UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE CIÊNCIAS DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

ANA CAMILA MENDES ANDRADE

METAGENÔMICA COMPARATIVA DE AMOSTRAS DE SOLO E DE ÁGUA DO BIOMA CAATINGA PARA BIOPROSPECÇÃO DE ENZIMAS RELACIONADAS AO METABOLISMO DE CARBOIDRATOS (CAZYMES)

Salvador
2015

ANA CAMILA MENDES ANDRADE

METAGENÔMICA COMPARATIVA DE AMOSTRAS DE SOLO E DE ÁGUA DO BIOMA CAATINGA PARA BIOPROSPECÇÃO DE ENZIMAS RELACIONADAS AO METABOLISMO DE CARBOIDRATOS (CAZYMES)

Dissertação apresentada ao curso de Pós-Graduação em Biotecnologia, Instituto de Ciências da Saúde, Universidade Federal da Bahia, como requisito para obtenção do título de Mestre em Biotecnologia.

Orientador: Prof. Dr. Thiago Bruce Rodrigues

Salvador
2015

A543 Andrade, Ana Camila Mendes

Metagenômica comparativa de amostras de solo e de água do bioma Caatinga para bioprospecção de enzimas relacionadas ao metabolismo de carboidratos (CAZymes)/ Ana Camila Mendes Andrade. – Salvador, 2015.

109 f.

Orientador: Prof. Dr. Thiago Bruce Rodrigues

Dissertação (Mestrado) – Universidade Federal da Bahia. Instituto de Ciências da Saúde, 2015.

1. Enzimas. 2. Caatinga. 3. Biocombustíveis. 4. CAZymes. I. Rodrigues, Thiago Bruce. II. Universidade Federal da Bahia. III. Título.

CDU 577.15

ANA CAMILA MENDES ANDRADE

Metagenômica comparativa de amostras do solo e de água do bioma caatinga para bioprospecção de enzimas relacionadas ao metabolismo de carboidratos

Dissertação apresentada como requisito para obtenção do grau de Mestre em Biotecnologia pelo Instituto de Ciências da Saúde da Universidade Federal da Bahia.

Aprovada em 24 de abril de 2015.

BANCA EXAMINADORA:

Thiago Bruce Rodrigues – Orientador _____
Doutor em Ciências Biológicas (Genética) pela Universidade Federal do Rio de Janeiro, UFRJ, Brasil.
Faculdade de Tecnologia e Ciências

Milton Ricardo de Abreu Roque – _____
Doutor em Ciências Biológicas pela Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil.
Universidade Federal da Bahia.

Alinne Pereira de Castro _____
Doutora em Biologia Molecular pela Universidade de Brasília, UNB, Brasil
Universidade Católica Dom Bosco

AGRADECIMENTOS

Aos meus pais, Ana Lúcia e Moacy, e meus irmãos, Ana Carolina e Rodrigo, por todo apoio e confiança sem os quais eu não teria conseguido. São minhas principais fontes de admiração e inspiração para dar sempre o melhor de mim. Cada um de vocês contribuiu à sua maneira para o que sou hoje e qualquer conquista minha é também uma conquista de vocês.

A Alexandre, por ter sido um grande companheiro em todas as etapas durante o mestrado e também fora dele, por ter estado ao meu lado nos momentos de entusiasmo e por me dar forças nos momentos difíceis.

Às minhas amigas "gêmeas" Luiza e Carol, e aos demais amigos biólogos por compartilharem incertezas, alegrias, por dividirem tantos momentos e se tornarem essenciais na minha vida.

À minha prima-irmã Dandara, por estar sempre tão perto, mesmo que de longe, me aconselhando, acalmando ou simplesmente ouvindo. Faltam palavras pra descrever o quanto é bom poder contar com você.

Aos meus colegas de laboratório e aos amigos que fiz durante o mestrado, pela convivência e por terem tornado tudo mais leve e proveitoso.

A Thiago Bruce, meu orientador, por ter sido sempre muito solícito e presente durante todo o desenvolvimento do trabalho. Obrigada por todas as palavras de incentivo, pelo conhecimento compartilhado, pelo voto de confiança e por ter possibilitado o engrandecimento da minha formação científica.

Ao Me. Fabyano Alvares e à Dr. Adriana Fróes pela imensa colaboração na análise dos dados, na solução de problemas e de todas as dúvidas que tive durante este tempo.

Ao Dr. Milton Roque e à Dra. Alinne Castro por terem aceitado participar da banca de avaliação do presente trabalho. Obrigada pela dedicação de tempo, esforços e pela contribuição oferecida.

Ao Programa de Pós-Graduação em Biotecnologia, pela oportunidade de desenvolver o projeto, pelo aprendizado durante todo o mestrado e pelo auxílio financeiro, na realização do treinamento no Rio de Janeiro, o qual foi de fundamental importância para a realização deste trabalho.

À FAPESB, pelo apoio financeiro no concedimento da bolsa de mestrado.

Ao Centro de Genômica de Alto Desempenho do DF, pelo sequenciamento das amostras.

À CEPLAC, pela colaboração nas análises físico-químicas das amostras de solo.

ANDRADE, Ana Camila Mendes. Metagenômica comparativa de amostras de solo e de água do bioma Caatinga para bioprospecção de enzimas relacionadas ao metabolismo de carboidratos (CAZymes). 109 f. 2015. Dissertação (Mestrado) – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2015.

RESUMO

A Caatinga é a única região natural exclusivamente brasileira, sendo, no entanto, a área menos conhecida dentre os demais biomas. Pouco se sabe sobre a diversidade microbiana da Caatinga e menos ainda sobre o potencial biotecnológico desta região, no que diz respeito, por exemplo, à bioprospecção enzimática. Um dos principais grupos de enzimas de interesse biotecnológico são as hidrolases, que catalisam a hidrólise de ligações covalentes da matéria orgânica e por isso podem ser aplicadas na conversão da biomassa vegetal, para a produção de biocombustíveis. Apesar das hidrolases representarem as principais enzimas com aplicações biotecnológicas para esse fim, outros grupos de enzimas envolvidas no metabolismo de carboidratos (CAZymes) também detém um papel importante neste processo. O presente trabalho se propõe a utilizar a abordagem metagenômica para analisar amostras de água do rio Paraguaçu e amostras de solo de uma localidade da Chapada Diamantina, quanto à presença de enzimas potencialmente aplicáveis na bioconversão de biomassa vegetal. O DNA metagenômico extraído das amostras foi sequenciado pelo método *shotgun* e foram realizadas duas estratégias de anotação: a anotação pela tecnologia de subsistemas e a anotação baseada em regiões conservadas das sequências de CAZymes. Observou-se que o solo e a água apresentaram diferenças nos seus perfis taxonômicos e na distribuição dos subsistemas e das famílias de CAZymes que predominaram em cada ambiente. O subsistema de carboidratos foi o mais abundante no solo e o segundo com maior contribuição na água. Os subsistemas clustering-based e de aminoácidos e derivados também estiveram dentre os mais representativos nos dois ambientes. Em relação às classes de CAZymes, as glicosil hidrolases foram dominantes no solo (~44%) enquanto que as glicosil transferases foram mais frequentes na água (~50%). Em relação aos principais táxons associados às CAZymes, a classe Planctomycetia apresentou contribuição de 29% nas amostras de solo e Alphaproteobacteria contribuiu com 27% nas amostras de água. O mesmo não aconteceu ao analisar a estrutura da comunidade microbiana total, na qual Actinobacteria foi a classe dominante no solo e Betaproteobacteria na água. Os resultados encontrados indicam o potencial biotecnológico da Caatinga. Determinados grupos de enzimas identificados no solo e na água podem desempenhar atividades na degradação de substratos de interesse industrial, como o amido, o xilano, a lignina e outros compostos lignocelulósicos, tornando este bioma uma interessante fonte para bioprospecção.

Palavras-chave: Abordagem metagenômica. Caatinga. Subsistemas. CAZymes. Biocombustíveis.

ANDRADE, Ana Camila Mendes. Comparative metagenomics of soil and freshwater samples from the Caatinga biome for bioprospecting of carbohydrate-active enzymes (CAZymes). 109 f. 2015. Master Dissertation – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2015.

ABSTRACT

The Caatinga biome is the only natural area exclusively Brazilian, however, it is the area with the lowest number of scientific studies among other Brazilian biomes. The available knowledge about microbial diversity of Caatinga is very limited and even less is known about the biotechnological potential of this region, with regards, for example, to the enzyme bioprospecting. One of the major enzyme groups of interest for biotechnological purposes are hydrolases, which catalyze the hydrolysis of covalent bonds existent in organic matter and, therefore, can be applied in the conversion of plant biomass to be used in biofuels production. Despite the fact that hydrolases are the main enzyme group with biotechnological application for this purpose, other groups of enzymes involved in carbohydrate metabolism (CAZymes) also have an important role in this process. The present study aims to use the metagenomic approach in order to analyze freshwater samples from Paraguaçu river and soil samples from one location of Chapada Diamantina, seeking to detect the presence of potentially applicable enzymes in bioconversion of plant biomass. The metagenomic DNA extracted from samples was sequenced through the shotgun sequencing method and two annotation strategies were performed: the annotation through subsystems technology and the annotation based on conserved domains of CAZyme sequences. It was observed that the soil and freshwater presented differences on their taxonomic profiles and in relation to the subsystems and CAZymes families prevailing in each environment. The Carbohydrates subsystem was the most abundant in soil and had great contribution in freshwater samples. Other subsystems such as Amino acids and Derivatives also had greater contribution in both sites. Regarding CAZymes classes, glycoside hydrolases were dominant in soil (~44%) and glycosiltransferases in freshwater (~50%). In relation to the main taxons associated with CAZymes, the Planctomycetia class had 29% contribution in soil samples and Alphaproteobacteria contributed with 27% in freshwater samples. Nevertheless, the same scenario was not observed when the structure of microbial community was analysed as a whole, in which Actinobacteria was the ruling class in soil and Betaproteobacteria in freshwater. This results indicate the biotechnological potential of Caatinga, since certain groups of enzymes found both in soil and freshwater samples may have activities in degrading substrates of industrial interest such as starch, xylan, lignin and other lignocellulosic compounds.

Keywords: Metagenomic approach. Caatinga. Subsystems. CAZymes. Biofuels.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AA – Atividades auxiliares
ANOVA – Análise de Variância
CAZymes – *Carbohydrate-Active Enzymes*
CBM – Módulos de ligação a carboidratos
CE – Carboidrato esterases
CEPLAC – Comissão Executiva do Plano da Lavoura Cacaueira
CONAMA – Conselho Nacional do Meio Ambiente
DbCAN - *DataBase for automated Carbohydrate-active enzyme ANnotation*
DNA – Ácido desoxirribonucléico
EDTA – Ácido etilenodiamino tetra-acético
FDR - *False Discovery Rate*
FIG-fam - *Fellowship Interpretation of Genomes Family*
GH – Glicosil hidrolases
GOS – Coleta Global dos Oceanos
GT – Glicosil transferases
HMM - *Hidden Markov Models*
ICMBio - Instituto Chico Mendes de Conservação Ambiental
ID – Identificador
LPMO - Monooxigenase lítica de polissacarídeo
LPS – Lipopolissacarídeo
MEGAN – *MetaGenome Analyzer*
MG-RAST – *MetaGenomic Rapid Annotation by Subsystem Technology*
MMA – Ministério do Meio Ambiente
NGS – Sequenciamento de nova geração
ORF - *Open reading frame*
PCR – Reação em cadeia da polimerase
pH – Potencial hidrogeniônico
PL – Polissacarídeo liases
PNCD – Parque Nacional da Chapada Diamantina
RAST – *MetaGenomic Rapid Annotation by Subsystem Technology*
RNA – Ácido Ribonucléico
rRNA – RNA ribossômico

SDS – Dodecil sulfato de sódio

STAMP – *Statistical Analyses of Metagenomic Profiles*

TAE - Tris-Acetato-EDTA

UV – Ultravioleta

SUMÁRIO

1. INTRODUÇÃO

A Caatinga é a única região natural exclusivamente brasileira (LEAL; TABARELLI, SILVA, 2003). O bioma representa a formação vegetal dominante na região semiárida do país, sendo o bioma semiárido com a maior biodiversade conhecida (COSTA; ARAÚJO, 2012; MINISTÉRIO DO MEIO AMBIENTE). Ainda assim, a Caatinga é o bioma brasileiro menos estudado (MINISTÉRIO DO MEIO AMBIENTE) e uma das lacunas está justamente no estudo da biodiversidade microbiana da Caatinga quanto ao seu potencial biotecnológico.

A necessidade de se propor soluções sustentáveis para diversos processos produtivos tem levado a um aumento na demanda por enzimas industriais (ADRIO; DEMAIN, 2014). Mais de 500 produtos abrangendo 50 aplicações utilizam enzimas produzidas por microorganismos em larga escala (CHERRY; FIDANTSEF, 2003).

Um dos principais grupos de enzimas de atual interesse para a indústria são enzimas de ação hidrolítica, mais conhecidas como hidrolases (KIRK; BORCHERT; FUGLSANG, 2002). Estas enzimas realizam a hidrólise de ligações covalentes da matéria orgânica e por isso apresentam importantes aplicações nos mais diferentes processos como, por exemplo, na conversão da biomassa vegetal para a produção de biocombustíveis (PARISUTHAM; KIM; LEE, 2014). Tal conversão requer a atuação de grupos de enzimas que sejam capazes de transformar eficientemente a biomassa vegetal em açúcares fermentáveis, possibilitando assim a produção de bioetanol. As celulases, por exemplo, são um grupo de enzimas que pertencem à classe das glicosil hidrolases e que, ao atuarem em conjunto, hidrolisam a celulose, que é o principal componente da biomassa vegetal e é um biopolímero de grande abundância na natureza (SADHU; MAITI, 2013). Muitos organismos, incluindo animais, plantas e microorganismos produzem celulases, entretanto, a maioria das que são conhecidas provém de origem microbiana (DUAN; FENG, 2010). Os sistemas especializados de decomposição dos polissacarídeos da parede celular vegetal incluem, além das celulases, hemicelulases e outras glicosil hidrolases relacionadas, assim como polissacarídeo liases e carboidrato esterases (HIMMEL et al., 2010). Estas classes de enzimas - glicosil hidrolases (GH), polissacarídeo liases (PL) e carboidrato esterases (CE) - juntamente às glicosil transferases (GT), atividades auxiliares (AA) e módulos de ligação a carboidratos (CBM) compõem as denominadas CAZymes (LOMBARD et al.,

2014). O termo CAZymes refere-se ao espectro de atividades catalíticas envolvidas na biotransformação de açúcares e seus derivados, para assegurar tanto a montagem de monossacarídeos em oligo e polissacarídeos quanto a clivagem de diversos tipos de açúcares, permitindo, ainda, a conjugação destes a compostos como proteínas e lipídeos, entre outros (ANDRÉ et al., 2014). As CAZymes representam portanto um vasto repertório de enzimas com alta relevância biotecnológica, em função, entre outras coisas, da sua potencial aplicação na produção de bioenergia a partir da biomassa vegetal residual (HORN et al., 2012).

Apesar de um grande número de microorganismos habitarem a biosfera, mais de 99% não são passíveis de serem cultivados através das técnicas de laboratório existentes (LI et al., 2015). No entanto, avanços recentes em genômica, metagenômica, técnicas emergentes de DNA recombinante, entre outras, têm facilitado a descoberta de novas enzimas microbianas na natureza, através, por exemplo, dos metagenomas (ADRIO; DEMAIN, 2014).

A metagenômica é o estudo de sequências genômicas obtidas diretamente do meio ambiente e que contorna a necessidade de isolar e cultivar os microorganismos (HANDELSMAN, 2004; WOOD; SALZBERG, 2014). O sequenciamento dos metagenomas permite explorar microorganismos recalcitrantes ao cultivo que habitam os mais diversos locais (PRAKASH; TAYLOR, 2012). A análise dos metagenomas pode ser feita através de ferramentas de bioinformática que permitem investigar uma variedade de aspectos das comunidades microbianas estudadas. A abordagem metagenômica permite responder às perguntas “Quem esta aí?” e “O que podem fazer?”, proporcionando elucidações sobre a história evolutiva dos microorganismos e o potencial metabólico da comunidade microbiana (SESHADRI et al., 2007). A metagenômica, além de permitir estudar as relações entre os microorganismos e os habitats nos quais eles vivem, pode também fornecer informações genéticas sobre novos biocatalisadores ou enzimas potenciais (THOMAS; GILBERT; MEYER, 2012; WOOLEY et al., 2010). A enorme quantidade e diversidade de dados genômicos gerados através da abordagem metagenômica representa uma importante fonte para prospecção de novos genes de interesse biotecnológico. A metagenômica torna acessível o potencial biotecnológico das bactérias recalcitrantes ao cultivo, permitindo a prospecção de genes com as mais diversas funções e o fornecimento único de moléculas bioativas para aplicação industrial (LORENZ; ECK, 2005).

O presente trabalho representa a primeira caracterização comparativa do potencial funcional da microbiota do solo e da água doce de uma região da Caatinga baiana. Neste sentido, o trabalho se propõe a utilizar a metagenômica como ferramenta, no intuito de bioprospectar grupos de enzimas relacionados à bioconversão lignocelulósica, com potencial aplicação da cadeia produtiva de bioetanol.

2. REVISÃO DA LITERATURA

2.1 O bioma Caatinga

A Caatinga ocupa uma área de aproximadamente 844.453km², representando cerca de 11% do território nacional, sendo o principal bioma da região nordeste e o bioma semiárido mais biodiverso do mundo (Figura 1) (MINISTÉRIO DO MEIO AMBIENTE). O bioma Caatinga é caracterizado por apresentar uma vegetação tropical seca e decídua, composta por pequenas árvores, arbustos, gramíneas e por deter altos níveis de insolação, altas temperaturas e recursos hídricos escassos, o que ocasiona grandes períodos de seca ao longo do ano (GORLACH-LIRA; COUTINHO, 2007).

A Caatinga representa um bioma exclusivamente brasileiro e, por isso, seu patrimônio biológico é único e de extrema importância. Segundo o Ministério do Meio Ambiente (MMA), o bioma detém um imenso potencial para uso sustentável e bioprospecção, pois a biodiversidade da Caatinga é capaz de amparar diversas atividades econômicas, incluindo as industriais.

No entanto, ainda que detenha um patrimônio biológico único e seja uma importante fonte de recursos naturais, a Caatinga é o bioma proporcionalmente menos estudado dentre as regiões naturais brasileiras, já que grande parte dos estudos científicos se concentra em torno de algumas das principais cidades da região (LEAL; TABARELLI; SILVA, 2003). Informações sobre a densidade e diversidade de microorganismos do solo são escassas na Caatinga (GORLACH-LIRA; COUTINHO, 2007) e ainda são incipientes os trabalhos neste bioma brasileiro com foco na prospecção de enzimas microbianas com aplicação industrial.

2.1.1 O Parque Nacional da Chapada Diamantina (PNCD)

O Parque Nacional da Chapada Diamantina encontra-se inserido na ecorregião do Complexo da Chapada Diamantina, dentro dos limites do bioma da Caatinga brasileira. O parque é uma Unidade de Conservação de cerca de 152.000ha, tendo sido criado pelo decreto nº 91.655 de 17 de setembro de 1985, estando localizado na região central do estado da Bahia (VELLOSO et al., 2002).

Figura 1. Cobertura do bioma Caatinga no Brasil. Fonte: www.cerratinga.org.br/caatinga.

A criação do parque tem como objetivo proteger amostras dos ecossistemas da Serra do Sincorá, na Chapada Diamantina, assegurando a preservação de seus recursos naturais e proporcionando, entre outras coisas, oportunidades para pesquisa científica. A preservação desses ecossistemas propicia a manutenção de um banco genético de grande importância tanto para a pesquisa, quanto para a manutenção da biodiversidade brasileira (GONÇALVES et al., 2005).

Assim como para o bioma Caatinga em sua totalidade, ainda são escassas as pesquisas científicas conduzidas dentro dos limites do PNCD. A maioria dos estudos realizados abarca aspectos da flora (BASTOS; STRADMANN; VILAS BÔAS-BASTOS, 1998; RIBEIRO-FILHO; FUNCH; RODAL, 2009) e fauna (PEREIRA; GEISE, 2009). Pesquisas sobre a biodiversidade microbiana e a bioprospecção de enzimas produzidas por esses microorganismos são realizadas de forma ainda mais limitada.

2.2 Abordagens de estudo dos microorganismos

Estima-se que existam cerca de 4–6 × 10^{30} células procarióticas habitando a terra (WHITMAN; COLEMAN; WIEBE, 1998). Tais procariotos comumente organizam-se em comunidades microbianas complexas, ocupando os mais diversos nichos existentes e desempenhando funções cruciais numa variedade de processos biológicos essenciais à manutenção dos ecossistemas (PRAKASH; TAYLOR, 2012).

As principais formas de se ter acesso a esta imensa diversidade de microorganismos são através de duas diferentes estratégias de estudo, as quais serão apresentadas a seguir: a abordagem dependente e a abordagem independente de cultivo.

2.2.1 Abordagem dependente de cultivo

No sentido de se estudar os mais diferentes processos bacterianos, o isolamento em culturas puras é, tradicionalmente, a etapa inicial a ser realizada (RIESENFELD; SCHLOSS; HANDELSMAN, 2004). Tais técnicas da microbiologia, baseadas na análise de culturas puras crescidas em laboratório, se reúnem na denominada *abordagem dependente de cultivo*. Esta abordagem permite a identificação os microorganismos obtidos através do cultivo – e, com isso, o conhecimento da sua diversidade – assim como permite triar determinadas características fisiológicas e bioquímicas de interesse (AMANN; LUDWIG; SCHLEIFER, 1995).

A precisão de tais estudos e a carência de outras metodologias disponíveis fizeram com que, durante muito tempo, grande parte do conhecimento gerado fosse oriundo de microorganismos estudados a partir de suas culturas puras crescidas em laboratório. De fato, a diversidade de microorganismos do solo foi por muitos anos explorada através de técnicas de cultivo e isolamento das espécies (DANIEL, 2004). Como resultado, muitos produtos naturais de valor econômico (como por exemplo, antibióticos) foram derivados de microorganismos do solo cultivados (DANIEL, 2004; STROHL, 2000).

Os maiores desafios que envolvem o cultivo dos microorganismos em laboratório estão relacionados à identificação e ao fornecimento, na sua correta concentração, dos nutrientes necessários para promover o crescimento microbiano (ALAIN; QUERELLOU, 2009). Avanços recentes nas estratégias de cultivo têm levado à superação destes e de outros desafios associados à obtenção das culturas, uma vez que as novas ferramentas têm proporcionado o melhoramento das condições que permitem o estabelecimento dos microorganismos.

Novas abordagens de cultivo estão sumarizadas nas revisões de Alain e Querellou (2009) e de Pham e Kim (2012), indicando os diversos estudos que embasam o desenvolvimento de cada uma das técnicas. Estas novas estratégias incluem: o refinamento dos meios e das condições de cultivo, através da diversificação da formulação dos meios, modificações no período de incubação e no tamanho do inóculo; o cultivo *in situ*, através da simulação das condições naturais em biorreatores ou em câmaras de difusão; o cultivo de alto rendimento pela microencapsulação combinada com citometria de fluxo ou pela utilização de *chips* de culturas microbianas e robotização; e a cultura baseada nas interações célula-célula, através da adição de compostos sinalizadores no meio, pelo cultivo de comunidades inteiras (a partir de um *mix* de bactérias que requerem cooperação) ou pelo cultivo em biofilmes.

No entanto, apesar dos métodos dependentes de cultivo permitirem a análise e a caracterização da diversidade microbiana em determinada amostra, e de ter havido um grande avanço nas estratégias de cultivo, existe um grande gargalo na diversidade vista através desta abordagem, em relação à diversidade microbiana real no ambiente. Estima-se que as técnicas de cultivo tradicionais sejam capazes de representar apenas 1% ou menos da diversidade bacteriana na maioria das amostras de ambientes naturais, ou seja, a grande maioria das bactérias presentes na natureza (≥ 99%), ainda não foi cultivada (AMANN; LUDWIG; SCHLEIFER, 1995; LI et al., 2015; RIESENFELD; SCHLOSS; HANDELSMAN, 2004; TORSVIK; ØVREÅS, 2002). Este fato pode se tornar particularmente crítico em se tratando de microorganismos em amostras de solo, por exemplo. Isso ocorre devido à estimativa de que 1g de solo pode conter mais de 10 bilhões de microorganismos, podendo pertencer a milhares de espécies diferentes (MYROLD; ZEGLIN; JANSSON, 2013; ROSSELLÓ-MORA; AMANN, 2001).

2.2.2 Abordagem independente de cultivo

Para contornar estas e outras limitações associadas à abordagem dependente de cultivo, diferentes técnicas de estudo das comunidades microbianas que independem do cultivo foram sendo desenvolvidas. As análises independentes de cultivo permitem o acesso ao genoma de toda a comunidade microbiana de uma amostra, possibilitando a análise de dezenas de espécies de maneira simultânea e não apenas o estudo de algumas espécies individualmente,

como a abordagem dependente de cultivo propõe (HANDELSMAN, 2004; WOOLEY; GODZIK; FRIEDBERG, 2010).

2.2.2.1 A metagenômica

Dentre as formas de abordagem independentes de cultivo, que permitem o acesso à informação genética de comunidades microbianas, está a metagenômica. O termo "metagenômica" foi pela primeira vez utilizado por Handelsman e colaboradores (1998), e refere-se à abordagem independente de cultivo baseada na investigação das moléculas de DNA de uma mistura de populações microbianas, ou seja, é baseado na análise genômica de DNA microbiano extraído diretamente do ambiente (HANDELSMAN et al., 1998). A metagenômica representa um conjunto de técnicas que incluem abordagens e métodos relacionados à genômica, como o sequenciamento e a clonagem. Tais técnicas proporcionam uma visão menos enviesada do que a oferecida pelas abordagens dependentes de cultivo, não somente em relação à estrutura da comunidade (riqueza e distribuição de espécies) como também sobre o potencial funcional (metabólico) desta comunidade (WOOLEY; GODZIK; FRIEDBERG, 2010). Sendo assim, a metagenômica se caracteriza por contornar a necessidade de cultivo e, em função da vasta diversidade microbiana existente, pode ser conduzida em grande escala (HANDELSMAN, 2005).

Os dados proporcionados pela análise de metagenomas permitem inferir uma série de características sobre o ambiente estudado, seja quanto à composição microbiana em determinada comunidade, à presença de microrganismos específicos ou dominantes, à existência de rotas metabólicas ou à identificação de genes de interesse (STEELE; STREIT, 2005). É possível, por exemplo, realizar um *screening* do DNA metagenômico estudado buscando alguma atividade enzimática específica (HUGENHOLTZ; TYSON, 2008). Por isso, esta abordagem representa uma poderosa ferramenta para prospecção de genes responsáveis pela síntese de moléculas com propriedades de interesse biotecnológico (BARONE et al., 2014; BERGMANN et al., 2014; COWAN et al., 2004; HANDELSMAN, 2004; STEELE; STREIT, 2005).

A idéia da clonagem do DNA diretamente de amostras ambientais foi primeiramente proposta por Pace, Stahl e Oisen, em 1985, que utilizaram a técnica de PCR para explorar a diversidade microbiana a partir de sequências de RNA ribossômico (PACE; STAHL; OISEN,

1985). Isso levou ao primeiro estudo de isolamento e clonagem de DNA em grande escala a partir de amostras ambientais, o qual foi realizado por Schimdt, Delong e Pace (SCHMIDT; DELONG; PACE, 1991). Neste estudo, a diversidade filogenética picoplanctônica marinha foi analisada a partir da clonagem de sequências de DNA, oriundas de uma região do oceano pacífico, em bacteriófagos λ. A análise do rRNA 16S amplificado detectou a presença de grupos filogenéticos de Proteobacteria distantes das sequências de rRNA 16S conhecidas até então. Estes achados fortaleceram a idéia de que a diversidade microbiana é muito mais complexa do que a conhecida através das técnicas de cultivo tradicionais.

A partir de então, a diversidade microbiana em ecossistemas terrestres e aquáticos foi sendo cada vez mais estudada através da abordagem metagenômica. Janssen (2006) analisou bibliotecas de rRNA 16S preparadas a partir de solos das mais diferentes localidades como Canadá, EUA, Brasil, Alemanha e Áustria. Os filos dominantes encontrados nas bibliotecas foram, na ordem decrescente: Proteobacteria, Acidobacteria, Actinobacteria, Verrucomicrobia, Bacteriodetes, Chloroflexi, Planctomycetes, Gemmatimonadetes e Firmicutes, que juntos compuseram mais de 92% das bibliotecas do solo.

Em relação às amostras aquáticas, Venter e colaboradores (2004) também realizaram um estudo metagenômico em larga escala no Mar de Sargaço. Foram gerados mais de 1.045 bilhão de pares de bases, que foram anotadas e analisadas. Em relação à diversidade filogenética, o grupo dominante foi Proteobacteria (das subdivisões α, β, e γ) e houve também uma contribuição moderada de Firmicutes e de Cyanobacteria. Este estudo faz parte do projeto Coleta Global dos Oceanos (GOS), que envolve a coleta, sequenciamento e análise do DNA de microorganismos de diversos ambientes aquáticos.

Pesquisas mais recentes têm ampliado o espectro de alcance da metagenômica, associando a diversidade microbiana existente em determinado ambiente com a sua diversidade funcional. Mendes e colaboradores (2015), por exemplo, realizaram um estudo no intuito de perceber as consequências da perturbação do solo na composição de uma comunidade microbiana e na sua capacidade funcional. Os autores encontraram que, dentre os quatro tipos de solo analisados – de uma floresta nativa, de um local desmatado, um solo utilizado para agricultura e um para pastagem -, os dois últimos estiveram entre os mais diversos e apresentaram maior redundância funcional. Tais resultados sugeriram que o equilíbrio funcional na floresta (ambiente não perturbado) se manteve baseado numa alta abundância dos microorganismos, enquanto que para os ambientes de solo perturbardos, uma maior diversidade microbiana parece ser a responsável pela manutenção deste equilíbrio.

Já Debroas e colaboradores (2009) utilizaram a metagenômica para estudar a diversidade funcional e taxonômica da comunidade microbiana no maior lago da França, o *Lac du Bourget*. Os autores encontraram que Actinobacteria, seguida de Alpha- e Betaproteobacteria e Bacteroidetes foram os grupos dominantes nas bibliotecas analisadas. Em relação aos perfis metabólicos, ao comparar os resultados encontrados com os obtidos em outros estudos de ambientes aquáticos, foi observado que o metagenoma da comunidade microbiana do *Lac du Bourget* é caracterizado por uma maior representatividade de genes envolvidos na degradação de xenobióticos e metabolismo de glicanos, com menor representatividade de genes relacionados ao metabolismo de alguns aminoácidos. Os autores ainda associaram o perfil metabólico encontrado com os grupos taxonômicos que foram mais abundantes no lago. Foi visto que a maioria dos genes anotados para Alphaproteobacteria pertencem à degradação de xenobióticos, enquanto que Actinobacteria esteve mais relacionada ao metabolismo de nucleotídeos, cofatores e vitaminas, energia, replicação e reparo; Betaproteobacteria distinguiu-se dos outros grupos por genes de transporte de membrana e transdução de sinal e Bacteroidetes teve uma maior representatividade de genes de biossíntese de glicanos e do metabolismo de carboidratos.

A partir dos estudos supracitados é possível perceber que a análise da diversidade microbiana através da metagenômica, em ambientes como o solo e a água, tem permitido uma caracterização taxonômica e funcional que até então não era possível através das técnicas tradicionais de cultivo, ilustrando o grande potencial dessa abordagem.

