**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Propagação semi-automática de termos Gene Ontology a proteínas com
potencial biotecnológico para a produção de bioenergia**

Lucas Mitsuo Taniguti

Dissertação apresentada para obtenção do título de Mestre
em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2014**

Lucas Mitsuo Taniguti
Bacharel em Biotecnologia

Propagação semi-automática de termos Gene Ontology a proteínas com potencial biotecnológico para a produção de bioenergia

Orientadora:
Prof. Dra. **CLAUDIA BARROS MONTEIRO-VITORELLO**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2014

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Taniguti, Lucas Mitsuo

Propagação semi-automática de termos Gene Ontology a proteínas com potencial biotecnológico para a produção de bioenergia / Lucas Mitsuo Taniguti. - - Piracicaba, 2014.

79 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2014.

1. Perfil-HMM 2. Anotação 3. Biocuradoria 4. Bioinformática 5. Bioenergia I. Título

CDD 576.119245

T164p

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

*À minha família, namorada e
amigos, sem os quais este mundo
seria um lugar demasiado cinza*

AGRADECIMENTOS

O desenvolvimento deste trabalho contou com importantes incentivos e ajudas, essenciais para que se chegasse ao ponto no qual se encontra neste documento.

À Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ/USP pela possibilidade de realizar o curso de mestrado e por oferecer um ambiente essencial para o desenvolvimento profissional.

À professora Dra. Claudia Barros Monteiro-Vitorello, pela orientação, apoio, críticas, palavras de incentivo, experiência em curadoria de proteínas e por manter um grupo de trabalho com o qual sou feliz por ter sido parte.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro investido.

Ao professor Dr. João Carlos Setubal, que é colaborador do projeto e que contribuiu com ideias, sugestões e recursos. À sua aluna de mestrado, Fernanda Orpinelli, por ter ajudado diretamente na curadoria das famílias estudadas.

Ao Dr. João Paulo Kitajima, pela exposição a outras linhas de pesquisa bastante relevantes para a carreira que pretendo seguir.

Ao técnico em informática Carlos R. Macedônio, pelo auxílio e otimismo diante de problemas aleatórios surgidos no hardware de nosso servidor.

À Nathália de Moraes por ter revisado cada página e auxiliado a formatar as 103 referências deste documento e a Suzane Saito, por ter prontamente atendido meu pedido de elaborar críticas sobre o texto deste trabalho.

A todos que me incentivaram estudar, principalmente temas um pouco diferentes de minha área de formação, como linguagens de programação, LATEX, Linux e softwares livres.

E por fim agradeço aos que indiretamente contribuíram para a conclusão deste trabalho.

“We are all apprentices in a craft where no one ever becomes a master.”

—ERNEST HEMINGWAY (1898 - 1961)

“Be curious. Read widely. Try new things. What people call intelligence just boils down to curiosity.”

—AARON SWARTZ (1986 - 2013)

SUMÁRIO

RESUMO	11
ABSTRACT	13
1 INTRODUÇÃO	15
2 REVISÃO BIBLIOGRÁFICA	19
2.1 Bioenergia microbiana.....	19
2.2 Anotação de sequências biológicas	22
2.3 Bancos de dados biológicos.....	24
2.4 Vocabulários controlados – Gene Ontology.....	27
2.5 Famílias de proteínas	29
2.6 Alinhamento múltiplo de sequências.....	31
2.7 Homólogos distantes e métodos computacionais	33
3 MATERIAL E MÉTODOS	39
3.1 Proteínas referência	39
3.2 Caracterização das sequências primárias.....	39
3.3 Plataforma de interação	40
3.4 Fluxo da análise	41
3.4.1 Alinhamento múltiplo de sequências.....	42
3.4.2 Geração dos modelos.....	43
3.4.3 Buscas pelos membros.....	44
3.4.4 Gráfico com grupos taxonômicos e espécies.....	46
3.4.5 Armazenamento no banco de dados	46
3.4.6 Análise por redes de similaridade.....	47
3.4.7 Validação dos modelos	48
4 RESULTADOS E DISCUSSÃO.....	49
4.1 Descrição das funcionalidades do sistema MENGOfams	49
4.1.1 Propagação dos termos GOs em genomas completamente sequenciados.....	53
4.1.2 Enriquecimento da cobertura de anotações Gene Ontology nos registros	55
UniProt	55
4.2 Análise por redes de similaridade.....	57
4.3 Validação dos Modelos	58
5 CONSIDERAÇÕES FINAIS	61
REFERÊNCIAS	63
ANEXOS	75

RESUMO

Propagação semi-automática de termos Gene Ontology a proteínas com potencial biotecnológico para a produção de bioenergia

O aumento no volume de dados biológicos, oriundos principalmente do surgimento de sequenciadores de segunda geração, configura um desafio para a manutenção dos bancos de dados, que devem armazenar, disponibilizar e, no caso de bancos secundários, propagar informações biológicas para sequências sem caracterização experimental. Tal propagação é crucial, pois o fluxo com que novas sequências são depositadas é muito superior ao que proteínas são experimentalmente caracterizadas. De forma análoga ao *EC number* (*Enzyme Commission number*), a organização de proteínas em famílias visa organizar e facilitar operações automáticas nos bancos de dados. Dentro desse contexto este trabalho teve como objetivos a geração de modelos computacionais para famílias de proteínas envolvidas em processos microbianos biotecnologicamente interessantes para a produção de bioenergia. Para a geração dos modelos estatísticos foram escolhidas proteínas referência analisadas a priori em colaboração com o projeto MENGO¹. A partir da proteína referência foram realizadas buscas no UniProtKB com o objetivo de encontrar proteínas representativas para cada família e descrições de função com base na literatura científica. Com a coleção de sequências primárias das proteínas selecionadas foram realizados alinhamentos múltiplos de sequências com o programa MUSCLE 3.7 e posteriormente com o programa HMMER foram gerados os modelos computacionais (perfis de cadeia oculta de Markov). Os modelos passaram por consecutivas revisões para serem utilizados na propagação dos termos do Gene Ontology com confiança. Um total de 1.233 proteínas puderam receber os termos GO. Dessas proteínas 79% não apresentavam os termos GO disponibilizados no banco de dados UniProtKB. Uma comparação dos perfis-HMM com a utilização de redes de similaridade a um E-value de 10⁻¹⁴ confirmou a utilidade dos modelos na propagação adequada dos termos. Uma segunda validação utilizando um banco de dados construído com sequências aleatórias com base nos modelos e na frequência de codons das proteínas anotadas do SwisProt permitiu verificar a sensibilidade da estratégia quanto a recuperar membros não pertencentes aos modelos gerados.

Palavras-chave: Perfil-HMM; Anotação; Biocuradoria; Bioinformática; Bioenergia

¹ Disponível em: <http://www.mengo.biochem.vt.edu/>

ABSTRACT

Semi-automatic propagation of Gene Ontology terms to proteins with biotechnology potential for bioenergy production

The increase of biological data produced mainly by the second generation technologies stands as a challenge for the biological databases, that needs to address issues like storage, data availability and, in the case of secondary databases, to propagate biological information to sequences with no experimental characterization. The propagation is important since the flow that new sequences are submitted into databases is much higher than proteins having their function described by experiments. Similarly to the EC. number (Enzyme Commission number), an organization of protein families aims to organize and help automatic processes in databases. In this context this work had as goals the generation of computational models for protein families related to microbial processes with biotechnology potential for production of bioenergy. Several proteins annotated by MENGO², a project in collaboration, were used as seeds to the statistic models. Alignments were made on UniProtKB, querying the seeds proteins, looking for representatives for each family generated and the existence of function descriptions referenced on the scientific literature. Multiple sequence alignment were made on each collection of seeds proteins, representatives of the families, thorough the MUSCLE 3.7 program, and after were generated the computational models (profile Hidden Markov Models) with the HMMER package. The models were consecutively reviewed until the curator consider it reliable for propagation of Gene Ontology terms. A set of 1,233 proteins from UniProtKB were classified in our families, suggesting that they could be annotated by the GO terms using MENGOfam families. From those proteins, 79% were not annotated by the MENGO specific GO terms. To compare the results that would be obtained using only BLAST similarity measures and using pHMMs we generated similarity networks, using an E-value cutoff of 10⁻¹⁴. The results showed that the classification results of pHMMs are valuable for biological annotation propagation because it identifies precisely members of each family. A second analysis was applied for each family, using the respective pHMMs to query a collection of sequences generated by a null model. For null model were assumed that all sequences were not homologous and could be represented just by the aminoacid frequencies observed in the SwissProt database. No non-homologous proteins were classified as members by the MENGOfam models, suggesting that they were sensitive to identify only true member sequences.

Keywords: Profile-HMM; Annotation; Biocuration; Bioinformatics; Bioenergy

² Available at: <http://www.mengo.biochem.vt.edu/>

1 INTRODUÇÃO

Atualmente o Brasil é o segundo maior produtor de bioetanol do mundo, atingindo uma produção de aproximadamente 27,5 bilhões de litros na safra de 2013/2014 (UNIÃO DA INDÚSTRIA DE CANA-DE-ACÚCAR, 2014). A produção somada do Brasil e dos EUA, o qual detém o primeiro lugar, representou 84% da produção mundial no ano de 2013. Nos EUA o álcool é obtido principalmente a partir do milho, enquanto que no Brasil produção advém da utilização de cana-de-açúcar (*Saccharum spp.*) (LICHT, 2013). Com a tecnologia utilizada hoje pelas usinas, apenas o caldo de açúcar da cana é utilizado para o processo de fermentação e produção de etanol, pois o bagaço e as folhas não são aproveitados e são queimados para a produção de energia térmica (MACRELLI; MOGENSEN; ZACCHI, 2012; BUCKERIDGE et al., 2012).

Pesquisas estão sendo realizadas com o objetivo de desenvolver um processo capaz de utilizar material vegetal rico em tecidos lignocelulósicos, por exemplo o bagaço da cana-de-açúcar, como fonte de açúcar para a fermentação alcoólica. Os microrganismos biotecnologicamente interessantes para a produção desse tipo de bioenergia possuem limitações que impedem sua aplicação na produção em escala industrial, como a inibição por subprodutos da degradação da celulose (QI et al., 2014) ou a falta de certas enzimas catalíticas (ATSUMI et al., 2010). Apenas com o conhecimento dos mecanismos moleculares que regem seus metabolismos é possível avançar no desenvolvimento de tecnologias alternativas para a produção de combustíveis renováveis, seja alterando o metabolismo de um microrganismo ou seja introduzindo novas características, ausentes originalmente na espécie – metodologia conhecida como engenharia de metabolismos (HALLENBECK, 2014).

Diante deste cenário, a organização de bancos de dados biológicos é importante para que pesquisadores possam armazenar e acessar informações sobre as sequências de nucleotídeos ou aminoácidos conhecidas, analisar padrões, inferir funções e homologias. Na última década, muitos projetos disponibilizaram ferramentas que almejam facilitar a predição de funções de sequências ainda não caracterizadas, entre eles estão: o Pfam que gera modelos para domínios protéicos utilizando um método automático conhecido como *Automatic Domain Decomposition Algorithm (ADDA)* (PUNTA 20 et al., 2012); o Gene Ontology que possui um vocabulário controlado (e estruturado) de termos para serem utilizados na anotação (CONSORTIUM, 2010); o HAMAP, que gera modelos para famílias de proteínas baseados em suas arquiteturas de domínios e que possui uma série de condições e regras que coordenam a propagação das anotações contidas em cada família (PEDRUZZI et al., 2013) e

o *Carbohydrate-Active enZymes* (CAZy) (CANTAREL et al., 2009), que classifica enzimas em famílias envolvidas na síntese, no metabolismo ou no transporte de carboidratos. Tais famílias são criadas com base em proteínas experimentalmente caracterizadas e populadas por sequências presentes nos bancos de dados públicos com similaridade significativa (CANTAREL et al., 2009). Cada um desses bancos apresentam competências complementares na predição de função e análise de sequência de proteínas.

O Gene Ontology, citado acima, é um projeto que visa a geração de vocabulários controlados para a anotação de funções moleculares, de componentes celulares e de processos biológicos a sequências de produtos gênicos (LEONELLI et al., 2011). As enzimas listadas no CAZy, por exemplo, podem ser descritas utilizando também tais termos. O projeto MENGO (*Microbial ENergy processes Gene Ontology Project*) está desenvolvendo novos termos para o Gene Ontology que descrevem processos microbianos relacionados a produção de bioenergia, possibilitando dessa forma o enriquecimento das informações sobre as enzimas associadas a esses processos e a biotecnologia.

A coleção de sequências protéicas passíveis de serem anotadas por termos GO cresce a cada dia, principalmente em decorrência da redução dos custos de sequenciamento. Muitos microrganismos com características desejáveis para a produção de biocombustíveis tiveram seus genomas completamente sequenciados e disponibilizados em bancos de dados públicos. As técnicas de predição das regiões codificantes nos genomas (ORFs) costumam ser bastante precisas, no entanto, a inferência de características funcionais a partir de sequências primárias de proteínas ainda é um processo laborioso e que tem se tornado cada vez mais necessário (LIMA et al., 2009). Métodos automáticos são muitas vezes empregados em inferências funcionais, porém a qualidade das anotações que são propagadas muitas vezes é inadequada, pois estes processos costumam considerar apenas a porcentagem de identidade/similaridade entre as sequências comparadas (BAKKE et al., 2009).

Dentro desse contexto, o presente trabalho foi dedicado à geração de famílias representadas por perfis-HMM gerados com base em alinhamentos curados para a propagação semi-automática de termos do Gene Ontology anotados pelo projeto MENGO. Com esses modelos foi possível ampliar a cobertura dos termos GO sobre os registros de bancos de dados, bem como facilitar a anotação de sequências de interesse. O método é chamado semi-automático por incluir etapas onde um biocurador decide quais proteínas devem ser incluídas nas famílias, tendo como base para suas decisões características tanto da própria sequência

primária, como domínios e/ou regiões conservadas importantes para a função, quanto descrições encontradas na literatura científica disponível.

Assim, sob a hipótese de que famílias construídas com base em perfis-HMM podem ampliar a anotação dos termos do Gene Ontology gerados pelo projeto MENGO nos bancos de dados disponíveis, tivemos os seguintes objetivos específicos:

1. Estabelecer um fluxo de trabalho para a geração de famílias e perfis-HMM;
2. Realizar a curadoria de proteínas biotecnologicamente interessantes no contexto de bioenergia;
3. Desenhar um banco de dados capaz de armazenar e disponibilizar os produtos deste trabalho.

2 REVISÃO BIBLIOGRÁFICA

Essa revisão foi organizada em tópicos que descrevem os principais embasamentos teóricos estudados para a execução deste trabalho. Em (2.1) Bioenergia microbiana foi descrito de forma geral as formas de utilização de bioenergia microbiana e algumas técnicas experimentais que são utilizadas; em (2.2) - Anotação de sequências biológicas - foi discutido o contexto do trabalho dos biocuradores em anotar sequências e foram citadas as abordagens que podem auxiliar neste processo. Em (2.3) - Bancos de dados biológicos - foi abordada a origem dos bancos, os desafios iniciais, as principais formas de organização e exemplificação com o UniProt.

Posteriormente, em (2.4) Vocabulários controlados: Gene Ontology – foi apresentado a importância de padronizar os termos para anotações biológicas e sobre a estrutura organizacional deste banco de dados. Uma definição para famílias protéicas e os conceitos de homologia, identidade e similaridade podem ser encontrados em (2.5) Famílias protéicas. Como os modelos estatísticos utilizados neste trabalho são fortemente ditados pela qualidade dos alinhamentos múltiplos que são fornecidos foi redigida uma sessão sobre (2.6) Alinhamentos múltiplos, que descreve de maneira resumida o desafio computacional para a execução dessa tarefa e as principais etapas da estratégia de alinhamento conhecida como progressiva. Por fim foi apresentado uma pequena revisão sobre a evolução de métodos para detecção de motivos/domínios e de sequências de homólogos distantes.

2.1 Bioenergia microbiana

Bactérias são seres vivos diversos e os mais antigos que existem na face da Terra, sua diversidade metabólica tornou possível a ocupação de nichos tão extremos quanto fontes hidrotermais (FROCK; KELLY, 2012), ambientes com elevado nível de radiação (SINGH; GABANI, 2011) e até ambientes árticos (HARDING et al., 2011). Com a crescente demanda por fontes alternativas aos combustíveis fósseis algumas dessas bactérias ganharam destaque por serem capazes de gerar biocombustíveis a altas temperaturas a partir da degradação de tecidos, como por exemplo os lignocelulósicos que são abundantes em restos vegetais (SOREK et al., 2014). A característica termofílica de algumas dessas bactérias é importante para os processos de geração de bioenergia por trazer vantagens como: maior solubilidade do açúcar em altas temperaturas; destilação direta; e diminuição das chances de contaminações por outros organismos (SINGH, 2012).

A biossíntese de combustível a partir desses tecidos tem como principais vantagens amenizar os problemas do efeito estufa, pois estima-se que o uso de álcool evite a emissão de 2,85 Kg CO₂/ L de etanol anidro e 1,97 KG CO₂/L de etanol hidratado, segundo boletim técnico da Copersucar (MACEDO; LEAL; SILVA, 2004), e de ser renovável, em contraste com os combustíveis derivados de petróleo. As fontes alternativas atualmente pesquisadas são: bioetanol, biobutanol, álcoois de cadeia longa, biohidrogênio e biodiesel (YOGESH; RICHARD; BAIKUN, 2011).

Dentre as espécies de bactérias mais promissoras para a produção de bioetanol estão as que sintetizam um complexo protéico especializado na degradação de tecidos lignocelulósicos. Primeiramente identificado em *Clostridium thermocellum*, foi nomeado como *Carbohydrate binding factors (CBF)* e apenas em estudos posteriores identificado como um complexo enzimático, passando a ser chamado de celulosomo. Os celulosomos foram também identificados em outras bactérias anaeróbicas dos gêneros *Clostridium*, *Acetivibrio*, *Bacteroides*, *Ruminococcus* e aeróbicas do gênero *Vibrio* (BAYER et al., 1998). Com relação a energias renováveis o Brasil já possui uma estrutura respeitável.

