Universidade de São Paulo

**Modelagem computacional de famílias de proteínas
microbianas relevantes para produção de bioenergia**

Fernanda Orpinelli Ramos do Rego

Dissertação apresentada para obtenção
do título de Mestre em Ciências. Área de
concentração: Bioinformática.

**São Paulo
2015**

Fernanda Orpinelli Ramos do Rego
Bacharel em Ciências Biológicas

**Modelagem computacional de famílias de proteínas
microbianas relevantes para produção de bioenergia**

Orientador:
Prof. Dr. **JOÃO CARLOS SETUBAL**

Dissertação apresentada para obtenção
do título de Mestre em Ciências. Área de
concentração: Bioinformática.

São Paulo
2015

*À minha família e meus amigos pelo apoio e
incentivo incondicionais.*

AGRADECIMENTOS

À Universidade de São Paulo – USP pela estrutura para realização do curso de Mestrado.

Ao professor Dr. João Carlos Setubal pela orientação, paciência, críticas, sugestões e pela experiência em Bioinformática essencial para minha formação nesta área.

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

À professora Dra. Cláudia Barros Monteiro-Vitorello, colaboradora do projeto, que compartilhou sua experiência em curadoria e contribuiu com sugestões sempre com muita atenção e carinho. Ao seu aluno de mestrado, Lucas Taniguti, por ter criado o site que abriga às HMMs geradas por este trabalho.

Aos membros do Setulab, especialmente Luiz Thibério Rangel, Rodrigo Guarischi, Leandro Lemos, Gianluca Major e George Condomitti pelas discussões enriquecedoras que ajudaram no desenvolvimento do trabalho, além da amizade que levarei para vida.

Ao Dr. Arthur Trancoso Lopo de Queiroz, por me incentivar a buscar o programa de pós-graduação em Bioinformática e por todo apoio e paciência ao longo desta caminhada.

Aos meus pais, que não mediram esforços para me ajudar na realização do sonho de seguir a carreira acadêmica. Ao meu irmão, por ter sido sempre o meu maior exemplo em dedicação.

Aos meus avós, Regina, Sylvino por me acolherem sempre com muito amor. E à vó Cida (*in memoriam*) pela sabedoria e pela fé que me inspiram até hoje.

Ao Felipe Failta, meu namorado, por tornar os momentos de ansiedade e angústia mais leves e felizes.

Aos meus amigos de Salvador, por estarem comigo sempre independente da distância.

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

“Não é na ciência que está a felicidade, mas na aquisição da ciência.”

(Edgar Allan Poe)

SUMÁRIO

RESUMO	1
ABSTRACT	2
1 INTRODUÇÃO	3
1.1 Famílias de proteínas	6
1.2 Formas de fazer propagação automática	7
2 HIDDEN MARKOV MODELS (HMMs)	11
2.1 O conceito de HMM	11
2.2 HMMs para representar famílias	12
2.3 Exemplos de conjuntos de HMMs representando famílias	14
2.3.1 PFAM	14
2.3.2 TIGRFams	14
2.3.3 PANTHER	15
2.3.4 dbCAN	16
2.4 Alternativa a HMMs	16
3 MENGO	18
4 MENGOFams	20
4.1 Pipeline para criação e curadoria de um HMM	21
4.2 Comparação com a literatura	27
5 RESULTADOS	29
5.1 Conjunto de HMMs	29
5.2 Validação	29
5.3 Estudos de caso: famílias complexas e simples	35
5.4 Tabela de proteínas candidatas a serem reanotadas	45

5.5 Aplicação dos modelos em dados de metagenômica	46
6 CONCLUSÃO	50
7 REFERÊNCIAS BIBLIOGRÁFICAS	51
APÊNDICE A - Resumo apresentado na Biocuration 2013	57
APÊNDICE B - Criação e manutenção do site	58
APÊNDICE C - Tabela com as famílias que foram estudadas	65

RESUMO

Modelagem computacional de famílias de proteínas microbianas relevantes para produção de bioenergia

Modelos ocultos de Markov (HMMs - hidden Markov models) são ferramentas essenciais para anotação automática de proteínas. Por muitos anos, bancos de dados de famílias de proteínas baseados em HMMs têm sido disponibilizados para a comunidade científica (e.g. TIGRfams). Muitos esforços também têm sido dedicados à geração automática de HMMs de famílias de proteínas (e.g. PANTHER). No entanto, HMMs manualmente curados de famílias de proteínas permanecem como o padrão-ouro para anotação de genomas. Neste contexto, este trabalho teve como principal objetivo a geração de cerca de 80 famílias de proteínas microbianas relevantes para produção de bioenergia, baseadas em HMMs. Para gerar os HMMs, seguimos um protocolo de curadoria manual, gerado neste trabalho. Partimos de uma proteína que tenha função experimentalmente comprovada, esteja associada a uma publicação e tenha sido manualmente anotada com termos da *Gene Ontology*, criados pelo projeto MENGO¹ (Microbial ENergy Gene Ontology). Os próximos passos consistiram na (1) definição de um critério de seleção para inclusão de membros à família; (2) busca por membros via BLAST; (3) geração do alinhamento múltiplo (MUSCLE 3.7) e do HMM (HMMER 3.0); (4) análise dos resultados e iteração do processo, com o HMM preliminar usado nas buscas adicionais; (5) definição de uma nota de corte (*cutoff*) para o HMM final; (6) validação individual dos modelos. As principais contribuições deste trabalho são 74 HMMs (manualmente curados) disponibilizados via *web* (<http://mengofams.lbi.iq.usp.br/>), onde é possível fazer buscas e o *download* dos modelos, um protocolo detalhado sobre a curadoria manual de HMMs para famílias de proteínas e uma lista com proteínas candidatas a reanotação.

