIFLL	RP	137
BAA06380.1-Aero	122	SIVRDAELIREQL	-----SP	HPWVSLGQSF	GFCSLTYLSLFPDSLHEVYLTGGVAP	IGR	177
D_prot_seq	87	RFRDEVEQVRAAL	-----GL	TGFYLVGHSWG	MLAIEYALKYQSHLKGGLIISNMTAS	LA	140
cons	253	: : * : *		: : * : *	: : *		315

d.



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Abhydrolase_6	Alpha/beta hydrolase family	Domain	CL0028	55	309	56	306	2	225	228	105.8	3e-30	130,296,269	Show

Suplemento 8. Análise de identificação de proteínas chaperonas de lipase.

Análise realizado em *String* (Franceschini et al., 2013) após submissão da sequência da proteína C (lipase). **Painel a.** Análise da vizinhança de genes associados a lipases em diferentes grupos taxonômicos relacionados. **Painel b.** Análise evidenciando maior associação da proteína C (lipase) com a sua chaperona de lipase do que com outras proteínas.