A produção de bioetanol com a safra de 2013/2014 atingiu 27,5 bilhões de litros (UNIÃO DA INDÚSTRIA DE CANA-DE-ACÚCAR, 2014), o que corresponde a aproximadamente a 30% da produção mundial. Sua produção poderia ser ainda maior com a utilização de todo o seu potencial vegetal. Estima-se que apenas 1/3 do material vegetal colhido seja utilizado para a produção de etanol, o resto corresponde a tecidos lignocelulósicos que geralmente são queimados para obtenção de energia térmica (BUCKERIDGE et al., 2012).

O tecido lignocelulósico é parte da parede celular das células vegetais e é composto principalmente por lignina, hemicelulose e celulose. Etapas de pré-tratamento são necessárias para que a celulose cristalina seja separada e possibilite a degradação por endo e exo-nucleases. Após esse processo, as moléculas de celobiose e glicose são produzidas. A glicose pode ser absorvida diretamente pela célula microbiana enquanto que a celobiose necessita de etapas adicionais para ser convertida em energia. Experimentos com a levedura *Saccharomyces cerevisiae* mostraram que é possível aumentar a conversão de celobiose em etanol otimizando a expressão de duas proteínas: uma de membrana, responsável pelo transporte de celobiose, e uma citosólica, responsável pela sua conversão em moléculas de glicose. O aumento da eficiência na produção de etanol reportado por esse experimento foi de 47% (ERIKSEN et al., 2013).

Atualmente, o processo para a produção do bioetanol de segunda geração está em execução apenas em instalações demonstrativas, sendo que a maior produção com essa tecnologia pertence a uma empresa na Noruega, chamada Borregaard (LENNARTSSON; ERLANDSSON; TAHERZADEH, 2014). Para a obtenção do bioetanol de segunda geração é preciso a realização de um pré-tratamento químico dos tecidos lignocelulósicos, seguido pela adição de várias enzimas necessárias para a despolimerização da celulose em monômeros. A grande quantidade de enzimas que devem ser produzidas e isoladas para a obtenção de bioetanol por este método é um dos fatores que ainda o torna comercialmente inviável (ZANNONI; PHILIPPIS, 2014).

Os processos de produção de bioenergia podem ser categorizados em três grupos: i) hidrólise e fermentação em etapas separadas, como o citado no parágrafo anterior, ii) uso de microrganismos capazes de hidrolisar e fermentar naturalmente e iii) engenharia de bioprocessos, no qual características desejáveis de uma espécie são transferidas para uma outra de forma a aumentar o potencial de degradação (HALLENBECK, 2014).

A grande vantagem da utilização de espécies como *Clostridium thermocellum*, que se encaixa na segunda categoria descrita por Hallenbeck (2014), é a de que esse microrganismo é capaz de produzir grande quantidade de enzimas que hidrolisam as cadeias de celulose e fermentam os monômeros gerados. Isso diminui os custos de produção e minimiza a complexidade do processo industrial, uma vez que o mesmo organismo que degrada o material seria, também, responsável pela fermentação, ou seja, diretamente de celulose pode se obter o biocombustível (STEPHEN; MABEE; SADDLER, 2012).

A terceira categoria, mais complexa, seria o chamado bioprocessamento obtido por engenharia (em inglês, *Engineered consolidated bioprocessing*), em que organismos seriam modificados para que fossem capazes de metabolizar substratos de interesse e produzir biocombustíveis em altas taxas, seja pela incorporação de novas vias metabólicas ou pela inibição de vias indesejáveis. Muitos trabalhos vem mostrando que a síntese de novos biocombustíveis é possível pela incorporação de enzimas e vias metabólicas originalmente ausentes nos organismos. Já foi demonstrado que *Escherichia coli* modificada geneticamente pode produzir isobutanol, n-butanol e isopropanol (ATSUMI et al., 2008; ATSUMI et al., 2010) e que alterando-se vias que normalmente funcionam na biossíntese dos aminoácidos treonina e norvalina pode-se obter uma variedade de derivados de butanol (CONNOR; LIAO, 2008; SHEN; LIAO, 2008).

Casos de transferência de genes entre espécies diferentes não acontecem apenas quando é aplicada a engenharia genética, mas podem também ocorrer naturalmente no ambiente. Recentemente um trabalho apresentou fortes evidências de que a maior extinção já ocorrida na Terra pode ter sido consequência da transferência de um gene de bactérias degradadoras de celulose para uma arqueia produtora de metano, o que teria provocado aumento na temperatura global como consequência do agravamento do efeito estufa (ROTHMAN et al., 2014).

2.2 Anotação de sequências biológicas

A anotação de sequências biológicas é o processo responsável por preencher a lacuna entre as sequências e a biologia dos organismos (STEIN, 2001). A anotação pretende definir os mecanismos biológicos de uma molécula em toda sua complexidade e mapear esse conhecimento às respectivas regiões genômicas combinando abordagens experimentais e computacionais (YEATS et al., 2008).

Com o rápido desenvolvimento das tecnologias de sequenciamento o número de sequências biológicas disponíveis nos bancos de dados públicos vem aumentando em uma proporção que torna impossível, no atual cenário tecnológico, a realização de experimentos em laboratórios para caracterizar funcionalmente cada uma delas, tornando a anotação de sequências uma importante frente de trabalho. Dada a dimensão com que novas sequências são depositadas em bancos de dados é indispensável o uso de métodos automáticos de anotação, porém até o momento estão cronicamente sujeitos a erros de interpretação e propagação (MARCELLIN et al., 2013).

Toda anotação realizada tem grande visibilidade após ter sido publicada e estará sujeita a avaliações por outros grupos, muitas vezes utilizando novos recursos, como dados de transcriptômica ou proteômica para auxiliar na verificação da acurácia de tais anotações, e erros poderão ser reportados. No entanto, o número de genes que devem ser anotados em cada genoma recém sequenciado é grande, dificultando o trabalho de anotação manual (KOONIN; GALPERIN, 2003).

Erros na própria literatura científica são também fontes de anotações equivocadas de sequências. Um trabalho realizado por Poux et al. (2014) ilustrou por um estudo de caso como uma anotação sobre a função de uma proteína (SIRT5) altamente conservada entre várias espécies foi erroneamente curada e como esse erro foi sucessivamente propagado pela literatura. Apenas posteriormente, com estudos independentes utilizando cristalografia foi possível identificar o erro, verificando que a cavidade na estrutura molecular de SIRT5 correspondente à interação com grupos acetil em SIRT2 se apresentava muito maior, grande o

suficiente para que a interação fosse com grupos negativamente carregados de acilo, e não como previamente descrito como possuindo interação com grupos acetil.

A qualidade das anotações presentes nos bancos de dados biológicos constantemente é avaliada por grupos de biocuradores. Schnoes et al. (2009) analisaram quatro bancos de dados, UniProtKB/Swissprot, GenBank NR, UniprotKB/TrEMBL e KEGG, com o intuito de identificar falhas no processo de anotação das sequências. Para isso, utilizou um conjunto de 37 famílias de enzimas para as quais existem muitas informações experimentais. Foi identificado que em bancos onde todo o processo de anotação é automático pode existir até 60% de anotações erradas sobre função enzimática, e que ao longo do tempo o número de anotações erradas aumentou no banco de dados NR do GenBank no período de 1993 – 2005. No mesmo trabalho também foi mostrado que no SwissProt, bancos de dados com biocuradores (etapas manuais de curadoria), a taxa de erro observada foi próxima de 0% para algumas das famílias de enzimas testadas, reafirmando assim a importância de curadoria manual das sequências biológicas.

Por mais que estejam propensos a geração de anotações errôneas, os métodos automáticos são necessários para tornarem informativas as milhares de entradas disponíveis nos bancos de dados biológicos, sem tais métodos o acesso seria principalmente a sequências de bases A, T, C e G. Porém, frequentemente o que se ganha com quantidade acaba se perdendo com a qualidade das anotações que se obtém. O protocolo padrão de anotações automáticas, conhecido como predição funcional por homologia, transferência de anotação ou anotação transitiva, envolve a utilização do BLAST, um método heurístico para buscar rapidamente por sequências similares em grandes bancos de dados. Se uma sequência com alta similaridade é encontrada, as anotações contidas nela podem ser transferidas para a sequência com a qual a busca foi realizada (SJÖLANDER, 2013).

Métodos mais eficientes de busca, descritos com mais detalhes no item “Homólogos distantes e métodos computacionais” desta revisão, pontuam de forma específica cada posição das sequências pesquisadas, utilizando para isso um modelo pré-computado ou disponibilizado em algum banco de dados. Esses métodos permitiram a criação de programas preditores, como por exemplo para peptídeo sinal (PETERSEN et al., 2011), porções transmembrana (KROGH et al., 2001) ou estrutura de *Coiled-coil* (LUPAS; DYKE; STOCK, 1991).

O signalP é um preditor automático que permite inferir se determinada sequência de aminoácidos possui em sua porção N-terminal algum padrão conhecido de peptídeo sinal. As

sequências de peptídeos sinais são responsáveis por direcionar a proteína recém produzida para o retículo endoplasmático que, em eucariotos são eventualmente secretados ou, no caso de procariotos, permitem a travessia da membrana plasmática (PETERSEN et al., 2011). O TMHMM é outro programa que, de forma similar ao signalP, permite predizer se determinada sequência de aminoácidos possui resíduos que são característicos de porções transmembrana, presentes em proteínas integrais ou periféricas de membranas celulares (KROGH et al., 2001).

A construção de redes de similaridade também é útil para a inferência e verificação de anotações. Diferentemente dos estudos filogenéticos, as redes de similaridade permitem analisar as relações entre milhares de sequências de forma mais simplificada, representando a rede como um grafo, onde cada sequência é um nó e as relações entre elas são os vértices, que variam em tamanho de acordo com a similaridade entre elas. Essa técnica não substitui a análise filogenética, pois é baseada apenas na similaridade entre as sequências, não adotando nenhum modelo evolutivo (ATKINSON et al., 2009; SJÖLANDER, 2013), no entanto ganha em rapidez, que muitas vezes é desejada devido ao grande número de sequências.

Uma outra estratégia empregada é a aplicação de conceitos de jogos à programas que possam ser úteis para o desenvolvimento científico. Desta forma, aproveita-se o engajamento de diversas pessoas na web que não possuem formação científica para resolver problemas apresentados de forma que seja uma atividade simples para o cérebro humano, como um quebra-cabeças. Diversos projetos exploram essa idéia, entre eles existe o *Phylo* (KAWRYKOW et al., 2012), que é um jogo para o refinamento de alinhamentos múltiplos de sequências; o *Foldit*, utilizado para modelagem de proteínas; o *Dizeez* (LOGUERCIO; GOOD; SU, 2013), que visa gerar anotações de genes conhecidos associando-os as doenças; e o *CAFA* (RADIVOJAC et al., 2013), este último exige formação acadêmica e é uma competição onde os grupos participantes recebem um conjunto de proteínas que devem ser anotadas e, ao final do ciclo, é avaliado qual grupo teve melhor desempenho.

2.3 Bancos de dados biológicos

O contínuo avanço nas tecnologias de sequenciamento possibilita a geração de um fluxo cada vez maior de dados, derivados de diversos focos de estudo, como genômica, transcriptômica, proteômica ou metagenômica. Dessa forma, foi necessário o desenvolvimento de bancos de dados biológicos para o armazenamento e recuperação dessas informações de forma simples e segura. Os primeiros bancos de sequências de nucleotídeos foram: o EMBL, administrado no Instituto Europeu de Bioinformática, o GenBank sediado no Centro Nacional de Informações Biotecnológicas (EUA) e o DDBJ administrado pelo Centro

pra Informações Biológicas (Japão). Os pré-requisitos para um bom banco de dados biológicos são: (1) anotações de boa qualidade, (2) atualizações periódicas, (3) ser bem estruturado, (4) ter referência cruzada com outros bancos, e (5) uma boa interface para recuperação de dados (SACCONE; PESOLE, 2005).

A existência de três bancos distintos para a realização de depósitos de sequências de nucleotídeos feria alguns dos pré-requisitos listados acima, pois cada banco utilizava uma estrutura diferente para armazenar e disponibilizar seus dados. Foi necessária uma colaboração internacional para que os bancos fossem sincronizados periodicamente, tornando possível a busca por sequências que originalmente estivessem disponíveis em apenas um dos bancos. Atualmente, tal colaboração ainda é ativa e está envolvida em capturar, preservar e disponibilizar globalmente todas as sequências de nucleotídeos de domínio público, além de estabelecer formatos de arquivos e convenções para anotações (NAKAMURA; COCHRANE; KARSCH-MIZRACHI, 2013).

Os bancos de dados biológicos empregam diferentes estratégias para o armazenamento de suas informações. A mais simples delas é o armazenamento em arquivos texto, conhecidos como flat-files, onde toda a informação referente a uma determinada entrada é armazenada em linhas (um exemplo pode ser visto nesta entrada do Uniprot para endoglucanase D (<http://goo.gl/MCX0fE>). Tal formato é composto por linhas que se iniciam com padrões, por exemplo: ID para identificador, OS para espécie e OC para taxonomia no Uniprot.

Existem soluções mais sofisticadas para armazenamento de dados, como os bancos de dados relacionais e os arquivos de formato XML (*Extensible Markup Language*). Na arquitetura de bancos de dados relacionais os dados são armazenados em diversas tabelas que se relacionam. Cada linha das tabelas representa um registro e cada coluna um atributo, que deve ser do tipo pré-declarado durante a construção da estrutura do banco. É possível também controlar os valores que podem ser preenchidos para um novo registro com tabelas relacionadas pré-existent, dessa forma tem-se os valores de autores, por exemplo, já definidos. A essência dos bancos de dados relacionais, bem como a sua capacidade, é consequência do desenho das tabelas e das definições de relações entre elas (BESSANT; OAKLEY; SHADFORTH, 2014).

Bancos relacionais são facilmente integrados com diferentes linguagens de programação, como *Perl* e *Python*, permitindo assim o desenvolvimento de interfaces locais

ou via web que permitam a interação de diversos usuários simultaneamente com o banco, por exemplo: adicionando, visualizando e deletando registros.

No entanto, os dados de sequências e suas respectivas anotações geralmente são armazenados nos arquivos de texto (*flat-files*), pois apesar de sua simplicidade oferecem algumas vantagens: (1) pode-se realizar alterações em seus registros sem a necessidade de um sistema complexo para administração de bancos e (2) a troca de informações entre pesquisadores se dá de forma mais simples. A maior desvantagem desse sistema é que a busca por palavras-chaves é mais complexa e pode levar muito tempo dependendo do tamanho do banco. Para contornar esse problema muitos sistemas de indexação para flat-files foram desenvolvidos, oferecendo um index para facilitar as buscas de forma semelhante aos sumários de livros (SACCONE; PESOLE, 2005).

Dentre os bancos de dados de sequências protéicas existe o UniProt, que é constituído por quatro conjuntos de dados: (1) o *UniProt knowledgebase* (UniProtKB), que armazena informações de proteínas, capturadas de outros bancos colaboradores, como os participantes do consórcio internacional (NAKAMURA; COCHRANE; KARSCH-MIZRACHI, 2013); (2) o *Uniprot Archive* (UniParc) que correlaciona proteínas com diversos outros bancos de dados de forma a facilitar a obtenção das sequências que deram origem a uma mesma proteína, ou seja, uma busca realizada no UniParc é equivalente a buscar em diversos bancos simultaneamente, (3) o *UniProt Reference Clusters* (UniRef) que agrupa proteínas ou fragmentos semelhantes (identidade 100 , 90 ou 50%) e disponibiliza uma sequência representativa, de forma evitar o trabalho com sequências redundantes; e o (4) *UniProt Metagenomic and Environmental Sequences* (UniMES) criado com finalidade de atender os crescentes desenvolvimentos em metagenômica (CONSORTIUM, 2013).

O UniProtKB é dividido em dois bancos de dados, sendo um deles com entradas referentes à proteínas que tiveram, ao longo do processo de curadoria, etapas manuais, nas quais um biocurador pesquisou na literatura científica as caracterizações experimentais importantes para a anotação da proteína. Esse banco é conhecido como SwissProt e, devido à curadoria manual ser um processo lento e laborioso, existem relativamente poucas entradas – na versão 2014_08 representavam 0,66% do total de registros (APWEILER et al., 2004).

O outro banco que compõe o UniProtKB chama-se TrEMBL e é composto por entradas de proteínas que passaram por processos automáticos de anotação, obtendo assim anotações como de prováveis domínios funcionais, termos do Gene Ontology e

presença/ausência de peptídeos sinais ou de porções transmembrana. O UniProt é atualizado mensalmente, sendo que em sua última versão (09-Jul-2014) existiam 546.000 entradas no SwissProt e 79.824.243 no TrEMBL³.

2.4 Vocabulários controlados – Gene Ontology

A falta de padronização de termos descritivos para sequências biológicas dificulta o trabalho de pesquisadores assim como os processos automáticos executados periodicamente nos bancos de dados. O principal objetivo do Gene Ontology (GO) é fornecer uma linguagem comum e estruturada para anotar produtos gênicos, facilitando assim tarefas como a comparação de genes entre espécies. Esse trabalho teve início em 1998 como um projeto colaborativo de três bancos de dados de organismos modelos, *Flybase* (*Drosophila*), o *Saccharomyces Genome Database* (SGD) e o *Mouse Genome Database* (MGD), e, desde então, vem crescendo e incorporando novos bancos (ASHBURNER et al., 2000).

Seus termos descritivos são organizados em três vocabulários estruturados e controlados (ASHBURNER et al., 2000), que permitem inferir três aspectos diferentes das anotações referentes aos produtos gênicos. Essas ontologias são intituladas como:

- Componente Celular, que se refere a localização celular ou complexo ao qual o produto gênico está associado (por exemplo, proteína presente no núcleo);
- Função Molecular, representando a atividade primária em questão (por exemplo ligação a carboidratos);
- Processo Biológico, que informa o objetivo biológico final do produto gênico (por exemplo, metabolismo de purinas).