Palavras-chave: HMM, Anotação, Biocuradoria, Bioenergia, Gene Ontology.

¹ <http://www.mengo.biochem.vt.edu>

ABSTRACT

Computational modeling of microbial protein families relevants to bioenergy production process

Hidden Markov Models (HMMs) are essential tools for automated annotation of protein sequences. For many years now protein family resources based on HMMs have been made available to the scientific community (e.g. TIGRfams). Much effort has also been devoted to the automated generation of protein family HMMs (e.g Panther). However, manually curated protein family HMMs remain the gold standard for use in genome annotation. In this context, this work had as main objectives the generation of approximately 80 protein families based on HMMs. We follow a standard protocol, that was generated in this work, to create the HMMs. At first, we start from a protein with experimentally proven function, associated to a publication and that was manually annotated with new terms from Gene Ontology provided by MENGO¹ (Microbial ENergy Gene Ontology). The next steps consists of (1) definition of selection criteria to capture members of the family; (2) search for members via BLAST; (3) generation of multiple alignment (MUSCLE 3.7) and the HMM (HMMER 3.0); (4) result analysis and iteration of the process, using the preliminary HMM; (5) cutoff definition to the final HMM; (6) individual validation of the models using tests against NCBI's NR database. The main deliverables of this work are 74 HMMs manually curated available in the site project (mengofams.lbi.iq.usp.br) that allows browsing and download of all HMMs curated so far, a standard protocol manual curation of protein families, a list with proteins that need to be reviewed.

Key-words: HMM, Annotation, Biocuration, Bioenergy, Gene Ontology.

¹ <http://www.mengo.biochem.vt.edu>

1 Introdução

Anotação: definição

Anotação é o processo pelo qual trechos de genomas, particularmente genes codificadores de proteínas, são anotados com informações sobre sua função no organismo ao qual pertencem (STEIN, 2001). Neste trabalho nos interessam apenas os genes codificadores de proteína, portanto, o termo anotação nesta dissertação aplica-se somente a esses elementos genômicos.

Após a anotação estrutural dos genes (determinação do início e fim da região codante do gene, ou seja, responder a pergunta “onde?”), segue-se a anotação das proteínas (responder a pergunta “o que?”) potencialmente codificadas por essas sequências (STEIN, 2001). Sabe-se que a função das proteínas é determinada por sua estrutura tridimensional, e consequentemente pela sequência de aminoácidos. A anotação dos produtos gênicos consiste em catalogar, nomear e associar às sequências suas prováveis funções.

O processo de anotação de sequências gênicas em genomas completos pode ser feito manual ou automaticamente. Com a redução dos custos de projetos de sequenciamento de DNA e o rápido desenvolvimento das tecnologias, o número de sequências biológicas disponíveis nos bancos de dados públicos tem aumentado exponencialmente. Dados do principal banco de dados de sequências, GenBank, apontam que 852.301 novas sequências foram adicionadas entre fevereiro e abril de 2015, onde inicialmente o total de sequências era 181.336.445 e em seguida 182.188.746 (dados disponíveis em <http://www.ncbi.nlm.nih.gov/genbank/statistics>, acesso maio/2015). Atualmente, estima-se que os bancos de dados de proteínas armazenam mais de 20 milhões de sequências (ORENGO e BATEMAN, 2013). Assim, pelo o crescente volume de dados, geralmente adota-se a anotação automática ao invés da caracterização funcional experimental de proteínas, a qual é difícil e envolve altos custos. Cabe mencionar o recente artigo de Anton e colaboradores (ANTON *et al.*, 2014), onde se salienta a enorme discrepância que existe atualmente entre o número de genes com função experimentalmente verificada e o número de genes com sequência determinada e função predita apenas computacionalmente.

Transferência de anotação

Para atender à necessidade de anotação confiável de genes codificadores de proteínas, existem bancos de dados de famílias de proteínas que facilitam a caracterização de sequências “desconhecidas”. Tais bancos serão comentados adiante bem como as diferenças entre eles.

O uso de tais bancos depende do conceito de *transferência de anotação*. Esta forma de anotação é uma metodologia amplamente utilizada e se baseia na possibilidade de proteínas serem agrupadas em famílias. Uma sequência proteica com função desconhecida *A* pode “herdar” a anotação de outra proteína *B* já anotada, desde que exista similaridade suficiente entre *A* e *B*. Esta busca por proteínas similares é feita primordialmente por BLAST (SJÖLANDER, 2013). Esta forma de transferência de anotação tem se tornado essencial na anotação de produtos gênicos. No entanto, esta metodologia, sem nenhuma curadoria, pode levar à propagação de erros.