A abordagem metagenômica é conduzida de forma que o DNA metagenômico seja analisado através da investigação funcional e/ou baseada em sequenciamento (Figura 2). As análises baseadas em função tem o potencial de detectar novos genes, proteínas e outros compostos bioativos, requerendo a expressão heteróloga dos genes clonados e transformados, seguida da triagem dos clones com expressão positiva (DANIEL, 2004). De acordo com Li e colaboradores (2009), um dos principais problemas deste tipo de análise é a observação de uma baixa expressão dos genes que são clonados nos hospedeiros. Muitos genes podem não ser expressos de maneira eficiente devido a diferenças na utilização de códons, modificações pós-traducionais, toxicidade da enzima produzida entre outros (LI et al., 2009).

As análises baseadas em sequenciamento, por sua vez, não dependem da expressão de genes clonados em hospedeiros (DANIEL, 2004). Este tipo de abordagem pode envolver o sequenciamento completo dos clones, na busca do grupo taxonômico originário da sequência,

assim como o sequenciamento aleatório para identificação de genes de interesse (HANDELSMAN, 2004).

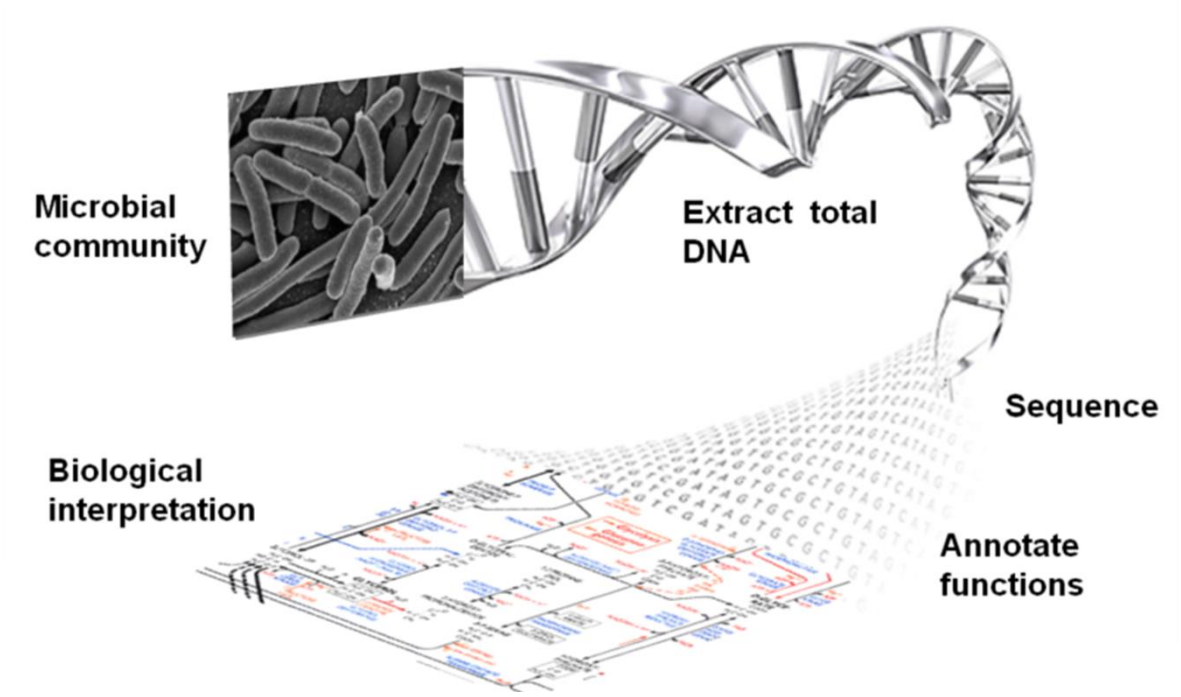


Figura 2. Esquema do fluxo de trabalho na abordagem metagenômica através do sequenciamento direto. Fonte: www.cardiovasculairegeneeskunde.nl.

Independente da estratégia de análise utilizada – seja ela baseada em função ou no sequenciamento - um grande número de clones é requerido para as análises. Em um estudo realizado por Henne e colaboradores (2000), foram necessários 730.000 clones para revelar apenas um clone positivo para a atividade lipolítica no meio ágar testado. Uma forma de eliminar a utilização destes sistemas de clonagem, e diminuir o tempo de execução, surgiu com o desenvolvimento de novas técnicas de sequenciamento, mais avançadas do que o método de Sanger (SANGER; COULSON, 1975).

2.2.2.2 Sequenciamento de nova geração (NGS)

As estratégias de sequenciamento de nova geração estão superando rapidamente o sequenciamento de Sanger, tanto no estudo de genomas pequenos quanto de grandes amostras ambientais (WOOLEY; GODZIK; FRIEDBERG, 2010). Segundo Metzker (2010), a capacidade de produzir com rapidez um enorme volume de dados (alcançando mais de um

bilhão de *reads* por corrida), representa o principal progresso proporcionado pelo sequenciamento de nova geração. Em relação à metagenômica, a maior eficiência que pode ser associada às técnicas de NGS advém da redução dos vieses relacionados à clonagem tradicional, o que é proporcionado pela realização da clonagem *in vitro* (CARVALHO; SILVA, 2010; SHAMSADDINI et al., 2014).

Dentre as técnicas de sequenciamento de nova geração pode-se citar a plataforma 454 da Roche, a SOLiD e a Ion Torrent da Life Technologies e a Illumina, da companhia homônima (Tabela 1). Tais plataformas podem englobar diferentes sistemas de sequenciamento (como por exemplo os sistemas Hiseq, Miseq e GAIIx, todos da Illumina), o que faz com que as estimativas dos tempos de execução e dos custos variem bastante dentro de uma mesma plataforma, a depender do sistema utilizado. A grande capacidade de geração de dados, a economia de tempo e custo são pontos em comum entre as diferentes técnicas (JI; NIELSEN, 2015). O que as difere e determina o tipo de dado produzido em cada uma das plataformas são as especificidades técnicas na realização do sequenciamento (METZKER, 2010). No presente estudo, o sistema de sequenciamento utilizado foi o Miseq, da Illumina, que será detalhado a seguir.

Tabela 1. Comparação entre as plataformas de sequenciamento de nova geração (NGS). Modificado de Glenn, 2011.

Plataforma	**Companhia**	**Tempo de execução**	**Milhões de *reads* por corrida**	**Custo de reagentes por corrida**
454	Roche	10-20h	0.1-1	U$1.100-6.200
Illumina	Illumina	26h-14 dias	3.4-3.000	U$750-23.470
SOLiD	Life Technologies	8-12 dias	>700	U$6.101-10.503
HeliScope	Helicos	Não disponível	800	Não disponível
Ion Torrent	Life Technologies	2h	0.1-8	U$500~925
PacBio	Pacific Biosciences	0.5-2h	0.01	U$110-900
Starlight	Life Technologies	Não disponível	~0.01	Não disponível

2.2.2.2.1 A plataforma Illumina

Na plataforma Illumina (assim como em outras plataformas de sequenciamento de nova geração) as bibliotecas metagenômicas são geradas a partir da fragmentação aleatória das sequências de DNA pelo método de sequenciamento *shotgun*. O DNA fragmentado é ligado a adaptadores em ambas as extremidades (Figura 3A) e, em seguida, este DNA é imobilizado na superfície de clonagem (*flow cell*) (Figura 3B). Isso ocorre a partir da hibridização entre o adaptador livre presente nas sequências do DNA ligado, com adaptadores presentes também na superfície das *flow cells*. A alta densidade de adaptadores nas *flow cells* faz com que ocorra esta hibridização, formando assim uma estrutura em "ponte" (Figura 3C). Dá-se então o processo de amplificação dos fragmentos de DNA fita simples ligados em ambas as extremidades à superfície, com a síntese da fita complementar ao DNA ligado, a partir da adição de nucleotídeos não marcados e da DNA polimerase ao sistema (Figura 3D). A DNA polimerase incorpora os nucleotídeos de maneira que são construídas "pontes" de DNA fita dupla no suporte sólido. Uma etapa de desnaturação dos fragmentos de fita dupla leva novamente à linearização das sequências de DNA fita simples ancoradas ao sistema (Figura 3E). A amplificação em fase sólida leva à formação de mais de 1.000 cópias idênticas de cada fragmento de DNA que permanecem próximos uns dos outros, formando *clusters* (Technology Spotlight: Illumina® Sequencing) (Figura 3F).

A tecnologia de sequenciamento utilizada na plataforma Illumina, assim como o sequenciamento de Sanger, é realizada por síntese, utilizando a DNA polimerase e quatro nucleotídeos terminadores marcados com diferentes fluoróforos (CARVALHO; SILVA, 2010). No entanto, o processo de sequenciamento da Illumina não requer a clonagem física de fragmentos, já que esta clonagem é feita *in vitro* na superfície das *flow cells*, o que caracteriza um avanço em relação ao método de Sanger, permitindo o sequenciamento de dezenas de milhares de *clusters* em paralelo (CARVALHO; SILVA, 2010). Durante cada ciclo de sequenciamento nos *clusters*, nucleotídeos terminadores marcados com fluorescência são adicionados à reação (Figura 3G).

O nucleotídeo marcado serve como terminador para a reação, então, durante cada incorporação de um nucleotídeo, é feita a leitura da fluorescência para identificação da base adicionada (Figura 3H). Em seguida, esse nucleotídeo é enzimaticamente clivado para remoção do terminal 3' bloqueado e do fluoróforo, permitindo a incorporação do próximo nucleotídeo (METZKER, 2010). A identificação das bases é feita diretamente através da

medida da intensidade do sinal em cada ciclo (Technology Spotlight: Illumina® Sequencing). Os ciclos são repetidos de maneira que, ao final, seja identificada a sequência de bases dos fragmentos de DNA em cada *cluster* (I-J), permitindo assim a geração dos dados que serão posteriormente analisados.

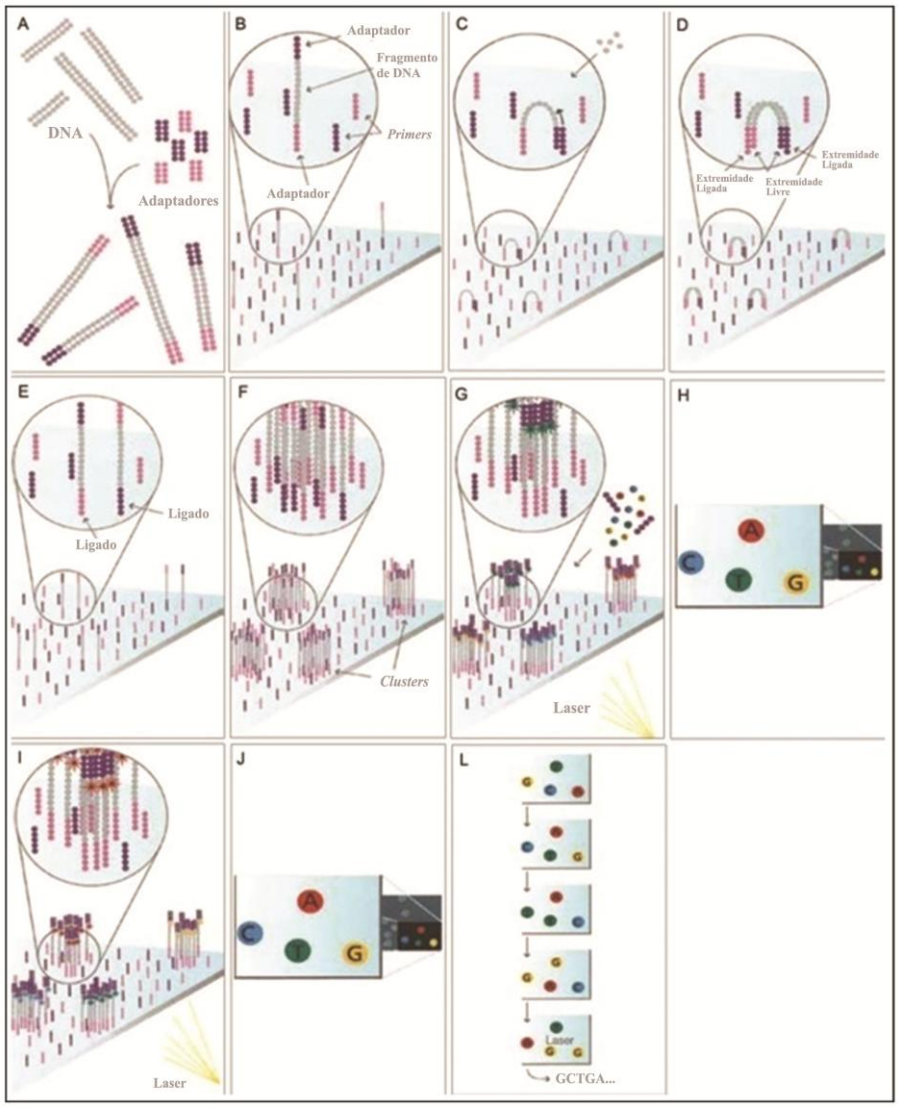


Figura 3: Modelo esquemático do sequenciamento na plataforma Illumina. (A) Fragmentação aleatória do DNA (*shotgun*) e ligação aos adaptadores em ambas extremidades. (B) Ligação dos fragmentos de DNA fita simples às suas sequências complementares na superfície das *flow cells*. (C) Etapa de anelamento, com a formação das "pontes", a partir da ligação do adaptador presente na extremidade livre do fragmento de DNA, com sua sequência complementar aderida à superfície. (D) Amplificação das cadeias de DNA a partir da incorporação de nucleotídeos pela DNA polimerase, formando "pontes" em fita dupla. (E) Etapa de desnaturação na qual as "pontes" são desfeitas. (F) Após diversos ciclos de amplificação, são formados *clusters* com milhares de sequências idênticas. (G) Adição dos nucleotídeos terminadores marcados e sua incorporação à cadeia pela DNA polimerase. (H) Após excitação à laser, a fluorescência em cada *cluster* é captada pelo sistema, identificando qual base foi incorporada à cadeia de DNA. (I e J) Repetição do ciclo de incorporação,

excitação à laser e detecção do nucleotídeo pelo sistema. (L) Determinação da sequência de bases em cada fragmento. Adaptado de Carvalho e Silva, 2010.

2.2.2.3 A Bioinformática na análise de dados metagenômicos

O desenvolvimento de técnicas avançadas de sequenciamento tem levado à produção de um grande volume de dados genômicos e metagenômicos, o que representa tanto uma oportunidade como um desafio para os cientistas. O crescimento na disponibilidade de sequências ocorreu concomitantemente ao desenvolvimento computacional, o que promoveu um surgimento cada vez maior de bases de dados, bem como de ferramentas de análises de dados de DNA/RNA e proteínas (REHM, 2001).

A Bioinformática é uma área interdisciplinar de pesquisa, desenvolvida no intuito de organizar, analisar e distribuir informações a fim de responder a questões biológicas complexas (ABD-ELSALAM, 2003; SINGH; KUMAR, 2001). Nos últimos anos, um grande número de ferramentas de bioinformática têm se tornado disponível, o que trouxe à tona as questões de quais ferramentas utilizar e como utilizá-las, no sentido de melhor responder às questões científicas propostas (REHM, 2001).

Estudos metagenômicos originam conjuntos de dados de grande volume e complexidade que, segundo Desai e colaboradores (2012), não podem ser vistos simplesmente como uma extensão da genômica, pois requerem diferentes ferramentas de bioinformática. Ainda segundo os autores, a análise dos dados é o fator limitante principal na metagenômica, tendo o aumento no volume dos dados trazido desafios às ferramentas existentes. A superação dessas dificuldades pode ser conseguida através da otimização das estratégias utilizadas em cada estudo metagenômico particular, a partir das ferramentas já disponíveis.

2.2.2.3.1 Estratégia de anotação pela tecnologia de subsistemas

A comparação de determinado conjunto de sequências, com bases de dados de sequências de referência, é o primeiro passo na análise dos metagenomas, o qual envolve um grande esforço computacional, mas fornece subsídios para diversas análises como comparações filogenéticas, anotações funcionais e reconstruções metabólicas (MEYER et al.,

2008). O servidor MG-RAST (MEYER et al., 2008) é um sistema automatizado de anotação para o processamento e análise de dados metagenômicos, capaz de promover tais análises. O servidor, até março de 2015, já contava com mais de 12.000 usuários e 163.729 metagenomas anotados. Seu fluxo de trabalho envolve o controle de qualidade das sequências, predição de proteínas, *clustering* e anotação baseada em similaridade. O servidor realiza uma busca por similaridade entre as proteínas preditas a partir das sequências submetidas (na etapa de *gene calling*) e as bases de dados de proteínas. A anotação pode ser visualizada em diversas categorias diferentes, entre elas a de subsistemas.

Um subsistema pode ser entendido como um conjunto de papéis funcionais que implementam um determinado processo biológico ou estrutural (OVERBEEK et al., 2005). Os subsistemas são classificados em níveis hierárquicos, de maneira que no nível 1 são incluídas funções gerais tanto catabólicas quanto anabólicas (por exemplo, o metabolismo do DNA) e nos níveis 2 e 3 constam vias mais específicas (como por exemplo, a assimilação do sulfato) (DINSDALE et al., 2008). Cada subsistema é curado por um grupo de cientistas especializados em vias metabólicas específicas, estabelecendo quais os principais genes envolvidos e sua arquitetura genômica. As proteínas que compõem os subsistemas formam famílias de homólogos isofuncionais (FIGfam-*Fellowship Interpretation of Genomes Family*) (GERLT; BABBITT, 2001). Sendo assim, existem três componentes principais que tornam esse método de anotação rápido e acurado: o servidor RAST, os subsistemas e as FIG-fam.

2.2.2.3.2 Anotação pela estratégia de domínios conservados

Outra estratégia para otimizar a busca por enzimas com atividades de interesse é triar as sequências-alvo, obtidas por sequenciamento de alta vazão, através das ferramentas de bioinformática. É possível identificar "assinaturas" nas sequências-alvo (através do seu alinhamento global) e utilizar tais assinaturas como sondas para efetuar a busca em bases de dados de sequências protéicas (FINN; CLEMENTS; EDDY, 2011). Essa abordagem permite a identificação de genes putativos para produção de enzimas de interesse, funcionando como um filtro para selecionar genes-alvo para clonagem molecular.

A busca por domínios catalíticos tem se mostrado mais eficiente do que a simples busca por similaridade de sequências para a identificação de novas proteínas (EDDY, 2011). Sequências que apresentam no mínimo 40% de identidade com uma cobertura de 70%, em

comparação à sequências depositadas em bases de dados de sequências homólogas, são utilizadas como parâmetros para uma inferência funcional relativamente segura, de acordo com evidências empíricas (EDDY, 2009; HINZ, 2010). Essa estratégia difere das tradicionais que realizam triagem funcional de bibliotecas metagenômicas de metagenomas inteiros, triagem a qual, em comunidades muito diversas não apresenta alta eficiência (DANIEL, 2004). A caracterização do potencial funcional baseado na análise de sequências conservadas no metagenoma da microbiota do solo e da água permite identificar previamente as sequências de interesse, otimizando, dessa forma, todo o processo.

2.3 Bioprospecção de enzimas e suas aplicações na indústria

As enzimas são proteínas com atividade catalítica e estão envolvidas em um grande número de transformações de moléculas com importância biológica (WHITESIDES; WONG, 1983). Essas proteínas apresentam características que favorecem sua utilização como catalisadores, como por exemplo, suas aceleradas taxas de reação e a seletividade específica ao substrato (WHITESIDES; WONG, 1983). Devido a estas propriedades catalíticas, as enzimas têm sido incorporadas nos mais diversos produtos e processos industriais (KIRK; BORCHERT; FUGLSANG, 2002), e o rápido desenvolvimento tecnológico atual tem aumentado cada vez mais o alcance das suas aplicações.

Microorganismos na natureza têm sido, ao longo dos anos, uma importante fonte de bioprospecção de diversos tipos de moléculas, incluindo as enzimas. Com os avanços da bioinformática e dos estudos genômicos e metagenômicos, tem se tornado cada vez mais eficiente o isolamento de genes, bem como de compostos de interesse no ambiente (KIRK; BORCHERT; FUGLSANG, 2002). Nesse sentido, é possível observar a crescente disponibilidade de estudos de bioprospecção de enzimas na literatura e, dentre os principais grupos de proteínas com interesse biotecnológico, estão as enzimas relacionadas ao metabolismo de açúcares. Isso se deve ao fato de que os carboidratos são moléculas que intercedem nos mais diversos processos biológicos, do reconhecimento e sinalização celulares à formação de reservas energéticas e de moléculas estruturais, sendo encontrados de maneira abundante no meio ambiente (LOMBARD et al., 2014). Sendo assim, as enzimas envolvidas no seu metabolismo também detém importância em diferentes funções biológicas e podem ser melhor compreendidas e exploradas dado o potencial biotecnológico que possuem.

2.3.1 Enzimas ativas em carboidratos (CAZymes)

De uma maneira geral, as enzimas envolvidas na clivagem de carboidratos complexos, bem como as relacionadas à sua biossíntese, controlam o metabolismo dos carboidratos e são denominadas de CAZymes (LOMBARD et al., 2014). As CAZymes são produzidas por um amplo espectro de organismos, que as utilizam das mais diferentes formas para modificar os carboidratos, seja na formação ou clivagem de ligações glicosídicas, o que permite, dentre outras coisas, a biossíntese e degradação da celulose e de outros polissacarídeos (MUNIR et al., 2014; OLIVEIRA et al., 2015).

As enzimas que compõem as CAZymes são classificadas com base na similaridade de sequências e de estruturas protéicas e, de acordo com a base de dados CAZy (www.cazy.org/), são ao todo 6 as suas principais classes. As glicosil hidrolases (GHs) compõem um amplo grupo de enzimas responsáveis pela hidrólise e/ou rearranjo de ligações glicosídicas; as carboidrato esterases (CEs) estão envolvidas na hidrólise de ésteres de carboidratos; as glicosiltransferases (GTs) catalisam a formação de ligações glicosídicas através da transferência de sacarídeos a uma variedade de aceptores; as polissacarídeo liases (PLs) promovem a clivagem de ligações glicosídicas através de um mecanismo de eliminação; as atividades auxiliares (AAs) são enzimas do tipo redox que agem em conjunto com outras CAZymes na degradação da parede celular vegetal; e os módulos de ligação a carboidratos (CBMs), que são domínios não catalíticos auxiliares, cuja função principal é a de reconhecer e ligar-se especificamente a polissacarídeos, permitindo a sua hidrólise pelo biocatalisador (CANTAREL et al., 2009; GUILLÉN; SÁNCHEZ; RODRIGUEZ-SANOJA, 2010; LEVASSEUR et al., 2013; LOMBARD et al., 2014; MUNIR et al., 2014; WEADGE; PALCIC, 2008; YIP; WITHERS, 2006).

Uma gama de estudos metagenômicos têm relatado a contribuição e a diversidade de CAZymes nos mais distintos ambientes, como no solo de uma plantação de pinheiros (UROZ et al., 2013), na resposta do bacterioplâncton marinho após um *bloom* de diatomáceas (TEELING et al., 2012) no rúmen de bovinos (BRULC et al., 2009; WANG et al., 2013;) e no intestino grosso de cupins (WARNECKE et al., 2007) e de humanos (KAOUTARI et al., 2013). Mais ainda, pesquisas recentes têm direcionado o foco à descoberta e identificação de novas CAZymes (BERGMANN et al., 2014; MATSUMURA et al. 2014; STROOBANTS; PORTETELLE; VANDENBOL, 2014).

2.3.1.1 CAZymes que possuem atividade hidrolítica

As hidrolases são enzimas que catalisam a hidrólise de ligações covalentes da matéria orgânica, geralmente convertendo uma molécula grande em duas unidades menores. A classe das hidrolases contém mais de 200 enzimas que são classificadas de acordo com o substrato alvo da sua ação, como por exemplo, o grupo das proteases ou peptidases, que agem sobre as ligações peptídicas e as glicosidases (ou glicosil hidrolases) que têm sua ação sobre as moléculas de açúcares em carboidratos (www.enzyme-database.org).

Na natureza, as hidrolases são produzidas por uma ampla diversidade de organismos, desde bactérias a eucariotos superiores (MORENO et al., 2013), e são responsáveis por funções degradativas importantes, como a clivagem de moléculas complexas que servirão como fonte de carbono para a produção de energia (DOUGHERTY et al., 2012).

Uma importante fonte de hidrolases no meio ambiente são os microorganismos, que as produzem e utilizam para a manutenção de diversas funções do seu metabolismo. Enzimas hidrolíticas de diferentes tipos produzidas por microorganismos vêm sendo estudadas e descritas na literatura. Na revisão de Moreno e colaboradores (2013), são mostrados os resultados de diversos estudos focados na diversidade de bactérias halófilas com atividade hidrolítica. Dentre os diferentes ambientes hipersalinos estudados, foram encontrados muitos microorganismos (como por exemplo, dos gêneros *Salicoca*, *Salinibacter* e *Pseudomonas*) produtores de lipases, proteases e amilases, entre outras enzimas hidrolíticas.

Na indústria, a maioria das enzimas utilizadas pertence ao grupo das hidrolases, as quais têm sido incorporadas nos processos de degradação de várias substâncias naturais (KIRK; BORCHERT; FUGLSANG, 2002). A revisão de Cherry e Fidantsef (2003) ressalta as aplicações na indústria de alguns destes principais grupos de enzimas.

Dentro da classificação das CAZymes, as enzimas hidrolíticas são encontradas em duas classes: na classe das carboidrato esterases (CEs) e das glicosil hidrolases (GHs). As carboidrato esterases reúnem-se atualmente em 16 famílias envolvidas na desacetilação de açúcares, reação que parece ter se desenvolvido para contornar a proteção conferida pela presença de ésteres na parede celular vegetal (BIELY, 2012). A ação das carboidrato esterases favorece a sacarificação enzimática - permitindo, portanto, a bioconversão da biomassa vegetal - ao facilitar a ação das glicosil hidrolases na clivagem das ligações glicosídicas

presentes nos polissacarídeos vegetais complexos, como por exemplo, as hemiceluloses (BIELY, 2012; CANTAREL et al., 2009; LI et al., 2009).

As glicosil hidrolases, por sua vez, agrupam-se em 133 famílias, constituindo a classe mais ampla, diversificada e melhor caracterizada dentre as CAZymes (CANTAREL et al., 2009; STROOBANTS; PORTETELLE; VANDENBOL, 2014). As famílias de glicosil hidrolases são distribuídas de maneira bastante variável entre os organismos, compreendendo uma ampla gama de atividades já descritas (SATHYA; KHAN, 2014). Dentre os principais grupos de glicosil hidrolases, que participam da decomposição de polissacarídeos da parede celular vegetal, estão as celulases, juntamente às hemicelulases e outras glicosil hidrolases relacionadas (HIMMEL et al., 2010). As celulases pertencem a 12 famílias de GHs e, dentre estas, as famílias GH9 e GH5 parecem ser as mais bem caracterizadas (SUKHARNIKOV et al., 2011). Segundo Cherry e Fidantsef (2003), a aplicabilidade das celulases ocorre principalmente na indústria têxtil, de papel e celulose, de detergentes e na alimentação humana e animal. No entanto, mais recentemente, grandes esforços têm sido direcionados ao estudo e prospecção destas enzimas para aplicação na conversão de biomassa vegetal em biocombustíveis (BARNARD et al., 2010; SADHU; MAITI, 2013; WILSON, 2009).

2.3.1.1.1 As celulases

As celulases são um conjunto de enzimas pertencentes à classe das glicosil hidrolases que são capazes de hidrolisar a celulose, o biopolímero mais abundante na natureza (LI et al., 2009). Tais enzimas estão amplamente distribuídas entre os microorganismos, incluindo fungos e bactérias, que são capazes de crescer em substratos celulósicos (KUHAD; GUPTA; SINGH, 2011). Dentre as bactérias conhecidas por terem atividade celulolítica estão as aeróbicas pertencentes aos gêneros *Bacillus*, *Pseudomonas*, *Cellulomonas* e *Acinetobacter*, e as anaeróbicas dos gêneros *Clostridium*, *Fibrobacter* e *Acetivibrio*, entre outros (KUHAD; GUPTA; SINGH, 2011; SUKUMARAN; SINGHANIA; PANDEY, 2005).

As enzimas celulolíticas têm demonstrado seu vasto potencial biotecnológico, evidenciado pelo crescente interesse na sua incorporação em processos industriais. Devido à sua multiplicidade de aplicações - como no processamento do algodão, na reciclagem de papel, na indústria de detergentes, na extração de sucos e na alimentação animal – as celulases têm representado o terceiro maior grupo de enzimas interesse industrial no mundo, em

volume de dólares (WILSON, 2009). Espera-se que ocorra uma expansão ainda maior do mercado de celulases, uma vez que estas podem ser utilizadas para converter a biomassa lignocelulósica em açúcares, os quais, por sua vez, podem ser fermentados para produzir bioetanol e outros produtos em larga escala (DUAN; FENG, 2010).

A aplicabilidade das celulases na produção de biocombustíveis está justamente no fato de que essas enzimas catalisam a hidrólise de resíduos vegetais, cujo principal componente é a celulose, produzindo assim açúcares fermentáveis, que podem ser utilizados na produção de bioetanol. Geralmente este rejeito vegetal é aproveitado como adubo ou ração animal. Atualmente, porém, tem se tornado uma fonte atrativa para a produção de bioetanol devido à sua disponibilidade, abundância, baixo custo e o fato de não ser utilizado como fonte de alimento humano (PHITSUWAN et al., 2012). Esses resíduos apresentam alto teor de material vegetal e, por isso, a conversão de biomassa em bioetanol pode ser realizada através da aplicação de hidrolases microbianas.

A despolimerização da celulose representa uma etapa de alto rendimento energético para a fermentação. A degradação completa da celulose requer um complexo formado principalmente por três classes enzimáticas, que devem atuar de maneira sinérgica (Figura 4).

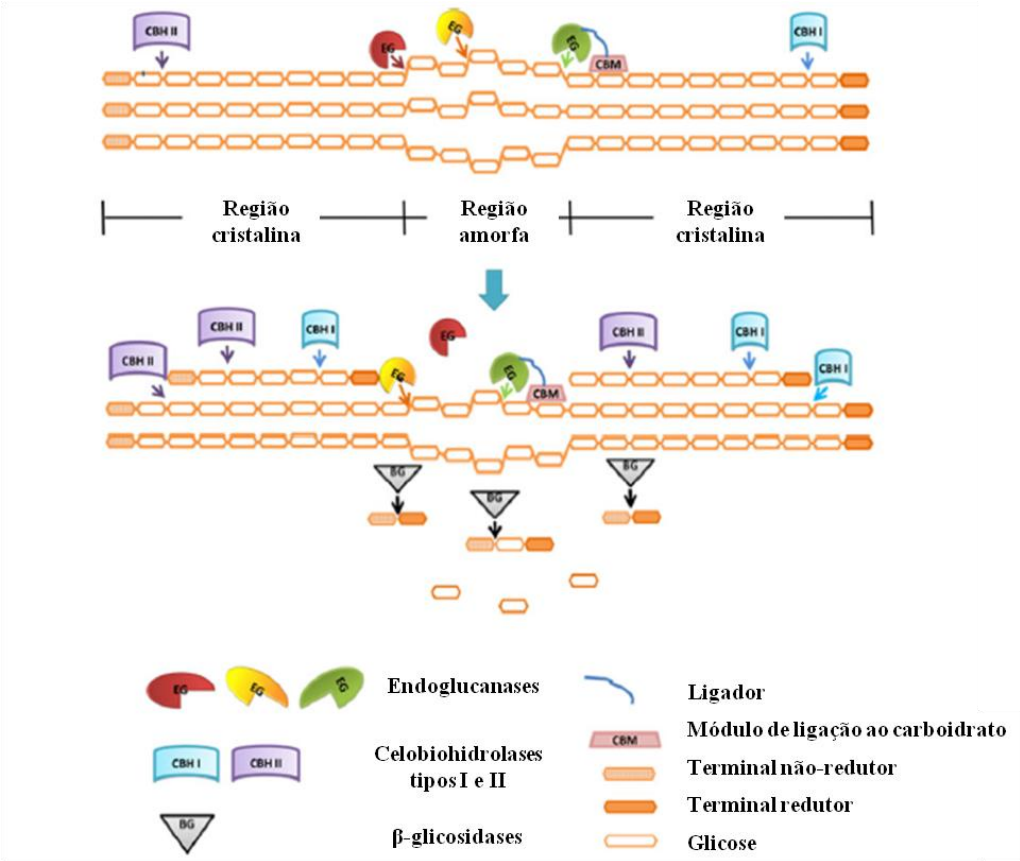


Figura 4. Modelo simplificado da hidrólise enzimática da celulose. As endoglucanases são as primeiras a clivarem as regiões amorfas da celulose, gerando celooligossacarídeos. As celobiohidrolases (exoglucanases) atuam nos terminais não-redutores, gerando principalmente unidades de celobiose livres. Por fim, as β-glicosidases hidrolisam os produtos resultantes – celobioses e oligossacarídeos – em glicose. Fonte: Phitsuwan et al., 2012.