O Gene Ontology é organizado em uma hierarquia do tipo *Directed Acyclic Graph* (DAG) e seus termos podem ter zero, um ou muitos descendentes que originalmente se relacionavam de duas formas: *is_a*, é um(a), ou *part_of*, faz parte do(a). Por exemplo, o cromossomo nuclear faz parte do (*part_of*) núcleo e é um (*is_a*) cromossomo, enquanto que o cromossomo mitocondrial é um (*is_a*) cromossomo, porém não faz parte do núcleo. Com o aumento do escopo do Gene Ontology, novas formas de relacionamento foram adicionadas para descrever de forma mais precisa as interações entre produtos gênicos, como por exemplo a inclusão de relações como *has_part* (é parte) e *regulates* (regula) (CONSORTIUM, 2010).

³ Dados disponíveis em: <ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt>

Em cada entrada do GO existe um indicador do tipo de evidência que levou o curador a criar o termo em questão, por exemplo: *IDA* – Inferido por Experimento Direto, *ISS* – Inferido por Similaridade de Sequência, *IMP* – Inferido por Fenótipo Mutante. A lista completa pode ser encontrada em <http://www.geneontology.org/GO.evidence.shtml>. Como a biologia está em constante desenvolvimento, muitas vezes é preciso alterar alguns termos, porém nunca nenhuma evidência é excluída do banco de dados do Gene Ontology. Os termos que passam por modificações ou que são removidos recebem um comentário informando o motivo dessas alterações e passam a pertencer ao conjunto obsoleto do GO.

O maior gargalo do Gene Ontology está na atribuição de termos GO aos produtos gênicos. Esse processo é feito manualmente por curadores que lêem e sintetizam conjuntos de publicações científicas com a caracterização funcional destes produtos gênicos. A atribuição manual é lenta e trabalhosa e soluções automáticas vem sendo empregadas para acelerar esse processo. Skunca, Altenhoff e Dessimoz (2012) avaliaram que a qualidade das anotações atribuídas de forma automática vem melhorando e afirmam que já é similar a precisão das anotações atribuídas a partir de evidências não-experimentais curadas (como inferências por similaridade de sequência ou estrutura).

Nesse contexto se insere o projeto MENGO, que é uma iniciativa colaborativa multi-institucional, financiada pelo Departamento de Energia dos EUA e que tem como objetivos criar novos termos para o Gene Ontology relacionados a processos microbianos envolvidos na produção de bioenergia e anotar manualmente genomas de interesse com esses novos termos (TORTO-ALALIBO et al., 2014). Os processos biológicos identificados como importantes pelo grupo MENGO são:

1. Degradação de material vegetal - produtos gênicos que possam substituir os processos termoquímicos atuais.
2. Sacarificação - enzimas que sejam ativas e estáveis a altas temperaturas e que não sejam inibidas pelos produtos finais.
3. Fermentação solventogênica - produtos gênicos que permitam a ocorrência dos processos de sacarificação e fermentação simultaneamente.
4. Produção fermentativa de H₂ - Produtos gênicos que permitam tolerância à inibição termodinâmica consequente do H₂ na degradação fermentativa de biopolímeros.

5. Metanogênese - Foco nos termos que ajudem a entender as interações responsáveis por esse processo na natureza para que possam ser implementados em processos industriais.
6. Síntese de hidrocarbonetos de cadeia longa - Produtos gênicos envolvidos em vias de biossíntese de ácidos graxos e isoprenóides que possam ser utilizados nas vias de síntese de biocombustíveis.
7. Bioconversão de lignina para combustível - Produtos gênicos que possibilitem a conversão de lignina aos seus monômeros ou em outros substratos que possam ser usados na produção de biocombustíveis.

Até o momento, os curadores do projeto MENGO geraram 660 novos termos (acesso 16/7/14) que visam cobrir os processos citados anteriormente. A lista completa destes termos pode ser acessada em <http://goo.gl/CksMUo>.

2.5 Famílias de proteínas

De forma semelhante ao que foi teorizado por Charles Darwin, as proteínas também estão sujeitas às forças da seleção natural, onde aquela com um bom “*fitness*” propiciaria reprodução diferencial do portador e, caso as características de sua sequência de aminoácidos conferissem vantagem sobre as demais presentes no ambiente, poderia vir a ser fixada em uma população. Os genes codificadores de proteínas presentes no DNA de todos os seres vivos estão sujeitos à mutações, inserções, deleções e duplicações. Quando uma proteína é duplicada, a nova cópia está livre para sofrer eventuais mutações, uma vez que a outra cópia continua funcional, caso essas alterações confirmem funcionalidades úteis ao organismo ela será mantida. Um bom exemplo sobre esse evento descrito é a hemoglobina. No genoma humano existem 5 genes que codificam para hemoglobinas que compartilham a mesma função, de transportar oxigênio, e que teriam surgido de um único gene ancestral ao longo da evolução dos animais, a cerca de 800 milhões de anos atrás (SJÖLANDER, 2013).

Ao conjunto de proteínas que surgiram a partir de um ancestral comum é dado o nome de família protéica. Dentro de uma família todas as sequências são consideradas homólogas, podendo ser ortólogas ou parálogas. No início da era das ômicas a aplicação incorreta dos termos ortólogo e parálogo foi tão frequente que seu uso foi criticado (PETSKO, 2001) e proposto então que em estudos de relação entre sequência, estrutura e função o uso dos termos ortólogo e parálogo deveriam ser descontinuado (com exceção de trabalhos com foco em eventos de duplicações e especiações), uma vez que na maioria das vezes as informações

obtidas pelos experimentos são insuficientes para determinar os eventos de especiação e duplicação que deram origem aos atuais genomas (JENSEN, 2001).

De forma geral proteínas homólogas em espécies diferentes, porém, com mesma função são chamadas de ortólogas enquanto que proteínas homólogas dentro de uma mesma espécie e com funções diferentes são chamadas de parálogas (SELZER; MARHÖFER; ROHWER, 2008).

A homologia não é um atributo de quantidade, diz-se simplesmente que uma sequência é ou não é homóloga a outra com base em sua identidade, similaridade ou pela comparação de suas estruturas tridimensionais. Identidade/similaridade pode ser utilizada para inferir homologia, porém o inverso não é verdadeiro: homologia nem sempre está relacionada com alta identidade/similaridade entre sequências (SJÖLANDER, 2013).

Dado que alta porcentagem de identidade/similaridade entre sequências está relacionada com semelhança entre as estruturas tridimensionais e consequentemente à conservação de funções, muitos trabalhos são desenvolvidos com o objetivo de agrupar sequências protéicas de mesma função (como Panther (MI et al., 2005), TigrFams (HAFT; SELENGUT; WHITE, 2003), Phylofacts (DATTA et al., 2009)) ou de identificar domínios funcionais como Pfam (PUNTA et al., 2012), CDD (MARCHLER-BAUER et al., 2005), ProDom (CORPET; GOUZY; KAHN, 1998) a fim de facilitar a propagação de anotações funcionais obtidas a partir de dados experimentais, que são poucos considerando a quantidade de sequências descritas, dando origem assim a bancos de dados de famílias de domínios/proteínas.

O CAZy é um exemplo de banco de dados de famílias protéicas que tem como foco a classificação de domínios funcionais presentes em enzimas relacionadas com a criação, modificação ou degradação de ligações glicosídicas. Por exemplo, neste banco existe a classe de enzimas *Glycosil Hydrolases* que é composta por 133 famílias (acesso 08 de setembro de 2014) e são agrupadas em clans de acordo com a ancestralidade, conservação da estrutura terciária, mecanismo de atuação ou resíduo catalítico e consequentemente independente de um alto valor de identidade entre as sequências (HENRISSAT; BAIROCH, 1996). O Pfam é um banco de dados de domínios funcionais conservados de proteínas que é dividido em dois conjuntos: o Pfam-A com registros de alta qualidade com etapas manuais de curadoria e o Pfam-B, onde os registros são gerados de maneira totalmente automática. Atualmente a

cobertura do Pfam sobre as proteínas presentes no UniProtKB é de aproximadamente 80% (PUNTA et al., 2012).

O sistema *High-quality Automated and Manual Annotation of Proteins (HAMAP)* foi desenvolvido diante da necessidade de propagar automaticamente informações como função protéica, atividades enzimáticas e localização subcelular, bem como domínios conservados, para as proteínas não-curadas presentes no UniProt/TrEMBL de bactérias e arqueias. O HAMAP possui uma coleção de famílias de proteínas manualmente curadas para classificação e regras para as anotações que devem ou não ser passadas para os membros da família (PEDRUZZI et al., 2013).

No contexto de biocuradoria, o uso de redes de similaridade para análise e ilustração de tendências funcionais entre conjuntos amplos de proteínas é uma técnica relativamente nova. Sua metodologia consiste em se obter valores indicativos de similaridade entre todas as sequências que se deseja estudar, valores que podem ser facilmente obtido com a utilização do programa BLAST para alinhamentos locais. Com os valores indicativos de similaridade é possível utilizar programas, como o Cytoscape (SHANNON et al., 2003), para agrupar as sequências mais próximas tornando cada sequência da análise um nó e o comprimento dos vértices entre elas proporcional à similaridade obtida no alinhamento. Pelo programa Cytoscape, atualmente, é possível trabalhar apenas com redes de aproximadamente 106 vértices, porém novas ferramentas, como o Pythoscape, estão sendo desenvolvidas (BARBER; BABBITT, 2012).

Para a validação dessa metodologia foram utilizadas algumas superfamílias. Superfamílias são grandes agrupamentos de proteínas relacionadas, podendo conter tanto ortólogos como parálogos, e com representantes de espécies muito diferentes, o que consequentemente proporciona diferenças estruturais nas proteínas e assim diferenças em suas funções (SJÖLANDER, 2013). A validação utilizando algumas superfamílias bem caracterizadas identificou algumas vantagens da representação em forma de rede, mostrando que, de forma geral, as redes de similaridade geradas por alinhamentos locais par-a-par (todos contra todos) de um grande número de proteínas divergentes possui boa correlação com as distâncias obtidas por métodos filogenéticos (ATKINSON et al., 2009).

2.6 Alinhamento múltiplo de sequências

Proteínas são moléculas complexas compostas por longas cadeias de aminoácidos, cada qual com suas próprias características físico/químicas. Essas cadeias se apresentam

naturalmente em conformações tridimensionais que lhes conferem funções específicas. As regiões importantes para a manutenção das funções protéicas são chamadas de domínios funcionais ou motivos, que tendem a ser mais conservados entre as sequências de uma mesma família. Os domínios existentes nas sequências de proteínas frequentemente possuem estruturas globulares que podem ter funções específicas (SJÖLANDER, 2013). O alinhamento de sequências é o procedimento mais comumente aplicado em análises de bioinformática que, de forma simplificada, pode ocorrer em dois cenários principais:

1. Buscas em um bancos de dados por homólogos a uma sequência de interesse;
2. Alinhamento entre várias sequências relacionadas para se inferir os efeitos da evolução ao longo de todas as sequências de uma família proteica (NOTREDAME, 2002).

Diversas análises exigem que sejam realizados alinhamentos múltiplos de sequências biológicas em alguma etapa da metodologia, como: inferência de relações filogenéticas, busca por homólogos de elementos funcionais, classificação de proteínas em famílias e desenho de marcadores para detecção. A aplicação de conceitos de programação dinâmica e métodos heurísticos é necessária para que seja possível realizar tais análises, principalmente em decorrência do grande número de sequências que vem sendo disponibilizadas nos bancos de dados (RUSSELL, 2014).

O problema de alinhamento múltiplo de sequências pode ser considerado como uma generalização do uso da metodologia de alinhamento de duas sequências, em que os principais algoritmos, como os de Needleman-Wunsch e Smith-Waterman para alinhamentos globais e locais respectivamente, aplicam conceitos de programação dinâmica para aumentar a eficiência com que realizam as comparações. De forma resumida, o método de Smith-Waterman compara todos os resíduos de uma sequência aos da outra e armazena a informação de quão bem as sequências se alinham a cada resíduo comparado, realizando assim todos os processamentos necessários para obtenção do melhor alinhamento local presente na comparação das duas sequências (SMITH; WATERMAN, 1981). Adicionalmente ao método de programação dinâmica, os programas da série FASTA e BLAST utilizam métodos heurísticos para se obter alinhamentos aproximados, ganhando tempo ao se reduzir o espaço a ser pesquisado (ALTSCHUL et al., 1990).

Os programas de alinhamento múltiplo mais populares são os que utilizam modelos de alinhamento baseados em árvore, pois sequências similares tendem a possuir um ancestral

comum mais próximo que as menos similares. O método adotado por programas como CLUSTALW, T-Coffee, MUSCLE e MAFFT é conhecido como progressivo e possui basicamente três etapas de processamento (RUSSELL, 2014):

1. Uma matriz de distância é calculada para cada par dentro de N sequências a serem analisadas. Existem vários métodos para se calcular essa distância sendo um deles a contagem de k-mers em comum entre elas (utilizado pelo programa MUSCLE).
2. Um algoritmo de clusterização é aplicado para a geração de uma árvore guia para as sequências. Algoritmos como UPGMA ou single linkage são bastante empregados.
3. Toda ‘folha’ da árvore gerada corresponde a uma sequência fornecida à análise e os nós representam um alinhamento, ou perfil, das duas sequências filhas. O alinhamento múltiplo começa das duas sequências mais próximas identificadas pela árvore e progressivamente vão sendo adicionadas sequências mais distantes.

Um problema desse método é que erros nas etapas iniciais de alinhamento par a par são propagados, isso se torna um grave quando as sequências dos alinhamentos iniciais são distantemente relacionadas. Para contornar esse problema métodos de refinamento iterativos pós processamento são adotados em programas como MUSCLE e MAFFT (PAIS et al., 2014).

Alinhamentos precisos são importantes para estudos com famílias protéicas, pois para se inferir as relações de determinada sequência à uma família costuma-se utilizar modelos probabilísticos ou, no caso dos perfis, modelos não probabilísticos, que são gerados a partir de um alinhamento múltiplo dos membros descritos como presentes na família. Esses modelos permitem que seja possível a identificação de homólogos mais remotos pertencentes à família por aplicar ao alinhamento um sistema de pontuação específica a cada posição, método que é conhecido como *positionspecific scoring matrix (PSSM)* (EDDY, 1998).

2.7 Homólogos distantes e métodos computacionais

Se existe um conjunto de proteínas catalogadas como pertencentes à uma mesma família é possível, por meio de alinhamento par a par, inferir se existem outras sequências homólogas depositadas em bancos de dados públicos. Para isso, a forma mais simples seria escolher uma das proteínas da família e realizar alinhamentos, como por exemplo utilizando o programa BLASTp. O processo poderia ser repetido para uma segunda proteína da família, e assim sucessivamente até ter realizado a análise com todas. Métodos como o empregado pela

série de programas BLAST assumem que todas as posições de uma sequência representam caracteres independentes uns em relação aos outros (EDDY, 1996).

Entretanto, a metodologia empregada pelo BLAST deixa de utilizar uma informação importante que o alinhamento múltiplo de sequências pode fornecer: posições específicas tendem a ser mais conservadas quando comparadas a outras. Por exemplo, nas hemoglobinas as regiões de loop são mais diversas quando comparadas as porções de alfa hélice (DURBIN, 1998). Para se fazer uso dessa informação foram propostos modelos computacionais para a representação do alinhamento de sequências referente a uma mesma família, como os perfis (profiles) (GRIBSKOV; MCLACHLAN; EISENBERG, 1987) e perfis de Cadeia Ocultas de Markov (daqui em diante será referenciado como perfil-HMM)(EDDY, 1998).

A aplicação de técnicas que empregam o uso de cadeias ocultas de Markov é bastante ampla na área de biologia molecular, estando presente na predição de genes em sequências de DNA (MAJOROS; PERTEA; SALZBERG, 2004), identificação de regiões transmembrana (KROGH et al., 2001), presença de peptídeo sinal, domínios funcionais (Pfam) e organização de famílias protéicas (TIGRfams (HAFT; SELENGUT; WHITE, 2003), Phylofacts (DATTA et al., 2009), PANTHER (MI et al., 2005)). Anteriormente ao uso de perfis-HMMs e profiles eram utilizadas expressões regulares, comumente empregadas em programas nativos de unix como grep, awk e em perl, na busca por padrões em sequências biológicas (KROGH, 1998). Essa busca consistia de se informar ao programa o padrão que se desejava encontrar, como no exemplo da

Figura 1 – Alinhamento hipotético representando um motivo de interesse

Poderia ser proposto uma expressão regular como:

[AN-][NF]V[FA]AN[QG]

Indicando que o primeiro resíduo poderia ser um A, N ou uma lacuna, seguido por um N ou F, logo em seguida seria, obrigatoriamente, um V, e dessa forma até o fim da expressão regular. Apesar das expressões regulares serem flexíveis elas não utilizam estatísticas na

busca. Para elaborar um método que possibilitasse a identificação de homólogos mais remotos de uma família Gribskov, McLachlan e Eisenberg (1987) propuseram uma metodologia que utiliza a matriz de substituição elaborada por Dayhoff (STRASSER, 2010) em conjunto com a frequência que cada resíduo é observada em uma dada coluna do alinhamento múltiplo para criar um profile.

O profile gerado a partir do alinhamento múltiplo permite atribuir uma pontuação específica para cada resíduo que possa ser encontrado em determinada posição de uma sequência a ser analisada. Assim, um aminoácido 'a' na posição 'p' de uma sequência de interesse recebe um valor no profile segundo a fórmula:

$$M(p, a) = \sum_{b=1}^{20} W(p, b)Y(a, b)$$

Onde $W(p, b)$ é a frequência com que o aminoácido b aparece na posição p do alinhamento e $Y(a, b)$ é o valor encontrado na matriz de substituição de Dayhoff. Para pontuar regiões de inserção/deleção é utilizado um método heurístico que considera a maior lacuna (gap) observada no alinhamento múltiplo para a coluna em questão (GRIBSKOV; MCLACHLAN; EISENBERG, 1987). A pontuação dada pelo profile gerado a partir do alinhamento da figura 1, para o caso de ocorrer um aminoácido b na posição 2 de uma dada sequência, seria:

$$M(2, b) = \frac{4}{5} Y("N", b) + \frac{1}{5} Y("F", b)$$

Na coluna 13 do alinhamento da figura 1 aparecem apenas N e F, todos os demais aminoácidos possíveis tem probabilidade 0 e por isso não foram exibidos na conta.