Erros de anotação e a Gene Ontology

O erro mais comum e mais problemático na transferência de anotações ocorre quando a anotação da proteína de referência está errada. Neste caso, todas as proteínas que herdarem esta anotação também ficarão com anotações erradas. Se estas, por sua vez, também forem usadas como referência, este erro se propagará com muita facilidade, e em curto espaço de tempo uma grande quantidade de genes em bancos de dados, como GenBank, poderá estar com anotações erradas. Segundo Brenner e colaboradores (1999), este cenário é real, embora seja difícil estimar a quantidade de anotações erradas que constam em um banco como GenBank. O problema apontado neste parágrafo é uma das principais motivações deste trabalho.

Outro problema de anotação em bancos como GenBank é a falta de padronização de termos descritivos para sequências biológicas. Muitos genes com mesma função são descritos com diferentes termos (não necessariamente errados); ou o mesmo termo pode ser usado para descrever funções que na verdade são diferentes. Esta falta de padronização dificulta a comparação entre conteúdos gênicos de genomas.

A *Gene Ontology* (GO) é um projeto que tem criado um vocabulário controlado para nomear funções moleculares, localização celular e processos biológicos para produtos gênicos

(ASHBURNER *et al.*, 2000). Essa padronização ocorre tanto nos termos descritivos em si quanto no relacionamento entre os termos. Este ocorre por meio de um grafo acíclico orientado, em que os nós contêm termos e as arestas que ligam os nós são relações do tipo ‘is a’, ‘is a part of’, ‘has part’, ‘regulates’, ‘positively regulates’, ‘negatively regulates’. O grafo é uma estrutura hierárquica, e os termos vão do nível mais geral para os níveis mais específicos (os termos-folha). Um exemplo de termo-folha é ‘starch binding’ (GO:2001070) relacionado a ligação ao amido, que tem como pais os termos ‘polysaccharide binding’ (GO:0030247), ‘pattern binding’ (GO:0001871), ‘carbohydrate binding’ (GO:0030246), ‘binding’ (GO:0005488), ‘molecular_function’ (GO:0003674) (figura 1).

Figura 1. Representação hierárquica dos termos da *Gene Ontology*.

Quando um termo é aplicado a um produto gênico, deve-se acrescentar uma série de outras informações, uma das quais sendo o código de evidência. Os códigos de evidência se dividem em cinco grupos, como o “Experimental Evidence Codes” e o “Automatically-assigned Evidence Codes”. Exemplos de códigos de evidência destes grupos são: ‘EXP, inferred from experiment’, ‘IDA, inferred from direct assay’ e ‘IEA, inferred from eletronic annotation’, sendo os dois primeiros pertencentes ao grupo de códigos de evidência baseados em dados experimentais e o último pertencente ao grupo de códigos usados para denotar anotações geradas automaticamente. Para efeitos deste trabalho, os códigos de evidência mais importantes são aqueles que denotam evidência experimental (p. ex. IDA, *inferred from direct assay*). As proteínas-semente de nosso trabalho (ver adiante) precisam necessariamente ter tido sua função comprovada experimentalmente, e, portanto sua anotação deverá ter feito uso de um código de evidência do grupo *Experimental Evidence Codes*.

1.1 Famílias de proteínas

Proteínas são constituídas por blocos estruturais que podem desempenhar papéis importantes na função destas macromoléculas. Estes blocos compreendem os domínios proteicos. Uma proteína pode ser constituída apenas por um domínio ou pode ser multidomínio, onde os aminoácidos interagem e formam estruturas organizadas e bem definidas. Os domínios de uma proteína podem ser classificados em famílias, separadamente ou não, e fornecem então o conhecimento sobre o dobramento e a função de uma proteína em estudo (HOLM e HEGER, 2013).

O conceito de família de proteínas é fundamental para o entendimento deste trabalho. Proteínas podem ser agrupadas a outras proteínas constituindo uma família de proteínas, a qual é definida como um grupo de produtos gênicos que compartilham uma origem evolutiva comum, que se reflete na manutenção da função na sequência apresentada por todos os seus membros (ORENGO e BATEMAN, 2013).

Conceitos de homologia

Em uma família como a definida acima, todos os membros são ditos homólogos e o tipo de homologia apresentada por eles pode ser ortologia ou paralogia. Em 1970, Fitch usou o termo ortólogo pela primeira vez para diferenciar eventos decorrentes de especiação de eventos decorrentes de duplicação gênica (SJÖLANDER, 2013). Ortólogos referem-se à proteínas homólogas em espécies diferentes que mantiveram a mesma função ao longo do processo evolutivo. As proteínas parálogas são aquelas resultantes de eventos de duplicação (KOONIN, 2005; SJÖLANDER, 2013), e que podem ou não ter a mesma função: uma das cópias pode apresentar o fenômeno de *neofuncionalização*, em que a menor pressão evolutiva sobre essa cópia permite o aparecimento de uma função diferente da função ancestral.

Curadoria de anotação de genes quanto à função

Na análise de sequências, devemos atentar para posições importantes para função, estabilidade e *folding*, pois estas determinam a conservação da função em relação à proteína-semente (ORENGO e THORNTON, 2005). Quando tais posições encontram-se conservadas, podemos classificar uma proteína em análise como pertencente à mesma família da proteína semente. Tais

regiões compreendem: sítios ativos, sítios de ligação, regiões repetitivas, peptídeo-sinal, e geralmente estão alojadas em domínios nas proteínas.