Este complexo é composto por endoglucanases, que hidrolisam randomicamente a fibra de celulose nas ligações β-1,4 internas reduzindo o tamanho do polímero; exoglucanases, que hidrolisam a celulose a partir dos terminais não-redutores, liberando unidades de celobiose; e β-glicosidases, que clivam os oligossacarídeos ou as celobioses livres, gerando glicose (HAN; YOO; KANG, 1995; LYND et al., 2002).

Segundo Duan e Feng (2010), apesar dos diversos estudos realizados acerca das celulases, continua a existir uma crescente demanda por novas enzimas com melhores propriedades catalíticas, como por exemplo, uma maior eficiência em substratos celulósicos insolúveis, uma maior estabilidade em determinado pH e sob condições de temperaturas elevadas.

Relatos como este reforçam a importância de pesquisas que envolvam a bioprospecção de celulases e demais CAZymes, revelando o potencial biotecnológico existente na descoberta de novos genes codificantes para estas enzimas, a partir de diferentes estratégias de estudo. A bioprospecção de enzimas envolvidas no metabolismo da matéria orgânica parece ser uma estratégia importante, tanto do ponto de vista da caracterização microbiana funcional, quanto do seu potencial alcance biotecnológico.

3. OBJETIVOS

3.1 Objetivo geral

Caracterizar a diversidade taxonômica e o potencial funcional da microbiota do solo e da água doce de uma localidade da Caatinga, para buscar genes com possível atividade hidrolítica e demais relacionadas à bioconversão lignocelulósica, que tenham potencial aplicação na cadeia produtiva de bioetanol de segunda geração, através da abordagem metagenômica**.**

3.2 Objetivos específicos:

1. Sequenciar o metagenoma da microbiota do solo e da água doce de uma região da Caatinga baiana;

2. Caracterizar a diversidade taxonômica e o potencial funcional a partir dos metagenomas da microbiota do solo e da água doce de uma região da Caatinga baiana;

3. Caracterizar a diversidade de hidrolases e demais enzimas relacionadas à bioconversão lignocelulósica na microbiota do solo e da água doce desta região de Caatinga;

4. Identificar possíveis novas sequências de CAZymes com atividade hidrolítica e outras também associadas à degradação da biomassa vegetal.

4. METODOLOGIA

4.1 Área de estudo e coleta das amostras de solo e de água

A coleta das amostras de solo e de água foi realizada na ecorregião da Chapada Diamantina, mais especificamente na localidade da Toca do Morcego (Figura 5A-C). A expedição ocorreu de 25 a 27 de fevereiro de 2013 e contou com o suporte técnico do Instituto Chico Mendes de Conservação Ambiental (ICMBio).

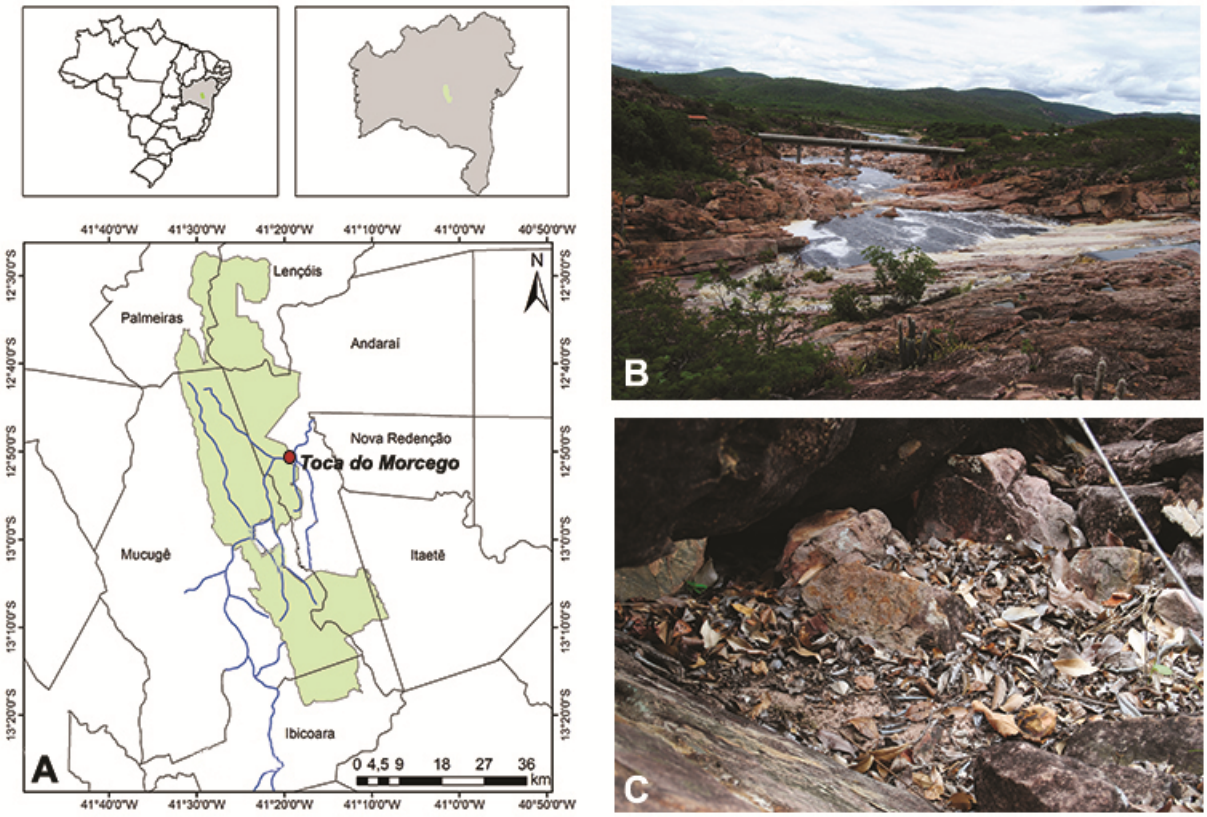


Figura 5: Área de estudo. (A) Mapa do Parque Nacional da Chapada Diamantina (em verde), destacando a localidade da Toca do Morcego, onde foram coletadas as amostras; (B) Vista do rio Paraguaçu na Toca do Morcego; (C) Aspecto geral de um dos pontos de coleta das amostras de solo.

A coleta das amostras ocorreu no período de seca. Na maior parte do mês de fevereiro, a temperatura na região da Chapada Diamantina ficou em torno dos 30ºC, tendo a umidade sofrido uma ampla variação, de 54 a 98% durante este mês (Figura 6).

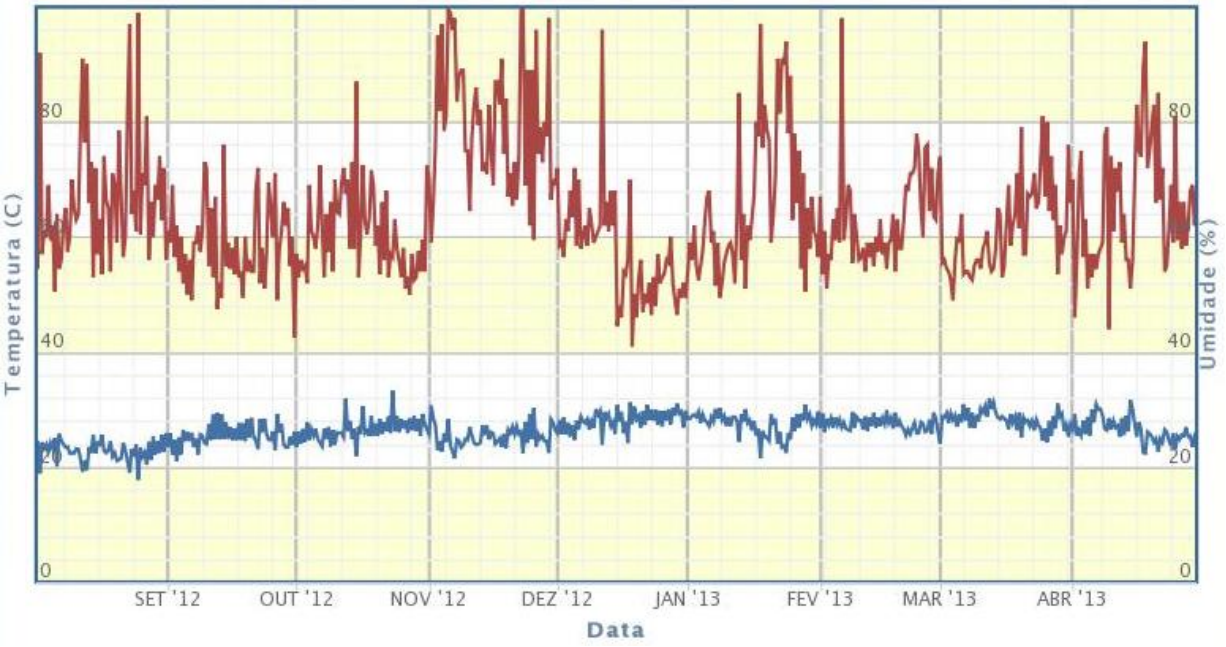


Figura 6. Variações de temperatura e de umidade referentes ao período de agosto/2012 a maio/2013 na Chapada Diamantina - Lençóis-BA, estação mais próxima dos pontos de coleta (Fonte: INMET).

Em relação à precipitação pluviométrica, o período compreendido entre 25/01/2013 e 21/03/2013 foi considerado como um dos piores períodos de seca da região, apresentando apenas cerca de 8mm de precipitação entre os dois meses considerados (Figura 7).

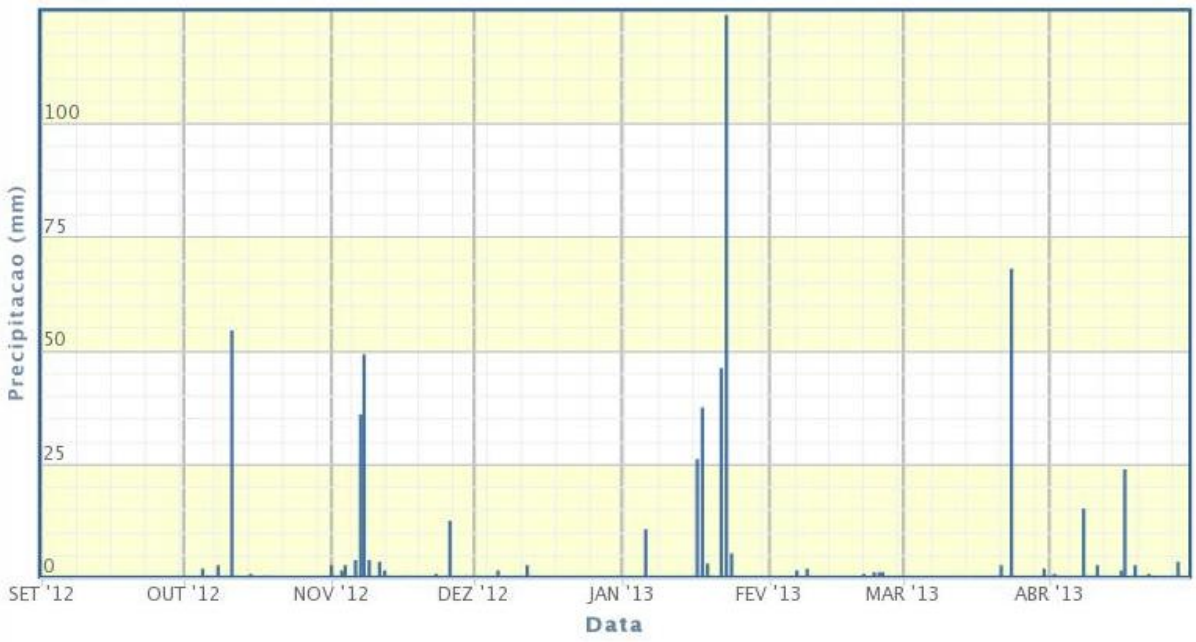


Figura 7. Variação da precipitação pluviométrica entre os meses de agosto/2012 e maio/2013 na Chapada Diamantina - Lençóis-BA, estação mais próxima dos pontos de coleta (Fonte: INMET).

Em relação à flora na região de coleta das amostras, esta possui características dos Campos Rupestres, com vegetação herbáceo arbustiva sobre afloramentos rochosos, com adaptações para resistir à flutuações acentuadas de temperatura e umidade. As plantas exibem características para resistir à perda de água, como folhas espessas e rígidas, cobertas pos cera ou pêlos e podem possuir raízes modificadas em órgãos subterrâneos para armazenamento de água e de nutrientes (PNCD).

Foram coletadas 3 amostras de solo, preservando uma distância de aproximadamente 100m entre elas. Cada amostra foi composta por solo recolhido em 3 pontos escolhidos de maneira aleatória, dentro de uma área de aproximadamente 1m², numa profundidade máxima de 8cm após a retirada da serrapilheira (Figura 6A-D). Foram coletados cerca de 200g de solo por ponto (totalizando 600g por amostra), e este material foi acondicionado em sacos plásticos estéreis e mantido sob refrigeração até o seu processamento. Parte das amostras de solo coletadas foi utilizada para extração de DNA metagenômico, enquanto que o restante das amostras foi encaminhado para análise físico-química. A caracterização físico-química do solo foi realizada em colaboração com o Laboratório de Solos da CEPLAC.

Figura 8. Coleta das amostras de solo e esquema de filtração das amostras de água. (A) Aspecto geral do ponto de coleta 1; (B e C) Coleta no ponto 2; (D) Armazenamento do solo; (E) Esquema de filtração Niskin com filtros acoplados.

Em relação às amostras de água, foram coletadas 2 amostras em pontos distintos do rio Paraguaçu, em uma profundidade de aproximadamente um metro, também na localidade da Toca do Morcego. Este material foi filtrado em uma rede de 20µm e novamente filtrado em três filtros de 0.22µm (Sterivex, Milipore, USA) por ponto de coleta. Filtrou-se 0.5 a 2L de água por filtro de 0.22µm. O material coletado foi preservado em solução-tampão (20% sacarose, 50 mM EDTA, 50 mM Tris-HCl, pH 8.0). A filtração foi realizada por pressão positiva aplicada ao sistema de garrafas Niskin (Figura 6E). Os parâmetros físico-químicos das amostras de água foram avaliados em colaboração com o Laboratório de Zooplâncton Marinho (UFRJ).

4.2 Extração do DNA metagenômico

Para a realização da extração do DNA metagenômico das amostras de solo foi utilizado um *kit* específico (PowerSoil DNA Isolation Kit da MO-BIO). O DNA foi extraído a partir de 0.25g de solo obtido de cada uma das 3 amostras, de acordo com as recomendações do fabricante. A eficiência da extração foi posteriormente analisada em gel de agarose a 1% em TAE 1X (p/v).

A extração do DNA metagenômico das amostras de água foi realizada em colaboração com o laboratório de Enzimologia, no Instituto de Biologia, da Universidade de Brasília, a partir de um protocolo já estabelecido no laboratório. O protocolo envolveu a extração do DNA utilizando lisozima (1mg/mL na concentração final) por 45 minutos à 37ºC. Em seguida foi adicionada proteinase K (0.2 mg/mL na concentração final) e SDS (1% na concentração final) e as amostras foram incubadas por 60 minutos à 60ºC sob leve agitação a cada cinco minutos. O lisado foi transferido para novos tubos e, em seguida, foi tratado com um volume de fenol : clorofórmio : álcool isoamílico (25:24:1, Sigma) e depois com clorofórmio : álcool isoamílico (24:1, Sigma). A precipitação foi realizada com etanol e acetato de sódio (3M final) a -20ºC *overnight*. Após a precipitação, o DNA foi purificado com Power Clean DNA Clean-Up Kit (MO-BIO) e submetido a um gel de agarose a 1% contendo brometo de etídeo (2µg/mL) com marcador molecular 1kb plus (Invitrogen).

4.3 Sequenciamento dos metagenomas

As amostras contendo o DNA metagenômico obtido através da extração foram encaminhadas ao Centro de Genômica de Alto Desempenho do DF, em Brasília, para realização do sequenciamento, o qual foi conduzido utilizando-se o sistema Illumina MiSeq de sequenciamento de nova geração (Illumina® Sequencing). Bibliotecas foram geradas utilizando o *kit* Nextera (Nextera DNA Sample Preparation Kit, Illumina), segundo o protocolo do fabricante. O DNA proveniente da fragmentação foi submetido às etapas de purificação, ligação a adaptadores e quantificação, para então ser sequenciado no sistema Illumina Miseq. A corrida foi *paired-end* 2x300pb para as amostras de solo e 2x250pb para as amostras de água.

4.4 Anotação dos metagenomas pela tecnologia de subsistemas e análise estatística

Os metagenomas foram anotados utilizando o servidor MG-RAST (*MetaGenomic Rapid Annotation by Subsystem Technology*), um sistema de genômica comparativa que permite estudar a função e a composição de comunidades microbianas (MEYER et al., 2008). O servidor oferece controle de qualidade, anotação, análise comparativa e arquivamento automatizados. O fluxo de informações, bem como a interação com os dados e ferramentas de análise, ocorre dentro de um ambiente virtual (SEED) (OVERBEEK et al., 2005). Este ambiente abriga o servidor RAST que contém um banco de genomas completos anotados, bem como a lista de subsistemas já caracterizados. Foram utilizados os parâmetros padrão de qualidade do servidor, com exceção do *lowest phred score* que foi alterado para 20.

4.4.1 Análise taxonômica e do potencial funcional pela abordagem de subsistemas

Os perfis taxonômicos e funcionais foram gerados pelo MG-RAST através de uma análise de similaridade entre as sequências submetidas contra um conjunto de bases de dados não-redundantes incluindo o SEED, de subsistemas. A análise taxonômica foi baseada na correlação entre as proteínas identificadas nos metagenomas e seu organismo de origem nas bases de dados, permitindo o cálculo da abundância relativa de cada táxon obtido. Os táxons foram identificados através do Best Hit Classification, utilizando como base de dados o GenBank, e com os parâmetros *E* value $<10^{-5}$, mínima identidade de 60% e comprimento mínimo do alinhamento de 15 como *cutoff*. A classificação do potencial funcional dos

metagenomas foi realizada no MG-RAST através da anotação pela tecnologia de subsistemas, utilizando a base de dados do SEED e também com os parâmetros *E* value $<10^{-5}$, mínima identidade de 60% e comprimento mínimo do alinhamento de 15 como *cutoff*.

Os dados taxonômicos e funcionais foram estatisticamente avaliados através da comparação entre dois grupos (solo e água), no programa STAMP (Statistical Analyses of Metagenomic Profiles) (PARKS et al., 2014). Para o perfil taxonômico, o teste estatístico aplicado foi o teste t de Welch bicaudal, o teste de Welch invertido como método de intervalo de confiança e o FDR de Storey (STOREY, 2002) como método de correção dos dados. Para a análise do potencial funcional, foi também aplicado do teste t de Welch e o teste de Welch invertido como método de intervalo de confiança. No entanto, como os valores de *p* não estavam distribuídos uniformemente, o método FDR de Benjam-Hochberg foi utilizado para correção. A diferença entre as abundâncias relativas foi considerada significativa para $q<0.05$. As *reads* não classificadas foram removidas das análises.

4.5 Anotação dos metagenomas pela estratégia de domínios conservados

A segunda estratégia de anotação utilizada no presente estudo envolveu a busca das CAZymes (Carbohydrate Active Enzymes) (LOMBARD et al., 2014) nos metagenomas do solo e da água, a partir da assinatura dos domínios conservados de tais grupos de enzimas. A anotação das CAZymes foi realizada localmente, a partir dos dados obtidos das etapas de anotação no MG-RAST e com os perfis HMM das CAZymes, baixados no servidor dbCAN (http://csbl.bmb.uga.edu/dbCAN/).

Foram acessadas as sequências submetidas à anotação no MG-RAST após a etapa de *Screening* (remoção de *reads* quase idênticas a sequências de genomas de organismos-modelo como por exemplo humanos e ratos), sendo obtido um arquivo multifasta contendo as sequências de DNA de todas as amostras. Estas sequências foram traduzidas nas 6 janelas de leitura utilizando o programa Transeq, do pacote EMBOSS (European Molecular Biology Open Software Suite) (RICE; LONGDEN; BLEASBY, 2000). Em seguida, foram baixados os perfis (domínios conservados) de todas as famílias de CAZymes já descritas no servidor dbCAN. Foi então utilizado o programa *hmmscan*, do pacote de programas HMMER (FINN; CLEMENTS; EDDY, 2011), que usou os perfis HMM das famílias de CAZymes baixados, pra realizar uma busca contra o multifasta obtido do MG-RAST e traduzido, o qual funcionou

como uma “base de dados”, permitindo a anotação. A ideia principal do *hmmscan* é a de utilizar perfis baseados nas cadeias ocultas de markov (HMM) como modelos na busca de regiões conservadas homólogas na base de dados de interesse. O resultado do *hmmscan* mostrou a quantidade de *hits* para CAZymes em cada ambiente, permitindo uma análise quanto à distribuição das diferentes classes e famílias destas enzimas. Para determinar a diferença entre a contribuição das CAZymes encontradas no solo e na água o teste ANOVA um-fator foi aplicado, considerando o resultado estatisticamente significativo para $p<0.05$.

4.6 Análise taxonômica das CAZymes

Para realizar a análise taxonômica das *reads* associadas às CAZymes, a partir do resultado do *hmmscan* foram recuperadas apenas as sequências que apresentaram *hit* com os perfis HMM das enzimas. Tais sequências foram recuperadas a partir dos seus identificadores (ID’s) no arquivo multifasta baixado do MG-RAST e traduzido. Posteriormente, essas sequências foram comparadas contra a base de dados não redundante NR, utilizando o algorítmo BLASTP (ALTSCHUL et al. 1997), para identificação de sua origem taxonômica.

O resultado do BLASTP foi então computado utilizando o programa MEGAN5.6.3 (*MetaGenome Analyzer)* (HUSON et al. 2007). Este programa utiliza-se do resultado do BLAST contra bases de dados de referência (como NR, NT, ENV-NR, ENV-NT) para calcular a classificação taxonômica das *reads*, possibilitando também uma classificação do potencial funcional. A estimativa do perfil taxonômico das amostras é conseguida através de algorítmos, que classificam as *reads* a partir da sua associação a um determinado grupo na taxonomia do NCBI (NCBI taxonomy), baseando-se nos *hits* obtidos para sequências conhecidas, os quais são registrados no arquivo de saída do BLAST.

4.7 Identificação de possíveis novas CAZymes nos metagenomas

O reconhecimento de possíveis novas sequências codificantes para CAZymes foi conseguido através do emprego de dois códigos (*scripts*). O primeiro utilizou o arquivo FASTA dos metagenomas (traduzido) e os dados provenientes do BLASTP para gerar uma tabela, contendo as porcentagens de cobertura da *query* (sequência de interesse) e as porcentagens da identidade do alinhamento. A cobertura indica a porcentagem da sequência

oriunda da base de dados que é coberta por resíduos correspondentes da sequência *query* (www.alphalyse.com). A identidade do alinhamento, por sua vez, aponta a porcentagem de resíduos idênticos nas mesmas posições no alinhamento entre duas sequências (www.ncbi.nlm.nih.gov/). O segundo código foi utilizado para plotar os dados resultantes da tabela no programa R (R Development Core Team, 2008), permitindo a visualização gráfica da cobertura da *query* (%) vs. a identidade do alinhamento (%).

As etapas *in silico* foram executadas durante um treinamento na Unidade Multiusuários de Genômica Funcional e Estrutural (UFRJ). Os dados referentes aos itens 4.5, 4.6 e 4.7 foram analisados através de expressões regulares que constam no Apêndice desta dissertação.

5. RESULTADOS

5.1 Parâmetros físico-químicos das amostras de solo e de água

A tabela 1 mostra o resultado da análise físico-química para as amostras de solo. Os parâmetros revelam que o solo desta região da Chapada Diamantina pode ser classificado como muito fortemente ácido (pH 4.5-5.0) (USDA - The U.S. Department of Agriculture) e possui características mais arenosas. Em grande parte da porção leste da Chapada Diamantina, onde foram coletadas as amostras de solo, existem áreas nas quais predominam os Espodossolos (medianamente profundos, bem drenados, textura argilosa e fertilidade média) e os Latossolos (profundos, bem drenados, ácidos e de fertilidade baixa) (VELLOSO et al., 2002). Uma das amostras de solo apresentou valores mais discrepantes em relação às outras duas amostras coletadas, o que acarretou nos altos desvios-padrão encontrados para alguns dos parâmetros analisados. Estes desvios já eram de certa forma esperados devido à hetereogeneidade do ambiente de solo amostrado, como pôde ser observado na figura 6, na seção 4.1 da Metodologia.

A tabela 2, por sua vez, apresenta o resumo dos parâmetros físico-químicos obtidos para as amostras de água. Pode-se observar que, para os dados analisados, nenhuma das concentrações dos nutrientes que provavelmente têm origem/atividade antropogênica está acima do permitindo pela resolução nº 357 do CONAMA que dispõe, dentre outras coisas, sobre a classificação dos corpos de água e diretrizes ambientais para seu enquadramento.

5.2 Características gerais da anotação dos metagenomas

O sequenciamento do DNA resultou em aproximadamente 6.3 milhões de sequências a partir das amostras de solo, com um comprimento médio de 242pb (pares de bases) e 1.5 milhão de sequências a partir das amostras de água, com comprimento médio de 246pb (Tabela 3). Cerca de 97% das sequências submetidas a partir das amostras de solo, e 84% das submetidas a partir das amostras de água, mantiveram-se após a etapa de controle de qualidade do servidor MG-RAST.

Tabela 2. Parâmetros físico-químicos das amostras de solo

	pH	Al	H + Al	Ca	Mg	K	Na	Sb	CTC	V	m	P	Fe	Zn	Cu	Mn	C	N
	cmol / dm³									%		mg/Kg					g/dm³	
Amostras de solo (média)	4,46	1,23	10,66	0,73	0,4	0,09	0,03	1,26	11,93	10	55	5,66	20,76	2,1	0,26	5,63	23,36	1,47
Desvio padrão	0,058	0,493	6,048	0,611	0,436	0,047	0,026	1,097	7,023	3,606	22,539	4,509	19,714	1,500	0,058	6,301	13,427	0,645

	Areia grossa	Areia fina	Silte	Argila total	Argila natural	Silte/ Argila	Floculação	Densidade real
	g/kg						%	g/cm³
Amostras de solo (média)	651,66	127	178,66	42,66	16,66	4,19	60,75	2,64
Desvio padrão	20,65	21,52	5,51	2,31	2,31	0,13	6,75	0,03

Tabela 3. Parâmetros físico-químicos das amostras de água

	Ortofosfato (µM)	Fosfato total (µM)	Silicato (µM)	Amônia (µM)	Nitrito (µM)	Nitrato (µM)	Nitrogênio total (µM)
Média das amostras de água	0,238	0,339	0,556	1,56	0,279	8,026	44,150
Desvio padrão	0,031	0,011	0,009	0,052	0,002	0,115	0,892

Condutividade	DO		Temp	pH
uS/cm	%	mg/L	°C	
35	77,6	6,21	27,1°	4,1

Foram preditas um total de 1.96 milhão de proteínas para o solo (41.6% do total de *reads*), sendo que cerca de 76% destas proteínas foram anotadas em categorias funcionais. Para as amostras de água, foram preditas 362.729 proteínas (37% do número total de sequências), tendo 85% das sequências sido identificadas funcionalmente. Dentre todas as sequências analisadas, para ambos ambientes, mais de 55% não possuem similaridade às bases de dados de proteínas associadas ao servidor MG-RAST.

Tabela 4. Dados gerais da anotação dos metagenomas no servidor MG-RAST.

Amostras	Solo			Água	
	1 (4566398.3)	2 (4566399.3)	3 (4567013.3)	1 (4566739.3)	2 (4566680.3)
Contagem das sequências	903.004	2.531.434	2.094.312	748.085	707.145
Número de sequências após o controle de qualidade	894.648 (99.1%)	2.466.863 (97.4%)	1.984.723 (94.8%)	696,806 (93.1%)	595.678 (84.2%)
Comprimento médio das sequências (pb)	262 ±106	259 ±118	204±117	261±106	231±112
Sequências protéicas identificadas	363.839 (42.6%)	961.753 (43.2%)	636.058 (39.25%)	202.091 (38%)	160.638 (36.3%)
Sequências sem similaridade à base de dados do SEED	490.549 (57.4%)	1.264.869 (56.8%)	988.482 (60.8%)	329.876 (62%)	281.965 (63.7%)
Sequências protéicas anotadas em categorias funcionais	279.477 (76.8%)	735.710 (76.5%)	484.191 (76.1%)	171.693 (85%)	136,958 (85.3%)

5.3. Análise comparativa dos perfis funcionais

Os perfis funcionais observados para o solo e para a água se assemelharam em alguns aspectos - quando considerados de maneira mais global - porém diferiram à medida que uma análise mais detalhada foi conduzida. Este fato pode ser observado através da análise das duas

estratégias de anotação utilizadas e que serão apresentadas a seguir: a anotação baseada na tecnologia de subsistemas e a anotação pela abordagem de domínios conservados.

5.3.1 Comparação entre o solo e a água quanto à distribuição dos subsistemas

A anotação dos metagenomas no servidor MG-RAST, utilizando a base de dados do SEED, de subsistemas, permitiu realizar uma análise do potencial funcional das sequências obtidas do solo e da água, plotada através do programa STAMP. Foram identificados um total de 28 subsistemas nos metagenomas provenientes dos ambientes estudados, e observou-se que existem diferenças entre os subsistemas de nível 1 mais abundantes nos dois (Figura 7A).

Nas as amostras de solo, o subsistema de Carboidratos foi o mais abundante, com uma média de 16.89%, seguido dos subsistemas *Clustering-based* (aqueles que possuem genes que são encontrados próximos uns dos outros no genoma de diversos táxons, porém sua função ainda não é bem compreendida), com 13.81%. O subsistema de Aminoácidos e Derivados foi o terceiro que mais contribuiu, apresentando uma abundância de 11.18%. Para as amostras de água, houve uma inversão na ordem entre os mais abundantes: primeiramente têm-se os subsistemas *Clustering-based* (14.57%), seguido do subsistema de Carboidratos (11.55%) e então de Aminoácidos e Derivados, com 9.62% de abundância relativa.

Comparativamente, o solo apresentou uma abundância significativamente maior do subsistema de Carboidratos, assim como o de Respiração, Metabolismo de Compostos Aromáticos, Metabolismo Secundário, Metabolismo de Enxofre e Resposta ao Stress (Figura 7B). A água, por sua vez, mostrou-se como um ambiente com uma abundância significativamente maior de subsistemas como o de Metabolismo de RNA, Ciclo e Divisão Celular, Parede Celular e os subsistemas *Clustering-based*.