Apesar de oferecer uma forma de utilizar estatísticas produzidas a partir de um alinhamento múltiplo, esse método ainda descarta a informação sobre quão alta é a evidência de um determinado resíduo ser conservado. Ou seja, um alinhamento com, por exemplo, uma coluna conservada em 100 sequências alinhadas terá o mesmo peso que se observado em um alinhamento com apenas 2 sequências (DURBIN, 1998). Dada a importância de modelos computacionais para famílias protéicas, constantemente os métodos vem sendo aprimorados. É possível concluir a partir da literatura que existe grande competição entre os defensores do uso de profiles contra os de perfis-HMMs (DURBIN, 1998; GRIBSKOV; VERETNIK, 1996).

Perfil-HMM é um tipo especial de cadeia oculta de Markov que possibilita a pontuação e penalização dependente da posição em um alinhamento. De forma geral, é calculada a probabilidade de cada resíduo ser observado em uma determinada posição do alinhamento – obtendo assim um valor chamado de probabilidade de emissão (KROGH, 1998). A figura 2 ilustra a organização do modelo.

Figura 2 – Esquema da composição de um perfil-HMM. Os quadrados são chamados de estados principais (*main states*), os losangos são estados de inserção (*insert states*) e os círculos são estados nulos (*null states*). As setas entre eles correspondem às probabilidades de transição (*transition probabilities*)

A frequência com que cada aminoácido é observado em determinada posição é armazenada nos estados principais e as probabilidades de transição são obtidas com a frequência em que cada sequência do alinhamento adota determinada transição/caminho. Os estados de inserção são utilizados para modelar regiões altamente variáveis e os estados nulos permitem que seja pulado um estado, como por exemplo para posições onde ocorrem lacunas no alinhamento múltiplo (KROGH, 1998).

Para evitar que a observação de que um determinado resíduo em uma posição em que está ausente no alinhamento múltiplo do modelo seja altamente penalizada, é utilizada uma técnica chamada de '*pseudocounts*', em que valores baixos de probabilidade são dados à todos os possíveis resíduos, como por exemplo, proporcional à frequência com que eles são observados em um banco de dados. Essa é uma informação a priori que é fornecida ao modelo (DURBIN, 1998). Os escores apresentados em buscas com a utilização de perfis-HMMs são valores logarítmicos da razão da probabilidade de se observar a sequência pontuada sobre a probabilidade dada à hipótese nula, de que cada aminoácido, por exemplo, tem probabilidade 1/20 de ocorrer ao longo de toda sequência. Dessa forma, quando é atribuído um valor negativo a uma sequência pode-se interpretar que o modelo nulo a descreve melhor.

Assim, para o alinhamento múltiplo do exemplo acima poderia se ter como pontuação de uma sequência:

$$Escore = \frac{P(\text{"NAVFANQ"} \mid \text{modelo})}{\left(\frac{1}{20}\right)^7}$$

Onde $P(\text{"NAVFANQ"} \mid \text{modelo})$ é a probabilidade de ser observada a sequência “NAVFANQ” dado o modelo perfil-HMM, a probabilidade $\left(\frac{1}{20}\right)^7$ de se observar uma sequência de tamanho 7 ao acaso assumindo que todos resíduos tem a mesma probabilidade de serem observados e o Log para tornar o valor da razão mais adequada a interpretação.

Os perfis-HMM utilizam uma base probabilística formal associada, por se tratar de um caso especial de cadeia oculta de Markov, e aplicam uma teoria consistente para ajustar os escores de inserção e deleção, além de serem robustos o suficiente para serem utilizados em grandes bancos de dados sem comprometer o tempo das análises (EDDY, 2011).

3 MATERIAL E MÉTODOS

Nesta sessão foram descritas as principais etapas para a criação do banco de dados deste trabalho, detalhes do website desenvolvido com formulário para criação de famílias e visualização de resultados, e sobre o processo de curadoria aplicado na criação das famílias para propagação dos termos que foram anotados pelo projeto colaborador, o MENGO.

3.1 Proteínas referência

O projeto MENGO é um trabalho em desenvolvimento na *Virginia Polytechnic Institute and State University*, nos EUA, e que tem como principais objetivos (i) a geração de novos termos relacionados a bioenergia microbiana para o Gene Ontology e (ii) a anotação manual de genomas de interesse. As proteínas que foram estudadas no presente trabalho foram as mesmas anotadas pela equipe MENGO. Esse projeto recebeu do colaborador acima descrito uma planilha com as seguintes informações:

- Código identificador (ID) NCBI ou UniProt da proteína referenciada.
- Breve descrição da atividade protéica (ex. *Cellulase activity*).
- Termo do Gene Ontology atribuído à proteína.
- Identificador único PubMed do artigo com a evidência para a atribuição do termo.
- Nível de evidência apresentado no artigo utilizado.

Foram utilizados identificadores de proteínas do UniProtKB para fazer referência a proteína anotada pelo MENGO, pois neste banco de dados já existem algumas informações sumarizadas (por exemplo: espécie, domínios, número EC e outras referências cruzadas). Alguns dos registros foram fornecidos com identificadores referentes a outros bancos, como por exemplo, do GenBank. Estes foram separados e submetidos ao ID Mapping, que possibilita rapidamente a obtenção do identificador UniProt correspondente. O *ID Mapping* é uma das ferramentas disponíveis no site UniProt e pode ser acessado pelo link: <http://www.uniprot.org/mapping/>. Posteriormente, para separar apenas as proteínas de bactérias ou arqueias foram utilizados os *Taxon identifier* (Taxid) 2 e 2157, respectivamente.

3.2 Caracterização das sequências primárias

Para a caracterização das sequências primárias das proteínas referência foram utilizados os seguintes preditores: SignalP, para presença/ausência de peptídeo sinal (PETERSEN et al., 2011) e TMHMM, para predição de porções protéicas de transmembrana (KROGH et al., 2001). Foram realizadas também buscas nos bancos de dados CAZy (CANTAREL et al., 2009), PANTHER (MI et al., 2005), TIGRfams (HAFT; SELENGUT;

WHITE, 2003), HAMAP (LIMA et al., 2009) e PROSITE (SIGRIST et al., 2002) a fim de coletar as informações relevantes já disponíveis.

A literatura científica foi consultada quando disponível, buscando-se evidências da função exercida pelas proteínas referência estudadas e informações experimentais sobre os membros que optamos por incluir na família.

3.3 Plataforma de interação

Como este projeto é em colaboração com o projeto MENGO, foi desenvolvido um banco de dados e um website especialmente para a execução deste trabalho com o intuito de facilitar a divulgação dos resultados com o grupo localizado nos EUA, bem como auxiliar no processo de curadoria com os colaboradores do Brasil. As informações de cada família produzida foram armazenadas em um banco de dados SQL e exibidas em páginas HTML únicas, para cada versão de família gerada. A tabela com as informações de famílias possui os seguintes itens:

- Identificador da família criada – Um padrão com o nome do banco de dados, seguido por hífen e três dígitos referentes a família, seguido por mais um hífen e dois dígitos referentes à versão da família, seguindo então: mengofam-ddd-dd.
- Pontuação de corte – Ao executar o processo de criação de famílias, a menor pontuação que uma proteína utilizada para gerar o modelo recebe é tido como a pontuação de corte. Abaixo dela todas as proteínas precisam de maior atenção para se propagar informações geradas pelo projeto MENGO.
- Proteína referência – O identificador UniProt da proteína referência anotada pela equipe do projeto MENGO.
- Nome da família – Um nome descritivo da anotação associada à família.
- Autor – Curador responsável.
- Descrição – Um texto que resume as principais características da família, como quais são os domínios obrigatórios, presença de peptídeo sinal e se faz parte de um complexo.
- Anotações GO atribuídas pelo MENGO – Termos GO que a família possibilita propagar.
- Membros – Todas as proteínas do UniProtKB que receberam uma pontuação superior à definida como linha de corte.

- Sementes – Identificadores das sequências de proteínas utilizadas como sementes para família.
- Referências cruzadas com os bancos Pfam, PANTHER, CAZy, HAMAP.

A tabela descrita acima foi relacionada com a tabela que armazena as informações referentes às proteínas referência, de forma que o campo “Proteína Referência” da tabela de famílias foi considerada como uma chave estrangeira da tabela de proteínas referência. A figura 3 mostra o diagrama de entidade-relacionamento da organização do banco.

Figura 3 – Diagrama da entidade-relacionamento do banco de dados MENGOFams. Relação entre a tabela Família, que armazena dados de cada versão de família produzida, com a tabela Proteína referência, que possui os dados das proteínas anotadas pelo projeto colaborador MENGO

3.4 Fluxo da análise

Foi elaborado um formulário, disponível no *website*⁴, que permitiu adicionar novas famílias ao banco de dados do MENGOfams, para isso foram necessárias as sequências das proteínas semente para a geração do perfil-HMM, bem como informações referentes a família a ser gerada. As etapas foram descritas a seguir e a Figura 4 ilustram o fluxograma de trabalho.

⁴ <http://mengofams.lbi.iq.usp.br/>

3.4.1 Alinhamento múltiplo de sequências

Partindo do pressuposto de que sequências semelhantes devam possuir funções semelhantes foram realizadas buscas por sequências de mesma função utilizando o programa BLASTP. As proteínas anotadas pelo projeto MENGO foram utilizadas como query contra a coleção de proteínas do UniProtKB (Trembl + SwissProt) e os resultados passaram por verificação manual. Para cada registro retornado da busca foram aplicadas as análises listadas no item 3.2 – Caracterização das sequências primárias.

Com as sequências selecionadas a partir dos alinhamentos com o BLASTP foram gerados alinhamentos múltiplos utilizando o programa MUSCLE, que foi bem avaliado em testes de acurácia realizados por Sedaghatinia et al. (2009). Os alinhamentos múltiplos consistem de uma etapa importante na construção de modelos, pois é a partir das frequências de aminoácidos observadas em cada coluna que são estimados os parâmetros para o perfil-HMM.

A etapa de seleção de sequências sementes para o modelo é repetida até que a família esteja bem definida, ou seja, que o alinhamento múltiplo se apresente consistente com os domínios conservados esperados para a família e que as proteínas preditas como membros pelo modelo possuam as mesmas características importantes anotadas na proteína referência.

Figura 4 – Etapas para a geração perfis-HMM. A etapa manual da geração dos perfis-HMM envolvem a busca por proteínas de mesmas características da referência no banco de dados UniProtKB e consulta da literatura científica a procura de informações que auxiliem na seleção, como por exemplo domínios característicos. Os processos automáticos envolvem a geração do alinhamento múltiplo das sequências sementes pelo programa MUSCLE, a geração do perfil-HMM com as observações do alinhamento

gerado e busca no SwissProt e TrEMBL por proteínas passíveis de receberem os termos GO da família gerada. Essas etapas são repetidas até que a família esteja suficientemente definida

3.4.2 Geração dos modelos

Foi utilizado o pacote de programas HMMER (EDDY, 1998) para a criação de perfis-HMM das famílias estudadas. Entre os programas desse pacote estão presentes o HMMbuild, que é responsável por gerar os modelos a partir de estatísticas inferidas do alinhamento múltiplo, o HMMsearch que permite utilizar um perfil HMM para realizar buscas por homólogos distantes em bancos de dados e o HMMscan, que computa qual a melhor família presente em um banco de perfil-HMMs para representar determinada sequência de interesse.

O alinhamento múltiplo de sequências, descrito no item 2.6, é utilizado para calcular as frequências de aminoácidos e gerar o modelo, pelo uso do programa HMMbuild. Com exceção das extremidades do alinhamento múltiplo, as colunas onde são observadas mais do que de 50% de lacunas não são utilizadas para a geração da sequência consenso do modelo. O programa aplica uma metodologia para que sejam dados pesos diferentes para as sequências, atribuindo uma pontuação diferente posição-específica (HENIKOFF; HENIKOFF, 1994), a fim de reduzir possíveis redundâncias no alinhamento múltiplo e atribuir maior pontuação às sequências que enfatizem a diversidade.

Pelo website MENGO_{fams} foi desenhado um formulário para criação de famílias que contém campos a serem preenchidos com as informações necessárias para disparar o processo automático de geração de família (Figura 5). O formulário é composto por duas etapas: a primeira envolve fornecer informações gerais e as sequências das proteínas sementes que geram o modelo e na segunda etapa os termos GO a serem propagados, bem como as referências cruzadas com PANTHER, CAZy, HAMAP e Pfam, são exibidos para que sejam confirmados ou modificados pelo biocurador (Figura 6).

O campo ‘Status’ foi proposto para que fosse fornecido o estado do processo de curadoria em que a família se encontra, os possíveis valores são: “revisada”, “preliminar” e “na lista”. Esse campo deve ser preenchido pelo biocurador responsável. De maneira geral, o estado “revisada” significa que a família já se encontra pronta para uso, ou seja, que pode se atribuir os termos do Gene Ontology anotados pelo MENGO para todas as proteínas que receberam um valor de escore superior ao do *cutoff* estabelecido.

Em “*reference protein*” o curador deve indicar qual a proteína a ser utilizada para a geração do modelo. Essa proteína é a que foi manualmente anotada com termo Gene Ontology pelo projeto MENGO. As opções desse botão *dropdown* são preenchidas com os

registros contidos na tabela de proteínas referências do banco SQL que, além do termo GO a ser propagado, possui as referências cruzadas capturadas do banco de dados UniProtKB.

“*Family name*” é o título que o biocurador atribui à família e que geralmente é uma breve descrição sobre os termos GO a serem propagado. O campo seguinte, “*Description*”, permite que seja incluído algumas características importantes da família, como presença/ausência de peptídeo sinal, porções transmembrana e domínios funcionais.

Figura 5 – Formulário para geração de famílias - Parte 1. Formulário para submissão de sequências sementes para geração de famílias

Por fim, em “*seed sequences*” devem ser inseridas as sequências protéicas no formato fasta para a geração do perfil-HMM. Uma vez que a primeira parte do formulário é preenchida uma nova página exibe um outro formulário (Figura 6), que permite ao curador modificar as referências cruzadas capturadas da proteína referência.

3.4.3 Buscas pelos membros

Uma cópia do UniprotKB foi mantida em um servidor localizado no IQ/USP para que seja possível utilizar localmente os perfis-HMM gerados pelo trabalho na busca dos membros

de cada família produzida. Sempre que há uma nova versão do UniProtKB disponível a cópia mantida no servidor é substituída pela nova.

Check information (A3DBF5)
Correct any incorrect information, but keep the same patterns

Family version

GO_

Pfam_

CAZy_

PANTHER_

HAMAP_

Figura 6 – Formulário para geração de famílias - Parte 2. Segunda etapa do processo de submissão de famílias. Neste as anotações do Gene Ontology e as respectivas referências cruzadas devem ser verificadas

O *output* do HMMsearch contém informações de E-valor das sequências capturadas, indicativo do número de hits com o respectivo escore que seria esperado de se obter em um banco de dados com o mesmo número de entradas, porém contendo apenas sequências randômicas não relacionadas.

Outra informação é o valor da pontuação, *Log-odds* escore, obtida por um cálculo da razão entre a probabilidade de observar a sequência dado o perfil-HMM sobre a probabilidade de observar a mesma sequência dada a hipótese nula. Essa pontuação é atribuída independentemente do número de registros presentes no banco de dados onde a busca foi realizada.

Sempre que um novo perfil-HMM é gerado por esse trabalho o programa HMMsearch é executado a fim de identificar os membros da família e de analisar proteínas que receberam uma pontuação próxima ao limite estabelecido como nota de corte. As proteínas próximas à nota de corte passaram por inspeção manual e, caso apresentem as características descritas

para a família, a sequência foi selecionada para a inclusão da mesma ao alinhamento múltiplo gerador do modelo e posteriormente uma nova versão da família foi gerada.

3.4.4 Gráfico com grupos taxonômicos e espécies

Um vez que os identificadores dos membros de uma determinada família foram definidos, uma busca foi feita para capturar as informações taxonômicas dos membros da família. Um script com essa função foi elaborado, onde para cada identificador fornecido é consultado o respectivo registro UniProt no formato texto e capturado apenas a linha relativa à espécie (OS) ou a classificação taxonômica (OC). Finalizada a lista de espécies/grupos foi realizada uma contagem e um gráfico de setores foi gerado utilizando uma Interface de Programação de Aplicativo (em inglês, *API*) do Google Charts.

3.4.5 Armazenamento no banco de dados

Ao longo do processo para a geração das famílias, algumas informações foram armazenadas em uma tabela utilizando o Sistema Gerenciador de Banco de Dados SQLite3, que utiliza a linguagem SQL, e outras informações foram organizadas na forma de arquivos texto, já formatadas para a geração dos gráficos apresentados. A estrutura do banco está descrita no item “Plataforma de interação” e os arquivos texto foram disponibilizados para cada família, como descritos a seguir:

- Sequências das proteínas sementes no formato fasta. As sequências que o curador selecionou para que fossem utilizadas para a geração do modelo foram armazenadas para que a família criada possa ser reproduzida, caso necessário. Todos os arquivos foram identificados com o identificador da família seguido pelo autor, ex. “mengofam-002-03_author.fa”.
- Arquivo contendo o alinhamento múltiplo das sequências sementes. O alinhamento múltiplo realizado pelo programa MUSCLE foi armazenado para que possa ser disponibilizado em conjunto com os dados da família. O alinhamento é disponibilizado no formato Clustal.
- Arquivo com o perfil de cadeia oculta de Markov (perfil-HMM). A matriz com as probabilidades de cada aminoácido em cada uma das posição do consenso do alinhamento das sementes foi armazenada. Esse arquivo foi gerado pelo programa HMMbuild. O arquivo contendo a matriz de probabilidades foi identificado com sufixo “.hmm”

- Arquivos com os dados taxonômicos. Esse arquivo foi criado com o objetivo de armazenar as informações de espécies e grupos presentes nos resultados de cada família. O arquivo contendo os dados de distribuição taxonômica foi identificado com o sufixo “.pie”.