A complexidade funcional de proteínas nem sempre permite uma classificação clara numa família bem delimitada. Por esse motivo, utilizamos os conceitos de *superfamília* e *subfamília*. Uma superfamília congrega diversas famílias com função relacionada. Uma subfamília subdivide uma família existente, para permitir uma anotação de aspectos mais sutis de sua função (NELSON e COX, 2009).

1.2 Formas de fazer a propagação automática

A transferência da anotação entre sequências pode ser feita das seguintes formas: 1. via BLAST (ALTSCHUL, 1997), 2. por perfis de alinhamentos múltiplos, 3. por árvores filogenéticas ou 4. via HMMs, sendo que os métodos (2), (3) e (4) têm como base a utilização de alinhamentos múltiplos.

BLAST: propagação automática baseada no uso do primeiro hit

Esta é a forma mais simples de realizar transferência de anotação. A sequência da proteína a ser anotada é fornecida como sequência consulta, e os resultados da execução de BLAST sobre um banco de sequências (geralmente NR) fornecem as sequências que são candidatas à anotação para sequência consulta (os *hits*), em geral juntamente com algum critério de e-value máximo e cobertura (do hit e da consulta) mínima (procurando minimizar falsos positivos, ou seja, hits que não sejam homólogos à consulta). Adota-se como anotação a mesma anotação do primeiro hit que satisfaz os critérios estabelecidos. Esta técnica tem como vantagens a simplicidade e a rapidez das buscas (BATEMAN e HAFT, 2002). O principal problema deste método é que a anotação do primeiro hit, mesmo que este seja homólogo da proteína consulta, pode estar errada, conforme já mencionado. A adoção desta anotação representa, portanto propagação do erro (BRENNER, 1999). Uma forma de tentar minimizar tais erros é acrescentar aos critérios a exigência de que a proteína-hit tenha função comprovada experimentalmente; isto garante que sua anotação está correta. No entanto, na maioria das buscas com BLAST tais genes não aparecem (pois são relativamente raros), ou é difícil identificar automaticamente quais eles sejam na lista de hits. Isto deriva do fato, já mencionado, de que a vasta maioria das proteínas, cujas

sequências foram determinadas por genômica, não teve sua função comprovada experimentalmente.

Perfis: uso de alinhamentos múltiplos para transferência de anotação

Neste caso, é necessário inicialmente montar um conjunto de sequências que sejam todas homólogas e com mesma função. A forma como isto é feito será explicada mais adiante. É essencial que pelo menos uma das sequências pertencente ao conjunto seja de um gene com função experimentalmente comprovada. Esta é uma característica comum a este método e aos dois outros que iremos descrever. De posse desse conjunto, cria-se um alinhamento múltiplo com ele, o qual, após curadoria, pode ser usado para transferência automática de anotações. Esta é a técnica usada pelo projeto HAMAP (LIMA, 2009). Tem como vantagem o fato de que o perfil procura encapsular as variações de sequência da família, tornando-o mais sensível do que uma busca baseada numa única sequência. Assim, uma sequência com bom alinhamento ao perfil de uma família receberá automaticamente a anotação que foi dada para a família. A desvantagem deste método é que o cálculo da nota atribuída à substituição de aminoácidos em cada sítio segue um esquema arbitrário, com pouco embasamento estatístico. (BATEMAN e HAFT, 2002).

Propagação automática através de análises filogenômicas e sua complexidade técnica

Outra forma de propagação automática é o uso de árvores filogenéticas, como é feito pelos sistemas PHYLOFACTS (KRISHNAMURTHY, 2006), SIFTER (ENGELHARDT, 2005) e PAINT (GAUDET, 2011). Neste caso, é preciso montar uma árvore filogenética com as sequências homólogas e de mesma função, pertencentes a uma mesma família. Essa árvore é então utilizada como base na busca de novos membros da família. Este método seria o ideal para propagação, uma vez que árvores filogenéticas dão uma confiabilidade alta para afirmações de homologia, em geral mais alta do que a simples similaridade entre sequências, que são a base dos dois métodos já descritos. Porém este método é consideravelmente mais complexo do que os anteriores (KRISHNAMURTHY, 2006). Originalmente, o SIFTER 1.1 (ENGELHARDT, 2005) se baseava no uso de redes bayesianas, no entanto, a dificuldade de lidar com famílias mais complexas ou genomas completos levou à reformulação do programa. Conforme publicado em 2011, o SIFTER 2.0 (ENGELHARDT, 2011) se baseia em cadeias de Markov de tempo contínuo, o que melhorou seu desempenho mantendo a equivalência dos resultados obtidos.

Mesmo após as mudanças, a implementação do SIFTER ainda é mais custosa computacionalmente, em tempo e memória, quando comparada a outras metodologias, como citaremos adiante.

Uso de modelos de Markov de Estados Ocultos

Por fim, podem-se utilizar modelos de Markov de estados ocultos (ou HMMs) para fazer a propagação automática de anotações. Esta foi a metodologia escolhida para este trabalho. HMMs são modelos bem conhecidos e estabelecidos na comunidade científica. Os HMMs têm diversas aplicações em Bioinformática e são mais comumente usados para: definição de regiões distintas dentro de genes ou em sequências proteicas, definição de famílias de proteínas através de alinhamentos múltiplos e descoberta de genes (CRISTIANINI e HAHN, 2006). Bancos de famílias de proteínas como PFAM (FINN, 2014), PANTHER (THOMAS, 2003) e TIGRFams (HAFT, 2013) utilizam tais modelos para prover fontes de anotação automática. O software AMPHORA, para criação de árvores filogenéticas baseadas em genes marcadores (WU e EISEN, 2008) também usou HMMs para caracterizar as famílias desses genes.