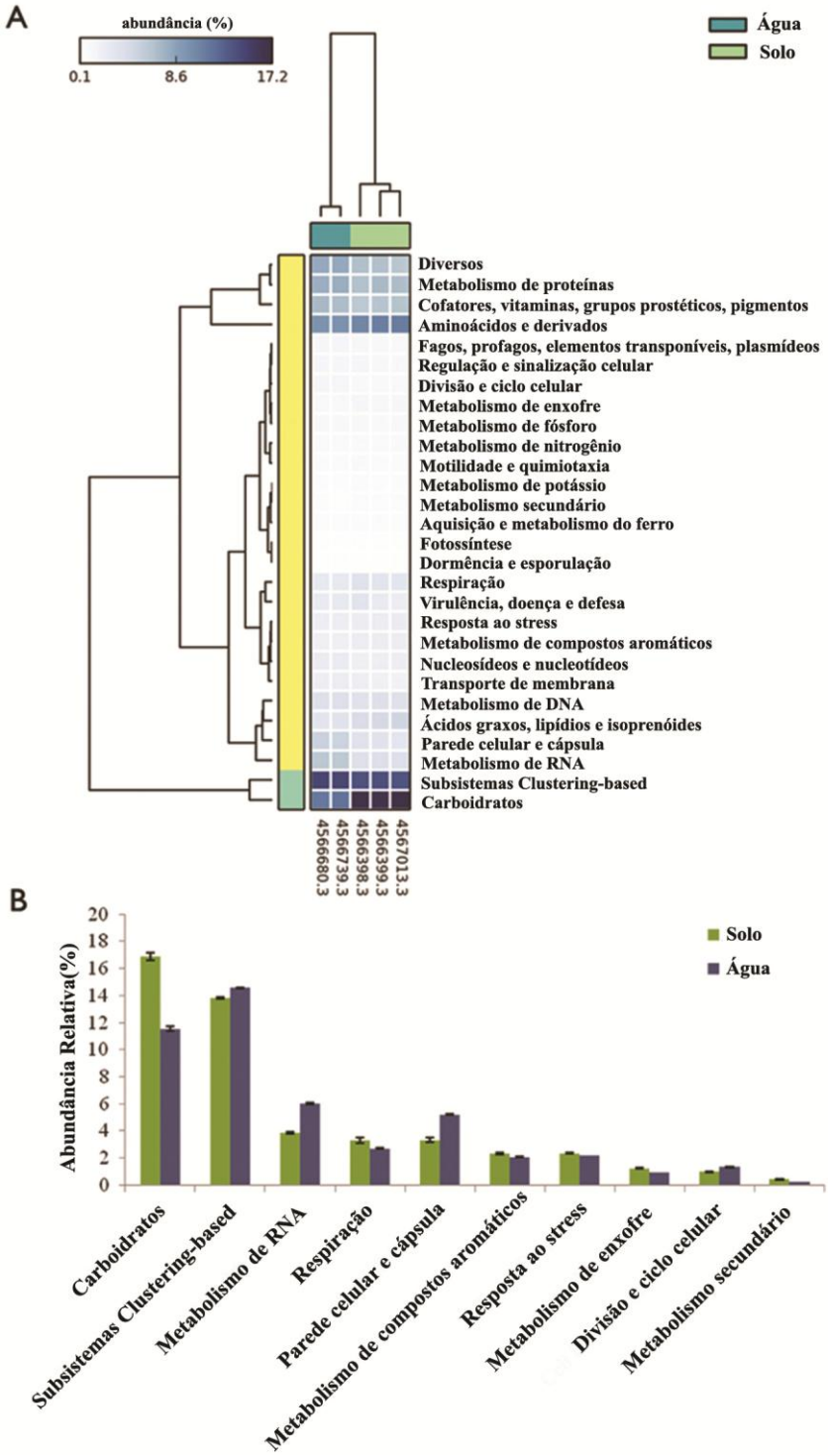


Figura 9. Comparação dos perfis funcionais do solo e da água pela abordagem de subsistemas. (A) O *heatmap* retrata os subsistemas de nível 1 encontrados nas amostras dos dois ambientes. As duas primeiras linhas verticais correspondem aos metagenomas da água e as três últimas, do solo. Os quadrados preenchidos em escala de azul representam a abundância de cada subsistema para cada metagenoma. O dendograma na parte superior mostra o agrupamento das amostras, de acordo com a distribuição dos subsistemas e o dendograma à esquerda mostra como os subsistemas estão agrupados, de acordo com sua coocorrência nos metagenomas. (B) Subsistemas de nível 1 significativamente diferentes entre o solo e a água ($p<0.05$), analisados pelo STAMP.

5.3.2 Comparação entre os dois ambientes quanto à contribuição das CAZymes

O subsistema de Carboidratos revelou-se como um componente de grande importância tanto para o ambiente de solo quanto para a água. A análise dos metagenomas quanto à distribuição de enzimas relacionadas ao metabolismo de carboidratos, as CAZymes, mostrou que existem diferenças na contribuição das classes e famílias dessas enzimas entre os dois ambientes. Foram encontradas um total de 10.081 *reads* associadas a CAZymes na água e 61.603 no solo. Nas figuras 8 e 9 é possível observar a representatividade de cada classe de CAZymes e das famílias pertencentes a estas classes, no solo e na água, respectivamente.

5.3.2.1 Análise comparativa entre o solo e a água para o nível classes

Em relação às classes, as glicosil hidrolases (GHs) foram as mais abundantes no solo, com 44.68% de abundância, seguida das glicosil transferases (GTs) com 32.15% (Figura 8). Com menor representatividade estão as classes das carboidrato esterases (CEs), com 9.77%, das atividades auxiliares (AAs), com 8.52% de abundância, dos módulos de ligação a carboidratos (CBMs), com 3.94% e, por fim, a classe das polissacarídeo liases (PLs), com 0.92% de abundância.

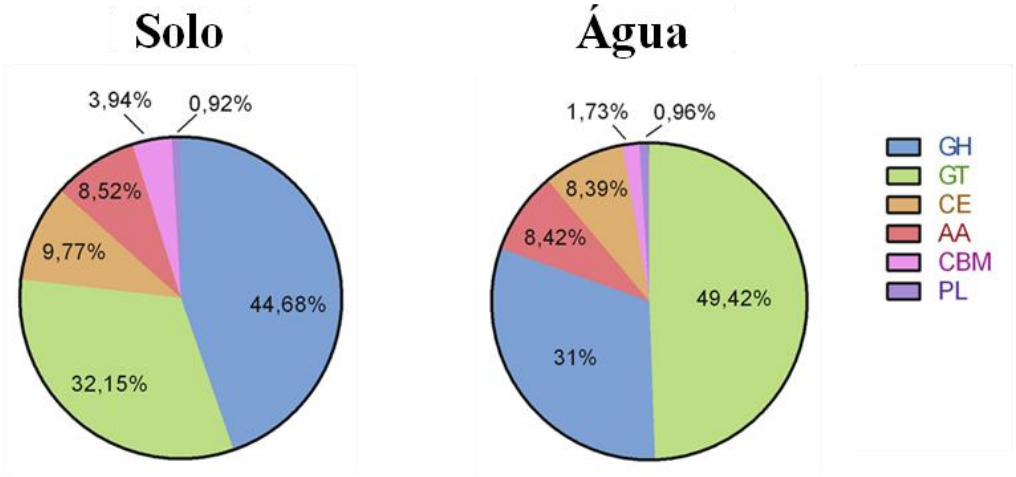


Figura 10. Composição das CAZymes no solo e na água. Os gráficos mostram a abundância relativa das 6 classes de CAZymes para cada um dos ambientes. GH (Glicosil Hidrolase); GT (Glicosil Transferase); CE (Carboidrato Esterase); AA (Atividades Auxiliares); CBM (Módulos de Ligação a Carboidratos) e PL (Polissacarídeo Liases). As classes GH e CBM foram significativamente mais abundantes no solo, enquanto que a classe GT teve representatividade maior na água ($p<0.05$), com aplicação da ANOVA um-fator (ambiente). Para as demais classes não houve diferença estatística.

Por outro lado, nas amostras de água, a classe de CAZymes que apresentou maior abundância foi a das GTs (49.42%), seguida das GHs (31%), AAs (8.42%), CEs (8.39%), CBMs (1.73%) e PLs (0.96%). As classes GH e CBM apresentaram abundância significativamente maior no solo em relação à água, enquanto que nesta, a representatividade das GTs foi significativamente maior ($p<0.05$). Para as demais classes, não houve diferença estatisticamente significativa entre os ambientes.

5.3.2.2 Comparação entre o solo e a água quanto às famílias de CAZymes encontradas

No que diz respeito às famílias de CAZymes, os gráficos de barras empilhadas da figura 9 mostram a fração das famílias encontradas, nos ambientes de solo e de água, que tiveram maior contribuição em termos de abundância relativa, para cada uma das 6 classes anteriormente descritas.

Em relação às CAZymes pertencentes à classe das glicosil hidrolases, foram encontradas um total de 110 famílias nos dois ambientes, com 27.676 *reads* identificadas no solo e 3.125 na água. Destas 110 famílias, estão sendo mostradas na figura 9 as 11 mais abundantes, que contribuem com aproximadamente 35% do total de GHs no solo e 45% na água. A família GH13 foi a que teve a maior representatividade dentre o total de GHs encontradas nas amostras de solo, cerca de 9.4% de abundância relativa, seguida de GH15 com 6.4% e GH3 com 5.1% de abundância. Já para as amostras de água, houve maior contribuição da família GH23 (com 9.3%), seguida de GH13, GH1 e GH3 com aproximadamente 6% cada.

Dentre as demais famílias de CAZymes encontradas, as pertencentes à classe das glicosil transferases estão mais relacionadas à biossíntese de carboidratos do que à sua degradação. Foram encontradas um total de 72 famílias de GTs nos ambientes (19.819 *reads* no solo e 4.977 na água), sendo que as 12 famílias mais frequentes (Figura 9), contribuíram com 80% do total de GTs no solo e 84% na água. As famílias mais abundantes em ambos foram GT4, GT2 e GT41. A família GT4 apresentou uma abundância de aproximadamente 18% do total de GTs no solo e 13% na água, enquanto que para GT2 a representatividade foi de 15% no solo e 12% na água e, por fim, GT41 com 14.1% no solo e 22.7% de abundância na água.

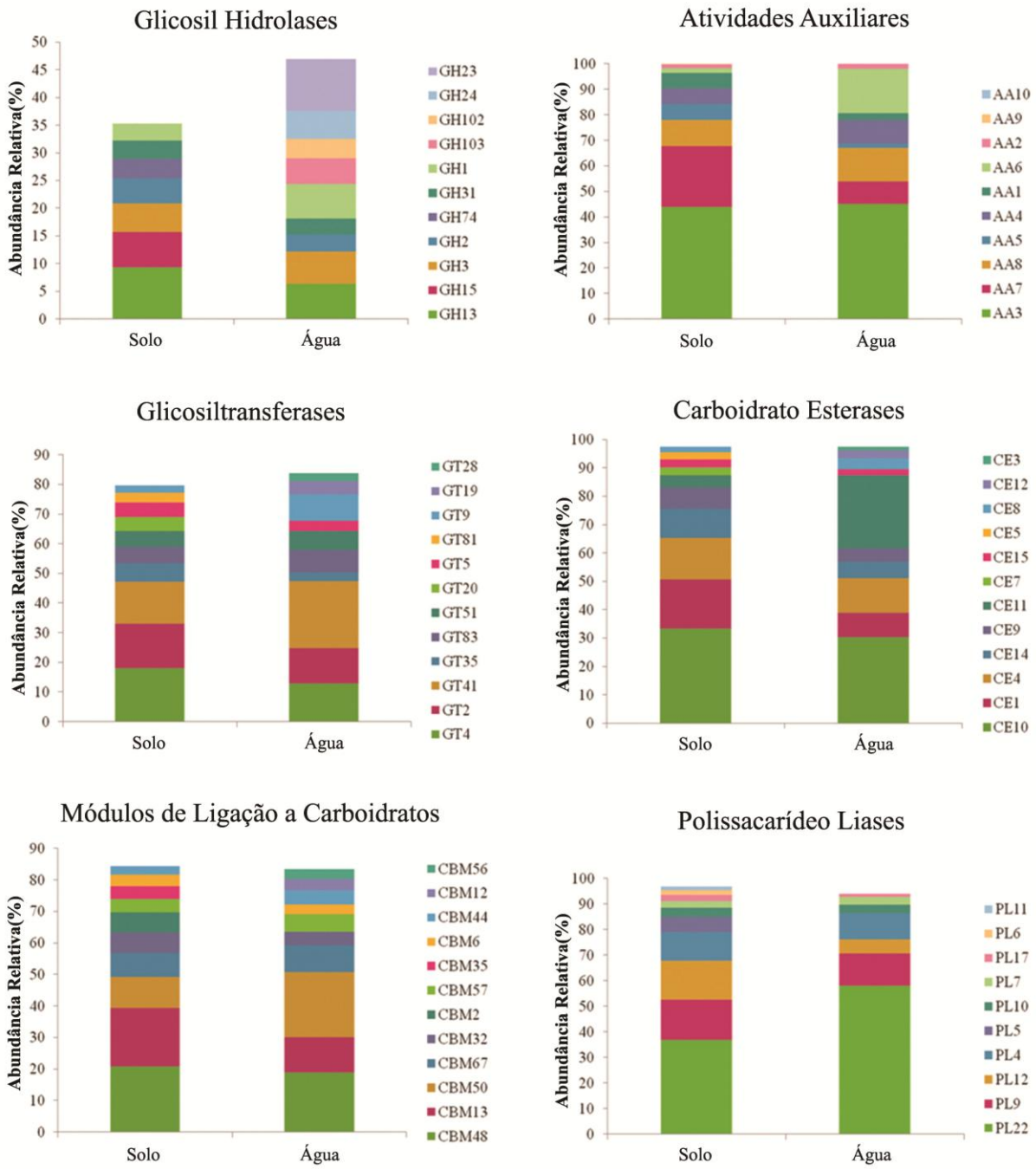


Figura 11. Contribuição das famílias mais abundantes de CAZymes, pertencentes a cada classe, para ambos os ambientes. Os gráficos de barras empilhadas mostram a média da abundância relativa encontrada para cada família, no solo e na água.

As carboidrato esterases são enzimas que catalisam a desacetilação de açúcares substituídos, ou seja, ésteres ou amidas nos quais o açúcar desempenha o papel do álcool e da amina (BIELY, 2012). Foram encontradas um total de 16 famílias de CEs (6.011 *reads* no solo e 852 na água) e, dentre estas, as 12 famílias mais frequentes somaram aproximadamente

95% do total das CEs identificadas. A família com maior contribuição nos dois ambientes foi a CE10 sendo que, no solo, sua contribuição dentre o total das carboidrato esterases foi de 33.4%, enquanto que na água foi de 30.3%. A família CE11 também teve grande representatividade na água, com uma abundância relativa de 25.7%.

As famílias da classe das atividades auxiliares compreendem enzimas que estão potencialmente envolvidas na conversão de material lignocelulósico, mas através de uma via oxidativa e não hidrolítica. Dez famílias de AAs foram encontradas no solo e na água (com a identificação de 5.079 e de 850 *reads*, respectivamente) e a que apresentou maior abundância em ambos os ambientes foi a família AA3, com 43.8% de abundância relativa no solo e 45% na água.

A classe dos complexos de ligação a carboidratos (CBM) compreende proteínas não catalíticas, que são capazes de se ligar a carboidratos cristalinos e solúveis, sendo geralmente encontradas com outras classes de CAZymes aumentando a atividade catalítica dessas enzimas. Os CBMs representaram um total de aproximadamente 4% das CAZymes no solo (com 2.457 *reads*) e 2% na água (174 *reads*), sendo distribuídos em 42 famílias. As 12 famílias mais abundantes contribuíram com 83.4 e 84.4% do total de CBMs no solo e na água, respectivamente. Dentre todas as famílias encontradas, as que obtiveram maior representatividade na água foram a CBM50, com 20.8% de abundância relativa, seguida de CBM48, com 18.9%. Já para o solo, a mais abundante foi a foi a CBM48, com 20.8%, seguida de CBM13, com 18.5%.

As polissacarídeo liases, que tiveram uma baixa representatividade tanto nos metagenomas de solo quanto da água (menos de 1% em cada ambiente), são enzimas que clivam certas ligações glicosídicas presentes em polissacarídeos acídicos (YIP; WITHERS, 2006). Dentre as 18 famílias de PLs encontradas (totalizando 549 *reads* no solo e 98 na água), as 10 mais abundantes representaram aproximadamente 97 e 94% das PLs detectadas, respectivamente, em cada ambiente. A família PL22 foi a mais abundante para os dois locais, com 36.8% de abundância no solo e 58% na água.

5.4 Comparação dos perfis taxonômicos

A caracterização taxonômica da estrutura da comunidade microbiana revelou, como esperado, que existem perfis distintos para o solo e para a água desta região de Caatinga, no

que diz respeito às classes de microorganismos que compõem estes ambientes. Mais ainda, a distinção entre os perfis encontrados pôde ser observada em dois níveis que serão apresentados a seguir: através da comparação entre as estruturas das comunidades microbianas totais e através da análise comparativa dos táxons que, especificamente, contribuíram para as sequências identificadas como CAZymes nos dois ambientes.

5.4.1 Análise comparativa da estrutura da comunidade microbiana total

Os perfis taxonômicos das comunidades microbianas do solo e da água, computados através do MG-RAST, utilizando como base de dados o GenBank, podem ser observados na figura 10. Estão sendo representadas as 11 classes mais abundantes em cada ambiente, que correspondem a aproximadamente 88% do total de classes encontradas no solo e 94% na água. Em relação ao solo, as classes com maior abundância foram Actinobacteria (pertencente ao filo homônimo) com 45.4%, seguida de Alphaproteobacteria, com 14.5%. Foram identificadas também as classes Solibacteres (6.2%), Acidobacteriia (5.9%), Betaproteobacteria (4.5%), Gammaproteobacteria (3.2%), Deltaproteobacteria (2.9%), Ktedonobacteria (1.6%), Planctomycetia (1.8%), Clostridia (1.32%) e Sphingobacteriia (0.28%).

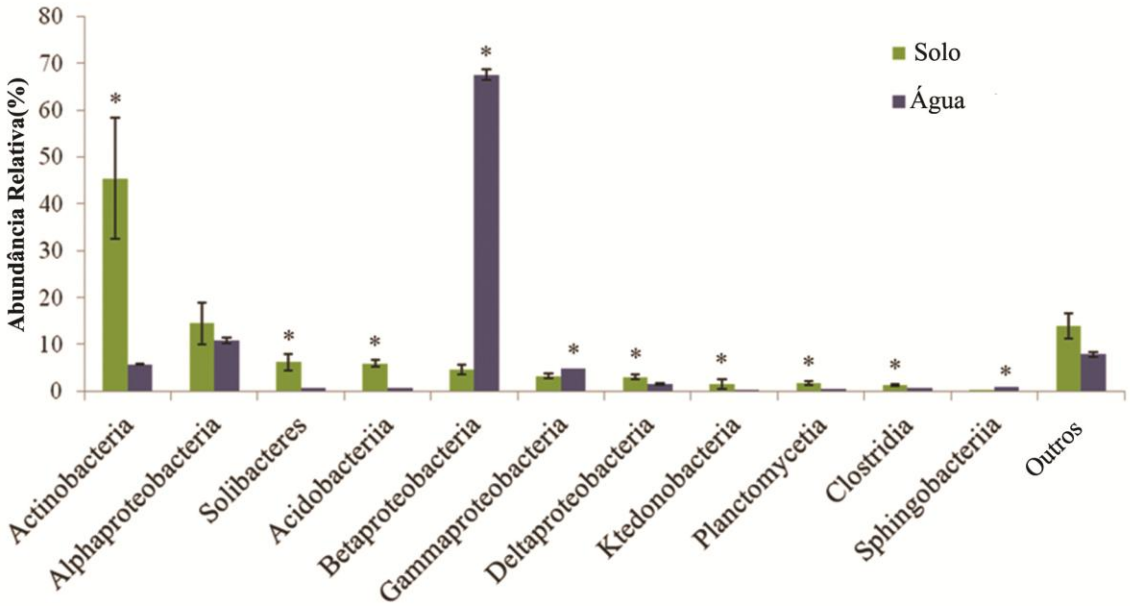


Figura 12. Perfil da comunidade microbiana total nos ambientes de solo e de água para o nível taxonômico Classe. As sequências foram anotadas através do servidor MG-RAST, utilizando o GenBank como base de dados, e analisadas pelo STAMP ($p<0.05$).

Já para as amostras de água, a classe com grande predominância foi Betaproteobacteria, pertencente ao filo Proteobacteria, com 67.5% de abundância relativa. Foram também encontradas nas amostras de água as classes Alphaproteobacteria (10.8%), Actinobacteria (5.7%), Gammaproteobacteria (4.8%), Deltaproteobacteria (1.5%), Sphingobacteriia (0.81%), Clostridia (0.75%), Acidobacteriia (0.7%), Solibacteres (0.7%), Planctomycetia (0.53%) e Ktedonobacteria (0.1%). Apenas a classe Aplhaproteobacteria não foi significativamente distinta entre os ambientes. Em relação às demais, o solo teve uma contribuição maior de Actinobacteria, Solibacteres, Acidobacteriia, Deltaproteobacteria, Ktedonobacteria, Planctomycetia e Clostridia. Na água, por sua vez, as classes Betaproteobacteria, Gammaproteobacteria e Sphingobacteriia foram significativamente mais representativas ($p<0.05$).

5.4.2 Comparação dos perfis das comunidades microbianas associadas às CAZymes

A realização do BLASTP entre as sequências de CAZymes e a base de dados NR (não-redundante) permitiu identificar a origem taxonômica destas enzimas, nas amostras de solo e de água (Figura 11). O táxon com maior abundância relativa no ambiente de solo foi Planctomycetia (29%), enquanto que para a água, houve uma maior representatividade de Alphaproteobacteria (27%). Foram também associadas às CAZymes as classes Cytophagia (com 17% de abundância no solo e 11% na água), Spirochaetia (16% no solo e 18% na água), Deltaproteobacteria (com 10% e 7% respectivamente), Betaproteobacteria (com 6% e 12%), Chloroflexia (com 3% e 1%), Flavobacteriia (com 1% e 2%) e Gammaproteobacteria, com 1% de abundância relativa em ambos os ambientes.

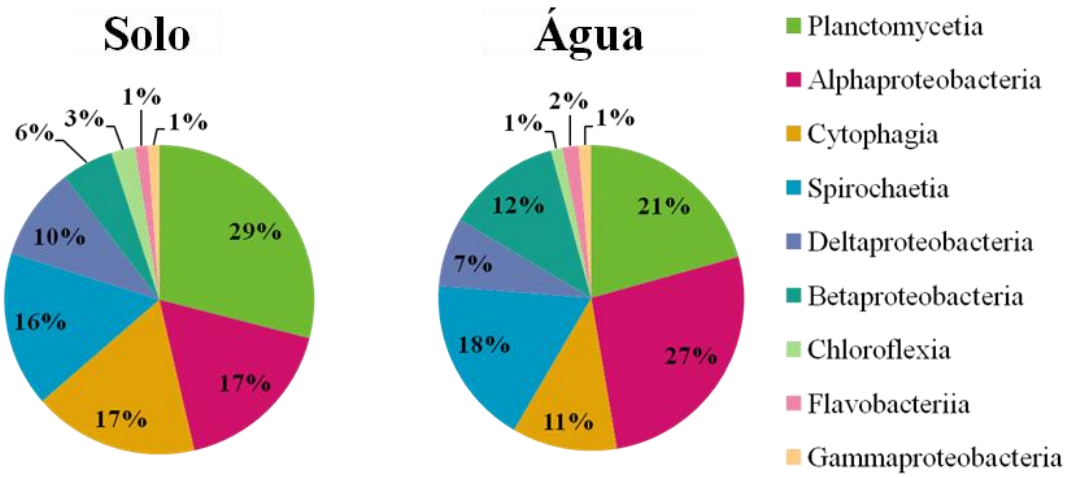
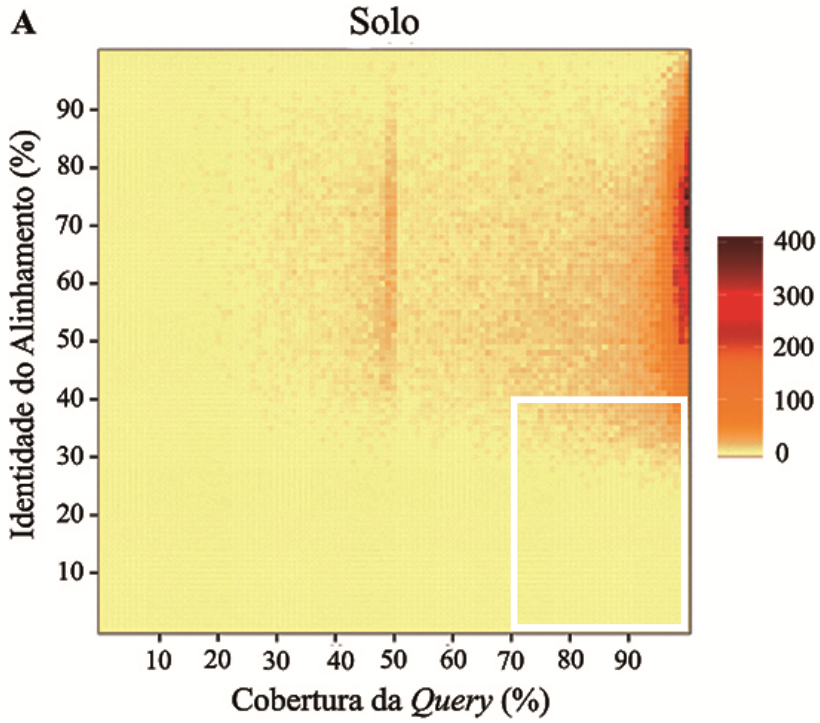


Figura 13. Contribuição dos táxons (Classes) para as sequências identificadas como CAZymes nos ambientes do solo e da água. Foi realizado um BLASTP entre as sequências de CAZymes encontradas e a base de dados NR (não-redundante), permitindo a identificação dos táxons. O resultado foi analisado utilizando o *software* MEGAN.

5.5 Identificação de possíveis novas sequências de CAZymes no solo e na água

A comparação dos dados das sequências metagenômicas do solo e da água, com as sequências protéicas de referência da base de dados NR, através do BLASTP, permitiu reconhecer possíveis novas CAZymes nos metagenomas (Figura 12A e B).

De uma maneira geral, para ambos os ambientes, a maioria das sequências de CAZymes recuperadas possuem uma alta identidade (maior do que 50%), em relação às sequências de aminoácidos da base de dados analisada. Entretanto, dentre este total de *reads* identificadas para as CAZymes, no solo e na água, foram encontradas sequências aparentemente pouco relacionadas às já existentes. Tais sequências, que quando comparadas àquelas depositadas na base de dados NR, apresentaram uma alta cobertura (>70%), porém uma baixa identidade (<40%), podem, possivelmente, representar novas CAZymes.



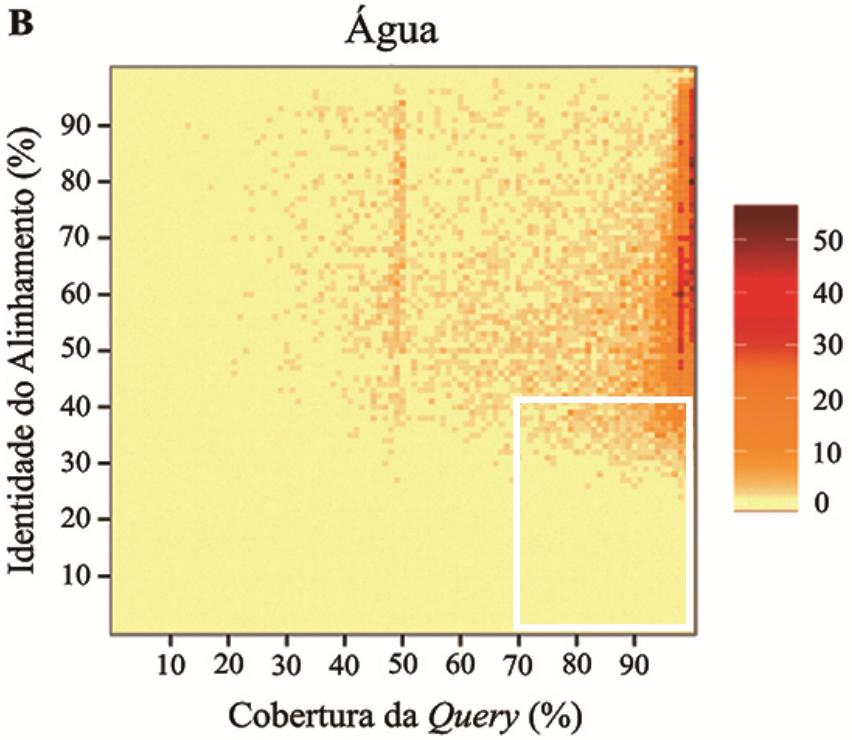


Figura 14. Possíveis novas sequências de Cazymes identificadas nos metagenomas do solo (A) e da água (B). A realização do BLASTP contra a base de dados NR permitiu inferir a existência de novas sequências putativas para CAZymes, com base na relação entre a identidade do alinhamento e a cobertura das sequências-alvo (*query*). As sequências que apresentaram uma alta cobertura (>70%), porém baixa identidade (<40%), foram atribuídas a possíveis novas sequências de CAZymes (região evidenciada em branco). A escala numérica corresponde à quantidade de sequências presentes em cada ponto no gráfico, de maneira que uma maior intensidade de cor significa um maior número de sequências, num determinado ponto considerado.

6. DISCUSSÃO

Embora o solo seja um dos principais focos da metagenômica funcional (HANDELSMAN, 2005), ainda são parcos os trabalhos que visam a bioprospecção de enzimas do solo da Caatinga brasileira. Um cenário ainda mais restrito ocorre para ambientes de água doce - incluindo os da Caatinga - nos quais até mesmo a caracterização da diversidade microbiana ainda é subexplorada. Segundo Wetzel (2000), menos de 10% dos dados limnológicos disponíveis dizem respeito aos vírus, bactérias, fungos e protistas e seus processos metabólicos nestes ambientes aquáticos. Uma das evidências de que o conhecimento sobre microorganismos de água doce é ainda inferior, em relação aos marinhos, reside na escassez de sequências depositadas em bases de dados para bactérias de água doce (GHAI et al., 2011).

No presente estudo, mais da metade das sequências obtidas tanto no ambiente do solo como da água foram classificadas como sequências desconhecidas, através da anotação no MG-RAST. Esta baixa similaridade das sequências com as bases de dados de referência do servidor já é um fato esperado, e deve-se provavelmente ao viés associado a estas, fator que em última instância reflete a escassez de sequências genômicas microbianas depositadas para diversos tipos de ambientes. O viés associado às bases de dados reside no fato de que estes dependem de sequências de origem e função gênicas já descritas, as quais, muitas vezes, pertencem a microorganismos modelo ou cultiváveis (SIMON; DANIEL, 2009). Sendo assim, por não encontrarem sequências similares nas bases de dados, até 90% das *reads* de um conjunto de dados metagenômicos podem permanecer desconhecidas (HUSON et al., 2007).