3.4.6 Análise por redes de similaridade

As redes de similaridades foram utilizadas para estudar, em um contexto abrangente, a relação entre cada um dos membros identificados como pertencentes às famílias criadas. Para a construção das redes de similaridade primeiramente foi necessária uma medida da relação entre todas as sequências a serem analisadas. Neste trabalho foram utilizadas as estatísticas de alinhamentos com o algoritmo do BLASTP para se inferir a relação entre elas (WITTKOP et al., 2011).

O resultado do BLASTP, com um *cutoff* de 10–14 foi salvo no formato tabulado, opção 6 no pacote *blast* 2.2.28, para que fosse possível importar as informações para o programa Cytoscape (SHANNON et al., 2003), utilizando o plug-in *blast2similarity*⁵. Na representação em forma de rede de similaridade, cada nó do grafo representou uma proteína e cada vértice uma relação entre elas. O comprimento dos vértices indica o grau de semelhança entre as sequências, que é calculado com base nas estatísticas obtidas pelo BLASTP, seguindo a fórmula aplicada pelo algoritmo contido no plug-in. O plug-in permite também optar entre observar apenas o melhor hit de cada sequência (*Best bidirectional hit*) ou um cálculo considerando todos os hits (*Sum of all hits*) identificados no alinhamento das proteínas (todas contra todas) (WITTKOP et al., 2011).

Após importar os resultados do BLASTP e definir os parâmetros para o cálculo da relação entre as proteínas foi necessário optar por uma forma de agrupar os nós. A opção do *layout>yfiles>Organic* foi feita para permitir o desenho da rede em duas dimensões com o menor número de vértices cruzantes possível. O algoritmo utilizado tem código fechado, no entanto é aplicado o conceito de grafos *Force-directed* (GMBH, 2013). Após todos os passos, com o programa Cytoscape, pôde ser observada a família de cada uma das proteínas/nós analisadas e dessa forma evitar famílias redundantes ou auxiliar na identificação de proteínas muito discrepantes que precisassem ser revisadas.

⁵ Disponível em <http://apps.cytoscape.org/apps/blast2similaritygraph>

3.4.7 Validação dos modelos

Com o intuito de fornecer evidências da precisão dos modelos criados foi elaborada a seguinte estratégia de validação: um banco de sequências de aminoácidos randômico - com 37 (número de famílias) x 2000 sequências e variando em tamanho conforme os perfis-HMM - foi utilizado como alvo para os modelos produzidos por esse trabalho, a validação consistiu em observar quantas sequências randômicas seriam capturadas pelos modelos, sendo que um alto número seria indicativo de imprecisão do modelo. Considerando que as frequências dos aminoácidos observados nas sequências de proteínas não são completamente aleatórias foi seguido o seguinte protocolo:

- Foi obtido o tamanho médio das sequências de cada família.
- Foram utilizadas as frequências de aminoácidos observadas em todas as proteínas do banco de dados SwissProt⁶.
- Para cada família, foram geradas 2.000 sequências aleatórias com as frequências observadas no SwissProt e número de aminoácidos segundo uma distribuição normal, utilizando a média da respectiva família com desvio padrão de 50.
- Os perfis-HMM criados foram utilizados na busca de novos membros e assim avaliar a suas respectivas precisões.

Todas as sequências geradas pelo protocolo descrito acima foram concatenadas em um único arquivo para servir como banco de dados de proteínas aleatórias, com o qual todas as famílias foram testadas.

⁶ Disponível em: <http://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html>

4 RESULTADOS E DISCUSSÃO

Nesta seção foram discutidas as formas de se utilizar o sistema MENOOfams bem como uma descrição, do ponto de vista da importância biológica, das famílias geradas por este trabalho. Também foram apresentadas as análises por redes de similaridade realizadas com as proteínas identificadas e os resultados do método de validação aplicado.

4.1 Descrição das funcionalidades do sistema MENOOfams

Até o momento foram produzidas 36 famílias revisadas que, com a utilização dos perfis-HMM gerados por este trabalho, identificaram 1.233 proteínas que receberam termos do Gene Ontology relacionados a biotecnologia para produção de bioenergia microbiana. Um total de 26 termos foram propagados neste contexto, divididos entre as ontologias *Cellular Component*, *Molecular Function* e *Biological Process* (Tabela 1).

Tabela 1 – Descrição dos termos GO propagados pelo MENOOfams. Cada termo apresenta uma descrição e uma ontologia (Ontolg.), seguidos pelo número de proteínas anotadas pelo grupo MENO e pelo número total de proteínas capturadas pela família. F: *Molecular function*, C: *Cellular component*, P: *Biological process*

(continua)				
Identificador	Descrição	Ontolg.	MENO	Família*
GO:0004022	alcohol dehydrogenase (NAD) activity	F	1	280
GO:0004029	aldehyde dehydrogenase (NAD) activity	F	1	280
GO:0004556	alpha-amylase activity	F	1	32
GO:0008810	cellulase activity	F	6	117
GO:0008911	lactaldehyde dehydrogenase activity	F	1	11
GO:0016162	cellulose 1,4-beta-cellobiosidase activity	F	3	37
GO:0019867	outer membrane	C	3	63
GO:0030245	cellulose catabolic process	P	4	54
GO:0030248	cellulose binding	F	4	99
GO:0031176	endo-1,4-beta-xylanase activity	F	3	22
GO:0031216	neopullulanase activity	F	1	66
GO:0033905	xylan endo-1,3-beta-xylosidase activity	F	2	24
GO:0033946	xyloglucan-specific endo-beta-1,4-glucanase activity	F	1	11
GO:0043263	cellulosome	C	3	38
GO:0043794	formate dehydrogenase (coenzyme F420) activity	F	1	91
GO:0043852	monomethylamine methyltransferase activity	F	2	26
GO:0044583 ^a	Cellotriose binding	F	1	16
GO:0044586 ^a	Cellotetraose binding	F	1	16
GO:0044587 ^a	Cellopentaose binding	F	1	16
GO:0044683	methylthiol:coenzyme M methyltransferase activity	F	1	4
GO:0045493	xylan catabolic process	P	1	11
GO:0050545	sulfofpyruvate decarboxylase activity	F	1	10

Tabela 1 – Descrição dos termos GO propagados pelo MENGINEfams. Cada termo apresenta uma descrição e uma ontologia (Ontolg.), seguidos pelo número de proteínas anotadas pelo grupo MENGINE e pelo número total de proteínas capturadas pela família. F: *Molecular function*, C: *Cellular component*, P: *Biological process*

(conclusão)				
Identificador	Descrição	Ontolg.	MENGINE	Família*
GO:0051912	CoB--CoM heterodisulfide reductase activity	F	1	15
GO:0071271	1-butanol biosynthetic process	P	1	280
GO:2000899 ^a	xyloglucan catabolic process	P	1	11
GO:2001070 ^a	starch binding	F	2	31

* Uma mesma proteína pode estar anotada com diferentes termos.

^a Novos termos gerados pelo projeto MENGINE.

O sistema desenvolvido por este trabalho, o MENGINEfams, foi responsável por administrar a porção automática do processo de criação de famílias e por disponibilizar os resultados alcançados. O acesso foi oferecido via navegador de internet, o código foi escrito em Python e o banco de dados operado pelo sistema gerenciador de bancos de dados SQLite3. Todo o sistema ficou hospedado em um servidor sob responsabilidade do professor João Carlos Setubal, no Instituto de Química da USP.

As famílias⁷ foram listadas pelo sistema (ANEXO 5), sendo possível realizar buscas por palavras que componham o nome da família (opção padrão) ou indicando o campo no qual se deseja realizar a busca. Os campos que podem ser pesquisados corresponderam as colunas do banco SQL e foram: domínios Pfam; famílias PANTHER, CAZy ou HAMAP; autor responsável pela construção da família; palavras na descrição; identificação do Gene Ontology, membros capturados ou membros que compõem o modelo. A caixa de diálogo facilitou a consulta utilizando uma camada de filtro que simplifica a linguagem SQL.

Por exemplo, para se listar todas as famílias geradas que foram classificadas como família *Glycosyl hidrolase 10* pelo CAZy é possível realizar uma busca com o texto: CAZy'GH10'. Os operadores da linguagem SQL 'AND' e 'OR' são válidos nesta busca, porém limitados a combinações de apenas duas colunas, exemplo: CAZy'GH10' AND author'Taniguti'.

Toda a versão de uma família apresenta um escore específico que representa a nota de corte para definir os novos membros, descrita no item 3.3. Esse valor foi indicado em um gráfico de pontos que apresenta a distribuição dos escores atribuídos pelo perfil-HMM a cada

⁷ Disponível em: <http://mengofams.lbi.iq.usp.br/families>

uma das proteínas que pertence a uma mesma família e as demais proteínas⁸, abaixo da linha de corte, que não foram consideradas membros (Figura 7A). Essa visualização foi importante para o processo de curadoria, pois facilitou a identificação de proteínas que recebiam pontuação ligeiramente abaixo à nota de corte e assim permitir a revisão da família, tomando a decisão sobre inseri-la ou não como um novo membro a família. Para uma decisão como a descrita acima foi necessário confrontar os termos GO que deveriam ser propagados pela família e as assinaturas/sinais presentes na sequência identificada. Como a literatura científica com validações experimentais, apesar de intensa, não acompanha o avanço das tecnologias de sequenciamento, muitas vezes foi preciso avaliar recursos complementares como descrito no item 3.2, Caracterização das sequências primárias.

Além do gráfico de pontos representando escores, o número de espécies e os grupos taxonômicos observados em cada família foi representado na forma de um gráfico de setores (Figura 7B). Essa funcionalidade permitiu ter uma primeira visão sobre a abrangência da família em questão. De maneira geral as proteínas do metabolismo primário são conservadas entre as espécies enquanto que as integrantes de metabolismos secundários são mais específicas (KARLOVSKY, 2008). Dentre as famílias geradas que agregam um maior número de membros está a família mengofam074 que propaga os termos GO “*alcohol dehydrogenase (NAD) activity*”, “*aldehyde dehydrogenase (NAD) activity*” e “*1-butanol biosynthetic process*”. Em experimentos com *Clostridium acetobutylicum* foi demonstrado que essa função está envolvida no mecanismo de produção de butanol e atua como uma segunda álcool/aldeído desidrogenase, presente quando essa espécie é cultivada em meio alcoolgênico, isto é, crescida em pH neutro com grande disponibilidade de *nicotinamide adenine dinucleotide phosphate-oxidase*, também conhecido como NAD(P)H (FONTAINE et al., 2002).

Entre as famílias mais específicas está a mengofam-071, que propaga o termo GO “*methylthiol:coenzyme M methyltransferase activity*”. Essa família foi gerada com base na literatura disponível sobre a *Methanosarcina barke*, particularmente interessante pois representa o único gênero de arqueia conhecido pela literatura com a habilidade de utilizar precursores diferentes para a produção de metano, entre eles: acetato, dióxido de carbono, trimetilamina, dimetilamina, monometilamina, metanol, dimetilsulfito, metanotiol, metilmercaptopropinado, ion tetrametilamônio e piruvato (PAUL; KRZYCKI, 1996; HOOK;

⁸ No máximo 5.000 identificadores

WRIGHT; MCBRIDE, 2010). Em 2012, uma arqueia de intestino humano foi sequenciada, a *Candidatus Methanomethylophilus alvus* (BORREL et al., 2012) e a família em questão classificou uma proteína predita de seu genoma como membro, exibida no gráfico de setores da família mengofam-071. Essa observação pode ser importante ao corroborar com outros recursos de estudos *in silico* e vir a auxiliar no planejamento de experimentos para validação experimental.

Figura 7 – Gráficos da distribuição de escores e da distribuição taxonômica de uma família hipotética. A. A primeira coluna representa a pontuação das proteínas utilizadas como semente para a família e a segunda todas as proteínas do UniProtKB que obtiveram alguma pontuação. B. Os grupos taxonômicos, ou amostras ambientais, dos membros que compõem famílias são exibidos em um gráfico de setores

Como já citado anteriormente, uma característica desejável em um banco de dados é a referência a outros bancos que apresentem informações complementares às apresentadas. Neste caso, os bancos de dados: PANTHER, CAZy, Pfam e HAMAP, assim como as definições dos termos do Gene Ontology foram escolhidos para complementar as descrições de cada uma das famílias MENGOfams. O Pfam, por exemplo, é um banco de dados de domínios funcionais que pode identificar domínios em 70–90% de todas as proteínas preditas de alguns genomas sequenciados, sendo assim uma base de dados de grande inserção (SJÖLANDER, 2013). O MENGOfams gerou famílias considerando toda a arquitetura de domínios das sequências protéicas estudadas, aspecto importante uma vez que cerca de 60% das proteínas do genoma de procariotos e 80% das proteínas em genomas de eucariotos

possuem mais de um domínio funcional em suas sequências (HAN et al., 2007), o que situa o MENGOfams em um “nicho” de análises *in silico* complementar ao Pfam.

MENGOfam families
Total of families: **84**

You can query our families by using terms like:
pfam_domains, panther, CAZy, hamap, members, seeds, author, description, family_name, CAZy, GeneOntology.

Example1: pfam_domains 'dock' OR author 'Taniguti'
Example2: family_name 'Endoglucanase+T' AND author 'Taniguti'

*Note: On your specific queries use '+' instead of white space'' (like family_name 'Endoglucanase+T')

e.g pfam_domains 'dock' AND author 'Orpinelli'

* You can click on the column header to sort this table.

Identifier ▲	Name	Status	Author
mengofam-001-03	Endoglucanase (F, N, Z)	Reviewed	Monteiro-Vitorello
mengofam-002-01	Extracellular solute-binding protein family 1	Reviewed	Orpinelli
mengofam-003-01	Dockerin type 1	Preliminary	Orpinelli
mengofam-004-01	Endoglucanase D	Reviewed	Monteiro-Vitorello
mengofam-005-01	Sugar ABC transporter	Preliminary	Orpinelli

Figura 8 – Exibição dos registros de famílias. Uma caixa de diálogo para realizar buscas nos registros do banco de dados. As versões preliminares e revisadas são indicadas em Status. Na tabela foi utilizado um plug-in de jQuery, o tSort, para permitir reorganização da ordem das linhas conforme a coluna de interesse

O banco de dados com o qual o sistema MENGOfams opera foi desenhado de forma a permitir a atualização das famílias criadas e, para evitar problemas de reprodutibilidade ou de perda de informações, todas as versões anteriores foram mantidas e continuam disponibilizadas, utilizando o padrão “mengofam-família-versão”, como descrito no item 3.3. Além disso, todos os procedimentos realizados para a criação de famílias foram registrados na forma de um relatório. Neste relatório informações detalhadas sobre os processos disparados para criação de cada família puderam ser acompanhados em tempo real, bem como o registro de famílias que foram excluídas devido a algum erro no processo, como por exemplo identificadores duplicados nas sequências inseridas. Ao atualizar as famílias os formulários de criação foram pré-preenchidos com informações recuperadas das versões anteriores, que puderam ser editadas.

4.1.1 Propagação dos termos GOs em genomas completamente sequenciados

Entre os objetivos da propagação de termos GO associados ao projeto MANGO está a anotação de genomas completamente sequenciados de interesse⁹. As espécies bacterianas que apresentam o maior número de estudos experimentais no que se refere aos processos de degradação de tecidos lignocelulósicos são do gênero *Clostridium* (PRAWITWONG et al.,

⁹ Lista de produtos gênicos anotados: <http://www.mengo.biochem.vt.edu/terms/GeneAnnotations.php>

2013; SCHELLENBERG et al., 2014; XIN; WU; HE, 2014). Algumas espécies possuem a habilidade de degradar tanto pentoses como hexoses (HALLENBECK, 2014), recurso esse que não se encontra na metodologia com linhagens de *Saccharomyces cerevisiae* nos processos de fermentação alcoólica (NANDA; DALAI; KOZINSKI, 2014). As bactérias do gênero *Clostridium* são conhecidas pelo seu alto potencial biotecnológico, a exceção das consideradas patogênicas como *Clostridium botulinicum* (CENTERS FOR DISEASE CONTROL AND PREVENTION, 2014). Existem hoje 34 genomas desse gênero cujas sequências estão completas e sem gaps, outros 48 genomas representados em *scaffolds* e 89 genomas em *contigs* (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2014).

Neste trabalho um grupo específico de proteínas que compõem um complexo enzimático altamente eficiente na degradação de celulose foi anotado com termos do Gene Ontology. Esse complexo é encontrado em alguns representantes do gênero *Clostridium*, bem como em *Acetovibrio*, *Bacteroides*, *Ruminococcus* e *Vibrio* (BAYER et al., 1998) e é conhecido como celulosomo.

O complexo celulosomo é composto por uma glicoproteína estrutural conhecida como escafoldina (do inglês, *scaffoldin*), a qual possui domínios do tipo coesina (do inglês, *cohesin*) e domínio de ligação a carboidratos (CBM, do inglês *Carbohydrate Binding Module*). As regiões presentes nas unidades catalíticas que compõem o celulosomo possuem domínios conhecidos como dockerina (do inglês, *dockerin*), que são responsáveis por conectar as unidades enzimaticamente ativas por meio da interação com os domínios coesina. Até recentemente acreditava-se que a montagem deste complexo era realizado de maneira aleatória, porém Borne et al. (2013) mostraram que as proteínas primeiramente ligadas a escafoldina ditam a afinidade das próximas subunidades a serem montadas, sendo assim um processo altamente específico e organizado. De forma semelhante a interação dockerina/coesina, o celulosomo é conectado a superfície da célula bacteriana por proteínas da membrana.