Os fatores que nos levaram a escolha deste método para construir as nossas famílias foram:

1. O uso bastante geral de HMMs para modelar famílias de proteínas, como evidenciado pelas referências acima.
2. A rapidez com que sequências de proteínas podem ser comparadas com um HMM (ver abaixo)
3. A existência do excelente pacote HMMER 3.0, mantido por Sean Eddy e colaboradores, o qual é amplamente utilizado na geração e busca de HMMs. É este o software utilizado pelos mantenedores do PFAM. Isto significa que nossos HMMs se beneficiam da existência de programas desse pacote, tais como *hmmsearch* (permite busca num banco de sequências usando como consulta um HMM) e *hmmScan* (permite busca num banco de HMMs usando como consulta uma sequência). Qualquer usuário no mundo pode fazer uso de nossos HMMs por meio do pacote HMMER.

Alternativamente ao HMMER 3.0, existe o HHblits (Hm-Hmm-based lightning-fast iterative sequence search), o qual converte a *query*, ou o alinhamento múltiplo, e o banco de sequências

em HMMs, proporcionando maior sensibilidade e qualidade do alinhamento nas buscas por sequências (REMMERT *et al.*, 2012). Esta metodologia poderá ser aplicada a estudos no futuro, dado que pela praticidade optamos pela utilização do HMMER.

Como exemplo da rapidez mencionada acima, podemos mencionar que uma execução de *hmmsearch* com algum de nossos HMMs contra o banco NR do Genbank leva em geral menos de um minuto, quando rodado em uma máquina com Intel Core i7 (2ª geração) 2.90 GHz e 8Gb de RAM. Uma execução de SIFTER usando como consulta a proteína Endoglucanase D (UniprotKB id: A3DDN1) de 649 aminoácidos, levou cerca de três minutos quando executado na mesma máquina descrita acima. Vale mencionar que os autores de SIFTER indicam que a complexidade computacional do algoritmo é estimada em $O(N((2^M)^{2.3}))$, onde N é o número de genes na família de proteínas e M o número de funções candidatas (ENGELHARDT, 2011).

HMMs curados para famílias de proteína precisam estar associados a *notas de corte* (em inglês, *cutoff*), as quais são um componente essencial para permitir classificação automática de uma nova sequência numa família representada por um HMM (HAFT, 2001). Ao se rodar *hmmsearch*, obtém-se como resultado, para cada alinhamento encontrado, uma nota (ou *score*). Essa nota é uma medida da similaridade entre o modelo e a sequência, e corresponde conceitualmente ao bitscore de BLAST (ALTSCHUL, 1997). A nota de corte é uma nota mínima que define o limiar acima do qual as sequências encontradas serão consideradas como membros da família.

A determinação de uma nota de corte para um dado HMM precisa ser feito por um curador. Este irá julgar os alinhamentos, e com base no seu conhecimento da família em questão, irá determinar uma nota de corte que caracterize a família. A determinação de notas de corte para as famílias que estudamos foi uma parte significativa do trabalho aqui apresentado, conforme se verá mais adiante.

Vale notar que TIGRFams (HAFT, 2013) usa os conceitos de *trusted cutoff* e *noise cutoff*. O *noise cutoff* pode ser entendido como nota mínima para incluir uma sequência numa família. Porém não necessariamente o valor do *noise cutoff* será o mesmo que o valor de *trusted cutoff* para uma dada família. Isso cria a possibilidade de se ter uma zona cinza de notas, que pode ser

usada no caso de análise de famílias complexas. Neste trabalho não adotamos esse esquema para simplificar o processo de curadoria.

2 HMMs para representação de famílias e propagação automática.

2.1. O conceito de HMMs.

O início do uso de HMMs em sequências de DNA é datado de 1989 por Gary Churchill para definição/segmentação de diferentes regiões das sequências. A partir desta data, o uso de HMMs em genômica tornou-se mais amplo e somado à técnica de alinhamento múltiplo, estes modelos são os mais representativos do campo da Bioinformática. HMMs são representações probabilísticas que correspondem ao consenso de um conjunto de sequências (sejam elas de nucleotídeos ou aminoácidos) (KROGH, 1998). Tais modelos permitem sejam encontrados padrões dentro de um conjunto de sequências de forma que não necessariamente, obedeçam a uma estrutura rígida (CRISTIANINI e HAHN, 2006). Estas representações escolhidas para análises de sequências biológicas são assim chamadas por serem cadeias de Markov: cadeias de estados, em que a probabilidade de transição de um estado E para um estado A depende apenas do estado atual. Um HMM procura modelar uma dada realidade (por exemplo, uma sequência), supondo que esta realidade tenha os estados presentes no modelo. No entanto, em um dado instante (ou posição, quando percorremos a sequência da esquerda para a direita) não sabemos em que estado se encontra essa realidade, daí o fato de dizermos que os estados estão ocultos. Probabilisticamente, iremos estimar o estado em que se encontra a realidade num dado instante ou posição. Os estados correspondem à rotulação (estados de correspondência ou *gaps*), que podem representar diferentes tipos de sequências, desde as mais simples, como: “regiões ricas em GC”, ou mais complexas, como: “intron” ou “região codificante”. Os símbolos do modelo são os nucleotídeos ou aminoácidos, no caso de moléculas de DNA ou proteínas respectivamente. Assim, para cada posição na sequência tem-se um estado *oculto* e o símbolo visível gerado que depende do estado atual.