6.1 Comparação dos perfis funcionais quanto à representatividade dos subsistemas

Para as sequências que foram atribuídas a categorias funcionais, a estratégia de anotação pela tecnologia de subsistemas permitiu observar um perfil do potencial funcional com algumas similaridades entre o solo e a água, mas também podem ser destacadas particularidades. Os subsistemas mais abundantes nos dois ambientes foram os de Carboidratos e os *Clustering-based* (sendo que o primeiro foi o mais representativo no solo, e o segundo, na água). Em relação ao solo, estes resultados se assemelham aos encontrados por Delmont e colaboradores (2012), que realizaram um estudo metagenômico numa região de

pastagem natural. Dentro do nível hierárquico mais abrangente dos subsistemas (nível 1), os subsistemas *Clustering-based* e o de Carboidratos foram os que obtiveram a maior quantidade de *reads* associadas, cerca de 17 e 14%, respectivamente. Pacchioni e colaboradores (2014) também encontraram os subsistemas *Clustering-based* e de Carboidratos como mais representativos, tanto nos solos da Mata Atlântica quanto da Caatinga que foram analisados, não tendo havido uma diferença discrepante entre as proporções desses subsistemas nos dois ambientes.

Para Delmont e colaboradores (2012), a relativamente alta abundância dos subsistemas *Clustering-based* nas amostras sequenciadas sugere que eles têm um papel chave nos ecossistemas de solo ao redor do mundo, e devem ser explorados no intuito de melhor entender a composição destes ecossistemas.

Dinsdale e colaboradores (2008), por sua vez, realizaram um estudo de metagenômica comparativa com quase 15 milhões de sequências de 9 biomas distintos, incluindo amostras marinhas, de lagoas hipersalinas e de água doce, entre outros. Os autores encontraram que a maior porcentagem das sequências, em média, estava associada ao subsistema de Carboidratos (17.218%), seguido do subsistema de Aminoácidos e Derivados (12.036%), numa análise combinada dos biomas. Resultado semelhante foi encontrado por Breitbart e colaboradores (2009), ao compararem as funções metabólicas dos microbialitos com as de ecossistemas de água doce e marinhos. O subsistema de Carboidratos seguido do de Aminoácidos e Derivados apresentaram a maior porcentagem de sequências associadas (aproximadamente 18 e 14%), tanto no metagenoma dos microbialitos quanto nos metagenomas provenientes da água doce e marinhos.

A análise comparativa entre os ambientes mostrou que os subsistemas *Clustering-based* e os relacionados ao Metabolismo de RNA, Ciclo e Divisão Celular e Parede Celular e Cápsula foram significativamente mais frequentes nas amostras de água, enquanto que os subsistemas de Carboidratos, de Metabolismo Secundário, de Resposta ao Stress, Metabolismo de Enxofre, Respiração e Metabolismo de Compostos Aromáticos foram mais abundantes no solo. As condições mais arenosas do solo desta região Caatinga - que leva a uma baixa retenção hídrica e de nutrientes - a alta incidência de radiação UV e a existência de períodos de estiagem podem explicar a maior representatividade, no solo, dos genes relacionados à resposta ao stress, ao metabolismo secundário e à respiração.

A atividade dos microorganismos no solo é também influenciada por outros fatores como por exemplo, as plantas (GRAYSTON et al., 1998). As raízes das plantas liberam uma

grande variedade de compostos que incluem etileno, açúcares, vitaminas, aminoácidos, ácidos orgânicos, polissacarídeos e enzimas (GARBEVA; VAN VEEN; VAN ELSAS, 2004). Estes e outros exsudatos podem ser utilizados como fonte de carbono pelos microorganismos (BULGARELLI et al., 2013), o que pode explicar a maior ocorrência do subsistema de Metabolismo de Compostos Aromáticos no solo em relação à água. A maior presença de matéria orgânica proveniente da vegetação pode explicar a maior contribuição do subsistema de Carboidratos no solo, no qual, ao realizar uma análise mais detalhada, sobressaíram-se os subsistemas de nível 2 relacionados à Fermentação, ao Metabolismo de Polissacarídeos e de outros açúcares (dados não mostrados).

No que diz respeito às amostras de água, segundo GHAI e colaboradores (2014), apesar da importância ecológica e de saúde pública da microbiota de água doce, estes microorganismos têm recebido relativamente pouca atenção dos microbiologistas. Tal fator pode estar relacionado à maior abundância dos subsistemas *Clustering-based* (que possuem genes acoplados funcionalmente, porém com função ainda desconhecida) nesse ambiente. No presente estudo, a classe predominante nas amostras de água foi Betaproteobacteria e, dentro desta, o gênero *Polynucleobacter* teve a maior representatividade, com abundância de 52% considerando apenas os gêneros pertencentes à classe e de 34.2% no total de gêneros encontrados (dados não mostrados). A grande contribuição deste gênero pode explicar a maior abundância dos subsistemas de Metabolismo de RNA (com genes envolvidos no processamento e modificação do RNA), Parede Celular e Cápsula (com genes relacionados aos componentes de paredes celulares gram-negativas) e Ciclo e Divisão Celular. De acordo com Wu e Hahn (2006), *Polynucleobacter* representa um grupo de distribuição cosmopolita em ambientes de água doce como rios e lagos. A espécie *Polynucleobacter necessarius,* por exemplo, possui organismos tanto endossimbiontes obrigatórios (*P. n. necessarius)* quanto bactérias de vida livre (*P. n. asymbioticus*) que são heterotróficas, gram-negativas e de habitat planctônico, podendo compreender de <1% a 70% do total do bacterioplâncton em ambientes de água doce (HAHN et al., 2012; WU; HAHN, 2006). Segundo HAHN (2003), uma das características necessárias para o estabelecimento de populações de bactérias de vida livre, que contribuem para uma grande fração do bacterioplâncton, é seu menor tamanho. Os autores descreveram que as bactérias da espécie *Polynucleobacter necessarius* de vida livre possuem um menor tamanho em relação a outras bactérias (<0.1 μm^3), o que permite minimizar sua vulnerabilidade frente a predadores e pode explicar a maior representatividade destes microorganismos de tamanho reduzido no bacterioplâncton. Num estudo conduzido

por Boscaro e colaboradores (2013), foi realizada uma análise comparativa entre as subespécies de *Polynucleobacter necessarius* e, para a subespécie de vida livre, características como o crescimento lento, um genoma compacto, especializado e de flexibilidade metabólica reduzida foram relatadas. Essas características podem explicar os resultados encontrados neste estudo: a maior representatividade, na água doce, de subsistemas de metabolismo de RNA (relacionados ao seu processamento e modificação para uma maior especialização metabólica); dos subsistemas de Parede Celular e Cápsula (paredes gram-negativas características do gênero) e dos subsistemas de Ciclo e Divisão Celular, devido ao crescimento lento destas bactérias que apresentam dominância populacional na água doce.

6.2 Comparação do potencial funcional quanto à contribuição das CAZymes no solo e na água

Os resultados da anotação pela tecnologia de subsistemas confirmaram que o metabolismo de carboidratos é um subsistema importante na manutenção das comunidades microbianas do solo e da água em diversos ecossistemas, evidência que pode ser observada pela sua representatividade tanto nas amostras analisadas quanto nos estudos comparados.

Considerando o metabolismo de carboidratos e dada a importância biotecnológica da degradação de matéria orgânica, em especial a de material lignocelulósico, a segunda estratégia de anotação realizada permitiu identificar a composição das CAZymes em dois níveis hierárquicos de organização que serão em seguida apresentados: as classes, que compõem o nível mais geral e abrangente e as famílias, que reúnem-se de maneira mais específica com base na similaridade de sequências e de estruturas protéicas (www.cazy.org).

6.2.1 Representatividade das diferentes classes de CAZymes nos ambientes

A análise de sequências putativas de CAZymes nos metagenomas revelou a maior presença da classe das glicosil hidrolases no solo (~45%) e das glicosiltransferases na água (~50%). As GHs já foram encontradas como a classe de CAZymes significativamente mais abundante no horizonte orgânico do solo (~33%), em relação ao mineral (~28%), em uma plantação de pinheiros, revelando a especialização dos microrganismos habitantes desse horizonte na degradação de moléculas de carboidratos complexas (UROZ et al., 2013).

De fato, as glicosil hidrolases compreendem um grande grupo de enzimas que, ao possibilitarem o metabolismo de polissacarídeos como o amido, a celulose, o xilano e a quitina, viabilizam o crescimento bacteriano no ambiente (BERLEMONT; MARTINY, 2015). Devido à abundância e ampla distribuição dos genes codificantes pra GHs nos genomas, a classe se tornou a mais bem caracterizada, dentre o conjunto de enzimas que compreendem as CAZymes (CANTAREL et al, 2009). Diversos estudos têm mostrado o potencial biotecnológico deste amplo grupo de enzimas, como a aplicação na indústria alimentícia, na produção de aditivos para alimentação animal, na fabricação de papel e fibras, de detergentes e, em especial, na degradação de biomassa lignocelulósica para produção de biocombustíveis (HIMMEL et al., 2010; LYND et al., 2002; PAËS; BERRIN; BEAUGRAND, 2012; PHITSUWAN et al., 2012; SATHYA; KHAN, 2014; STROOBANTS; PORTETELLE; VANDENBOL, 2014; WILSON, 2009).

Além das GHs, os CBMs também foram significativamente mais frequentes no solo em relação à água. Esta classe de CAZymes possui módulos não-catalíticos, os quais interagem e ligam-se a polissacarídeos como o amido, a celulose e o xilano (OH et al., 2015). O principal papel dos CBMs é o de reconhecer e ligar-se de maneira específica aos carboidratos, o que pode resultar em diferentes funções como aumentar a hidrólise de substratos insolúveis, aproximar o domínio catalítico do substrato e promover a ruptura da estrutura polissacarídica (BORASTON et al., 2004; GUILLÉN; SÁNCHEZ; RODRÍGUEZ-SANOJA, 2010). A maior presença desta classe de CAZymes no solo em detrimento à água faz sentido dada à sua ação conjunta com as glicosil hidrolases, que também foram dominantes no solo. Os CBMs podem associar-se aos carboidratos complexos presentes no solo, permitindo um aumento da eficiência catalítica das glicosil hidrolases na degradação destas moléculas.

As glicosiltransferases representaram aproximadamente 50% do total de CAZymes encontradas na água, sendo o único grupo cuja contribuição foi significativamente maior nos metagenomas da água em relação ao solo. Esta classe de CAZymes compreende enzimas responsáveis pela biossíntese de carboidratos complexos a partir de doadores de açúcares ativados, os quais podem ser açúcares ligados a nucleotídeos (a maioria), a fosfatos ou fosfatos lipídicos (LAIRSON et al., 2008). Os aceptores da reação podem ser carboidratos, proteínas, lipídios, DNA e diversas moléculas menores como antibióticos e esteróides, o que faz com que seus produtos compreendam a maior classe de compostos naturais encontrados na natureza (PALCIC, 2011; WEADGE; PALCIC, 2008). Os resultados obtidos no presente trabalho indicam que o ambiente de água doce desta região de Caatinga possui um maior

potencial de biossíntese de compostos sacarídicos e de glicosilação de pequenas moléculas, em detrimento ao potencial de clivagem e degradação da matéria orgânica, que foi melhor observado no solo, devido à predominância de GHs.

Juntamente com as GHs, as polissacarídeo liases são enzimas que também participam do processo de degradação de polissacarídeos (HIMMEL et al., 2010). No entanto, as PLs clivam os polissacarídeos por um mecanismo de eliminação ao invés da hidrólise, ou seja, a quebra catalisada pelas PLs ocorre sem a intervenção de moléculas de água (LOMBARD et al., 2010). Apenas uma diminuta fração de PLs foi encontrada no solo e na água (menos de 1%), o que indica a pequena contribuição deste mecanismo de degradação em ambos os ambientes analisados.

As carboidrato esterases compõem, por sua vez, um grupo de CAZymes que remove substituintes ésteres das cadeias de glicano e assim facilitam a ação das GHs e PLs (CANTAREL et al., 2009; KAOUTARI et al., 2013). Estas enzimas parecem ter evoluído no sentido de contornar a complexidade da parece celular vegetal e por isso podem auxiliar as glicosil hidrolases na conversão de biomassa (BIELY, 2012). No presente estudo, as CEs representaram o 3º e o 4º grupo de CAZymes com maior contribuição no solo e na água, respectivamente, não havendo, no entanto, diferença estatística entre eles.

As atividades auxiliares compõem uma classe de CAZymes mais recentemente descrita, que agrupa módulos catalíticos envolvidos na degradação de material lignocelulósico por uma via oxidativa (LEVASSEUR et al., 2013). Integram as AAs as famílias de enzimas redox e de LPMOs (monooxigenases líticas de polissacarídeos), com habilidade em potencial para auxiliar as GHs, PLs e CEs no acesso aos carboidratos presentes na parede celular vegetal (LEVASSEUR et al., 2013; STROOBANTS; PORTETELLE; VANDENBOL, 2014). Este grupo apresentou entre 8 e 8.5% do total de CAZymes encontrado no solo e na água, respectivamente, o que não garante diferença significativa entre os ambientes, mas aponta a contribuição destas enzimas no processo degradativo lignocelulósico.

6.2.2 Contribuição das CAZymes no solo e na água em termos de famílias

Numa análise mais detalhada do perfil de CAZymes encontrados no solo e na água, foi possível identificar as famílias que mais contribuíram para o metabolismo de carboidratos nestes ambientes. Em relação à classe das glicosil hidrolases, as famílias mais abundantes nas

amostras de solo foram, em ordem decrescente, GH13, GH15 e GH3, enquanto que nas de água foram GH23, GH13, GH1 e GH3.

A família GH13, também conhecida como a família das α-amilases, é uma das mais amplas famílias e contém enzimas com diferentes atividades e especificidades ao substrato, que agem, por exemplo, sobre o amido, o glicogênio e oligo e polissacarídeos relacionados (como as α-amilases, α-glicosidases, pululanases, entre outras) (KUMAR, 2010; STAM et al., 2006). Já foi relatada na literatura a importância biotecnológica desta família de enzimas dada a sua aplicação em diversas indústrias, como por exemplo, a de alimentos (SÁNCHEZ; CARDONA, 2008; SATHYA; KAHN, 2014), têxtil (HAQ et al., 2010) e também na produção de bioetanol de primeira geração (ELLEUCHE et al., 2014;). Segundo Li e colaboradores (2009), o fato da família GH13 ser a mais ampla dentre as famílias de GHs, e englobar diversas atividades enzimáticas e distintas especificidades ao substrato, pode explicar o fato de essa ser a família dominante em diversos metagenomas de diferentes ambientes.

As enzimas da família GH15 possuem atividades de glicoamilase, glicodextranase e α,α-trealase (www.cazy.org), e podem também estar envolvidas no processamento do amido (BERTOLDO; ANTRANIKIAN, 2002; MARÍN-NAVARRO; POLAINA, 2011). As glicosil hidrolases da família GH23 possuem como atividades conhecidas as lisozimas tipo G, peptideoglicano liases e quitinases. As lisozimas e quitinases representam importantes grupos de enzimas que hidrolizam polissacarídeos, estando a quitinase responsável pela quebra da quitina e a lisozima pela ruptura do peptideoglicano da parede celular bacteriana (MONZINGO et al., 1996; WOHLKÖNIG et al., 2010).

Enquanto as lisozimas da família GH23 possuem origem eucariótica e atividade sobre a quitina e quitinoligossacarídeos (além do peptideoglicano), as peptideoglicano liases são de origem bacteriana e bacteriófaga, não sendo reconhecida outra atividade, além da clivagem do peptideoglicano, a qual ocorre sem a intervenção de moléculas de água (www.cazy.org). Apesar da família GH23 conter reconhecidamente apenas as lisozimas e peptideoglicano liases (também chamadas de transglicosilases líticas), um estudo recente identificou uma celulase pertencente ao grupo. A proteína até então hipotética *Ct*Cel124, de *Clostridium thermocellum*, foi caracterizada como uma endoglucanase, cujo sítio ativo exibiu uma estrutura conservada com enzimas da família GH23 (BRÁS et al., 2011). Este estudo ratifica o quanto apenas uma porção da capacidade celulolítica microbiana tem sido estudada e

anotada, e que a busca por potenciais celulases eficientes - pertencentes a outras famílias além das até então conhecidas - deve ser aumentada (SUKHARNIKOV et al., 2011).

As enzimas da família GH1 possuem como atividades enzimáticas mais comuns a β-glicosidase e β-galactosidase, sendo reconhecidas também as atividades β-manosidase, β-D-fucosidase e outras (WARNECKE et al., 2007). Membros desta família são responsáveis por clivar ligações β-glicosídicas em celooligossacarídeos e outros substratos para gerar, por exemplo, glicose (HILL; RAILLEY, 2008). As β-glicosidases são frequentemente componentes dos celulossomas, mas podem também ser secretadas, exercendo função na conversão de biomassa vegetal (CAIRNS; ESEN, 2010). Além do potencial lignocelulolítico da família, sua aplicabilidade na indústria alimentícia também já foi reportada: as β-galactosidases da família GH1 podem ser aplicadas, por exemplo, para produção de leite com baixo teor de lactose e na produção de galacto-oligossacarídeos (SATHYA; KAHN, 2014).

A família GH3 compreende membros que possuem atividades enzimáticas distintas, incluindo além das β-D-glicosidases, as β-D-xilopiranosidases, α-L-arabinofuranosidases e N-acetil-β-D-glicosaminidases (www.cazypedia.org). As enzimas deste grupo realizam uma série de funções que envolvem, por exemplo, a degradação da biomassa celulósica, a renovação de componentes da parede celular e a modificação de algumas moléculas como antibióticos (FAURE, 2002). Li e colaboradores (2009) conduziram uma pesquisa com 46 metagenomas publicamente disponíveis e, ao realizarem um *blast* destes metagenomas contra as sequências de CAZymes, conseguiram recuperar 7.338 homólogos putativos de glicosil hidrolases. A família GH3 apareceu dentre as 5 mais abundantes famílias de GHs na maioria dos metagenomas analisados, incluindo o microbioma do intestino humano (11%), de ratos (10%) e de cupins (8%); a microbiota de comunidades marinhas (7%), de um lodo australiano (7%) e até em uma amostra de ar (9%). A família GH3 apareceu também dentre as mais abundantes em amostras metagenômicas do microbioma do rúmen de bovinos, com >700 sequências associadas (BRULC et al., 2009) e com contribuição de 14.9% do total de GHs detectadas (WANG et al., 2013). Resultado semelhante foi encontrado para o intestino grosso de cupins, nos quais foram identificados 69 módulos correspondentes aos domínios catalíticos de GH3, sugerindo que o metabolismo final de oligossacarídeos ocorre através da atividade dos módulos de glicosidase desta família (WARNECKE et al., 2007). No presente estudo, a GH3 foi a 3ª família de glicosil hidrolases mais abundante no solo e a 4ª na água, o que pode indicar o potencial de degradação lignocelulósica destes locais, se comparado a metagenomas de ambientes caracterizados por uma constante entrada e ciclagem de biomassa vegetal

complexa, como é o caso dos metagenomas de comunidades microbianas derivadas do intestino de cupins, humanos e ratos e do rúmen de bovinos, por exemplo (LI et al., 2009).

Devido à ampla variedade de funções dos carboidratos, várias proteínas atuantes nestas moléculas adquiriram os CBMs, que são módulos não-catalíticos capazes de interagir especificamente com mono, oligo e polissacarídeos (GUILLÉN; SÁNCHEZ; RODRÍGUEZ-SANOJA, 2010). De acordo com Boraston e colaboradores (2004), apesar de muitos destes módulos associarem-se à parede celular da planta, existem diversas famílias de CBM relacionadas à ligação a polissacarídeos de armazenamento tais como o glicogênio, o que é o caso da família CBM48. Tal família, que foi a mais abundante nas amostras de solo, contém módulos de aproximadamente 100 resíduos, que possuem a função de ligação ao glicogênio e são anexos aos módulos GH13 (www.cazy.org). Tal resultado é consistente com o fato de a família GH13 ter sido a família de glicosil hidrolases mais abundante nas amostras de solo.

A família de CBMs com maior contribuição nas amostras de água, por sua vez, foi a CBM50, cujos módulos têm a propriedade de ligar-se ao peptideoglicano e à quitina. Os CBM50 podem ser encontrados associados a 6 famílias de GHs, incluindo a GH23. Esta família foi a mais abundante dentre as GHs nas amostras de água, o que pode também explicar a predominância dos módulos de CBM50 neste ambiente.

A família CBM13 foi a segunda mais representativa no solo e terceira na água. A família, apesar de promover ligações fracas a mono, di e oligossacarídeos, tem, na verdade, grande afinidade pelo xilano (BORASTON et al., 2000). Segundo Boraston e colaboradores (2004), CBM13 parece ser mais frequente em toxinas bacterianas ou enzimas (glicosil hidrolases e glicosiltransferases) que atacam superfícies de células eucarióticas ou a matriz de glicanos. A representatividade de CBM13 nas amostras de solo e de água também pode estar associada à maior contribuição de uma das famílias de glicosil hidrolases, a GH3, cujas atividades descritas envolvem a hidrólise do xilano.

Ao contrário do potencial de quebra de ligações glicosídicas associado às famílias de glicosil hidrolases, podendo, para isso, ter o auxílio dos CBMs, as glicositransferases são enzimas que estão mais relacionadas à formação destas ligações glicosídicas. Os resultados obtidos no presente trabalho revelaram que o ambiente de água possui maior potencial de biossíntese de carboidratos do que o solo. A família que mais contribuiu para este cenário foi a família GT41. Esta família possui duas atividades conhecidas, a UDP-GlcNAc: peptídeo β-N-acetilglicosaminatransferase e UDP-Glc: peptídeo β-N-glicosiltransferase. Até 2013, apenas 3 estruturas haviam sido descritas para a família e as atividades já encontradas estão

relacionadas, por exemplo, à glicosilação de proteínas envolvidas na sinalização, resposta ao stress e metabolismo de energia (GÓMEZ-CASATI; MARTÍN; BUSI, 2013; HART; AKIMOTO, 2009). Outras duas famílias que tiveram maior abundância no solo e na água foram GT4 e GT2, que são as famílias de GTs mais numerosas e que catalisam uma vasta gama de reações, com diferentes doadores e receptores de açúcares (RUANE; DAVIES; MARTINEZ-FLEITES, 2008). Conhece-se o papel destas enzimas por estarem envolvidas na síntese de celulose, de quitina e de ácido hialurônico (GT2) e também de sacarose (GT4), dentre as muitas outras atividades.

As polissacarídeo liases tiveram a menor representatividade dentre todas as classes de CAZymes encontradas no presente estudo. Lombard e colaboradores (2010), no intuito de promover uma classificação hierárquica das PLs, analisaram genomas de diversos organismos distintos (desde Archea a plantas e animais superiores) e viram que a ocorrência de PLs é geralmente baixa e consiste entre 3 e 5% do número de GHs encontradas nos genomas. Segundo os autores, essa baixa frequência se deve provavelmente ao fato de que o substrato das PLs (polissacarídeos contendo ácidos urônicos) compõe apenas uma pequena fração de todos os polímeros de carboidratos existentes. No presente trabalho, as PLs detectadas pertencem principalmente à família PL22, a qual compreende enzimas que são geralmente encontradas em bactérias fitopatogênicas ou entéricas, participando do metabolismo da pectina. Estes microrganismos atacam a rede de pectina através de suas enzimas pectinolíticas, provocando a ruptura e perda da integridade da parede celular vegetal (LOMBARD et al, 2010).

Juntamente às glicosil hidrolases e polissacarídeo liases, as carboidrato esterases também participam da sacarificação de polissacarídeos da parede celular vegetal, já que a desacetilação dos resíduos de açúcares expõe as ligações glicosídicas à hidrólise enzimática (BIELY, 2012; HIMMEL et al., 2010). Em relação às carboidrato esterases encontradas no presente trabalho, a família com maior contribuição tanto no solo quanto na água foi a CE10, cujas atividades conhecidas envolvem aril-esterase, carboxilesterase, colinesterase, entre outras. Como a grande maioria (se não todos) os membros desta família são esterases que agem em substratos que não são carboidratos, suas informações não são mais atualizadas na base de dados do CAZy. A segunda família que teve maior representatividade na água foi a CE11, que compreende enzimas com atividade de UDP-3-0-acil N-acetilglicosamina desacetilase, que podem estar relacionadas, por exemplo, à biossíntese de LPS (lipopolissacarídeo) da parede de bactérias gram-negativas (COGGINS et al., 2003;

WHITTINGTON, et al., 2003). As outras duas famílias que mais contribuíram com as CEs no solo e na água foram as CE1 e CE4, que são conhecidas por exibirem atividade acetil xilano esterase, podendo agir em sistemas multifuncionais com as xilanases (presentes em determinadas famílias de GHs) participando na degradação das hemiceluloses vegetais (BIELY, 2012; HIMMEL et al., 2010;). De fato, as famílias CE1 e CE4 foram as que tiveram o maior número de sequências nos metagenomas do microbioma do rúmen de bovinos e do intestino grosso de cupins (BRULC et al., 2009; WARNECKE et al., 2007), evidências que reforçam o potencial destas famílias na bioconversão de biomassa vegetal.

O último grupo de CAZymes que envolve o sistema de bioconversão da lignocelulose compõe a classe das atividades auxiliares, a qual possui famílias de módulos catalíticos associados com a degradação da parede celular vegetal (MATSUMURA et al., 2014). A família AA3, que foi a mais representativa nos dois ambientes analisados, compreende enzimas com diversas atividades, incluindo celobiose desidrogenase, glicose oxidase, aril-álcool oxidase, álcool oxidase e piranose oxidase. As celobiose desidrogenases, por exemplo, são conhecidas por serem produzidas por diversos fungos decompositores de madeira, sendo geralmente secretadas no espaço extracelular durante o crescimento em substratos celulósicos (ZAMOCKY et al., 2006). As aril-álcool oxidases também são secretadas por fungos e participam, por sua vez, da degradação da lignina (HERNÁNDEZ-ORTEGA; FERREIRA; MARTÍNEZ, 2012). A família AA7, segunda mais abundante no solo, possui atividades glicooligossacarídeo oxidase e quitooligossacarídeo oxidase, podendo oxidar uma variedade de carboidratos (glicose, maltose, celobiose, lactose entre outros) e estão potencialmente envolvidas na biotransformação ou detoxificação de compostos lignocelulósicos (FAN; OGUNTIMEIN; REILLY, 2000; LEVASSEUR et al., 2013). A segunda família de AA mais abundantes na água, por sua vez, foi a AA6, cujas enzimas possuem atividade 1,4-benzoquinona redutase, que são enzimas intracelulares envolvidas na biodegradação de compostos aromáticos (www.cazy.org).

Até o momento, as famílias de AAs descritas estão fortemente enviesadas para o reino Fungi, o que pode explicar, por exemplo, o fato de não terem sido encontrados genes de degradação de lignina no metagenoma da microbiota do intestino grosso do cupim *Nasutitermes* (BUGG et al., 2011; LEVASSEUR et al., 2013; WARNECKE et al., 2007). No entanto, na revisão de BUGG e colaboradores (2011), constam relatos de bactérias cuja capacidade de romper a lignina foi mostrada *in vitro* e, interessantemente, muitas destas bactérias já foram isoladas do intestino grosso de cupins. Estes fatos reforçam a necessidade

de mais estudos acerca do papel enzimático das bactérias na conversão de lignina, o que vai ajudar a preencher a lacuna nas famílias de AAs descritas para bactérias, permitindo uma anotação funcional confiável e uma melhor compreensão de todo o processo lignolítico relacionado a essas enzimas bacterianas (LEVASSEUR et al., 2013).

6.3 Comparação taxonômica quanto à estrutura da comunidade microbiana total

A análise da composição da comunidade microbiana da região de Caatinga estudada mostrou que existem perfis taxonômicos com diferenças e semelhanças entre o solo e a água: as classes Actinobacteria (45%) seguida de Alphaproteobacteria (14%) foram dominantes no solo, enquanto que Betaproteobacteria (67%) seguida também de Alphaproteobacteria (10,8%), foram as classes mais abundantes na água.

Resultados semelhantes em relação ao solo foram encontrados por Pacchioni e colaboradores (2014). Actinobacteria e Alphaproteobacteria foram também as classes dominantes no solo de Caatinga estudado através da abordagem metagenômica (36.4% e 12.41% de abundância, respectivamente). Uma grande abundância do filo Actinobacteria já foi encontrada na rizosfera de espécies de cactos no semiárido mexicano (AGUIRRE-GARRIDO et al., 2012) assim como foi o filo dominante nas amostras obtidas, durante a estação seca, no solo e na rizosfera associada a *Cereus jamacaru* (mandacaru), numa região de Caatinga brasileira (KAVAMURA et al., 2013).

De fato, já foi reportado na literatura que Actinobacteria é um dos filos que mais contribui para a comunidade microbiana em solos áridos e semiáridos (BACHAR et al., 2010; CHANAL et al., 2006; FIERER et al., 2012), estando sua maior presença nestes ambientes relacionada à existência de diversas espécies resistentes ao estresse hídrico, que são tolerantes a altas temperaturas e radiação (CONNON et al., 2007; SINGLETON et al., 2003). Acredita-se que seja um grupo que possui um papel importante na ciclagem de nutrientes, decomposição e formação do húmus no solo (GOODFELLOW; WILLIAMS, 1983), já tendo sido encontrada sua potencial significância na decomposição de materiais lignocelulósicos (VĚTROVSKÝ; STEFFEN; BALDRIAN, 2014).

A alta abundância de Actinobacteria em solos áridos e semiáridos contrasta com o que já foi descrito para solos de florestas, como por exemplo, da Mata Atlântica (BRUCE et al., 2010; FAORO et al., 2010) e do cerrado (ARAÚJO et al., 2012). Nestes ambientes, a

frequência de Actinobacteria foi mais baixa (<10%, 1.2% e 4-7% nos respectivos estudos) em detrimento de Acidobacteria, cuja contribuição para as sequências de rRNA 16S obtidas foi dominante (29–54%, 63% e 40–47%, respectivamente). O filo Acidobacteria, apesar de ser um dos filos mais abundantes no solo (CATÃO et al., 2014; JANSSEN, 2006) incluindo no ambiente semiárido (AGUIRRE-GARRIDO et al., 2012; BACHAR et al. 2010; CHANAL et al., 2006) teve apenas 6% de abundância no presente estudo. Pacchioni e colaboradores (2014) também encontraram uma pequena contribuição de Acidobacteria (2.5%) no solo de Caatinga e da Mata Atlântica analisados. Os autores atribuíram esse achado, entre outros fatores, ao maior teor de areia das suas amostras, que pode ter tornado menos propícia a presença de Acidobacteria, já que eles observaram que este grupo parece ser mais abundante na fração argilosa do solo, em relação à siltosa e à arenosa (LILES et al. 2010; RUSSO et al. 2012).

No que diz respeito às amostras de água, a classe Betaproteobacteria - que foi o grupo taxonômico dominante - parece ser também frequente em outros ambientes de água doce. Ghai e colaboradores (2011) realizaram um estudo metagenômico da microbiota do rio Amazonas, por meio de duas estratégias distintas: utilizando as sequências de rRNA 16S recuperadas dos metagenomas e analisando o total de *reads*, através da anotação no servidor MG-RAST. Os autores encontraram Betaproteobacteria como a segunda classe mais abundante nas amostras analisadas, atrás apenas da classe Actinobacteria, levando em conta as sequências de rRNA 16S. No entanto, ao considerar o total de *reads*, a classe Betaproteobacteria foi a dominante, o que sustenta os achados do presente trabalho, no qual também foram utilizadas as *reads* totais para anotação. Os dados obtidos também corroboram com os encontrados por Zwart e colaboradores (2002), que realizaram um estudo com sequências de rDNA 16S depositadas em bases de dados e provenientes do plâncton de água doce em diferentes localidades (América do Norte, Europa e Ásia). Os autores encontraram que Proteobacteria (das subdivisões Alphaproteobacteria e Betaproteobacteria) estiveram dentre as principais divisões bacterianas representadas em todos os ambientes de água doce analisados. Ainda segundo os autores, existe uma separação evolutiva entre bactérias que crescem na água doce e em oceanos, fato que é evidenciado, por exemplo, pela ampla distribuição de Betaproteobacteria em ambientes de água doce, contrastando com sua ausência em ambientes oceânicos.