Foram analisadas algumas famílias protéicas identificadas como unidades em potencial de celulosomos, dentre elas às enzimas conhecidas como endoglucanases, para as quais foi atribuído o termo GO:0008810 (Tabela 1). Este termo representa a ontologia função molecular (*molecular function*) *cellulase activity*, cuja descrição é a catálise da endohidrólise de ligações (1->4)-beta-D-glicanos de polissacarídeos como celulose; armazenados de amido de líquen (polissacarídeo com estrutura semelhante a celulose encontrado como substância de

armazenamento de liquens); e fibras da parede celular de cereais, conhecidas como B-glucanos ((1->3, 1->4)-B-d-glucan)¹⁰.

Essas proteínas foram classificadas pelo CAZy como pertencentes a 7 grupos funcionais (glicosil hidrolases: GH5, GH8, GH9, GH10, GH26, GH44, GH74) e com a presença de 8 módulos de ligação a carboidratos (CBM3, CBM4, CBM6, CBM11, CBM13, CBM22, CBM30 e CBM44). As enzimas pertencentes ao grupo GH9 são mais abrangentes em relação a sua presença entre as espécies de bactérias de vários grupos taxonômicos, mesmo aquelas que não formam celulosomo. No Anexo B, Termos propagados e predições GO, são listadas todas as famílias estudadas, com os respectivos domínios CAZy, termos GO, número de membros e presença ou ausência de peptídeo sinal predito.

Uma descrição detalhada da atividade enzimática e subdivisões estabelecidas para esse grupo de enzimas pode ser encontrada no banco de dados CAZy¹¹. Já as endoglucanases da família GH5 são menos conservadas¹² e a proteína Q05332, utilizada como referência para a geração da família mengofam-039, é parte do celulosomo em *Clostridium thermocellum* (LEMAIRE; BÉGUIN, 1993). Assim, para a família mengofam-39 foi atribuído o termo GO:0043263, que representa a ontologia componente celular (*cellular component*) cuja descrição é um complexo multi-enzimático com função de hidrólise de celulose contendo até 11 enzimas diferentes agrupadas sobre a glicoproteína estrutural escafoldina¹³, além do termo GO:0008810 descrito anteriormente.

A presença de termos GO que definem claramente o que se conhece funcionalmente das proteínas referência foi determinante para a construção de famílias que, apesar de apresentarem a mesma ontologia em função molecular, podem não compartilhar a mesma ontologia em componente celular. Assim, na construção das famílias foi considerada a arquitetura dos domínios das proteínas referências, uma vez que é necessária precisão para a propagação dos termos GO para novos membros.

4.1.2 Enriquecimento da cobertura de anotações Gene Ontology nos registros

UniProt

O UniProt é um banco de dados muito utilizado por pesquisadores que buscam informações sobre sequências biológicas, principalmente devido à anotações curadas de alta

¹⁰ <http://amigo.geneontology.org/amigo/term/GO:0008810>

¹¹ http://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_9

¹² http://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_5

¹³ <http://amigo.geneontology.org/amigo/term/GO:0043263>

qualidade. No entanto um número considerável de proteínas são automaticamente anotadas. Esse banco é atualizado a cada quatro semanas e, entre as informações que são oferecidas, encontram-se anotações do Gene Ontology, informando o respectivo projeto que atribuiu a anotação ao registro UniProt.

Com o intuito de avaliar a contribuição das famílias MENGOfam geradas por este trabalho na propagação de termos GO, foram realizadas buscas nos registros UniProtKB, procurando identificar quais das proteínas membros de famílias MENGOfam já possuíam o termo GO proposto para propagação, mesmo que atribuído de maneira automática (por exemplo, via InterPro). A Tabela 2 mostra o resultado dessa busca, bem como o projeto e o nível de evidência encontrado para definir a anotação.

Tabela 2 – Proteínas no UniProtKB anotadas com os termos GO analisados. Foram buscados nos registros do UniProtKB os termos anotados pelo MENGO às proteínas identificadas neste projeto como candidatos a anotação pelos perfis-HMM

Projeto responsável	Evidência	Proteínas MENGOfam (%)
*Sem termo	-	79,10
CACAO	Experimento direto	0,05
MENGO	Experimento direto	1,97
UniProtKB	Experimento direto	0,20
InterPro	Experimento direto	13,41
UniProtKB-EC	Automática	4,76
UniProtKB-KW	Automática	0,30
UniProtKB-UniPathway	Automática	0,05
MENGO	Estudo com mutante	0,15

Dentre os projetos anotadores que aparecem na busca realizada estão: InterPro (HUNTER et al., 2011), que possui diversos bancos de dados colaboradores que fornecem modelos para prever automaticamente famílias, domínios e sítios importantes nas sequências; anotação manual do SwissProt, associando evidências na literatura científica ao registro de proteína; anotações automáticas do UniProt, propagando informações para a porção trEMBL do banco de dados; MENGO, projeto colaborador já descrito anteriormente; e CACAO (BERARDINI et al., 2012), uma competição entre grupos formados por alunos de graduação de universidades por todo o mundo, que são treinados a ler criticamente artigos científicos e realizar anotações corretamente, amenizando assim o *deficit* de biocuradores.

Como resultado dessa análise, foi observado que 79,10% das proteínas contidas nas famílias deste trabalho não possuem os respectivos termos do Gene Ontology anotados nos registros da versão 2014_08 do UniProt. Uma forma de enriquecer os dados contidos no

UniProt seria tornar este trabalho, o MENGOFams, colaborador de algum outro banco já bem estabelecido, fornecendo a eles os modelos e informações estudadas.

4.2 Análise por redes de similaridade

Pela análise por redes de similaridade, metodologia descrita no item 3.4.6, foi possível observar que as proteínas das famílias classificadas pelos perfis-HMM possuem um padrão diferente do atribuído por relações de semelhança entre sequências, obtidas a partir do resultado de alinhamentos por BLASTP. Na Figura 9, cada nó (na forma de círculos) representa uma proteína e os vértices representam uma relação de semelhança entre elas.

Figura 9 – Rede de similaridade. Agrupamentos estabelecidos utilizando estatísticas do resultado do BLASTP e plug-in do Cytoscape. Os números correspondem aos identificadores das famílias MENGOFams

Pela análise da rede é possível notar uma discordância entre o grau de similaridade (E-valor 10^{-14}) entre as sequências e as classificações dos modelos. Essa diferença foi observada por consequência das particularidades no sistema de pontuação empregados por cada metodologia. O algoritmo do BLASTP adota uma matriz de substituição, como as PAM ou BLOSUM, e um sistema de abertura de lacunas para estimar o melhor alinhamento local entre duas sequências de aminoácidos e assim atribuir uma pontuação (ALTSCHUL et al., 1990).

Os perfis-HMM são modelos estatísticos gerados a partir do alinhamento de sequências representantes de uma mesma família, o que permitem pontuar sequências de acordo com a posição de ocorrência de cada aminoácido (EDDY, 1998) – mais detalhes no item 2.7, Homólogos distantes e métodos computacionais.

A Figura 9 representa as relações de semelhança entre as proteínas que compõem algumas das famílias MENGOfams. Essa ilustração deixa claro que os termos GO dificilmente poderiam ser propagados corretamente utilizando apenas medidas de similaridade como as obtidas por BLASTP. O agrupamento indicado na Figura 9 pela seta “A” representa proteínas da família das escafoldinas, em rosa, e proteínas da família das Endoglucanases Y, em roxo. Apesar da análise por alinhamentos locais sugerir que essas proteínas possuem funções semelhantes, as conclusões obtidas pelos estudos realizados com perfis-HMM indicam importantes diferenças. As escafoldinas pertencem a família MENGOfams nomeada “*Cellulosomal-scaffolding protein A*” e possuem apenas função estrutural (termo GO propagado: *cellulose binding*), atuando como um esqueleto para o acoplamento de unidades celussomais catalíticas (BORNE et al., 2013). As endoglucanases Y, no entanto, são classificadas pelo CAZy como pertencentes a família GH48 e existem evidências experimentais na literatura científica que descrevem a proteína semente geradora da família, a Q5TIQ4, como sendo uma enzima não celulosomal, parte de um segundo sistema celulolítico presente em *Clostridium thermocellum* (BERGER et al., 2007) em que foi propagado o termo GO *cellulose 1,4-beta-cellobiosidase activity e cellulose catabolic process*.

4.3 Validação dos modelos

Os perfis-HMM gerados por este trabalho foram testados contra um banco de 72.000 sequências¹⁴ geradas por um modelo de hipótese nula via o programa HMMSearch, do pacote HMMER. Foi verificado que não houveram sequências desse banco classificadas como membros de famílias MENGOfams, fornecendo assim um indício de que os modelos são seletivos.

A hipótese nula refere-se a não-homologia entre as sequências comparadas. Para este trabalho foi considerado que um banco de sequências não relacionadas poderia ser produzido utilizando as frequências com que os aminoácidos são observados nas sequências protéicas do banco de dados SwissProt. Esse é o mesmo postulado empregado pelos desenvolvedores do

¹⁴ Número de famílias x 2.000

pacote HMMER para o cálculo do *Log-odd* score, obtido ao utilizar um perfil-HMM para encontrar sequências homólogas (EDDY, 2010).

Esse cuidado deve ser tomado pois caso não se tenha rigor ao selecionar apenas sequências relacionadas para a geração do perfil-HMM as probabilidades de emissão do modelo serão próximas às probabilidades da hipótese nula, ou seja, de que não existe aminoácido mais ou menos frequente em cada uma das posições. Ao se confrontar um modelo erroneamente construído utilizando sequências não homólogas com um banco de dados de sequências geradas a partir de um modelo nulo é esperado que sejam identificados membros. A Tabela 3 mostra que algumas sequências aleatórias receberam pontuação, porém muito abaixo do cutoff necessário para serem classificadas como membros.

Tabela 3 – Sequências geradas a partir de uma hipótese nula identificadas com baixa pontuação. Após concatenar as sequências geradas com tamanho médio no comprimento de cada perfil-HMM, os modelos de famílias foram utilizados para buscar por membros nesse banco de dados isento de homólogos. É possível notar que poucas sequências receberam pontuação por alguns perfis-HMM (famílias), porém a pontuação que lhes foram atribuídas (Escore) é muito inferior à que foi estabelecida como nota de corte (*cutoff*)

Família	Hit	Escore	Cutoff
mengofam-012-06	mengofam-004-01-165 Length-622	10,1	1371,5
mengofam-007-05	mengofam-011-07-357 Length-590	12,6	646,2
mengofam-046-04	mengofam-014-04-1608 Length-805	11,6	1285,1
mengofam-044-09	mengofam-015-04-820 Length-772	14,2	858,9
mengofam-014-04	mengofam-032-05-487 Length-379	12,3	400,6
mengofam-056-02	mengofam-036-02-1894 Length-394	11	1349,2
mengofam-032-05	mengofam-042-03-1790 Length-842	11,5	400,6
mengofam-045-03	mengofam-044-09-1789 Length-463	10,6	982,9
mengofam-043-02	mengofam-047-03-1021 Length-868	11,2	1317,1
mengofam-079-04	mengofam-050-04-289 Length-633	10,6	1644,4
mengofam-043-02	mengofam-050-04-677 Length-691	11,2	1317,1
mengofam-047-03	mengofam-059-02-92 Length-244	10,2	1696,7
mengofam-044-09	mengofam-059-02-1154 Length-228	15,6	858,9
mengofam-046-04	mengofam-072-03-665 Length-1286	11,1	1285,1
mengofam-059-02	mengofam-072-03-1272 Length-1311	12,1	423,4

5 CONSIDERAÇÕES FINAIS

As conclusões que podem ser extraídas diante dos resultados deste trabalho são:

- A organização de um banco de dados com o objetivo de armazenar e disponibilizar dados biológicos sobre famílias protéicas requer planejamento e conhecimentos de outras áreas. Existem diversas opções para cumprir esse objetivo, cada uma com suas vantagens e desvantagens.
- A curadoria manual para a propagação de termos GO é uma atividade extremamente importante para o enriquecimento das informações acumuladas nos bancos de dados, porém sem outros processos complementares o alcance dessa curadoria é limitado. Assim, a construção de famílias utilizando perfis-HMM para a propagação de termos é uma estratégia conveniente.
- Métodos automáticos de anotação de proteínas baseados apenas em similaridade podem ser ineficientes em alguns casos, como no caso deste trabalho em que o objetivo foi a propagação dos termos GO específicos. Bancos de dados de motivos conservados, domínios funcionais ou de famílias de proteínas são complementares uns aos outros e podem promover o enriquecimento de informações disponíveis para grande parte das sequências, assim como o MENGO_{fams}.
- Até o momento, este trabalho elaborou 36 famílias curadas que oferecem um meio para propagar os termos do Gene Ontology anotados pelo projeto MENGO. Essas famílias englobam 1.233 proteínas do UniProtKB, das quais 79% puderam receber os termos GO contidos nas famílias MENGO_{fams}. A propagação dos termos GO no UniProtKB utilizando MENGO_{fams} pode enriquecer a propagação automática no TrEMBL deste banco de dados.

REFERÊNCIAS

- ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, London, v. 215, n. 3, p. 403 – 410, 1990.
- APWEILER, R.; BAIROCH, A.; WU, C. H.; BARKER, W. C.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M.; MARTIN, M. J.; NATALE, D. A.; O'DONOVAN, C.; REDASCHI, N.; YEH, L.-S. L. Uniprot: the universal protein knowledgebase. **Nucleic Acids Research**, Oxford, v. 32, n. 1, p. 115 – 119, 2004.
- ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature genetics**, New York, v. 25, n. 1, p. 25 – 29, 2000.
- ATKINSON, H. J.; MORRIS, J. H.; FERRIN, T. E.; BABBITT, P. C. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. **PLoS ONE**, San Francisco, v. 4, n. 2, 2009. Disponível em: <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004345>>. Acesso em: 15 ago. 2014.
- ATSUMI, S.; CANN, A. F.; CONNOR, M. R.; SHEN, C. R.; SMITH, K. M.; BRYNILDSEN, M. P.; CHOU, K. J.; HANAI, T.; LIAO, J. C. Metabolic engineering of *Escherichia coli* for 1-butanol production. **Metabolic Engineering**, Waltham, v. 10, n. 6, p. 305 – 311, 2008.
- ATSUMI, S.; WU, T. Y.; ECKL, E. M.; HAWKINS, S. D.; BUELTER, T.; LIAO, J. C. Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. **Applied Microbiology Biotechnology**, Berlin, v. 85, n. 3, p. 651 – 657, 2010.
- BAKKE, P.; CARNEY, N.; DELOACHE, W.; GEARING, M.; INGVORSEN, K.; LOTZ, M.; MCNAIR, J.; PENUMETCHA, P.; SIMPSON, S.; VOSS, L.; WIN, M.; HEYER, L. J.; CAMPBELL, A. M. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. **PLoS ONE**, San Francisco, v. 4, n. 7, 2009. Disponível em: <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0006291>>. Acesso em: 9 jan. 2014.
- BARBER, A. E.; BABBITT, P. C. Pythoscape: a framework for generation of large protein similarity networks. **Bioinformatics**, Oxford, v. 28, n. 21, p. 2845 – 2846, 2012.
- BAYER, E. A.; CHANZY, H.; LAMED, R.; SHOHAM, Y. Cellulose, cellulases and cellulosomes. **Current Opinion in Structural Biology**, London, v. 8, n. 5, p. 548 – 557, 1998.

- BERARDINI, T. Z.; LI, D.; MULLER, R.; CHETTY, R.; PLOETZ, L.; SINGH, S.; WENSEL, A.; HUALA, E. Assessment of community-submitted ontology annotations from a novel database-journal partnership. **Database**, Oxford, v. 2012, n. 1, p. 1 – 13, 2012.
- BERGER, E.; ZHANG, D.; ZVERLOV, V. V.; SCHWARZ, W. H. Two noncellulosomal cellulases of *Clostridium thermocellum*, cel9i and cel48y, hydrolyse crystalline cellulose synergistically. **FEMS Microbiology Letters**, Amsterdam, v. 268, n. 2, p. 194 – 201, 2007.
- BESSANT, C.; OAKLEY, D.; SHADFORTH, I. **Building Bioinformatics Solutions**. London: OUP Oxford, 2014. p. 87 - 102.
- BORNE, R.; BAYER, E. A.; PAGES, S.; PERRET, S.; FIEROBE, H.-P. Unraveling enzyme discrimination during cellulosome assembly independent of cohesin/dockerin affinity. **FEBS Journal**, Chichester, v. 280, n. 22, p. 5764 – 5779, 2013.
- BORREL, G.; HARRIS, H. M. B.; TOTTEY, W.; MIHAJLOVSKI, A.; PARISOT, N.; PEYRETAILLADE, E.; PEYRET, P.; GRIBALDO, S.; TOOLE, P. W. O.; BRUGERE, J.-F. Genome sequence of *Candidatus Methanomethylophilus alvus* mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens. **Journal of Bacteriology**, Washington, v. 194, n. 24, p. 6944 – 6945, 2012.
- BUCKERIDGE, M. S.; SOUZA, A. P. D.; ARUNDALE, R. A.; ANDERSON-TEIXEIRA, K. J.; DELUCIA, E. Ethanol from sugarcane in Brazil: a "midway" strategy for increasing ethanol production while maximizing environmental benefits. **GCB Bioenergy**, Oxford, v. 4, n. 2, p. 119 – 126, 2012.
- CANTAREL, B. L.; COUTINHO, P. M.; RANCUREL, C.; BERNARD, T.; LOMBARD, V.; HENRISSAT, B. The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. **Nucleic Acids Research**, Oxford, v. 37, p. 233–238, 2009.
- CENTERS FOR DISEASE CONTROL AND PREVENTION. **Pathogens and Protocols**. Atlanta, 2014. Disponível em: <<http://www.cdc.gov/pulsenet/pathogens/clostridium.html>>. Acessado em: 17 set. 2014.
- CONNOR, M. R.; LIAO, J. C. Engineering of an *Escherichia coli* strain for the production of 3-methyl-1-butanol. **Applied and Environmental Microbiology**, Washington, v. 74, n. 18, p. 5769 – 5775, 2008.
- CONSORTIUM, T. G. O. The Gene Ontology in 2010: extensions and refinements. **Nucleic Acids Research**, Oxford, v. 38, n. suppl 1, p. 331 – 335, 2010.
- CONSORTIUM, T. U. Update on activities at the universal protein resource (uniprot) in 2013. **Nucleic Acids Research**, Oxford, v. 41, n. D1, p. D43 – D47, 2013.
- CORPET, F.; GOUZY, J.; KAHN, D. The ProDom database of protein domain families. **Nucleic Acids Research**, Oxford, v. 26, n. 1, p. 323 – 326, 1998.
- DATTA, R. S.; MEACHAM, C.; SAMAD, B.; NEYER, C.; SJÖLANDER, K. Berkeley phog: Phylofacts orthology group prediction web server. **Nucleic Acids Research**, Oxford, v. 37, n. 1, p. 84 – 89, 2009.
- DURBIN, R. Probabilistic approaches to phylogeny. In: _____. **Biological Sequence Analysis**. Cambridge: Cambridge University Press, 1998. chap. 8, p. 192 – 231.