Os parâmetros do HMM são as probabilidades de emissão E (símbolos observados) e as probabilidades de transição T entre estados. A probabilidade de estar no estado oculto l , dado o

estado anterior k , é dado por $T(k, l)$ na matriz de transição, e a probabilidade de emissão do símbolo b é determinado pelo modelo multinomial associado ao estado k , o qual é dado pela matriz de emissão por $E(k, b)$, sendo h : sequências de estados ocultos gerados pelo processo de Markov e s : sequência de símbolos gerados pela HMM.

$$T(k, l) = P(h_i = l | h_{i-1} = k).$$

$$E(k, b) = P(s_i = x | h_i = k).$$

2.2 HMMs para representar famílias.

Conforme mencionado no capítulo anterior, HMMs têm sido amplamente utilizadas no modelamento de dados biológicos, especialmente para o modelamento de famílias de proteínas. No caso de famílias proteicas, partimos da premissa que sequências com a mesma função podem diferir na estrutura primária, mas podem pertencer a uma mesma família, uma vez que devem manter os domínios conservados. Assim, o uso de representações probabilísticas confere maior importância a estas regiões conservadas, atribuindo uma pontuação de acordo com a frequência do símbolo naquela posição, e gera um padrão que é utilizado na busca por novos membros.

Para trabalhar com famílias proteicas é feita uma modificação na estrutura geral de um HMM. Esta nova estrutura é chamada de *profile*-HMM (figura 2). *Profile*-HMMs são modelos estatísticos de alinhamentos múltiplos de sequências homólogas, os quais capturam informações de posições específicas sobre a conservação de cada coluna do alinhamento, bem como a frequência de inserções e deleções, além do grau de variação permitido para gaps e inserções (CRISTIANINI e HAHN, 2006; FINN, 2011).

Para cada coluna do alinhamento há os estados de inserção e deleção além do estado de *match* (correspondência) e, para cada estado de inserção e *match* há uma probabilidade específica de emitir qualquer um dos 20 aminoácidos existentes com diferentes probabilidades. No entanto, não há probabilidade de emissão associada ao estado de deleção, uma vez que este não emite nenhum símbolo.

Figura 2. Representação de uma *profile-HMM*. I: estado de inserção; M: estado de *match*, ou correspondência; D: estado de deleção.

Usando uma sequência de aminoácidos como exemplo, a inicialização de um HMM ocorre com dados de alinhamento múltiplo fornecidos como entrada. De acordo com um determinado estado no alinhamento múltiplo, a observação é emitida para o símbolo mais provável naquela posição e naquele estado. A troca de um estado para outro pode ocorrer a depender do valor das probabilidades de transição; assim, em um dado momento um novo estado pode ser escolhido. Ou seja, para cada estado há a probabilidade de permanecer nele mesmo ou de mudar para o próximo estado. E em cada estado, existe a probabilidade de emissão de cada símbolo (um nucleotídeo ou um aminoácido).

O alinhamento é percorrido até o final e são identificados os símbolos que aparecem com maior frequência. Assim, ao final, tem-se uma *matriz* de valores de probabilidades que geram um *padrão* de símbolos. Este padrão será utilizado como *isca* para capturar mais sequências que sejam semelhantes ao modelo. Portanto, uma proteína analisada que apresente o padrão, ou a nota correspondente à família, pode ser rotulada como pertencente ou não.

As notas dos HMMs podem ser interpretadas de duas formas: 1. quando HMM é a *query* na busca em um banco de sequências; 2. quando a sequência é a *query* na busca em um banco de HMMs. O primeiro caso corresponde à busca de HMM em um banco de sequências, e haverá uma nota para cada *hit* do banco, que será a soma dos valores de correspondência (*match*) entre a HMM e a sequência. Estas notas são importantes para a definição e refinamento da nota de corte final do HMM. Esta nota final, aplicada ao segundo caso, será a nota apresentada pelo HMM, acima da qual as proteínas que pertencem a família devem estar para serem classificadas como

tal. Assim, ao fazer a busca por uma sequência em um banco de HMMs, o ideal é que a sequência seja classificada em apenas uma família com base na nota obtida, a qual deve ser superior à nota de corte do HMM a que pertence.

A vantagem do modelo consiste em calcular o quanto uma dada sequência se ajusta ao *pHMM* no alinhamento entre uma sequência e o modelo. Isto é, alinhar uma sequência a um dado modelo corresponde a alinhar a mesma sequência às centenas de sequências utilizadas na construção da representação estatística (CRISTIANINI e HAHN, 2006).