Betaproteobacteria é um grupo relativamente amplo de bactérias, que frequentemente pertence aos membros mais abundantes em amostras de bacterioplâncton de água doce

(LINDSTRÖM; KAMST-VAN; ZWART, 2005; SALCHER, 2014; WU; HAHN, 2006), possuindo um amplo espectro de modos de vida que incluem simbiontes, patógenos e organismos de vida livre. A maioria das sequências associadas ao grupo pertence ao *cluster Polynucleobacter*, cujos membros já foram detectados numa variedade de ecossistemas de lagos e rios (HAHN, 2003; ZWART et al., 2002). A presença deste grupo em habitats tão diferentes quanto a características físico-químicas e em diferentes zonas climáticas pode resultar de uma diversificação ecológica dentro do táxon, e não de uma adaptação generalista das cepas, embora os detalhes sobre suas adaptações específicas ainda sejam desconhecidos (HAHN et al., 2012; JEZBERA et al., 2011; JEZBEROVÁ et al., 2010).

Polynucleobacter foi o gênero mais abundante nas amostras de água analisadas no presente estudo (>50%; dados não mostrados) e a alta representatividade deste grupo nas amostras de água pode indicar a presença de características ambientais únicas nesta região de água doce da Caatinga, que favoreceram o estabelecimento do grupo neste ambiente. Segundo HAHN e colaboradores (2012), o fato de existirem poucas informações sobre as adaptações ecológicas das bactérias pertencentes ao grupo dificulta uma comparação intra-habitats das suas características. Nesse sentido, mais estudos precisam ser realizados a fim de se compreender o perfil metabólico e o papel ecológico destes microoorganismos nos mais diversos ambientes de água doce que eles ocupam.

6.4 Perfil taxonômico da comunidade microbiana associada às CAZymes

Em relação ao solo, o táxon que apresentou maior contribuição para as sequências anotadas como CAZymes foi um dos que obteve a menor representatividade, dentre os táxons dominantes encontrados na comunidade microbiana como um todo. Planctomycetia teve uma abundância de apenas 1.8% dentre o total de classes que foram encontradas no solo, enquanto que, em relação às CAZymes, sua contribuição foi de 29%, a maior frequência identificada. Resultado semelhante foi encontrado para as amostras de água. Planctomycetia foi a segunda classe com mais sequências pertencentes às CAZymes (21%) tendo, no entanto, uma pequena contribuição para a comunidade microbiana total (0.5%).

A classe Planctomycetia pertence ao filo Planctomycetes, o qual era originalmente encontrado em ambientes de água doce, mas que se mostrou amplamente distribuído em habitats marinhos, hipersalinos e em solos terrestres (FUERST, 1995). De fato,

Planctomycetes tem sido encontrado em bibliotecas de rRNA 16S provenientes de diversos tipos de solos (BORNEMAN; TRIPLETT, 1997; KUSKE; BARNS; BUSCH 1997; ZHOU et al., 2003). A onipresença geográfica e ampla diversidade filogenética de Planctomycetes sugerem que a capacidade metabólica do grupo seja bastante versátil (BUCKLEY et al., 2006; ELSHAHED et a. 2007). No entanto, todos os Planctomycetes cultivados até o momento parecem ser aeróbios especializados no metabolismo de açúcar (ELSHAHED et al. 2007). Esta especialização metabólica já verificada para os representantes cultivados pode explicar o fato de Planctomycetia ter sido o grupo taxonômico mais abundante em relação às CAZymes encontradas no solo e o segundo grupo mais frequente na água.

A classe Alphaproteobacteria também teve grande contribuição para as CAZymes encontradas, principalmente na água (27%). Os microorganismos pertencentes a esta classe são encontrados nos mais diversos ambientes e exibem uma enorme plasticidade nos seus genomas e estilos de vida (NEWTON et al., 2011). São ubíquos em lagos de água doce, apesar de serem menos numerosos do que em ambientes marinhos, nos quais parecem ser microorganismos dominantes (MORRIS et al., 2002; NEWTON et al., 2011).

De uma maneira geral, Alphaproteobacteria de água doce é uma classe pouco estudada, mas os dados disponíveis para os membros dominantes em ambientes lacustres sugerem que sejam bactérias resistentes, competitivas em circunstâncias de baixa disponibilidade de nutrientes e substrato, mas também capazes de degradar compostos orgânicos complexos (NEWTON et al., 2011). A predominância de Alphaproteobacteria, em relação aos outros táxons associados às Cazymes, evidencia a importância das enzimas relacionadas ao metabolismo de carboidratos na manutenção das características que garantem o sucesso de Alphaproteobacteria nestes ambientes de água doce. Uma estratégia diferente parece ocorrer para ambientes marinhos. Segundo Dittami e colaboradores (2014), as Alphaproteobacteria marinhas são conhecidas por conterem poucas CAZymes, apesar de estas bactérias geralmente não serem capazes de realizar fotossíntese e dependerem, portanto, de uma fonte externa de carbono e de energia.

As outras duas classes com maior abundância no total de CAZymes encontradas foram Cytophagia (solo) e Spirochaetia (água). A classe Cytophagia pertence ao filo Bacteroidetes, cujos membros mais conhecidos são especializados no processamento da matéria orgânica, especialmente em solos (gênero *Cytophaga*) ou no intestino de mamíferos (FERNÁNDEZ-GÓMEZ et al., 2013). A classe já compreendeu um dos grupos de Bacteroidetes significativamente mais frequentes no horizonte orgânico do solo em relação ao mineral,

evidenciando que o filo, juntamente com Verrucomicrobia e Proteobacteria, fazem parte da comunidade adaptada a substratos de carbono facilmente acessíveis (UROZ et al., 2013).

A classe Spirochaetia, por sua vez, pertence ao filo Spirochaetes, amplamente distribuído na natureza, cujos microorganismos ocorrem como formas de vida livre em ambientes de água doce, marinhos e hipersalinos, podendo associar-se também a invertebrados e hospedeiros vertebrados (BREZNAK, 2002). Um dos habitats em que Spirochaetes representa um dos principais grupos é no microbioma do intestino grosso de cupins (BREZNAK, 2002) e constituem, no caso do cupim *Nasutitermes*, um dos principais contribuintes para as enzimas degradativas de matéria orgânica (WARNECKE et al., 2007). Segundo Margulis e Chapman (2009), as espiroquetas provavelmente não possuem celulases, porém, produzem enzimas hábeis na digestão dos produtos da sua quebra inicial, sendo observadas em ambientes degradadores de celulose proveniente de algas ou plantas.

6.5 Reconhecimento de potenciais novas sequências de CAZymes

A anotação pela assinatura de domínios conservados permitiu reconhecer sequências de CAZymes recuperadas dos metagenomas do solo e da água. No entanto, uma fração destas sequências apresentou baixa identidade em relação àquelas depositadas na base de dados não-redundante NR. Este fato permitiu supor que se tratam de sequências distantemente relacionadas às já existentes, o que pode também significar que sejam novas sequências de CAZymes.

Stroobants, Portelle e Vandenbol (2014) identificaram novas CAZymes derivadas de um solo agrícola, ao realizarem um *screening* das bibliotecas metagenômicas obtidas. Os autores detectaram nos clones sete β-glicosidases putativas e duas glicosiltransferases putativas. As duas ORFs codificantes para as glicosiltransferases apresentaram identidades de 25 e 27% em relação às GTs já conhecidas, o que permitiu aos autores inferir que fossem enzimas pertencentes a novas famílias de CAZymes.

De uma maneira geral, na denominada "zona cinzenta ou de incerteza" (*twilight zone*) – intervalo de identidade entre 20-35% – não é possível inferir homologia com base apenas na quantidade de resíduos idênticos no alinhamento entre duas sequências (ROST, 1999). Outros parâmetros, como por exemplo, a porcentagem de cobertura, podem ser utilizados em paralelo para auxiliar nesta inferência.

A estratégia de anotação realizada no presente estudo, baseada na busca de domínios protéicos conservados, permitiu identificar a totalidade de sequências de CAZymes nos metagenomas do solo e da água. No entanto, as sequências de baixa identidade podem ser homólogas distantemente relacionadas às CAZymes já existentes e com isso, possivelmente representam novas sequências e até mesmo novas famílias destas enzimas relacionadas ao metabolismo de carboidratos. A recuperação destas sequências nos metagenomas pode ser futuramente realizada no intuito de se analisar e compreender melhor suas propriedades catalíticas, bem como as relações filogenéticas com as CAZymes até então conhecidas.

6.6 Importância dos estudos de metagenômica comparativa

Segundo Simon e Daniel (2009), "A metagenômica comparativa é útil para a identificação de diferenças na habilidade de comunidades microbianas de se adaptarem a condições ambientais variáveis". O estudo do perfil funcional de comunidades microbianas distintas permite, além de caracterizá-las individualmente, perceber tais diferenças que indicam que existem alguns metabolismos que são dominantes em cada um dos ambientes. São estes metabolismos dominantes que sugerem a existência de perfis funcionais que são particulares a cada metagenoma (DINSDALE et al, 2008).

Tringe e colaboradores (2005), realizaram um estudo de metagenômica comparativa entre diferentes comunidades microbianas que permitiu deduzir genes habitat-específicos e perfis funcionais característicos nos ambientes amostrados. Para Simon e Daniel (2009), ambientes com demandas compatíveis podem significar perfis funcionais similares, e é isso que uma "abordagem centrada em genes" parece sugerir. Provavelmente, a maior presença de determinado gene em uma comunidade se deve ao fato de ele ser de alguma forma benéfico neste ambiente (HUGENHOLTZ; TYSON, 2008).

Sendo assim, as proteínas codificadas por uma comunidade microbiana parecem ser mais elucidativas sobre o ambiente do que os próprios táxons que as codificam; de maneira que é possível até mesmo distinguir locais a partir dos perfis funcionais encontrados (TRINGE, 2005). Mais ainda, a maioria dos genes que são expressos diferencialmente em cada ambiente não têm função conhecida, o que os tornam interessantes focos de pesquisa tanto do ponto de vista ecológico quanto biotecnológico (HUGENHOLTZ; TYSON, 2008).

7. CONCLUSÕES

- As comunidades microbianas que habitam o solo e a água desta região de Caatinga apresentaram diferentes perfis taxonômicos e potenciais funcionais, indicando que existem distintos metabolismos dominantes e táxons que são mais frequentes em cada ambiente;
- A estratégia de anotação pela tecnologia de subsistemas permitiu realizar uma caracterização geral dos potenciais funcionais do solo e da água. Foi possível observar que o metabolismo de carboidratos compõe um subsistema importante na manutenção das comunidades microbianas em ambos os locais;
- A maior contribuição dos subsistemas *Clustering-based* na água indica que este ainda é uma ambiente menos explorado do ponto de vista científico, em relação, por exemplo, ao solo;
- A segunda estratégia de anotação util*izada* - *baseada* em domínios conservados de sequências protéicas - permitiu caracterizar de maneira mais precisa e específica o solo e a água, quanto à presença de enzimas relacionadas ao metabolismo de carboidratos (CAZymes);
- No presente estudo o solo mostrou-se como um ambiente potencialmente mais especializado na degradação de moléculas de carboidratos complexas, fato evidenciado pela predominância da classe de glicosil hidrolases (GHs) neste ambiente;
- A água, por sua vez, foi caracterizada por conter, significativamente, mais enzimas relacionadas à biossíntese de carboidratos e transferência de glicosídeos (GTs);
- O solo aparentou ser, inicialmente, mais vantajoso do ponto de vista da bioprospecção de enzimas lignocelulolíticas, dada à contribuição significativamente maior de GHs, em relação à água. Porém, nas amostras de água também foram encontradas famílias de CAZymes com tal potencial biotecnológico, mesmo que em menor proporção;
- Dentre as CAZymes mais abundantes, a família de glicosil hidrolases GH13, encontrada em ambos os ambientes e GH15, no solo, possui enzimas com possível aplicação na cadeia de bioetanol de primeira geração, devido à sua atividade na bioconversão do amido. Já as famílias GH1 e GH23 (mais frequentes na água), e GH3 (presente em ambos os ambientes), contém enzimas com capacidade de degradação da biomassa

lignocelulósica e com isso potencial aplicação na produção de bioetanol de segunda geração;

- Duas das mais frequentes famílias de carboidrato esterases encontradas no solo e na água, as famílias CE1 e CE4, têm capacidade de auxílio na conversão de biomassa vegetal através da degradação da hemicelulose;
- Para as famílias de atividades auxiliares, a família AA3 possui papel importante na decomposição fúngica da lignocelulose, enquanto que a família AA7 está potencialmente envolvida na biotransformação ou detoxificação de compostos lignocelulósicos, sendo grupos interessantes do ponto de vista biotecnológico;
- Dentre os CBMs, a família CBM13 possui afinidade pelo xilano, podendo ser encontrada em associação às glicosil hidrolases no ataque à matriz de glicanos da parede celular vegetal, auxiliando na sua degradação;
- A estratégia de anotação funcional que envolveu a busca de domínios protéicos conservados também permitiu inferir a existência de potenciais novas sequências de CAZymes, no solo e na água desta região de Caatinga. Estudos posteriores são necessários para se compreender melhor estas enzimas e sua ligação com as CAZymes já existentes.
- Em relação à caracterização taxonômica, a maior contribuição de Actinobacteria no solo sugere que esta seja uma classe relevante em solos com características mais áridas, provavelmente pela existência de diversas espécies resistentes ao *stress* hídrico;
- Já para as amostras de água, a maior frequência de Betaproteobacteria veio reforçar sua ubiquidade em ambientes de água doce;
- O táxon Planctomycetia foi o que mais contribuiu para as CAZymes identificadas nos metagenomas do solo, sendo o segundo grupo mais abundante na água. Entretanto, Planctomycetia esteve dentre os táxons mais raramente encontrados, em ambos os ambientes, levando em conta a estrutura da comunidade microbiana como um todo;
- Os resultados encontrados apontam para a relevância dos grupos menos abundantes na manutenção de determinadas atividades nas comunidades microbianas, assim como indicam a importância de abordagens como a metagenômica na identificação, caracterização e descoberta do potencial biotecnológico de tais grupos.

REFERÊNCIAS

ABD-ELSALAM, K. A. Bioinformatic tools and guideline for PCR primer design. **African Journal of Biotechnology**, v. 2, p. 91-95, 2003.

ADRIO, J. L.; DEMAIN, A. L. Microbial Enzymes: tools for biotechnological processes. **Biomolecules**, v. 4, p. 117-139, 2014.

AGUIRRE-GARRIDO, J. F. et al. Bacterial community structure in the rhizosphere of three cactus species from semi-arid highlands in central Mexico. **A. van Leeuw**, v. 101, p. 891–904, 2012.

ALAIN, K.; QUERELLOU, J. Cultivating the uncultured: limits, advances and future challenges. **Extremophiles**, v. 13, p. 583-594, 2009.

ALPHALYSE. **What is sequence coverage?** Disponível em: http://www.alphalyse.com/faq19.html. Acesso em: 11 mar. 2015.

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res**., v. 25, p. 3389–3402, 1997.

AMANN, R. I.; LUDWIG, W.; SCHLEIFER, K-H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. **Microbiological Reviews**, v. 59, p. 143–169, 1995.

ANDRÉ, I. et al. CAZyme discovery and design for sweet dreams. **Current Opinion in Chemical Biology**, v. 19, p. 17–24, 2014.

ARAÚJO, J. F. et al. Characterization of soil bacterial assemblies in Brazilian savanna-like vegetation reveals acidobacteria dominance. **Microb. Ecol.**, v. 64, p. 760–770, 2012.

BACHAR, A. et al. Soil microbial abundance and diversity along a low precipitation gradient. **Microb. Ecol.**, v. 60, p. 453–461, 2010.

BARNARD, D. et al. Extremophiles in biofuel synthesis. **Environmental Technology**, v. 31, p. 871–888, 2010.

BARONE, R. et al. Marine metagenomics, a valuable tool for enzymes and bioactive compounds discovery. **Frontiers in Marine Science**, v. 1, 2014.

BASTOS, C. J. P.; STRADMANN, M. T. S.; VILAS BÔAS-BASTOS, S. B. Additional contribution to the bryophyte flora of Chapada Diamantina National Park, State of Bahia, Brazil. **Tropical Bryology**, v. 15, p. 15-20, 1998.

BERGMANN, J. C. et al. Discovery of two novel b-glucosidases from an Amazon soil metagenomic library. **FEMS Microbiol Lett**, v. 351, p. 147–155, 2014.

BERLEMONT , R.; MARTINY, A. C. Genomic potential for polysaccharides deconstruction in bacteria. **Appl. Environ. Microbiol.**, v. 81, p. 1513-1519, 2015.

BERTOLDO, C.; ANTRANIKIAN, G. Starch-hydrolyzing enzymes from thermophilic Archaea and Bacteria. **Current Opinion in Chemical Biology**, v. 6, p. 151–160, 2002.

BIELY, P. Microbial carbohydrate esterases deacetylating plant polysaccharides. **Biotechnology Advances**, v. 30, p. 1575–1588, 2012.

BORASTON, et al. A novel mechanism of xylan binding by a lectin-like module from *Streptomyces lividans* xylanase 10A. **Biochem. J.**, v. 350, p. 933–941, 2000.

BORASTON, A. B. et al. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. **Biochem. J.**, v. 382, p. 769–781, 2004.

BORNEMAN, J.; TRIPLETT, E. W. Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. **Appl. Environ. Microbiol.**, v. 63, p. 2647–2653, 1997.

BOSCARO, V. et al. *Polynucleobacter necessarius*, a model for genome reduction in both free-living and symbiotic bacteria. **Proc. Natl. Acad. Sci. USA**., v. 110, p. 18590–18595, 2013.

BRÁS, J. L. A. et al. Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis. **Proc. Natl. Acad. Sci. U.S.A**., v. 108, p. 5237–5242, 2011.

BRASIL. Decreto nº 91.655 de 17 de setembro de 1985. Cria o Parque Nacional da Chapada da Diamantina. **Diário Oficial da União**, Poder Executivo, Brasília, DF, 18 set 1985. Seção 1, p. 13593.

BRASIL. Resolução nº 357, de 17 de março de 2005. Dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões de lançamento de efluentes, e dá outras providências. **Diário Oficial da União**, Poder Executivo, Brasília, DF, 18 mar 2005. p. 58-63.

BREITBART, M. et al. Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico. **Environmental Microbiology**, v. 11, p. 16–34, 2009.

BREZNAK, J. A. Phylogenetic diversity and physiology of termite gut spirochetes. **Integ. and Comp. Biol**., v. 42, p. 313–318, 2002.

BRUCE, T. et al. Bacterial community diversity in the brazilian atlantic forest soils. **Microbial Ecology**, v. 60, p. 840–849, 2010.

BRULC, J. M. et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. **Proc. Natl. Acad. Sci. USA**, v. 106, p. 1948–1953, 2009.

BUCKLEY, D. H. et al. Diversity of Planctomycetes in soil in relation to soil history and environmental heterogeneity. **Applied and Environmental Microbiology**, v. 72, p. 4522–4531, 2006.

BUGG, T. D. et al. The emerging role for bacteria in lignin degradation and bio-product formation. **Curr. Opin. Biotechnol**., v. 22, p. 394–400, 2011.
BULGARELLI, D. et al. Structure and Functions of the Bacterial Microbiota of Plants. **Annu. Rev. Plant Biol.**, v. 64, p. 807–38, 2013.

CAIRNS, J. R. K.; ESEN, A. β-Glucosidases. **Cell. Mol. Life. Sci.**, v. 67, p. 3389-3405, 2010.

CANTAREL, B. L. et al. The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics, **Nucleic Acids Res**., v. 37, p. 233–238, 2009.

CARVALHO, M. C. C. G.; SILVA, D. C. G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, v. 40, p. 735-744, 2010.

CATÃO, E. C. P. et al. Soil Acidobacterial 16S rRNA gene sequences reveal subgroup level differences between savanna-like Cerrado and Atlantic Forest brazilian biomes. **International Journal of Microbiology**, v. 2014, 2014.

CAZY – **Carbohydrate-active enzymes**. Disponível em: http://www.cazy.org/. Acesso em 20 fev. 2015.

CAZYPEDIA - **Carbohydrate-active enzymes**. Disponível em: http://www.cazypedia.org/index.php/Main_Page. Acesso em 15 fev. 2015.

CERRATINGA. Produção sustentável e consumo consciente. Disponível em: www.cerratinga.org.br/caatinga. Acesso em 23 mai. 2014.

CHANAL, A. et al. The desert of Tataouine: an extreme environment that hosts a wide diversity of microorganisms and radiotolerant bacteria. **Environ. Microbiol.**, v. 8, p. 514–525, 2006.

CHERRY, J. R.; FIDANTSEF, A. L. Directed evolution of industrial enzymes: an update. **Current Opinion in Biotechnology**, v. 14, p. 438–443, 2003.

COGGINS, B. E. et al. Structure of the lpxC deacetylase with a bound substrate analog inhibitor. **Nat. Struct. Biol.**, v. 10, p. 645-651, 2003.

CONNON, S. A. et al. Bacterial diversity in hyperarid Atacama desert soils. **J. Geophy**, v. 112, 2007.

COSTA, R. C.; ARAÚJO, F. S. Physiognomy and structure of a caatinga with *Cordia oncocalyx* (Boraginaceae), a new type of community in Andrade-Lima's classification of caatingas. **Rodriguésia**, v. 63, p. 269-276, 2012.

COWAN, D. A. et al. Metagenomics, gene discovery and the ideal biocatalyst. **Biochemical Society Transactions**, v. 32, p. 298-302, 2004.

DANIEL, R. The soil metagenome - a rich resource for the discovery of novel natural products. **Current Opinion in Biotechnology**, v. 15, p. 199-204, 2004.

DBCAN: **A webserver and database for Carbohydrate-active enzyme annotation**. Disponível em: <http://csbl.bmb.uga.edu/dbCAN/>. Acesso em 18 fev. 2015.

DEBROAS, D. et al. Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget - France). **Environmental Microbiology**, v. 11, p. 2412-2424, 2009.

DELMONT, T. O. et al. Structure, fluctuation and magnitude of a natural grassland soil metagenome. **The ISME Journal**, v. 6, p. 1677–1687, 2012.

DESAI, N. et al. From genomics to metagenomics. **Current Opinion in Biotechnology**, v. 23, p. 72–76, 2012.

DINSDALE, E. A. et al. Functional metagenomic profiling of nine biomes. **Letters, Nature**, v. 452, 2008.

DITTAMI, S. M. et al. Genome and metabolic network of "*Candidatus Phaeomarinobacter ectocarpi*" Ec32, a new candidate genus of Alphaproteobacteria frequently associated with brown algae. **Front. Genet.**, v. 5, 2014.

DOUGHERTY, M. J. et al. Glycoside Hydrolases from a targeted compost metagenome, activity-screening and functional characterization. **BMC Biotechnology**, v. 12, 2012.

DUAN, C-J.; FENG, J. X. Mining metagenomes for novel cellulase genes. **Biotechnol. Lett.**, v. 32, p. 1765–1775, 2010.

EDDY, S. R. Accelerated Profile HMM Searches. **PLoS Comput. Biol.**, v. 7, 2011.

EDDY, S. R. A new generation of homology search tools based on probabilistic inference. **Genome informatics**. International Conference on Genome Informatics. v. 23, p. 205–211, 2009.

ELLEUCHE, S. et al. Extremozymes - biocatalysts with unique properties from extremophilic microorganisms. **Current Opinion in Biotechnology**, v. 29, p. 116–123, 2014.

ELSHAHED, M. S. et al.Phylogenetic and metabolic diversity of Planctomycetes from anaerobic, sulfide- and sulfur-rich zodletone spring, Oklahoma. **Applied and Environmental Microbiology**, v. 73, p. 4707–4716, 2007.

EXPLORENZ – **The enzyme database**. Disponível em http://www.enzyme-database.org/. Acesso em 3 jun. 2014.

FAN, Z.; OGUNTIMEIN, G. B.; REILLY, P. J. Characterization of kinetics and thermostability of *Acremonium strictum* glucooligosaccharide oxidase. **Biotechnol Bioeng**, v. 68, p. 231–237, 2000.

FAORO, H. et al. Influence of soil characteristics on the diversity of bacteria in the southern brazilian atlantic forest. **Applied and Environmental Microbiology**, v. 76, p. 4744–4749, 2010.

FAURE, D. The family-3 Glycoside Hydrolases: from housekeeping functions to host-microbe interactions. **Applied and Environmental Microbiology**, v. 68, p. 1485–1490, 2002.

FERNÁNDEZ-GÓMEZ, B. et al. Ecology of marine Bacteroidetes: a comparative genomics approach. **ISME J.**, v. 7, p. 1026–1037, 2013.

FIERER, N. et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. **Proc. Natl. Acad. Sci. USA**, v. 109, p. 21390–21395, 2012.

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. HMMER web server: interactive sequence similarity searching. **Nucleic Acids Research**, v. 39, p. W29-W37, 2011.

FUERST, J. A. The Planctomycetes: emerging models for microbial ecology, evolution and cell biology. **Microbiology**, v. 141, p. 1493-1506, 1995.

GARBEVA, P.; VAN VEEN, J. A.; VAN ELSAS, J. D. Microbial diversity in soil: selection of microbial populations by plant and soil type and implications for disease suppressiveness. **Annu. Rev. Phytopathol**., v. 42, p. 243–70, 2004.

GERLT, J. A.; BABBITT, P. C. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. **Annual Review of Biochemistry**, v. 70, p. 209-246, 2001.

GHAI, R. et al. Metagenomics of the water column in the pristine upper course of the Amazon river. **PLoS ONE**, v. 6, 2011.

GHAI, R. et al. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. **Molecular Ecology**, v. 23, p. 6073–6090, 2014.

GLENN, T. C. Field guide to next-generation DNA sequencers. **Molecular Ecology Resources**, v. 11, p. 759–769, 2011.

GÓMEZ-CASATI, D. F.; MARTÍN, M.; BUSI, M. V. Polysaccharide-synthesizing glycosyltransferases and carbohydrate binding modules: the case of starch synthase III. **Protein & Peptide Letters**, v. 20, p. 856-863, 2013.

GONÇALVES, C. N. et al. **Plano de prevenção aos incêndios florestais: Parque Nacional da Chapada Diamantina**. Palmeiras, 2005.

GOODFELLOW, M.; WILLIAMS, S. T. Ecology of Actinomycetes. **Ann. Rev. Microbial.**, v. 37, p. 189-216,1983.

GORLACH-LIRA, K.; COUTINHO, H. D. M. Population dynamics and extracellular enzymes activity of mesophilic and thermophilic bacteria isolated from semi-arid soil of northeastern Brazil. **Brazilian Journal of Microbiology**, v. 38, p. 135-141, 2007.

GRAYSTON, S. J. et al. Selective influence of plant species on microbial diversity in the rhizosphere. **Soil Biol. Biochem.**, v. 30, p. 369±378, 1998.

GUILLÉN, D.; SÁNCHEZ, S.; RODRÍGUEZ-SANOJA, R. Carbohydrate-binding domains: multiplicity of biological roles. **Appl. Microbiol. Biotechnol**., v. 85, p. 1241–1249, 2010.

HAN, S. J.; YOO, Y. J.; KANG, H. S. Characterization of a Bifunctional Cellulase and Its Structural Gene. **Journal of Biological Chemistry**, v. 270, p. 26012-26019, 1995.

HAHN, M. W. Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones. **Applied and Environmental Microbiology**, p. 5248–5254, 2003.

HAHN, M. W. et al. The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-Living *Polynucleobacter* population. **PLoS ONE**, v. 7, 2012.

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. **Chemistry & Biology**, v. 5, p. R245-R249, 1998.

HANDELSMAN, J. Metagenomics: Application of genomics to uncultured microorganisms. **Microbiol. Mol. Biol. Rev.** v. 68, p. 669-685, 2004.

HANDELSMAN, J. Metagenomics or Megagenomics? **Nat. Rev. Micro**., v. 3, p. 457-458, 2005.

HAQ, I. et al. Production of alpha amylase from a randomly induced mutant strain of *Bacillus amyloliquefaciens* and its application as a desizer in textile industry. **Pak. J. Bot.**, v. 42, p. 473-484, 2010.

HART, G. W.; AKIMOTO Y. The O-GlcNAc Modification. In: VARKI, A. et al. **Essentials of Glycobiology**. 2ª edição. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, Capítulo 18, 2009. Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK1954/.

HENNE, A. et al. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. **Applied and Environmental Microbiology**, p. 3113–3116, 2000.

HERNÁNDEZ-ORTEGA, A.; FERREIRA, P.; MARTÍNEZ, A. T. Fungal aryl-alcohol oxidase: a peroxide-producing flavoenzyme involved in lignin degradation. **Appl. Microbiol. Biotechnol.**, v. 93, p. 1395–1410, 2012.

HILL, A. D.; REILLY, P. J. Computational Analysis of Glycoside Hydrolase Family 1 Specificities. **Biopolymers**, v. 89, p. 1021-1031, 2008.

HINZ, U. From protein sequences to 3D-structures and beyond: the example of the UniProt knowledgebase. **Cellular and molecular life sciences: CMLS**, v. 67, p. 1049–1064, 2010.

HIMMEL, M. E. Microbial enzyme systems for biomass conversion: emerging paradigms. **Biofuels**, v. 1, p. 323-341, 2010.

HORN, S. J. et al. Novel enzymes for the degradation of cellulose. **Biotechnol Biofuels**, v. 5, 2012.

HUGENHOLTZ, P.; TYSON, G. W. Metagenomics. **Nature News and Views**. v. 455, 2008.

HUSON, D. H. et al. MEGAN analysis of metagenomic data. **Genome Res.**, v. 17, p. 377–386, 2007.

INMET. Instituto Nacional de Meteorologia. Disponível em: http://www.inmet.gov.br/portal/. Acesso em: 06 mai. 2015.

JANSSEN, P. H. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. **Appl. Environ. Microbiol.**, v. 72, p. 1719–1728, 2006.

JEZBERA, J. et al. Ubiquity of *Polynucleobacter necessarius* subspecies asymbioticus results from ecological diversification. **Environmental Microbiology**, v. 13, p. 922–931, 2011.

JEZBEROVÁ, J. et al. Ubiquity of *Polynucleobacter necessarius* subsp *asymbioticus* in lentic freshwater habitats of a heterogenous 2000 km^2 area. **Environ. Microbiol.**, v. 12, p. 658–669, 2010.

JI, B.; NIELSEN, J. New insight into the gut microbiome through metagenomics. **Advances in Genomics and Genetics**, v. 5, p. 77-91, 2015.

KAOUTARI A. E. et al. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. **Nat. Rev. Microbiol.**, v. 11, p. 497–504, 2013.

KAVAMURA, V. N. et al. Water Regime Influences Bulk Soil and Rhizosphere of *Cereus jamacaru* Bacterial Communities in the Brazilian Caatinga Biome. **PLoS ONE**, v. 8, 2013.

KIRK, O.; BORCHERT, T. V.; FUGLSANG, C. C. Industrial enzyme applications, **Current Opinion in Biotechnology**, v. 13, p. 345–351, 2002.

KUHAD, R. C.; GUPTA, R.; SINGH, A. Microbial Cellulases and their Industrial Applications. **Enzyme Research**, v. 10, 2011.

KUMAR, V. Analysis of the key active subsites of glycoside hydrolase 13 family members. **Carbohydrate Research**, v. 345, p. 893-898, 2010.

KUSKE, C. R.; BARNS, S. M.; BUSCH, J. D. Diverse uncultivated bacterial groups from soils of the arid southwestern United States that are present in many geographic regions. **Appl. Environ. Microbiol.**, v. 63, p. 3614– 3621, 1997.