EDDY, S. R. Hidden markov models. **Current Opinion in Structural Biology**, London, v. 6, n. 1, p. 361 – 365, 1996.

_____. Profile hidden markov models. **Bioinformatics**, Oxford, v. 14, n. 9, p. 755 – 763, 1998.

_____. **HMMER Users Guide**. Washington: Howard Hughes Medical Institute, 2010. p. 93 - 103.

_____. Accelerated profile hmm searches. **PLoS Comput Biol**, San Francisco, v. 7, n. 10, 2011. Disponível em: <<http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002195>>. Acesso em: 17 set. 2014.

ERIKSEN, D.; HSIEH, P. C.; LYNN, P.; ZHAO, H. Directed evolution of a cellobiose utilization pathway in *Saccharomyces cerevisiae* by simultaneously engineering multiple proteins. **Microbial Cell Factories**, London, v. 12, n. 1, p. 1 – 61, 2013.

FONTAINE, L.; MEYNIAL-SALLES, I.; GIRBAL, L.; YANG, X.; CROUX, C.; SOUCAILLE, P. Molecular characterization and transcriptional analysis of adhe2, the gene encoding the nadh-dependent aldehyde/alcohol dehydrogenase responsible for butanol production in alcohologenic cultures of *Clostridium acetobutylicum* atcc 824. **Journal of Bacteriology**, Washington, v. 184, n. 3, p. 821 – 830, 2002.

FROCK, A. D.; KELLY, R. M. Extreme thermophiles: moving beyond single-enzyme biocatalysis. **Current Opinion in Chemical Engineering**, Amsterdam, v. 1, n. 4, p. 363 – 372, 2012.

GMBH yWorks. **Automatic Graph Layout**. Vor dem Kreuzberg, 2013. 1 - 13 p. Disponível em <<http://docs.yworks.com/yfiles/doc/developers-guide/>>. Acesso em: 17 set. 2014.

GRIBSKOV, M.; MCLACHLAN, A. D.; EISENBERG, D. Profile analysis: detection of distantly related proteins. **Proceedings of the National Academy of Sciences**, Washington, v. 84, n. 13, p. 4355 – 4358, 1987.

GRIBSKOV, M.; VERETNIK, S. Identification of sequence patterns with profile analysis. In: _____. **Computer Methods for Macromolecular Sequence Analysis**. San Diego: Academic Press, 1996. v. 266, p. 198 – 212.

HAFT, D. H.; SELENGUT, J. D.; WHITE, O. The TIGRFAMs database of protein families. **Nucleic Acids Research**, Oxford, v. 31, n. 1, p. 371 – 373, 2003.

HALLENBECK, P. Bioenergy from microorganisms: An overview. In: _____. **Microbial BioEnergy: Hydrogen Production**. Amsterdam: Springer Netherlands, 2014. v. 38, chap. 1, p. 1 – 6.

HAN, J.-H. H.; BATEY, S.; NICKSON, A. A.; TEICHMANN, S. A.; CLARKE, J. The folding and evolution of multidomain proteins. **Nature reviews**, London, v. 8, n. 4, p. 319 – 330, 2007.

HARDING, T.; JUNGBLUT, A. D.; LOVEJOY, C.; VINCENT, W. F. Microbes in high arctic snow and implications for the cold biosphere. **Applied and Environmental Microbiology**, Washington, v. 77, n. 10, p. 3234 – 3243, 2011.

HENIKOFF, S.; HENIKOFF, J. G. Position-based sequence weights. **Journal of Molecular Biology**, London, v. 243, n. 4, p. 574 – 578, 1994.

HENRISSAT, B.; BAIROCH, A. Updating the sequence-based classification of glycosyl hydrolases. **The Biochemical Journal**, London, v. 316, n. 2, p. 695 – 696, 1996.

HOOK, S. E.; WRIGHT, A.-D. G.; MCBRIDE, B. W. Methanogens: Methane producers of the rumen and mitigation strategies. **Archaea**, New York, v. 2010, n. 4, p. 11 – 14, 2010.

HUNTER, S.; JONES, P.; MITCHELL, A.; APWEILER, R.; ATTWOOD, T. K.; BATEMAN, A.; BERNARD, T.; BINNS, D.; BORK, P.; BURGE, S.; CASTRO, E. de; COGGILL, P.; CORBETT, M.; DAS, U.; DAUGHERTY, L.; DUQUENNE, L.; FINN, R. D.; FRASER, M.; GOUGH, J.; HAFT, D.; HULO, N.; KAHN, D.; KELLY, E.; LETUNIC, I.; LONSDALE, D.; LOPEZ, R.; MADERA, M.; MASLEN, J.; MCANULLA, C.; MCDOWALL, J.; MCMENAMIN, C.; MI, H.; MUTOWO-MUELLENET, P.; MULDER, N.; NATALE, D.; ORENGO, C.; PESSEAT, S.; PUNTA, M.; QUINN, A. F.; RIVOIRE, C.; SANGRADOR-VEGAS, A.; SELENGUT, J. D.; SIGRIST, C. J. A.; SCHEREMETJEV, M.; TATE, J.; THIMMAJANARTHANAN, M.; THOMAS, P. D.; WU, C. H.; YEATS, C.; YONG, S.-Y. InterPro in 2011: new developments in the family and domain prediction database. **Nucleic Acids Research**, Oxford, v. 40, p. 1 – 7, 2011.

JENSEN, R. Orthologs and paralogs - we need to get it right. **Genome Biology**, London, v. 2, n. 8, p. 1 – 3, 2001.

KARLOVSKY, P. **Secondary Metabolites in Soil Ecology**. Amsterdam: Springer, 2008. p. 2 – 4.

KAWRYKOW, A.; ROUMANIS, G.; KAM, A.; KWAK, D.; LEUNG, C.; WU, C.; ZAROOUR, E.; SARMENTA, L.; BLANCHETTE, M.; WALDISPÖHL, J.; PLAYERS, P. Phylo: A citizen science approach for improving multiple sequence alignment. **PLoS ONE**, San Francisco, v. 7, n. 3, 2012. Disponível em: <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0031362>>. Acesso em: 19 dez. 2013.

KOONIN, E.; GALPERIN, M. Genome annotation and analysis. In:____. **Computational Approaches in Comparative Genomics**. Norwell: Springer, 2003. chap. 5. p. 130-138.

KROGH, A. An introduction to hidden markov models for biological sequences. In:____. **Computational methods in molecular biology**. Amsterdam: Elsevier, 1998. chap. 4, p. 45 – 63.

KROGH, A.; LARSSON, B.; HEIJNE, G. von; SONNHAMMER, E. L. L. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. **Journal of Molecular Biology**, London, v. 305, n. 2, p. 567 – 580, 2001.

LEMAIRE, M.; BÉGUIN, P. Nucleotide sequence of the celg gene of *Clostridium thermocellum* and characterization of its product, endoglucanase celg. **Journal of Bacteriology**, Washington, v. 175, n. 11, p. 3353 – 3360, 1993.

LENNARTSSON, P. R.; ERLANDSSON, P.; TAHERZADEH, M. J. Integration of the first and second generation bioethanol processes and the importance of by-products. **Bioresource Technology**, Amsterdam, v. 165, n. 6, p. 3 – 8, 2014.

LEONELLI, S.; DIEHL, A.; CHRISTIE, K.; HARRIS, M.; LOMAX, J. How the gene ontology evolves. **BMC Bioinformatics**, London, v. 12, n. 1, p. 325, 2011.

LICHT, F. Ethanol Industry Outlook 2008-2013 reports. Washington: **Renewable Fuels Association**, 2013. Disponível em: <http://ethanolrfa.3cdn.net/dc207800043a5aa5aa_y5im6rokb.pdf>. Acessado em: 15 mar. 2014.

LIMA, T.; AUCHINCLOSS, A. H.; COUDERT, E.; KELLER, G.; MICHOD, K.; RIVOIRE, C.; BULLIARD, V.; CASTRO, E. de; LACHAIZE, C.; BARATIN, D.; PHAN, I.; BOUGUELERET, L.; BAIROCH, A. Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. **Nucleic Acids Research**, Oxford, v. 37, n. 1, p. 471 – 478, 2009.

LOGUERCIO, S.; GOOD, B. M.; SU, A. I. Dizzez: An online game for human gene-disease annotation. **PLoS ONE**, San Francisco, v. 8, n. 8, 2013. Disponível em: <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0071171>>. Acesso em: 15 ago. 2013.

LUPAS, A.; DYKE, M. V.; STOCK, J. Predicting coiled coils from protein sequences. **Science**, New Jersey, v. 252, n. 5010, p. 1162 – 1164, 1991.

MACEDO, I.; LEAL, M.; SILVA, J. da. **Balanço das emissões de gases do efeito estufa na produção e no uso do etanol no Brasil**. São Paulo: Copersucar, 2004. 32 p. Relatório com dados do Centro de Tecnologia Copersucar.

MACRELLI, S.; MOGENSEN, J.; ZACCHI, G. Techno-economic evaluation of 2 nd generation bioethanol production from sugar cane bagasse and leaves integrated with the sugar-based ethanol process. **Biotechnology for Biofuels**, London, v. 5, n. 1, p. 22 – 25, 2012.

MAJOROS, W. H.; PERTEA, M.; SALZBERG, S. L. Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. **Bioinformatics**, Oxford, v. 20, n. 16, p. 2878 – 2879, 2004.

MARCELLIN, E.; LICONA-CASSANI, C.; MERCER, T.; PALFREYMAN, R.; NIELSEN, L. Re-annotation of the *Saccharopolyspora erythraea* genome using a systems biology approach. **BMC Genomics**, London, v. 14, n. 1, p. 699 – 702, 2013.

MARCHLER-BAUER, A.; ANDERSON, J. B.; CHERUKURI, P. F.; DEWEESE-SCOTT, C.; GEER, L. Y.; GWADZ, M.; HE, S.; HURWITZ, D. I.; JACKSON, J. D.; KE, Z.; LANCZYCKI, C. J.; LIEBERT, C. A.; LIU, C.; LU, F.; MARCHLER, G. H.; MULLOKANDOV, M.; SHOEMAKER, B. A.; SIMONYAN, V.; SONG, J. S.; THIESSEN, P. A.; YAMASHITA, R. A.; YIN, J. J.; ZHANG, D.; BRYANT, S. H. Cdd: a conserved domain database for protein classification. **Nucleic Acids Research**, Oxford, v. 33, n. 1, p. D192 – D196, 2005.

MI, H.; LAZAREVA-ULITSKY, B.; LOO, R.; KEJARIWAL, A.; VANDERGRIF, J.; RABKIN, S.; GUO, N.; MURUGANUJAN, A.; DOREMIEUX, O.; CAMPBELL, M. J.; KITANO, H.; THOMAS, P. D. The PANTHER database of protein families, subfamilies, functions and pathways. **Nucleic Acids Research**, Oxford, v. 33, n. 1, p. 284 – 288, 2005.

NAKAMURA, Y.; COCHRANE, G.; KARSCH-MIZRACHI, I. The international nucleotide sequence database collaboration. **Nucleic Acids Research**, Oxford, v. 41, n. 3, p. 21 – 24, 2013.

NANDA, S.; DALAI, A. K.; KOZINSKI, J. A. Butanol and ethanol production from lignocellulosic feedstock: biomass pretreatment and bioconversion. **Energy Science and Engineering**, West Sussex, v. 2, n. 3, p. 138 – 148, 2014.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Assembly database**. Bethesda, 2014. Disponível em: <<http://www.ncbi.nlm.nih.gov/assembly>>. Acessado em: 16 set. 2014.

NOTREDAME, C. Recent progress in multiple sequence alignment: a survey. **Pharmacogenomics**, Marseille, v. 3, n. 1, p. 131 – 144, 2002.

PAIS, F. S.-M.; RUY, P.; OLIVEIRA, G.; COIMBRA, R. Assessing the efficiency of multiple sequence alignment programs. **Algorithms for Molecular Biology**, London, v. 9, n. 1, p. 1 – 4, 2014.

PAUL, L.; KRZYCKI, J. A. Sequence and transcript analysis of a novel *Methanosarcina barkeri* methyltransferase ii homolog and its associated corrinoid protein homologous to methionine synthase. **Journal of Bacteriology**, Washington, v. 178, n. 22, p. 6599 – 607, 1996.

PEDRUZZI, I.; RIVOIRE, C.; AUCHINCLOSS, A. H.; COUDERT, E.; KELLER, G.; CASTRO, E. de; BARATIN, D.; CUCHE, B. A.; BOUGUELERET, L.; POUX, S.; REDASCHI, N.; XENARIOS, I.; BRIDGE, A.; CONSORTIUM the U. Hamap in 2013 , new developments in the protein family classification and annotation system. **Nucleic Acids Research**, Oxford, v. 41, n. 1, p. 584 – 589, 2013.

PETERSEN, T. N.; BRUNAK, S.; HEIJNE, G. von; NIELSEN, H. Signalp 4.0: discriminating signal peptides from transmembrane regions. **Nature Methods**, New York, v. 8, n. 10, p. 785 – 786, 2011.

PETSKO, G. Homologuephobia. **Genome Biology**, London, v. 2, n. 2, p. 1 – 2, 2001.

POUX, S.; MAGRANE, M.; ARIGHI, C. N.; BRIDGE, A.; DONOVAN, C. O.; LAIHO, K.; CONSORTIUM, T. U. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. **Nucleic Acids Research**, Oxford, v. 2014, p. 1 – 9, 2014.

PRAWITWONG, P.; WAEONUKUL, R.; TACHAAPA KOON, C.; PASON, P.; RATANAKHANOKCHAI, K.; DENG, L.; SERMSATHANASWADI, J.; SEPTININGRUM, K.; MORI, Y.; KOSUGI, A. Direct glucose production from lignocellulose using *Clostridium thermocellum* cultures supplemented with a thermostable beta-glucosidase. **Biotechnology for Biofuels**, London, v. 6, n. 1, p. 184 – 192, 2013.

PUNTA, M.; COGGILL, P. C.; EBERHARDT, R. Y.; MISTRY, J.; TATE, J.; BOURSNEILL, C.; PANG, N.; FORSLUND, K.; CERIC, G.; CLEMENTS, J.; HEGER, A.; HOLM, L.; SONNHAMMER, E. L. L.; EDDY, S. R.; BATEMAN, A.; FINN, R. D. The pfam protein families database. **Nucleic Acids Research**, Oxford, v. 40, n. D1, p. 290 – 301, 2012.

QI, B.; CHEN, X.; YI, S.; WAN, Y. Inhibition of cellulase, b-glucosidase, and xylanase activities and enzymatic hydrolysis of dilute acid pretreated wheat straw by acetone-butanol-ethanol fermentation products. **Environmental Progress and Sustainable Energy**, Hoboken, v. 33, n. 2, p. 497 – 503, 2014.

RADIOVOJAC, P.; CLARK, W. T.; ORON, T. R.; SCHNOES, A. M.; WITTKOP, T.; SOKOLOV, A.; GRAIM, K.; FUNK, C.; VERSPOOR, K.; BEN-HUR, A.; PANDEY, G.; YUNES, J. M.; TALWALKAR, A. S.; REPO, S.; SOUZA, M. L.; PIOVESAN, D.; CASADIO, R.; WANG, Z.; CHENG, J.; FANG, H.; GOUGH, J.; KOSKINEN, P.; TORONEN, P.; NOKSO-KOIVISTO, J.; HOLM, L.; COZZETTO, D.; BUCHAN, D. W. A.; BRYSON, K.; JONES, D. T.; LIMAYE, B.; INAMDAR, H.; DATTA, A.; MANJARI, S. K.; JOSHI, R.; CHITALE, M.; KIHARA, D.; LISEWSKI, A. M.; ERDIN, S.; VENNER, E.; LICHTARGE, O.; RENTZSCH, R.; YANG, H.; ROMERO, A. E.; BHAT, P.; PACCANARO, A.; HAMP, T.; KASZNER, R.; SEEMAYER, S.; VICEDO, E.; SCHAEFER, C.; ACHTEN, D.; AUER, F.; BOEHM, A.; BRAUN, T.; HECHT, M.; HERON, M.; HONIGSCHMID, P.; HOPF, T. A.; KAUFMANN, S.; KIENING, M.; KROMPASS, D.; LANDERER, C.; MAHLICH, Y.; ROOS, M.; BJORNE, J.; SALAKOSKI, T.; WONG, A.; SHATKAY, H.; GATZMANN, F.; SOMMER, I.; WASS, M. N.; STERNBERG, M. J. E.; SKUNCA, N.; SUPEK, F.; BOSNJAK, M.; PANOV, P.; DZEROSKI, S.; SMUC, T.; KOURMPETIS, Y. A. I.; DIJK, A. D. J. van; BRAAK, C. J. F. t.; ZHOU, Y.; GONG, Q.; DONG, X.; TIAN, W.; FALDA, M.; FONTANA, P.; LAVEZZO, E.; CAMILLO, B. D.; TOPPO, S.; LAN, L.; DJURIC, N.; GUO, Y.; VUCETIC, S.; BAIROCH, A.; LINIAL, M.; BABBITT, P. C.; BRENNER, S. E.; ORENGO, C.; ROST, B.; MOONEY, S. D.; FRIEDBERG, I. A large-scale evaluation of computational protein function prediction. **Nature Methods**, New York, v. 10, p. 221 – 227, 2013.