2.3 Exemplos de conjuntos de HMMs representando famílias.

2.3.1 PFAM

O PFAM é um banco de dados de famílias de proteínas, as quais são reunidas em famílias conforme o domínio proteico que possuem (PUNTA, 2012). Este banco possui uma divisão entre registros curados, depositados no PFAM-A, e registros gerados automaticamente, depositados no PFAM-B. A versão atual do PFAM contém 14831 famílias no PFAM-A (FINN, 2014).

Cada família existente no PFAM é definida por dois alinhamentos e um HMM, a qual é utilizada na busca por todos os possíveis membros, mesmo os menos relacionados (mais distantes quanto a similaridade de sequência), da família. O HMM é acompanhado por uma nota (*gathering threshold*), a qual é definida pelos curadores da família, e consiste na soma de todos os *scores* das correspondências entre os aminoácidos das sequências e do HMM (*match*). Esta nota é atribuída a cada família, individualmente, e é útil para discriminar os falsos positivos, uma vez que, apenas aqueles que possuírem a nota superior ou igual a nota da família, serão capturados como novos membros (PUNTA, 2012). Do ponto de vista deste projeto, PFAM não oferece os HMMs necessários, visto que esse sistema está organizado em torno de domínios. De uma forma geral, as proteínas nas quais estamos interessados contêm múltiplos domínios, e, portanto cada um deles pertenceria a uma família diferente de PFAM.

2.3.2 TIGRFams

O TIGRFams é um banco de dados de famílias proteicas manualmente curadas. Para cada família há um HMM, alinhamento múltiplo de sequências, comentários, termos da *Gene*

Ontology, referências e outros bancos de dados associados (HAFT *et al.*, 2013). Para a construção dos modelos que representam cada família de proteínas, assim como neste trabalho, a sequência de aminoácidos da proteína inteira é utilizada e são considerados fatores como: a similaridade entre as sequências, evidências de função compartilhada, inferências filogenéticas, alinhamento de sequências, contexto de cada espécie e genes relacionados. TIGRFams tem o conceito de subfamílias, tal como descrito na introdução desta dissertação. Além disso, apresenta o conceito de *equivalogs*. Famílias consideradas como *equivalogs* contém membros cuja função é considerada conservada desde o último ancestral comum. Em outras palavras, são as famílias “bem comportadas” do ponto de vista da relação filogenética entre os membros. É interessante notar que a função pode não ser conhecida; para uma família ser um *equivalog*, basta que a filogenia dos membros obedeça a certos critérios (<http://www.jcvi.org/cgi-bin/tigrfams/Terms.cgi>). O TIGRFams permite que sejam feitas consultas as famílias de proteínas, em que uma sequência de aminoácidos dada como *query* será analisada computacionalmente para cada HMM e família proteica disponível no banco. Assim, como cada HMM gerado possui *trusted-cutoff* e *noise-cutoff*, é possível determinar se tal sequência fornecida para consulta pertence ou não a determinada família (HAFT *et al.*, 2013).

2.3.3 Panther

Outro importante banco de dados é o PANTHER (*Protein Analysis Through Evolutionary Relationships*). A principal meta do PANTHER é a classificação de proteínas de acordo com suas funções moleculares. Pode-se entender este projeto como criando um conjunto de HMMs complementar àqueles disponibilizados por TIGRFams. Como forma de agrupar as proteínas, o PANTHER classifica proteínas relacionadas em famílias, com base em análises de alinhamentos de sequências (MI *et al.*, 2013). Assim como neste trabalho, proteínas-semente servem para capturar outras proteínas aparentadas, utilizando o critério de similaridade para inferir o parentesco entre elas. Inicialmente, é feita uma busca, via BLAST, para capturar os membros dos *clusters* de cada família. Posteriormente, é construído o HMM para recuperar mais membros do *cluster*, resultando em um conjunto de proteínas potencialmente aparentadas. Ao *cluster* final, é atribuído um *score* mínimo, o qual permite a identificação de novos membros, quando novas proteínas são consultadas no banco de dados (THOMAS *et al.*, 2003).

Tanto TIGRFams quanto PANTHER disponibilizam os HMMs de cada família. HMMs do PANTHER foram geradas automaticamente (não possuem nota de corte), e isto faz com que este banco tenha grande cobertura com 1.026.421 genes em famílias contendo árvores filogenéticas, alinhamento múltiplo e HMMs publicados (acesso em maio/2015). Já o TIGRFams apresenta um número mais limitado de modelos, (a versão atual datada de janeiro/2014 conta com 4.424 famílias), uma vez que a criação manual dos alinhamentos demanda intensa atividade curatorial e conhecimento específico para cada proteína de interesse.

2.3.4 dbCAN

O banco dbCAN (*DataBase for automated Carbohydrate-active enzyme ANnotation*) é um recurso que foi desenvolvido com o objetivo principal de permitir que sejam transferidas automaticamente anotações de assinaturas de proteínas específicas (CAZyme - *carbohydrate-active enzymes*) para qualquer conjunto de proteínas (YIN *et al.*, 2012). Este banco de dados mistura informações de função e similaridade de sequência para classificar os registros. CAZymes são proteínas relacionadas a síntese, degradação e modificação de todos os carboidratos existentes, e estão classificadas dentro de famílias no banco de dados CAZy que alimenta o dbCAN. Neste, mais recursos são disponibilizados, como: alinhamentos de sequências, HMMs e filogenias para cada uma das famílias disponíveis no CAZydb, classificações em subfamílias, entre outros. No entanto, esses HMMs não possuem nota de corte associada e não são curados.