LAIRSON, L. L. et al . Glycosyltransferases: structures, functions, and mechanisms. **Annu. Rev. Biochem**., v. 77, p. 521–555, 2008.

LEAL, I. R.; TABARELLI, M.; SILVA, J. M. C. Ecologia e conservação da caatinga: uma introdução ao desafio. In: LEAL, I. R.; TABARELLI, M.; SILVA, J. M. (Edit.) **Ecologia e Conservação da Caatinga**. Recife : Ed. Universitária da UFPE, 2003. Introdução, p. XIII-XVI.

LEVASSEUR, A. et al. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. **Biotechnol. Biofuels**, v. 6, 2013.

LI, H. et al. Screening and characterization of a highly active chitosanase based on metagenomic technology. **Journal of Molecular Catalysis B: Enzymatic**, v. 111, p. 29–35, 2015.

LI, L. et al. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. **Biotechnology for Biofuels**, p. 2-10, 2009.

LILES, M. R. et al. A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. **Soil Biol. Biochem.**, v. 42, p. 739–747, 2010.

LINDSTRÖM, E. S.; AGTERVELD, M. P. K-V.; ZWART, G. Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. **Appl. Environ. Microbiol.**, v. 71, p. 8201-8206, 2005.

LOMBARD, V. et al. A hierarchical classification of polysaccharide lyases for glycogenomics. **Biochem. J.**, v. 432, p. 437–444, 2010.

LOMBARD, V. et al. The Carbohydrate-active enzymes database (CAZy) in 2013. **Nucleic Acids Res.**, v. 42, p. 490–495, 2014.

LORENZ, P.; ECK, J. Metagenomics and industrial applications. **Nature Reviews Microbiology**, v. 3, p. 510-516, 2005.

LYND, L. R. et al. Microbial Cellulose Utilization: fundamentals and biotechnology. **Microbiol. Mol. Biol. Rev.**, v.66, p. 506-577, 2002.

MARGULIS, L.; CHAPMAN, M. J. **Kingdoms and Domains: An Illustrated Guide to the Phyla of Life on Earth**. 4ª Edição. Estados Unidos: Academic Press, 864p, 2009.

MARÍN-NAVARRO, J.; POLAINA, J. Glucoamylases: structural and biotechnological aspects. **Appl. Microbiol. Biotechnol.**, v. 89, p. 1267–1273, 2011.

MATSUMURA, H. et al. Discovery of a eukaryotic pyrroloquinoline quinone-dependent oxidoreductase belonging to a new Auxiliary Activity family in the database of Carbohydrate-Active Enzymes. **PLoS ONE**, v. 9, 2014.

MENDES, L.W. et al. Soil-Borne Microbiome: Linking Diversity to Function. **Microb. Ecol**., 2015.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, v. 11, 2010.

MEYER, F. et al.The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. **BMC Bioinformatics**, v. 9, 2008.

MINISTÉRIO DO MEIO AMBIENTE. **Biodiversidade e conservação da Chapada Diamantina.** Disponível em: http://www.mma.gov.br/estruturas/chm/_arquivos/Bio13_chapada_diamantina.pdf. Acesso em: 1 jun. 2014.

MINISTÉRIO DO MEIO AMBIENTE. **Caatinga - Contexto, características e estratégias de conservação**. Disponível em: <http://www.mma.gov.br/biomas/caatinga/item/191>. Acesso em: 22 fev. 2015.

MONZINGO, A. F. et al. Chitinases, chitosanases, and lysozymes can be divided into prokaryotic and eukaryotic families sharing a conserved core. **Nature. Struct. Biol.**, v. 3, p. 133–140, 1996.

MORENO, M. L. et al. Halophilic bacteria as a source of novel hydrolytic enzymes. **Life**, v. 3, p. 38-51, 2013.

MORRIS, R. M. et al. SAR11 clade dominates ocean surface bacterioplankton communities. **Nature**, v. 420, p. 806-810, 2002.

MUNIR, R. I. et al. Comparative analysis of Carbohydrate Active Enzymes in *Clostridium termitidis* CT1112 reveals complex carbohydrate degradation ability. **PLoS ONE**, v. 9, 2014.

MYROLD, D. D.; ZEGLIN, L. H.; JANSSON, J. K. The potential of metagenomic approaches for understanding soil microbial processes. **Soil Science Society of America Journal**, p. 3-10, 2013.

NCBI. **BLAST Glossary**. Disponível em: www.ncbi.nlm.nih.gov/books/NBK62051/. Acesso em: 11 mar. 15.

NEWTON, R. J. et al. A guide to the natural history of freshwater lake bacteria. **Microbiol. Mol. Biol. Rev.**, v. 75, p. 14 –49, 2011.

OH, I. N. et al. Novel characteristics of a carbohydrate‐binding module 20 from hyperthermophilic bacterium. **Extremophiles**, 2015.

OLIVEIRA, C. et al. Recombinant CBM-fusion technology - Applications overview. **Biotechnol Adv**, 2015. Disponível em: http:// dx.doi.org/10.1016/j.biotechadv.2015.02.006. Acesso em: 25 fev. 2015.

OVERBEEK, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. **Nucleic Acids Res.**, v. 33, p. 5691–5702, 2005.

PACE, N. R.; STAHL, D. A.; OISEN, G. J. Analyzing natural microbial populations by rRNA sequences. **ASM News**, v. 51, p. 4–12, 1985.

PACCHIONI, R. G. et al. Taxonomic and functional profiles of soil samples from Atlantic forest and Caatinga biomes in northeastern Brazil. **Microbiology Open**, v. 3, 2014.

PAËS, G.; BERRIN, J.C.; BEAUGRAND, J. GH11 xylanases: Structure/function/properties relationships and applications. **Biotechnology Advances**, v. 30, p. 564–592, 2012.

PALCIC, M. M. Glycosyltransferases as biocatalysts. **Current Opinion in Chemical Biology**, v. 15, p. 226-233, 2011.

PARISUTHAM, V.; KIM, T. Y.; LEE, S. K. Feasibilities of consolidated bioprocessing microbes: from pretreatment to biofuel production. **Bioresource Technology**, v. 161, p. 431–440, 2014.

PARKS, D. H. et al. STAMP: Statistical analysis of taxonomic and functional profiles. **Bioinformatics**, v. 30, p. 3123-3124, 2014.

PEREIRA, L. P; GEISE, L. Non-flying mammals of Chapada Diamantina (Bahia, Brazil). **Biota Neotrop.**, v. 9, 2009.

PHAM, V. H. T.; KIM, J. Cultivation of unculturable soil bacteria. **Trends in Biotechnology**, v. 30, p. 475-484, 2012.

PHITSUWAN, P. et al. Present and potential applications of cellulases in agriculture, biotechnology, and bioenergy. **Folia Microbiol.**, v. 58, p. 163-176, 2012.

PNCD. Parque Nacional da Chapada Diamantina. Disponível em http://parnachapadadiamantina.blogspot.com.br/p/mapas.html. Acesso em: 06 mai. 2015.

PRAKASH, T.; TAYLOR, T. D. Functional assignment of metagenomic data: challenges and applications. **Briefings in Bioinformatics**, 2012.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0. 2008. Disponível em: http://www.R-project.org. Acesso em: 01 mar. 2015.

REHM, B. H. Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. **Appl. Microbiol. Biotechnol.**, v. 57, p. 579–592, 2001.

RIBEIRO-FILHO, A. A.; FUNCH, L. S.; RODAL, M. J. N. Composição florística da floresta ciliar do rio Mandassaia, Parque Nacional da Chapada Diamantina, Bahia, Brasil. **Rodriguésia**, v. 60, p. 265-276, 2009.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. **Trends in Genetics**, v. 16, p. 276-277, 2000.

RIESENFELD, C. S.; SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics: Genomic analysis of microbial communities. **Annu. Rev. Genet.**, v. 38, p. 325-352, 2004.

RUANE, K. M.; DAVIES, G. J.; MARTINEZ-FLEITES, C. Crystal structure of a family GT4 glycosyltransferase from *Bacillus anthracis* ORF BA1558. **Proteins: Structure, Function, and Bioinformatics**, v. 73, p. 784–787, 2008.

ROSSELLÓ-MORA, R.; AMANN, R. The species concept for prokaryotes. **FEMS Microbiology Reviews**. v. 25, p. 39-67, 2001.

ROST, B. Twilight zone of protein sequence alignments. **Protein Engineering**, v. 12, p. 85–94, 1999.

RUSSO, S. E. et al. Bacterial community structure of contrasting soils underlying Bornean rain forests: inferences from microarray and next-generation sequencing methods. **Soil Biol. Biochem.**, v. 55, p. 48–59, 2012.

SADHU, S.; MAITI, T. K. Cellulase production by bacteria: a review. **British Microbiology Research Journal**, v. 3, p. 235-258, 2013.

SALCHER, M. M. Same same but different: ecological niche partitioning of planktonic freshwater prokaryotes. **J. Limnol.**, v. 73, p. 74-87, 2014.

SÁNCHEZ, O. J.; CARDONA, C. A. Trends in biotechnological production of fuel ethanol from different feedstocks. **Bioresour. Technol.**, v. 99, p. 5270-5295, 2008.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **J. Mol. Biol.**, v. 94, p. 441-448, 1975.

SATHYA, T. A.; KHAN, M. Diversity of Glycosyl Hydrolase enzymes from metagenome and their application in food industry. **Journal of Food Science**, v. 79, p. 2149-2156, 2014.

SCHMIDT, T. M.; DELONG, E. F.; PACE, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. **J. Bacteriol.**, v. 173, p. 4371-4378, 1991.

SESHADRI, R. et al. CAMERA: A Community Resource for Metagenomics. **Plos Biology**, v. 5, p. 0394-0397, 2007.

SHAMSADDINI, A. et al. Census-based rapid and accurate metagenome taxonomic profiling. **BMC Genomics**, v. 15, 2014.

SIMON, C.; ROLF, D. Achievements and new knowledge unraveled by metagenomic approaches. **Appl. Microbiol. Biotechnol.**, v. 85, p. 265–276, 2009.

SINGH, V. K.; KUMAR, A. PCR Primer Design. **Molecular Biology Today**, v. 2, p. 27-32, 2001.

SINGLETON, D. R. et al. *Solirubrobacter pauli* gen. nov., sp. nov., a mesophilic bacterium within the Rubrobacteridae related to common soil clones. **Int. J. Syst. Evol. Microbiol.**, v. 53, p. 485–490, 2003.

STAM, M. R. et al. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of a-amylase-related proteins. **Protein Engineering, Design & Selection**, v. 19, p. 555–562, 2006.

STEELE, H. L.; STREIT, W. R. Metagenomics: Advances in ecology and biotechnology. **FEMS Microbiology Letters**, v. 247, p. 105–111, 2005.

STOREY, J. D. A direct approach to false discovery rates. **J. R. Statist. Soc.** v. 64, p.

479–498, 2002.

STROHL, W. R. The role of natural products in a modern drug discovery program. **Drug Dicovery Today**, v. 5, p. 39-41, 2000.

STROOBANTS, A.; PORTETELLE, D.; VANDENBOL, M. New carbohydrate-active enzymes identified by screening two metagenomic libraries derived from the soil of a winter wheat field. **Journal of Applied Microbiology**, v. 117, p. 1045-1055, 2014.

SUKHARNIKOV, L. O. et al. Cellulases: ambiguous nonhomologous enzymes in a genomic perspective. **Trends in Biotechnology**, v. 29, p. 473-479 , 2011.

SUKUMARAN, R. K.; SINGHANIA, R. R.; PANDEY, A. Microbial cellulases- production, applications and challenges. **Journal of Scientific & Industrial Research**, v. 64, p. 832-844, 2005.

TECHNOLOGY Spotlight. **Illumina sequencing technology**: highest data accuracy, simple workflow, and a broad range of applications. Disponível em: http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf. Acesso em: 05 jun. 2014.

TEELING, H. et al. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. **Science**, v. 336, p. 608-611, 2012.

THE ENZYME LIST CLASS 3 — **Hydrolases**. Disponível em: http://www.enzyme-database.org/downloads/ec3.pdf. Acesso em: 23 fev. 2015.

THOMAS, T.; GILBERT, J.; MEYER, F. Metagenomics - a guide from sampling to data analysis. **Microbial Informatics and Experimentation**, v. 2, 2012.

TORSVIK, V.; ØVREÅS, L. Microbial diversity and function in soil: from genes to ecosystems. **Current Opinion in Microbiology**, v. 5, p. 240-245, 2002.

TRINGE, S. G. et al. Comparative Metagenomics of Microbial Communities. **Science**, v. 308, p. 554-557, 2005.

UROZ, S. et al. Functional assays and metagenomic analyses reveals differences between the microbial communities inhabiting the soil horizons of a Norway spruce plantation. **PLoS ONE**, v. 8, 2013.

USDA: **Unites States Department of Agriculture**. Disponível em: http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_052208.pdf. Acesso em: 4 fev. 2015.

VELLOSO, A. L. et al. **Ecorregiões propostas para o Bioma Caatinga**. Recife: Associação Plantas do Nordeste; Instituto de Conservação Ambiental - The Nature Conservancy do Brasil, 76p., 2002.

VENTER, J. et al. Environmental genome shotgun sequencing of the Sargasso Sea. **Science**, v. 304, p. 66 – 74, 2004.

VĚTROVSKÝ, T.; STEFFEN, K. T.; BALDRIAN, P. Potential of cometabolic transformation of polysaccharides and lignin in lignocellulose by soil *Actinobacteria*. **PLoS ONE**, v. 9, 2014.

WANG, L. et al. Metagenomic Insights into the Carbohydrate-Active Enzymes Carried by the Microorganisms Adhering to Solid Digesta in the Rumen of Cows. **PLoS ONE**, v. 8, 2013.

WARNECKE, F. et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. **Nature**, v. 450, p. 560–565, 2007.

WEADGE, J. T.; PALCIC, M. M. Chemistry of glycosyltransferases. **Wiley Encyclopedia of Chemical Biology**, p. 1–13, 2008.

WETZEL, R. G. Freshwater ecology: changes, requirements, and future demands. **Limnology**, v. 1, p. 3–9, 2000.

WHITESIDES, G. M.; WONG, C-H. Enzymes as catalysts in organic synthesis. **Aldrichimica Acta**, v. 16, 1983.

WHITMAN, W. B.; COLEMAN, D. C.; WIEBE, W. J. Prokaryotes: The unseen majority. **Proc. Natl. Acad. Sci. USA**, v. 95, p. 6578–6583, 1998.

WHITTINGTON, D. A. et al. Crystal structure of LpxC, a zinc-dependent deacetylase essential for endotoxin biosynthesis. **Proc. Natl. Acad. Sci. USA**, v. 100, p. 8146-8150, 2003.

WILSON, D. B. Cellulases and biofuels. **Current Opinion in Biotechnology**, v. 20, p. 295–299, 2009.

WOHLKÖNIG, A. et al. Structural relationships in the lysozyme superfamily: significant evidence for glycoside hydrolase signature motifs. **PLoS ONE**, v. 5, 2010.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, v. 15, 2014.

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A primer on Metagenomics. **Computational Biology**, v.6, 2010.

WU, Q. L.; HAHN, M. W. High predictability of the seasonal dynamics of a species-like *Polynucleobacter* population in a freshwater lake. **Environmental Microbiology**, v. 8, p. 1660–1666, 2006.

YIP, V. L.; WITHERS, S. G. Breakdown of oligosaccharides by the process of elimination. **Current Opinion in Chemical Biology**, v. 10, p. 147–155, 2006.

ZAMOCKY, M. et al. Cellobiose dehydrogenase – A flavocytochrome from wood-degrading, phytopathogenic and saprotropic fungi. **Current Protein and Peptide Science**, v. 7, p. 255-280, 2006.

ZHOU, J. et al. Bacterial phylogenetic diversity and a novel candidate division of two humid region, sandy surface soils. **Soil Biol. Biochem.**, v. 35, p. 915–924, 2003.

ZWART, G. et al. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. **Aquat. Microb. Ecol.**, v. 28, p. 141–155, 2002.

APÊNDICE

Tutorial: Anotação de sequências de CAZymes

Sistema operacional: Linux
Distribuição: Ubuntu 14.04 LTS

O presente tutorial se utiliza das sequências metagenômicas obtidas conforme descrito no item 4.3, submetidas ao MG-RAST, para realizar a anotação funcional das CAZymes. É preciso recuperar tais sequências da plataforma de anotação, após a etapa de controle de qualidade, pois estas servirão como base de dados na anotação das CAZymes pelo pacote HMMER3.

PASSOS

1. Acessar a página do MG-RAST (http://metagenomics.anl.gov/?page=Home);
2. Acessar os itens: Browse metagenomes -> Available for analysis -> selecionar o metagenoma a ser baixado -> Download;

Na página "Downloads" existem duas alternativas:

- Baixar as sequências na etapa do SCREENING, que é a etapa logo após o controle de qualidade. Nesta etapa, as sequências ainda são de nucleotídeos.
- Baixar as sequências já na etapa do GENE CALLING, pois são de aminoácidos e, portanto, não será necessário traduzir as sequências posteriormente.

3. Após o *download* dos metagenomas na etapa escolhida, é necessário concatená-los em apenas um arquivo (reunindo os metagenomas das réplicas das diferentes amostras).

Comando:
cat arquivo1 arquivo2 arquivo3 > nome.do.arquivo.final

Ex: cat TCM_agua1.fna TCM_agua2.fna > TCM_agua.todos.fna
cat TCM_solo1.fna TCM_solo2.fna TCM_solo3.fna > TCM_solo.todos.fna

4. Se as sequências forem baixadas na etapa do "SCREENING", é necessário traduzi-las nos seis quadros de leitura. Isso pode ser realizado utilizando o programa TRANSEQ, do pacote EMBOSS (European Molecular Biology Open Software Suite). Antes de traduzir, no entanto, é necessário alterar os cabeçalhos das sequências no arquivo recém concatenado, para poder realizar a identificação de cada metagenoma posteriormente. Isso porque o programa TRANSEQ altera o cabeçalho dos arquivos durante o processo de tradução.

Comando para editar o cabeçalho utilizando o programa vim:

vim nome.do.arquivo.concatenado(enter)
(shift):%s///g

* No meio das primeiras // colocar o que se quer incluir e no próximo // o que quer substituir.

Ex: (shift):%s/:/_/g (substitui tudo que tem ":" por "_")

Após a modificação, teclar: esc(shift): wq! (Vai sair e salvar o arquivo).

5. Tradução das sequências. Após a edição do cabeçalho, os arquivos fastas podem ser traduzidos utilizando o programa TRANSEQ.

Para baixar o programa TRANSEQ: No Ubuntu sofwtare center, digitar "emboss" na janela. Clicar em "procurar" e depois clicar em "install" ou "instalar".

Comando:

transeq -sequence nome.do.arquivo -outseq nome.do.arquivo.de.saída -frame X (número de janelas de leitura que deseja)

Ex: transeq -sequence TCM_solo_todos.fna -outseq TCM_solo_todos.traduzido.faa -frame 6

A segunda etapa a ser realizada envolve a utilização do programa *hmmscan*, presente no pacote HMMER3, para a busca de homólogos das CAZymes, o que permite sua anotação a partir das sequências metagenômicas recuperadas do MG-RAST.

PASSOS

1. Acessar o banco de dados CAZydatabase (http://csbl.bmb.uga.edu/dbCAN/);

2. Acessar a seção "Download" (http://csbl.bmb.uga.edu/dbCAN/download.php);

3. Seguir as instruções do README (readme.txt), na página de downloads do dbCAN (http://csbl.bmb.uga.edu/dbCAN/download.php), transcritas abaixo:

Para rodar a anotação no dbCAN CAZyme localmente no Linux, realizar os seguintes passos:

****1**. Abrir a tela do Terminal e baixar os seguintes arquivos: dbCAN-fam-HMMs.txt, hmmscan-parser.sh e all.hmm.ps.len;

Comandos:

wget http://csbl.bmb.uga.edu/dbCAN/download/dbCAN-fam-HMMs.txt.

wget http://csbl.bmb.uga.edu/dbCAN/download/hmmscan-parser.sh.

wget http://csbl.bmb.uga.edu/dbCAN/download/all.hmm.ps.len.

****2**. Fazer o *download* do pacote HMMER 3.0 (hmmer.org) e instalar apropriadamente;

Comando para baixar o pacote HMMER 3.0:

No Ubuntu software center, digitar "hmmer" na janela. Aparecerá a opção "hmmer3". Clicar em "install" ou "instalar". Uma outra maneira é acessar o site do *hmmer* (http://hmmer.janelia.org/) e baixar a versão mais nova do HMMER para Linux, de acordo com a configuração do computador. Esta forma é menos simples já que depois será necessário colocar o *hmmer* no diretório usr/local/bin.

****3**. Formatar os perfis HMM que foram carregados (arquivo dbCAN-fam-HMMs.txt):

Comando:

hmmpress dbCAN-fam-HMMs.txt

****4.** Rodar o programa *hmmscan* do pacote HMMER3:

Comando geral:

hmmscan [-options] <hmmdb> <seqfile>

Comando com as configurações utilizadas:

hmmscan -E 0.00001 --cpu 5 -o TCM_solo_agua.traduzido.hmmscan.out --tblout TCM_solo_agua.traduzido.faa.hmmscan.table.out dbCAN-fam-HMMs.txt TCM_solo_agua_todos.traduzido.faa

****5.** Rodar o código (*script*) obtido, conforme descrito no item 1 (hmmscan-parser.sh):

Comando:

sh (ou bash) hmmscan-parser.sh yourfile.hmmscan.out (arquivo de saída do *hmmscan*) > nome.do.arquivo.de.saida.(yourfile)

Obs: se o alinhamento for > 80aa, utilizar *E-value* < 1e-5, se não, utilizar *E-value* < 1e-3; fração coberta do HMM > 0.3)

Análise dos dados obtidos

No comando para utilizar o programa *hmmscan* foi adicionada a opção "formato tabular" dos resultados (--tblout). Assim, a partir desta tabela que foi gerada, serão separadas as colunas referentes às sequências, aos domínios encontrados e ao *E-value* (colunas 1, 3 e 5).

Para obter os resultados correspondentes às colunas 1, 3 e 5 da tabela gerada pelo *hmmscan* (tblout), usa-se o seguinte comando:

Comando

Ex: grep 'TCM' TCM_solo_agua.traduzido.faa.hmmscan.table.out | awk -r '{print $1, $3, $5}' > TCM_solo_agua.traduzido.faa.hmmscan.txt

Depois, é preciso separar este novo arquivo gerado (.txt), de acordo com as categorias e as réplicas se for o caso (por exemplo, diferentes amostras de solo e de água). Outra maneira é rodar o programa *hmmscan* com os arquivos já separados desde o início.

Comando:

Ex: grep "agua_1" TCM_solo_agua.traduzido.faa.hmmscan.txt > TCM_agua1.hmmscan.txt

grep "solo_2" TCM_solo_agua.traduzido.faa.hmmscan.txt > TCM_solo2.hmmscan.txt

Agora, a partir de cada arquivo de réplica (que possui 3 colunas), é preciso separar apenas a coluna dos "domínios" (segunda coluna), para poder contar quantas CAZymes foram encontradas em cada amostra.

Comando:

Ex: grep "TCM" TCM_agua_1.hmmscan.txt | awk '{print $2}' > TCM_agua_1.hmm.txt

Tendo este novo arquivo é possível contar quantos domínios únicos existem dentro de cada réplica. Para isso utiliza-se o comando "uniq" descrito abaixo:

Comando:

uniq -c nome.do.arquivo

Ex: uniq -c TCM_agua_1.hmm.txt > TCM_agua_1.hmm.count.uniq.domain.txt

Obs: É possível concatenar os dois comandos citados acima em um só. Isso se faz utilizando o "|" (pipe).

Ex: grep "TCM" TCM_agua_1.hmmscan.txt | awk '{print $2}' | uniq -c > TCM_agua_1.hmm.txt

---

Alinhamento local utilizando o algoritimo BLAST (Basic Local Alignment Search Tool)

Finalidades para utilização do BLAST neste tutorial:

- Identificar as origens taxonômicas das sequências de CAZymes encontradas;
- Gerar o gráfico "dot plot" que indica possíveis novas sequências.

PASSOS

1. Baixar o BLAST;

Como proceder: Abra o Ubuntu sofwtare center e digite “blast+” na janela de procura. Aperte “enter” e ele vai listar os programas relacionados a essa palavra. Provavelmente o programa BLAST será o primeiro. Clique nele e, ao lado, aperte o botão “install”.

Para instalar a versão mais nova do BLAST (blast+ última versão), pode-se também ir até o site ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ e baixar a versão de acordo com o sistema operacional e configuração do computador. Ex: Para um computador mais novo, o qual deve ser 64 bits, com o Ubuntu instalado, baixe o ncbi-blast-2.2.30+-x64-linux.tar.gz.

Uma segunda forma de fazer isso é abrir o terminal, dentro do diretório onde deseja instalar o blast, e digitar: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.2.30+-x64-linux.tar.gz.

2. Criar um arquivo apenas com os ID's das sequências que deram “hits” com as CAZymes;

Esta etapa é realizada a partir do arquivo de saída tblout do *hmmscan* (ex: TCM_solo_agua.traduzido.faa.hmmscan.table.out)

Comando:

```
grep 'TCM' TCM_solo_agua.traduzido.faa.hmmscan.table.out | awk -r '{print $1}' > TCM_solo_agua_todos.hmmscan.txt
```

3. Formatar o arquivo multifasta recuperado do MG-RAST e traduzido nas seis janelas de leitura;
Comando do blast para formatar a base de sequências metagenômicas (multifasta) para buscar os fastas a partir do ID.

Comando:

Blast versão nova (blast+): Primeiro digite todo o caminho até o diretório onde o programa blast+ se encontra instalado, pelo terminal, ou entre nesse diretório onde o programa está instalado.

Se estiver já no diretório onde o programa está instalado o comando será como o citado abaixo.

Comando:

bin/makeblastdb -in TCM_solo_agua_todos.traduzido.faa -dbtype prot -parse_seqids

Se não estiver no diretório onde o programa está instalado, e ele estiver instalado em uma pasta chamada "Programas", dentro do diretório "Documentos" o comando será o seguinte:

Comando:

/home/Documentos/Programas/ncbi-blast/bin/makeblastdb -in TCM_todos.traduzido.faa -dbtype prot -parse_seqids

4. Utilizar o arquivo (.txt) gerado com os ID's para puxar somente os fastas correspondentes a estes ID's, no arquivo multifasta recém-formatado;

Comando:

Blast versão nova (blast+): blastdbcmd -db caminho.do.multifasta.formatado -entry_batch arquivo.com.os.ids -out arquivo.de.saida

5. Rodar o BLAST utilizando o arquivo multifasta gerado (fastas das sequências que deram "hits" contra os perfis de CAZymes) como arquivo de entrada ("input"), e a base de dados de proteínas curadas NR.

Comando:

Blast versão nova (blast+): -query Ana_Camila/TCM_solo_agua.hmmscan.fastacmd -db /home/work/db/nr.db/nr -outfmt '6 std qlen slen' -evalue 0.00001 -num_threads 30 –out Ana_Camila/TCM_solo_agua.hmmscan.fastacmd.blastp.out -max_target_seqs 1 &

No qual: -num_threads são os núcelos ("threads"), -outfmt 6 (formato tabular)

6. O Resultado tblout do blast pode então ser utilizado pelo programa MEGAN5.6.3 (MetaGenome Analyzer) para analisar a contribuição dos táxons nos metagenomas.

Visualização e análise dos resultados do BLASTP pelo programa MEGAN5

A versão mais recente do MEGAN pode ser instalada a partir do link: http://www-ab.informatik.uni-tuebingen.de/software/megan5. Para realizar a análise taxonômica é necessário baixar ainda dois arquivos extras, o "gi_taxid_nucl" e o "gi_taxid_prot", os quais podem ser obtidos nos seguintes endereços:

gi_taxid_nucl: http://www-ab.informatik.uni-tuebingen.de/data/software/megan5/download/gi_taxid_nucl-4March2015.zip

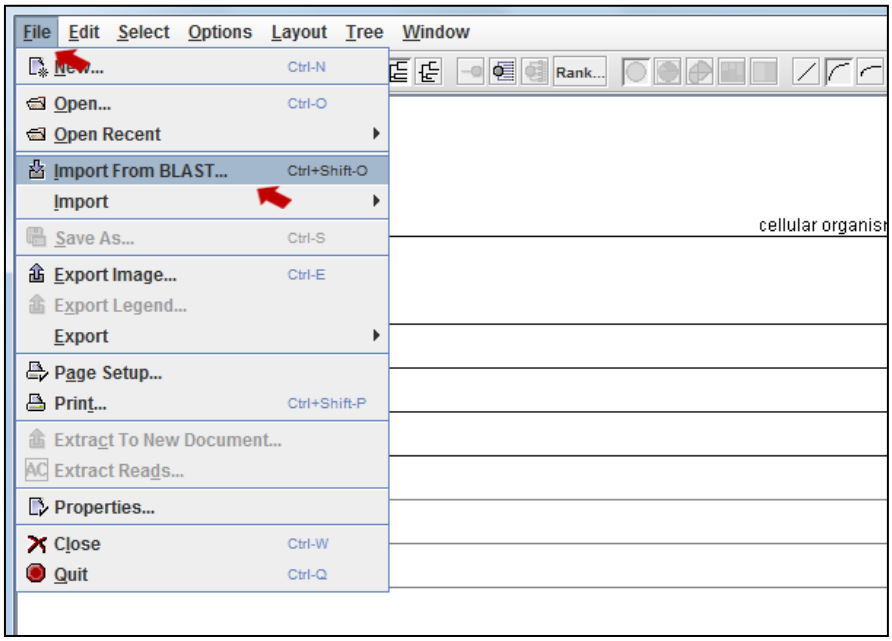
gi_taxid_prot : http://www-ab.informatik.uni-tuebingen.de/data/software/megan5/download/gi_taxid_prot-4March2015.zip

Para iniciar a utilização do programa MEGAN é preciso carregar, além do arquivo ".tblout" do BLAST, o arquivo fastacmd (multifasta com as sequências de CAZymes recuperadas pelo seu ID).