ROTHMAN, D. H.; FOURNIER, G. P.; FRENCH, K. L.; ALM, E. J.; BOYLE, E. A.; CAO, C.; SUMMONS, R. E. Methanogenic burst in the end-permian carbon cycle. **Proceedings of the National Academy of Sciences**, Washington, v. 111, n. 15, p. 5462 – 5467, 2014.

RUSSELL, D. J. **Multiple Sequence Alignment Methods**. New York: Humana Press, 2014. p. 143 - 150.

SACCONE, C.; PESOLE, G. Biological databases in the genomic era. In: _____. **Handbook of comparative Genomics**. Hoboken: John Wiley and Sons, 2005. chap. 5, p. 159 –186.

SCHELLENBERG, J.; VERBEKE, T.; MCQUEEN, P.; KROKHIN, O.; ZHANG, X.; ALVARE, G.; FRISTENSKY, B.; THALLINGER, G.; HENRISSAT, B.; WILKINS, J.; LEVIN, D.; SPARLING, R. Enhanced whole genome sequence and annotation of *Clostridium stercorarium* dsm8532t using rna-seq transcriptomics and high-throughput proteomics. **BMC Genomics**, London, v. 15, n. 1, p. 567 – 570, 2014.

SCHNOES, A. M.; BROWN, S. D.; DODEVSKI, I.; BABBITT, P. C. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. **PLoS Comput Biol**, San Francisco, v. 5, n. 12, 2009. Disponível em: <<http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000605>>. Acesso em: 13 set. 2013.

SEDAGHATINIA, A.; ATAN, R.; ARIFIN, K.; MURAD, M. Comparison and evaluation of multiple sequence alignment tools in bioinformatics. **International Journal of Computacional Science Networking Security**, Seoul, v. 9, p. 51 – 56, 2009.

SELZER, P.; MARHÖFER, R.; ROHWER, A. **Applied Bioinformatics: An Introduction**. Amsterdam: Springer, 2008. p. 57 - 64.

SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N. S.; WANG, J. T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. **Genome Research**, Woodbury, v. 13, n. 11, p. 2498 – 2504, 2003.

SHEN, C.; LIAO, J. Metabolic engineering of *Escherichia coli* for 1-butanol and 1-propanol production via the keto-acid pathways. **Metabolic Engineering**, Waltham, v. 10, n. 6, p. 312 – 320, 2008.

SIGRIST, C. J. A.; CERUTTI, L.; HULO, N.; GATTIKER, A.; FALQUET, L.; PAGNI, M.; BAIROCH, A.; BUCHER, P. Prosite: A documented database using patterns and profiles as motif descriptors. **Briefings in Bioinformatics**, London, v. 3, n. 3, p. 265 – 274, 2002.

SINGH, O. **Extremophiles: Sustainable Resources and Biotechnological Implications**. Hoboken: Wiley, 2012. p. 43 - 64.

SINGH, O.; GABANI, P. Extremophiles: radiation resistance microbial reserves and therapeutic implications. **Journal of Applied Microbiology**, Washington, v. 110, n. 4, p. 851 – 861, 2011.

SJÖLANDER, K. Phylogenomic databases and orthology prediction. In:____. **Protein Families: Relating Protein Sequence, Structure, and Function**. Hoboken: Wiley, 2013. chap. 5, p. 99 – 125.

SKUNCA, N.; ALTENHOFF, A.; DESSIMOZ, C. Quality of computationally inferred gene ontology annotations. **PLoS Comput Biol**, San Francisco, v. 8, n. 5, 2012. Disponível em: <<http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002533>>. Acesso em: 17 set. 2014.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of Molecular Biology**, London, v. 147, n. 1, p. 195 – 197, 1981.

SOREK, N.; YEATS, T. H.; SZEMENYEI, H.; YOUNGS, H.; SOMERVILLE, C. R. The implications of lignocellulosic biomass chemical composition for the production of advanced biofuels. **BioScience**, Irvine, v. 80, p. 1 – 13, 2014.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews Genetics**, London, v. 2, n. 7, p. 493 – 503, 2001.

STEPHEN, J. D.; MABEE, W. E.; SADDLER, J. N. Will second-generation ethanol be able to compete with first-generation ethanol? opportunities for cost reduction. **Biofuels, Bioproducts and Biorefining**, Hoboken, v. 6, n. 2, p. 159 – 176, 2012.

STRASSER, B. J. Collecting, comparing, and computing sequences: The making of margaret o. dayhoffs atlas of protein sequence and structure, 1954–1965. **Journal of the History of Biology**, Dordrecht, v. 43, p. 623 – 660, 2010.

TORTO-ALALIBO, T.; PURWANTINI, E.; LOMAX, J.; SETUBAL, J. C.; MUKHOPADHYAY, B.; TYLER, B. M. Genetic resources for advanced biofuel production described with the gene ontology. **Frontiers in Microbiology**, Lausanne, v. 5, n. 528, 2014.

UNIÃO DA INDÚSTRIA DE CANA-DE-ACÚCAR. **Moagem de cana-de-açúcar e produção de açúcar e etanol - safra 2013/2014**. São Paulo, 2014. Disponível em: <<http://www.unica.com.br/>>. Acessado em: 14 ago. 2014.

WITTKOP, T.; EMIG, D.; TRUSS, A.; ALBRECHT, M.; BÖCKER, S.; BAUMBACH, J. Comprehensive cluster analysis with Transitivity Clustering. **Nature Protocols**, London, v. 6, n. 3, p. 285 – 295, 2011.

XIN, F.; WU, Y.; HE, J. Simultaneous fermentation of glucose and xylose to butanol by *Clostridium* species strain boh3. **Applied and Environmental Microbiology**, Washington, v. 14, p. 53 – 59, 2014.

YEATS, C.; ORENGO, C.; VEUTHEY, A. L.; BOECKMANN, B.; JENSEN, L. J.; VALENCIA, A.; RAUSELL, A.; BORK, P. Modern genome annotation: the BioSapiens network. **Modern Genome Annotation**, Milan, p. 213 – 238, 2008.

YOGESH, S.; RICHARD, P.; BAIKUN, L. Bioenergy production from glycerol in hydrogen producing bioreactors (hpbs) and microbial fuel cells (mfcs). International. **Journal of Hydrogen Energy**, Oxford, v. 36, p. 3853 – 3861, 2011.

ZANNONI, D.; PHILIPPIS, R. D. **Microbial BioEnergy: Hydrogen Production**. Amsterdam: Springer Netherlands, 2014. p. 215 - 173.

ANEXOS

ANEXO A - FAMÍLIAS REVISADAS E PRELIMINARES

Tabela 1 – Lista completa das famílias estudadas, com versões revisadas ou preliminares disponíveis

(continua)			
MENGOfam	Nome da família	Status	Última análise
001-03	Endoglucanase (F, N, Z)	Reviewed	Monteiro-Vitorello
002-01	Extracellular solute-binding protein family 1	Reviewed	Orpinelli
003-01	Dockerin type 1	Preliminary	Orpinelli
004-01	Endoglucanase D	Reviewed	Monteiro-Vitorello
005-01	Sugar ABC transporter	Preliminary	Orpinelli
006-01	Acetyl-CoA decarbonylase/synthase complex subunit gamma 1	Preliminary	Orpinelli
007-05	Starch-utilization system protein F	Reviewed	Monteiro-Vitorello
008-01	ABC-type sugar transport system periplasmic component-like protein	Reviewed	Orpinelli
009-01	Extracellular solute-binding protein family 1	Preliminary	Orpinelli
011-07	Endo-1,3-beta-xylanase C	Reviewed	Monteiro-Vitorello
012-06	Cellulose 1,4-beta-cellobiosidase	Reviewed	Monteiro-Vitorello
013-01	Acetyl-CoA decarbonylase/synthase complex subunit delta 1	Reviewed	Orpinelli
014-04	Endoglucanase E	Reviewed	Monteiro-Vitorello
015-03	Endo-1,4-beta-xylanase Z	Reviewed	Taniguti
016-01	Methyl-coenzyme M reductase I subunit alpha	Preliminary	Orpinelli
017-01	Methyl-coenzyme M reductase I subunit beta	Preliminary	Orpinelli
018-01	Methyl-coenzyme M reductase I subunit gamma	Preliminary	Orpinelli
019-01	Butyrate–acetoacetate CoA-transferase subunit B	Preliminary	Orpinelli
020-01	Aldehyde-alcohol dehydrogenase	Preliminary	Orpinelli
021-01	Butyrate–acetoacetate CoA-transferase subunit A	Preliminary	Orpinelli
022-06	Glucan 1,4 - alphaglucosidase SusB	Reviewed	Monteiro-Vitorello
023-01	Phosphate butyryltransferase	Preliminary	Orpinelli
024-01	ACDS epsilon 1	Preliminary	Orpinelli
025-01	Phosphate acetyltransferase	Reviewed	Orpinelli
026-01	Acetate kinase	Preliminary	Orpinelli
027-03	Endoglucanase J	Reviewed	Monteiro-Vitorello
028-01	ACDS complex subunit beta 1	Preliminary	Orpinelli
029-01	Scaffoldin dockerin binding protein A	Preliminary	Orpinelli
031-01	NADH-dependent butanol dehydrogenase A	Preliminary	Orpinelli
032-05	Endoglucanase C	Reviewed	Monteiro-Vitorello
033-01	NADH-dependent butanol dehydrogenase B	Preliminary	Orpinelli
034-01	Scaffolding protein	Preliminary	Orpinelli
035-01	Cell surface glycoprotein 1	Preliminary	Orpinelli
036-02	L-sulfolactate dehydrogenase	Reviewed	Monteiro-Vitorello
037-02	Formate dehydrogenase subunit beta	Reviewed	Monteiro-Vitorello
038-01	Butyrate kinase 1	Preliminary	Orpinelli
039-03	Endoglucanase G	Reviewed	Monteiro-Vitorello

Tabela 1 – Lista completa das famílias estudadas, com versões revisadas ou preliminares disponíveis

(continuação)

MENGOfam	Nome da família	Status	Última análise
040-01	Putative substrate binding protein	Preliminary	Orpinelli
041-01	Carbohydrate-binding protein (cttA)	Preliminary	Orpinelli
042-03	Endoglucanase H	Reviewed	Monteiro-Vitorello
043-02	CoB–CoM heterodisulfide reductase iron-sulfur subunit A	Reviewed	Monteiro-Vitorello
044-09	Lactaldehyde dehydrogenase	Reviewed	Monteiro-Vitorello
045-03	Cellulose 1,4-beta-cellobiosidase (cbhA)	Reviewed	Monteiro-Vitorello
046-04	Endo-1,4-beta-xylanase D	Reviewed	Monteiro-Vitorello
047-03	Xyloglucan-specific endo-beta-1,4-glucanase activity	Reviewed	Monteiro-Vitorello
048-01	ScaC cellulosomal scaffoldin protein	Preliminary	Orpinelli
049-01	SusR - regulatory protein	Preliminary	Orpinelli
050-04	Neopullulanase SusA	Reviewed	Monteiro-Vitorello
051-01	Starch-utilization system protein C	Preliminary	Orpinelli
052-02	Starch-utilization system protein D	Reviewed	Monteiro-Vitorello
053-05	Alpha-amylase SusG	Reviewed	Monteiro-Vitorello
054-02	3-hydroxybutyryl-CoA dehydrogenase	Preliminary	Orpinelli
055-01	MalR - regulator protein	Preliminary	Orpinelli
056-02	Endoglucanase T	Reviewed	Monteiro-Vitorello
057-01	ScaA cellulosomal scaffoldin protein	Preliminary	Orpinelli
059-02	F420-dependent methylenetetrahydromethanopterin dehydrogenase	Preliminary	Monteiro-Vitorello
060-01	Acetyl-CoA decarboxylase/synthase complex subunit delta 2	Preliminary	Orpinelli
061-01	Acetyl-CoA decarboxylase/synthase complex subunit epsilon 2	Preliminary	Orpinelli
062-02	ACDS complex subunit alpha 2	Preliminary	Orpinelli
063-04	Cellulose 1,4-beta-cellobiosidase (celO)	Reviewed	Monteiro-Vitorello
064-01	Cellulosomal element P (cseP)	Preliminary	Orpinelli
065-01	Acetyl-CoA decarboxylase/synthase complex subunit gamma 2	Preliminary	Orpinelli
066-01	Acetyl-CoA decarboxylase/synthase complex subunit beta 2	Preliminary	Orpinelli
067-01	Basic membrane lipoprotein	Preliminary	Orpinelli
068-03	Monomethylamine methyltransferase MtmB1	Reviewed	Monteiro-Vitorello
069-02	Monomethylamine methyltransferase	Reviewed	Monteiro-Vitorello
071-03	Methylated-thiol–corrinoid protein MtsB	Reviewed	Monteiro-Vitorello
072-03	Endo-1,4-beta-xylanase Y	Reviewed	Monteiro-Vitorello
073-01	Cellulosomal-scaffolding protein A	Reviewed	Orpinelli
074-04	Aldehyde/alcohol dehydrogenase	Reviewed	Monteiro-Vitorello
077-02	Endoglucanase A	Reviewed	Monteiro-Vitorello
078-01	Endo-1,4-glucanase	Preliminary	Orpinelli
079-04	Endoglucanase Y	Reviewed	Monteiro-Vitorello
082-05	arabinoxylan-specific endo-b-1,4-xylanase	Queue	Monteiro-Vitorello

Tabela 1 – Lista completa das famílias estudadas, com versões revisadas ou preliminares disponíveis

(conclusão)			
MENGOfam	Nome da família	Status	Última análise
089-01	Pectin acetylesterase	Preliminary	Orpinelli
092-01	Fe-hydrogenase beta subunit	Preliminary	Orpinelli
093-01	Fe-hydrogenase alpha subunit	Preliminary	Orpinelli
098-01	Methenyltetrahydromethanopterin cyclohydrolase	Preliminary	Orpinelli
099-01	3-hydroxybutyryl-CoA dehydrogenase	Preliminary	Orpinelli
100-01	3-hydroxybutyryl-CoA dehydratase	Preliminary	Orpinelli
109-01	Hydrogenase small beta subunit	Preliminary	Orpinelli
110-03	Hydrogenase large alpha subunit	Preliminary	Orpinelli
112-01	Diacetylchitobiose deacetylase	Preliminary	Orpinelli
113-01	N,N'-diacetylchitobiose phosphorylase	Preliminary	Orpinelli

ANEXO B – Termos propagados e predições GO

Tabela 1 – Descrição dos domínios CAZy, número de membros, predição de peptídeo sinal e termos do Gene Ontology referentes a cada uma das famílias estudadas

(continua)				
MENGOfam	CAZy	Membros	Pep. Sinal	Gene Ontology
001-03	CBM3-GH9	54	Sim	- cellulase activity
002-01	--	18	Sim	cellodextrin binding
004-01	GH9	11	Sim	- NA
007-05	--	16	Sim	- outer membrane starch binding
008-01	--	16	Sim	- cellotriose binding cellotetraose binding cellopentaose binding
011-07	CBM22-GH10	11	Sim	- xylan endo-1,3-beta-xylosidase activity xylan catabolic process
012-06	CBM4-GH9	24	Sim	- NA
013-01	--	28	Não	- CoB–CoM heterodisulfide reductase activity
014-04	GH5	9	Sim	- cellulose binding
015-04	CBM6-GH10	14	Sim	- endo-1,4-beta-xylanase activity xylan endo-1,3-beta-xylosidase activity
022-06	--	164	Sim	- alpha-1,4-glucosidase activity
025-01	--	105	Não	- acetate biosynthetic process
027-03	CBM30-CBM44-GH44-GH9	9	Não	- cellulase activity cellulose catabolic process
032-05	GH5	76	Não	- cellulose binding
036-02	--	10	Não	- sulfopyruvate decarboxylase activity
037-02	--	91	Não	- formate dehydrogenase (coenzyme F420) activity
039-03	GH5	15	Sim	- cellulase activity cellulosome

Tabela 1 – Descrição dos domínios CAZy, número de membros, predição de peptídeo sinal e termos do Gene Ontology referentes a cada uma das famílias estudadas

(conclusão)				
MENGOfam	CAZy	Membros	Pep. Sinal	Gene Ontology
042-03	CBM11-GH26-GH5	8	Sim	- cellulase activity
043-02	--	15	Não	- CoB--CoM heterodisulfide reductase activity
044-09	--	11	Não	- lactaldehyde dehydrogenase activity
045-04	GH9	21	Não	- cellulose 1,4-beta-cellobiosidase activity cellulose catabolic process
046-05	CBM22-GH10	8	Não	endo-1,4-beta-xylanase activity
047-03	GH74	11	Sim	xyloglucan-specific endo-beta-1,4-glucanase activity cellulosome xyloglucan catabolic process
050-04	--	66	Sim	neopullulanase activity
052-02	--	15	Sim	outer membrane starch binding
053-05	--	32	Sim	alpha-amylase activity outer membrane
056-02	GH9	17	Não	cellulase activity
063-04	CBM3-GH5	6	Não	cellulose 1,4-beta-cellobiosidase activity cellulose binding
068-03	--	13	Não	monomethylamine methyltransferase activity
069-02	--	13	Não	monomethylamine methyltransferase activity
071-03	--	4	Não	methylthiol:coenzyme M methyltransferase activity
072-03	CBM22-GH10	12	Sim	cellulosome
073-01	CBM3	8	Sim	cellulose binding
074-04	--	280	Não	alcohol dehydrogenase (NAD) activity aldehyde dehydrogenase (NAD) activity 1-butanol biosynthetic process
077-02	GH8	14	Sim	cellulase activity cellulose catabolic process
079-05	CBM3-GH48	8	Não	cellulose 1,4-beta-cellobiosidase activity cellulose catabolic process