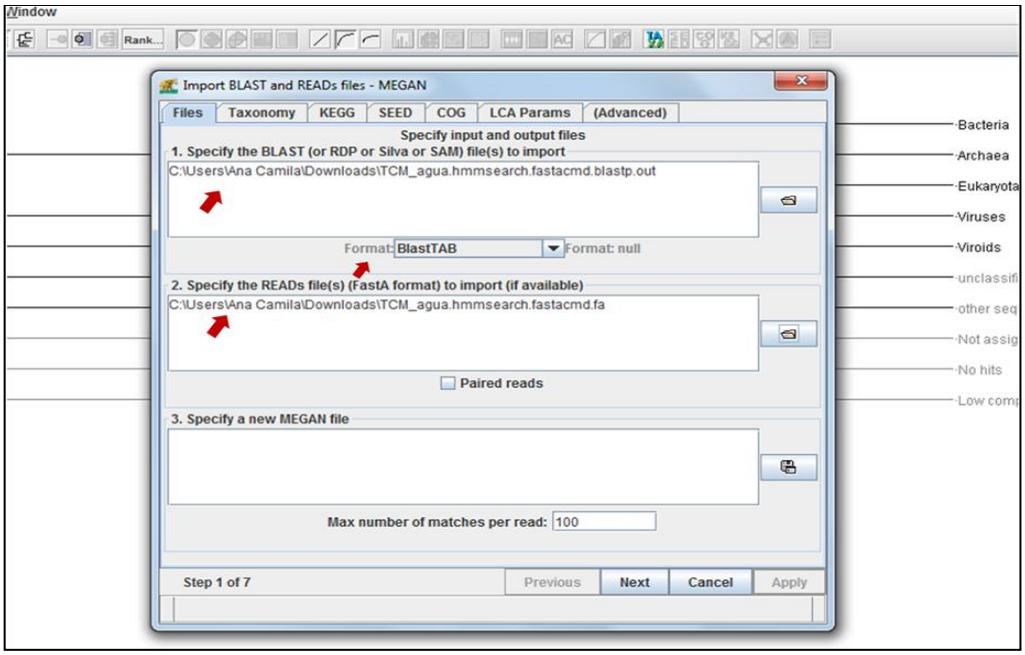
PASSOS

1. Primeiramente, carrega-se o resultado do BLASTP e o arquivo fastacmd (o mesmo utilizado no BLASTP):

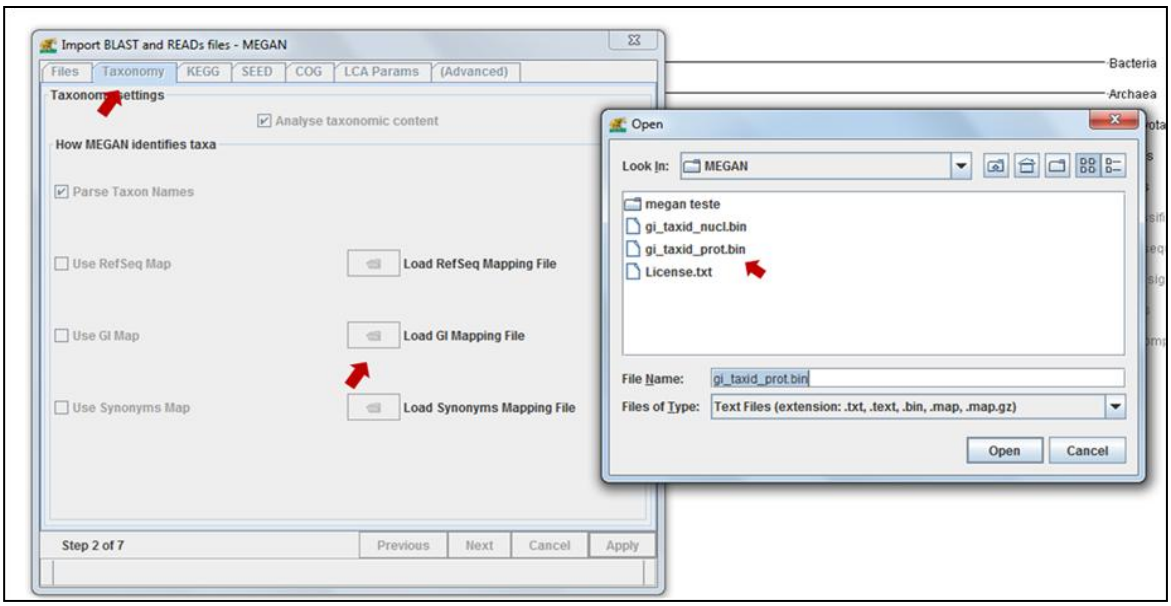
Clicar em "File" -> "Import from Blast":



Abrirá uma janela solicitando a indicação do diretório onde está o arquivo de saída do BLASTP e depois onde está o arquivo .fasta (fastacmd) utilizado para realizar o BLASTP. Após adicionar os arquivos, é necessário selecionar um formato. Ao invés de BlastP, selecionar a opção BlastTAB.

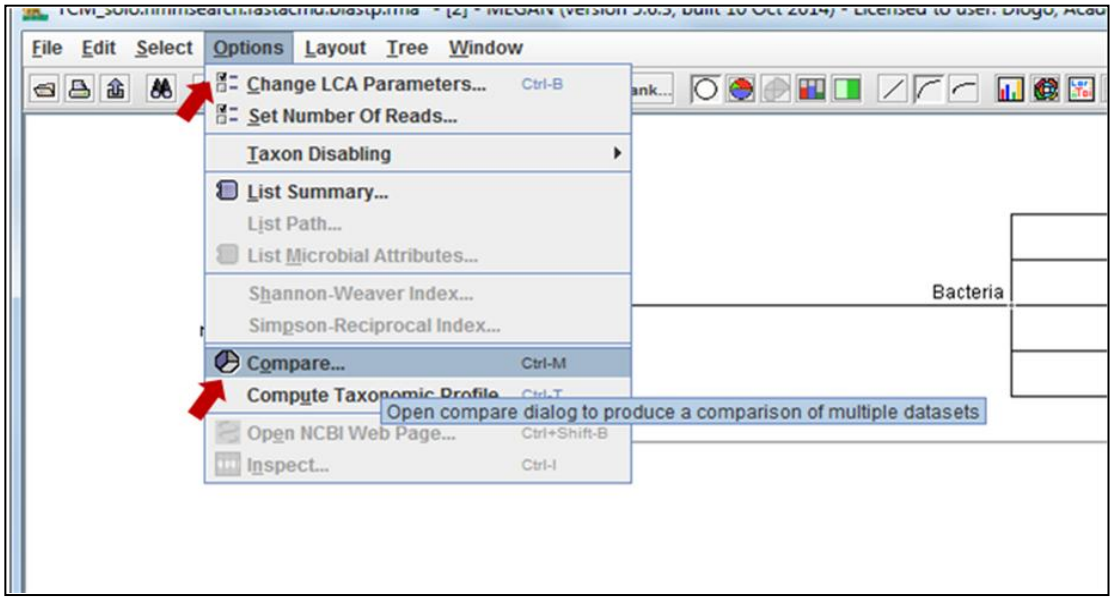


Em seguida, ainda nesta janela, abrir a aba “Taxonomy”, selecionar a opção “Load GI Mapping File” e colocar o caminho para o arquivo “gi_taxid_prot” (no caso de proteínas) ou “gi_taxid_nucl” (para nucleotídeos), que foram previamente baixados. Por fim, clicar em “Open” e depois em “Apply”.



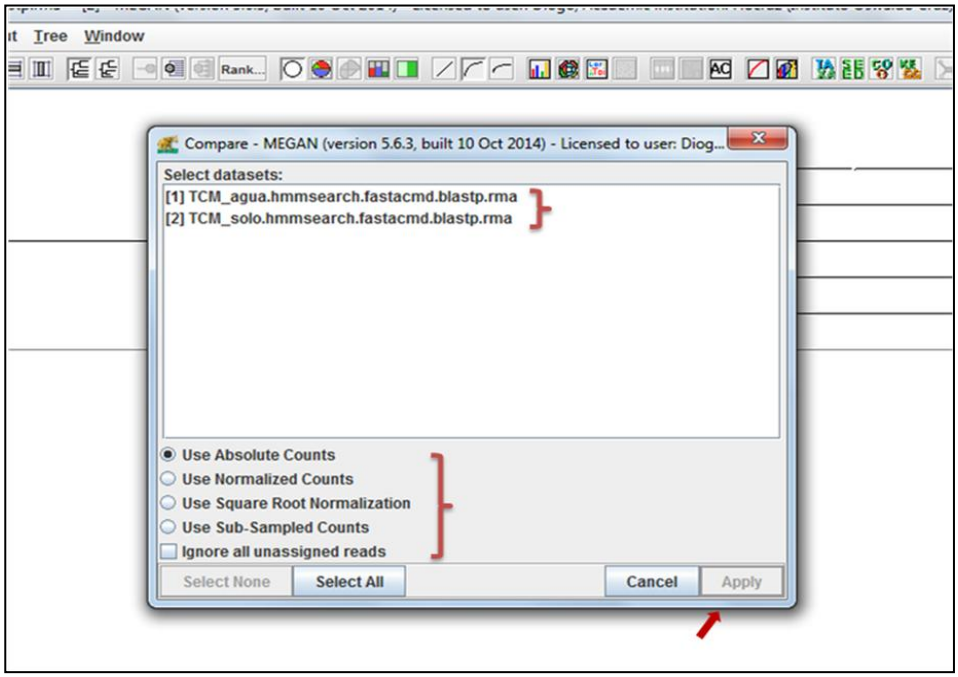
*Realizar estas etapas para todos os metagenomas que se deseja analisar, carregando um de cada vez, a partir da opção “File” -> “Import from Blast”.

2. Para realizar uma análise comparativa entre diferentes conjuntos de dados, com os metagenomas já carregados no MEGAN, selecionar, na barra de tarefas principal, “Options” e depois a opção “Compare”.

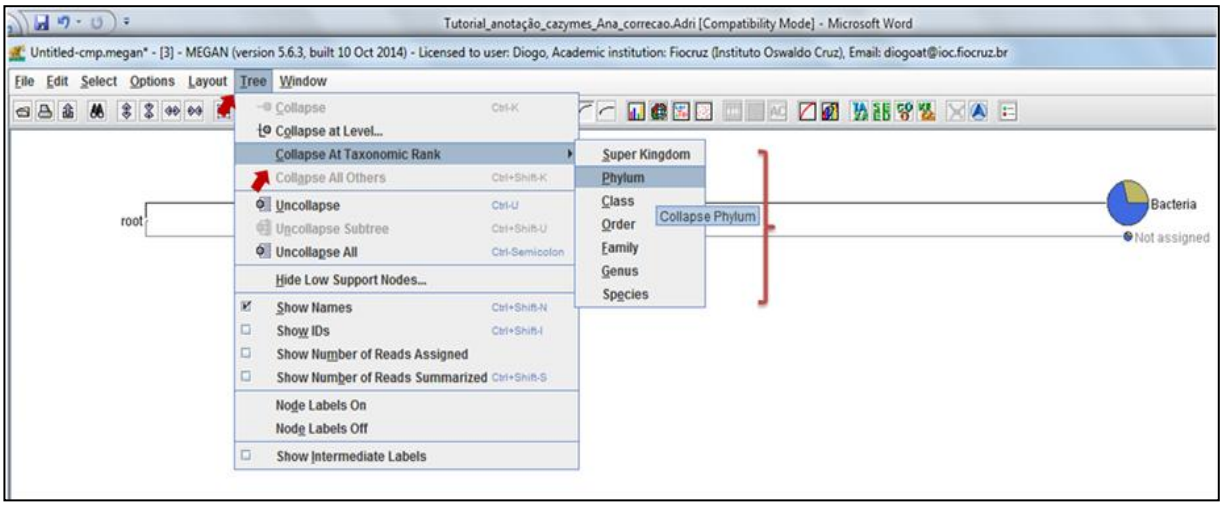


3. Existe a possibilidade de normalizar os dados no MEGAN para realizar a análise comparativa, de maneira que diferentes tamanhos de amostras serão contabilizados de forma proporcional. Uma vez que os metagenomas a serem comparados estejam escolhidos (para selecionar mais de um metagenoma manter pressionada a tecla “Ctrl”), pode-se selecionar uma das opções de normalização ou trabalhar com os dados de maneira absoluta.

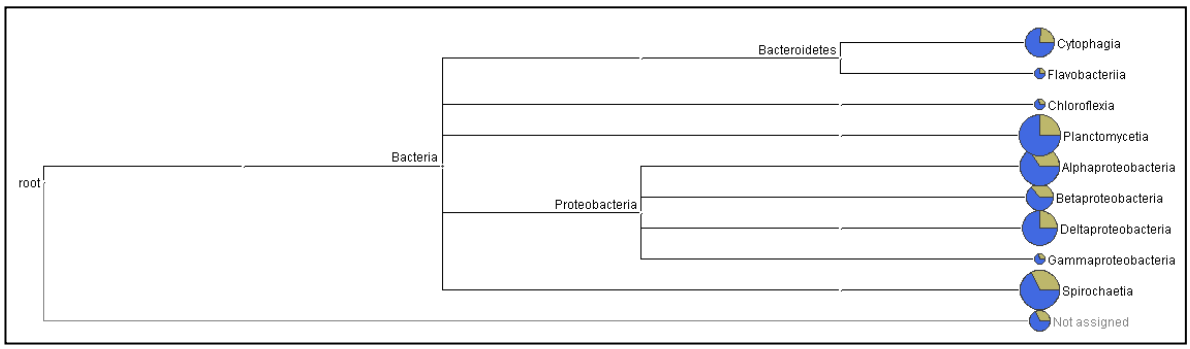
Uma vez que os metagenomas e a forma de utilização dos dados estejam escolhidos, clicar na opção “Apply”.

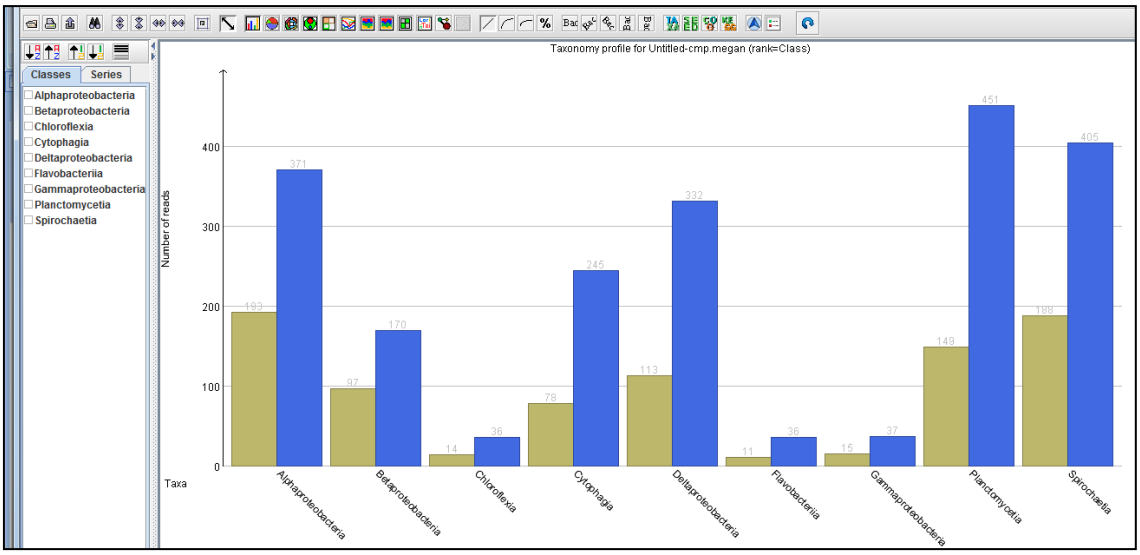


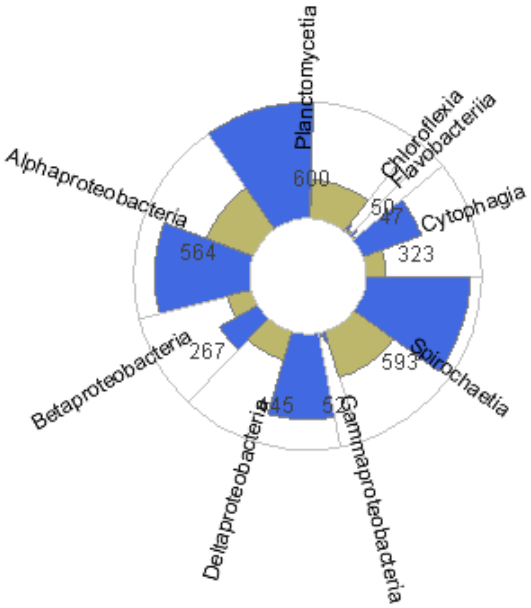
4. O resultado da comparação aparecerá, inicialmente, para o nível taxonômico "Reino", e será possível perceber a contribuição do nível hierárquico para cada conjunto de dados utilizados na comparação. Para observar os resultados em outros níveis (desde os filos até as espécies), deve-se selecionar a opção "Tree", que aparece na aba superior do programa e então selecionar a opção "Collapse At Taxonomic Rank", escolhendo o nível taxonômico desejado.



5. É possível então realizar as mais diferentes análises, com diversos tipos de gráficos ilustrando os resultados obtidos.







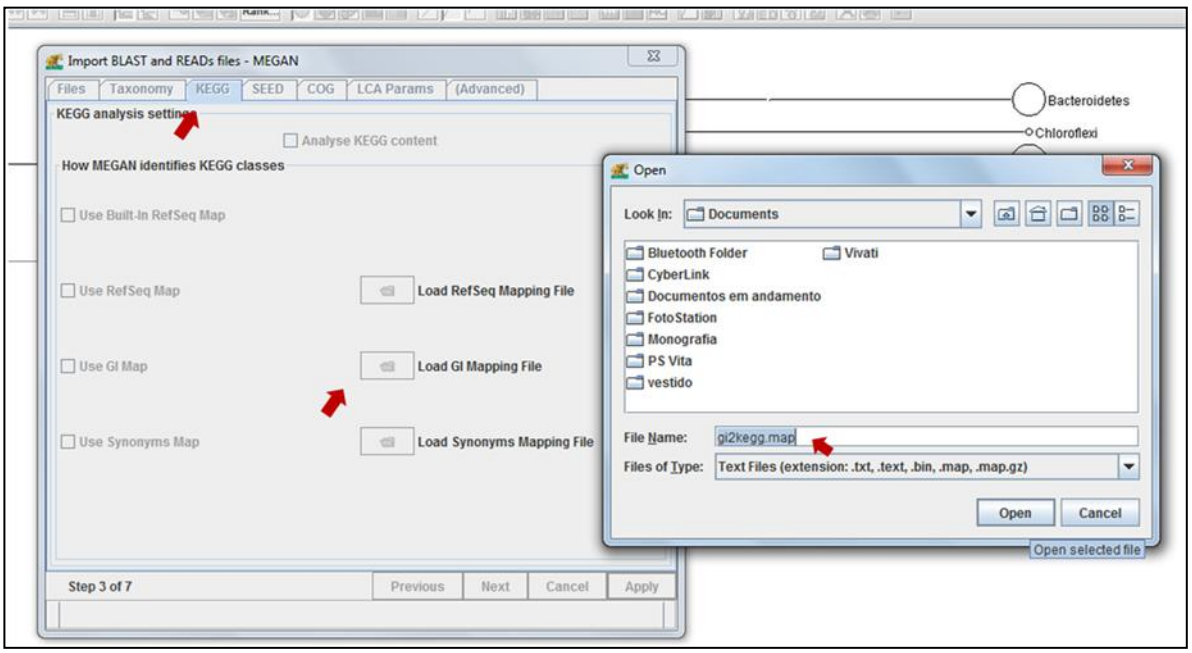
Obs: Para realizar uma análise funcional, existe a possibilidade de se mapear os resultados utilizando, por exemplo, as classificações KEGG e SEED. Para tal, é necessário baixar os seguintes arquivos: “gi2kegg.map.gz” e “gi2seed.map.gz”.

Os arquivos podem ser encontrados nos seguintes endereços:

gi2kegg.map.gz: http://www-ab.informatik.uni-tuebingen.de/data/software/megan5/download/gi2kegg.map.gz

gi2seed.map.gz: http://www-ab.informatik.uni-tuebingen.de/data/software/megan5/download/gi2seed.map.gz

Estes arquivos devem ser inseridos no momento inicial de carregamento dos dados (“File” -> “Import from Blast”). Na mesma janela onde são indicados os diretórios dos arquivos de saída do BLASTP e do arquivo FASTA (utilizado para rodar o BLAST), existem também as abas “KEGG” e “SEED”. Ao selecionar a aba “KEGG”, por exemplo, é possível indicar o arquivo “gi2kegg.map.gz” para realizar o mapeamento. O mesmo processo pode ser feito na aba “SEED” para mapear por esta classificação.



Identificação de possíveis novas sequências de CAZymes

Códigos (“scripts”) que foram utilizados:

1. Código para gerar uma tabela com os valores de cobertura da “query” vs. identidade do alinhamento, a partir do arquivo de saída do BLASTP e do arquivo FASTA (utilizado no BLASTP):

```
#!/usr/bin/perl

use strict;
use warnings;
use Data::Dumper;

$| = 1; # Do not buffer output

my ($file1,$file2) = @ARGV;

die "Please specify a FASTA file and a BLAST output file.\n" unless(defined $file1 &&
defined $file2);

open(FILE1,"perl -p -e 's/\r/\n/g;s/\n\n/\n/g' < $file1 |") or die "ERROR: Could not open file
$file1: $! \n";
if($file2 =~ /\.(gz|Z)$/) {
   open(FILE2,"gzip -dc $file2 | perl -p -e 's/\r/\n/g;s/\n\n/\n/g' |") or die "ERROR: Could not
open file $file2: $! \n";
} else {
   open(FILE2,"perl -p -e 's/\r/\n/g;s/\n\n/\n/g' < $file2 |") or die "ERROR: Could not open file
$file2: $! \n";
}

my ($progress,$counter,$part,$numlines);
$progress = $counter = $part = 1;
print STDERR "Estimate size of input data for status report (this might take a while for large
files)\n";
$numlines = &getLineNumber($file1);
print STDERR "\tdone\n";
#for progress bar
$progress = 0;
$counter = 1;
$part = int($numlines/100);

my (%lengths,$id);
print STDERR "Reading in FASTA file: $file1\n";
while(<FILE1>) {
   chomp();
   if(/^>(\S+)/) {
      $id = $1;
   } else {
      $lengths{$id} += length($_);
```

```
    }
    #progress bar stuff
    $counter++;
    if($counter > $part) {
        $counter = 1;
        $progress++;
        $progress = 99 if($progress > 99);
        print STDERR "\r\tstatus: ".int($progress)." \%";
    }
}
print STDERR "\r\tdone          \n";
print STDERR "Read in information for ".scalar(keys %lengths)." sequences\n";

$progress = $counter = $part = 1;
print STDERR "Estimate size of input data for status report (this might take a while for large
files)\n";
$numlines = &getLineNumber($file2);
print STDERR "\tdone\n";
#for progress bar
$progress = 0;
$counter = 1;
$part = int($numlines/100);
my @args;
my %hits;
my ($cov,$evalue);
print STDERR "Reading in BLAST output file: $file2\n";
while(<FILE2>) {
    chomp();
    @args = split(/\t/);
    $cov = int(100*((abs($args[6]-$args[7])+1)/$lengths{$args[0]}));
    if(!exists $hits{$args[0]}->{ci} || (exists $hits{$args[0]}->{ci} && $hits{$args[0]}->{ci}-
>{c} <= $cov && $hits{$args[0]}->{ci}->{i} <= int($args[2]))) {
        $hits{$args[0]}->{ci}->{c} = $cov;
        $hits{$args[0]}->{ci}->{i} = int($args[2]);
        $hits{$args[0]}->{ci}->{h} = $args[1];
    }
    foreach my $x (0..51) {
        if($args[10] < 10**(-1*($x-1))) {
            $evalue = $x;
        }
    }
    if(!exists $hits{$args[0]}->{ce} || (exists $hits{$args[0]}->{ce} && $hits{$args[0]}-
>{ce}->{c} <= $cov && $hits{$args[0]}->{ce}->{e} <= $evalue)) {
        $hits{$args[0]}->{ce}->{c} = $cov;
        $hits{$args[0]}->{ce}->{e} = $evalue;
        $hits{$args[0]}->{ce}->{r} = $args[10];
        $hits{$args[0]}->{ce}->{h} = $args[1];
    }
    #progress bar stuff
    $counter++;
```

```
    if($counter > $part) {
        $counter = 1;
        $progress++;
        $progress = 99 if($progress > 99);
        print STDERR "\r\tstatus: ".int($progress)." \%";
    }
}
print STDERR "\r\tdone          \n";
close(FILE1);
close(FILE2);
print STDERR "Found ".scalar(keys %hits)." sequences with hits using no thresholds\n";

my (%vals,@tmp,$out);
print STDERR "Calculating output values\n";
foreach my $id (keys %hits) {
    foreach my $t (keys %{$hits{$id}}) {
        $vals{$t}->{$hits{$id}->{$t}->{c}}->{$hits{$id}->{$t}->{($t eq 'ci' ? 'i' : 'e')}}++;
    }
}
$out = $file2.'_CItable.xls';
open(OUT, ">$out") or die "ERROR: could not write to output file $out: $! \n";
print OUT "#ID\tCoverage\tIdentity\tDBID\n";
foreach my $id (keys %hits) {
    print OUT join("\t",$id,$hits{$id}->{ci}->{c},$hits{$id}->{ci}->{i},$hits{$id}->{ci}-
>{h})."\n";
}
close(OUT);

$out = $file2.'_CEtable.xls';
open(OUT, ">$out") or die "ERROR: could not write to output file $out: $! \n";
print OUT "#ID\tCoverage\tEvalue\tDBID\n";
foreach my $id (keys %hits) {
    print OUT join("\t",$id,$hits{$id}->{ce}->{c},$hits{$id}->{ce}->{r},$hits{$id}->{ce}-
>{h})."\n";
}
close(OUT);

$out = $file2.'_CImatrix.xls';
open(OUT, ">$out") or die "ERROR: could not write to output file $out: $! \n";
print OUT "#Coverage vs. Identity values\n";
print OUT join("\t",",map {'C='.$_} (0..100))."\n";
foreach my $i (0..100) {
    @tmp = ('I='.$i);
    foreach my $c (0..100) {
        push(@tmp,(exists $vals{ci}->{$c}->{$i} ? $vals{ci}->{$c}->{$i} : 0));
    }
    print OUT join("\t",@tmp)."\n";
}
close(OUT);
```

```
$out = $file2.'_CEmatrix.xls';
open(OUT, ">$out") or die "ERROR: could not write to output file $out: $! \n";
print OUT "#Coverage vs. Evalue values\n";
print OUT join("\t",",map {'C='.$_} (0..100))."\n";
foreach my $e (0..51) {
   @tmp = ('E<='.(10**(-1*($e-1))));
   foreach my $c (0..100) {
      push(@tmp,(exists $vals{ce}->{$c}->{$e} ? $vals{ce}->{$c}->{$e} : 0));
   }
   print OUT join("\t",@tmp)."\n";
}
close(OUT);
print STDERR "\tdone\n";

sub getLineNumber {
   my $file = shift;
   my $lines = 0;
   open(FILE,"perl -p -e 's/\r/\n/g;s/\n\n/\n/g' < $file |") or die "ERROR: Could not open file
$file: $! \n";
   $lines += tr/\n/\n/ while sysread(FILE, $_, 2 ** 16);
   close(FILE);
   return $lines;
}
```

2. Código para plotar o gráfico "dot plot" no programa R, a partir da tabela gerada com a utilização do código anterior:

Para os dados das amostras de solo

```
#get required R packages

doInstall <- TRUE # Change to FALSE if you don't want packages installed.
toInstall <- c("ggplot2", "reshape2", "RColorBrewer")
if(doInstall){install.packages(toInstall, repos = "http://cran.us.r-project.org")}
lapply(toInstall, library, character.only = TRUE)

#load installed packages
require(ggplot2)
require(reshape2)
require(RColorBrewer)

#set plot color gradient
#myPalette <- colorRampPalette(brewer.pal(9, "GnBu"))
#myColors <- c("#FFFFFF", myPalette(99))
myColors <-colorRampPalette(c("#fff7bc","orange","#fe9929","dark orange",
"#e31a1c","red","#cb181d", "dark red"))(150)
#different color gradient (remove #-sign to uncomment)
```

```
#myPalette <- colorRampPalette(c("#DEEBF7","#3182BD"), space = "Lab")
#myColors <- c("#FAFAFA", myPalette(99))

#set path to the matrix.xls files
setwd("/home/afroes/Documents/AnaCamila/Ana_Camila/megan/");

#set file name prefix
filename <- "TCM_solo"
database <- ".blastp.out_" #  nr
matrix <- "CI" # CI or CE
charttitle <- paste(filename,"X NR")

#read matrix file
table <- read.delim(paste(filename, database,matrix,"matrix.xls",sep=""),header=FALSE);

#get data to plot
data <-
matrix(c(as.numeric(as.matrix(table[3:nrow(table),2:ncol(table)]))),nrow=101,ncol=101,byrow=F);
data.melt <- melt(data)
data.melt$Var1 <- data.melt$Var1-1
data.melt$Var2 <- data.melt$Var2-1

#create plot
p1 <- ggplot(data.melt, aes(y = Var1, x = Var2, fill = value))
p1 <- p1 + geom_tile() + labs(x = "Query Coverage [%]", y = "Alignment Identity [%]", title = charttitle)
p1 <- p1 + theme_bw()
p1 <- p1 + theme(legend.position = "right", legend.title=element_blank())
p1 <- p1 + scale_fill_gradientn(colours = myColors)
p1 <- p1 + coord_equal()
p1 <- p1 + scale_x_continuous(expand=c(0,0),breaks=c(10,20,30,40,50,60,70,80,90))
p1 <- p1 + scale_y_continuous(expand=c(0,0),breaks=c(10,20,30,40,50,60,70,80,90))
p1 <- p1 + theme(axis.line = element_blank(), panel.grid.minor = element_blank(),
panel.grid.major = element_blank())
print(p1)
```

Para os dados das amostras de água

```
#get required R packages

doInstall <- FALSE # Change to FALSE if you don't want packages installed.
toInstall <- c("ggplot2", "reshape2", "RColorBrewer")
if(doInstall){install.packages(toInstall, repos = "http://cran.us.r-project.org")}
lapply(toInstall, library, character.only = TRUE)

#load installed packages
```

```
require(ggplot2)
require(reshape2)
require(RColorBrewer)

### Show all the colour schemes available
display.brewer.all()
#display a divergent palette
display.brewer.pal(7,"BrBG")

#set plot color gradient
#myPalette <- colorRampPalette(brewer.pal(9, "GnBu"))
#myColors <- c("#FFFFFF", myPalette(99))
#myColors <- c("#C1D7E9", myPalette(100))
#myColors <- c("#88419D", "#EDF8FB", "#8B80BB", myPalette(5))
#myColors <- c("white", "orange", "yellow","red", "green", myPalette(10))
#myColors <- c("#000000", myPalette(99))
#myColors <-colorRampPalette(c("beige","dark white","dark
orange","orange","yellow","red","dark red"))(99)
myColors <-colorRampPalette(c("#fff7bc","orange","#fe9929","dark orange",
"#e31a1c","red","#cb181d", "dark red"))(150)
#heatmap.2(as.matrix(your data),col =colorRampPalette(c("light grey","green","green4",
"purple"))(100))
#myColors <- volcano.colors(10)
# Function to plot color bar

#set path to the matrix.xls files
setwd("/home/afroes/Documents/AnaCamila/Ana_Camila/megan/");

#set file name prefix
filename <- "TCM_agua"
database <- ".blastp.out_" #  nr
matrix <- "CI" # CI or CE
charttitle <- paste(filename,"X NR")

#read matrix file
table <- read.delim(paste(filename, database, matrix,"matrix.xls",sep=""),header=FALSE);

#get data to plot
#data <-
matrix(c(as.numeric(as.matrix(table[3:nrow(table),2:ncol(table)]))),nrow=101,ncol=101,byro
w=F);
data <-
matrix(c(as.numeric(as.matrix(table[3:nrow(table),2:ncol(table)]))),nrow=101,ncol=101,byro
w=F);
data.melt <- melt(data)
data.melt$Var1 <- data.melt$Var1-1
data.melt$Var2 <- data.melt$Var2-1

#create plot
p1 <- ggplot(data.melt, aes(y = Var1, x = Var2, fill = value))
```

```
p1 <- p1 + geom_tile() + labs(x = "Query Coverage [%]", y = "Alignment Identity [%]", title = charttitle)
p1 <- p1 + theme_bw()
p1 <- p1 + theme(legend.position = "right", legend.title=element_blank())
p1 <- p1 + scale_fill_gradientn(colours = myColors)
p1 <- p1 + coord_equal()
p1 <- p1 + scale_x_continuous(expand=c(0,0),breaks=c(10,20,30,40,50,60,70,80,90))
p1 <- p1 + scale_y_continuous(expand=c(0,0),breaks=c(10,20,30,40,50,60,70,80,90))
p1 <- p1 + theme(axis.line = element_blank(), panel.grid.minor = element_blank(), panel.grid.major = element_blank())
print(p1)
```