2.4 Alternativas a HMMs

Ma e colaboradores propuseram em 2014 um novo modelo para detecção de homologia. Este novo modelo consiste em três componentes principais: 1) representação de uma família de proteínas por Campos Markovianos Aleatórios (*Markov Random Fields* - MRF); 2) um *score* capaz de mensurar a similaridade entre dois MRFs; e 3) um algoritmo capaz de alinhar dois MRFs. Campos markovianos aleatórios podem ser definidos como representações probabilísticas, através de grafos e funções estatísticas, de um alinhamento múltiplo de sequências (MA *et al.*, 2014).

A principal vantagem desta metodologia em relação aos HMMs, segundo o artigo, é a capacidade de modelar padrões de interação dos resíduos em uma proteína, bem como fornecer informações a cerca da estrutura tridimensional da macromolécula. Conforme citado anteriormente, a estrutura tridimensional das proteínas é determinada pela estrutura primária que afetará diretamente as estruturas secundárias e terciárias de um polipeptídeo, sendo esta última a responsável por conferir à proteína uma função biológica. Assim, a adição de informações estruturais à um modelo pode ampliar a capacidade de detectar homologias remotas com maior sensibilidade, uma vez que apesar de divergências na sequência de aminoácidos, o padrão de interação entre resíduos pode ser similar (MA *et al.*, 2014).

Apesar de o artigo demonstrar que esta nova metodologia pode ser mais sensível para detecção de homologias remotas, optamos por manter nossa abordagem utilizando HMMs. Isto porque nossas famílias apresentam características particulares devido à área de interesse: bioenergia. As famílias são, em grande parte, restritas a determinados gêneros, o que pode ser facilmente detectado utilizando-se nossa metodologia. Além disto, existência do pacote HMMER confere maior praticidade na utilização da nossa metodologia. Acreditamos que não estamos deixando de capturar novos membros para as famílias, visto que utilizamos um protocolo de curadoria com critérios bem definidos e fazemos a validação dos modelos com ajustes às notas de corte.

3 MENGO

O aumento na demanda de recursos energéticos, dado o crescimento populacional e a dependência de fontes não renováveis de energia, mais precisamente o petróleo, somado a fatores como poluição, aquecimento global, trazem à tona a necessidade de desenvolvimento de novas tecnologias para agenda organizacional e social (KALIA e PUROHIT, 2008). Neste cenário, surge a geração de energia limpa, ou bioenergia, que consiste na produção de biocombustíveis a partir de degradação da biomassa por microrganismos especializados, com alto interesse biotecnológico. A possibilidade de produção de biocombustíveis em larga escala, torna tais microrganismos cada vez mais foco de estudos, dado que para promoverem as vias metabólicas de degradação, possuem proteínas específicas que atuam nos mais diversos compostos utilizados como matéria-prima, como celulose, hemicelulose, hidrocarbonetos, dentre outros (GIEG et al., 2008).

A expansão dos estudos nesta área de interesse revelou alguns problemas, como a falta de padronização dos termos utilizados para descrever os processos bioenergéticos microbianos na Gene Ontology. Assim, o projeto MENGO (*Microbial ENergy processes Gene Ontology* – ver <http://www.mengo.biochem.vt.edu>) foi concebido como um esforço de aprimorar as informações acerca das proteínas envolvidas no metabolismo de bioenergia microbiana.

Este projeto foi desenvolvido nos Estados Unidos e teve o orientador deste trabalho como um dos participantes. Os objetivos do projeto MENGO foram: (1) gerar novos termos da *Gene Ontology* para processos e funções de genes em procariotos que sejam relevantes para bioenergia e (2) anotar manualmente produtos gênicos de genomas de interesse com esses novos termos. Este conjunto de produtos gênicos manualmente anotados é a anotação-base (proteína-semente) para criação dos nossos modelos. No artigo de Balakrishnan e colaboradores (2013) há a descrição do processo de anotação manual de produtos gênicos, a qual é feita em dois passos principais: (1) a leitura de artigos relevantes para identificação da proteína de interesse, e (2) inferência de função manualmente por análises de motivos, estrutura e relações filogenéticas (BALAKRISHNAN *et al.*, 2013).

Com a missão de gerar termos GO e anotar proteínas associadas com processos microbianos úteis na produção de bioenergia, o projeto MENGO colaborou com a equipe da *Gene Ontology* atuando nos temas:

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 19/02/2015) que visam cobrir os processos citados anteriormente. A lista completa com estes termos pode ser acessada em http://www.mengo.biochem.vt.edu/terms/all_terms.php. Também foram publicados dois artigos referentes ao projeto MENGO (PURWANTINI, *et al.*, 2014 e TORTO-ALALIBO *et al.*, 2014).

4 MENGOfams

O projeto MENGO, embora tenha representado um avanço significativo na geração de novos termos para uso de anotação em bioenergia microbiana, no entanto não disponibilizou modelos computacionais das famílias com as quais trabalhou. Tais modelos são essenciais para que os novos termos GO criados possam ser aproveitados nas anotações de genomas e metagenomas relacionados à bioenergia microbiana. Esta, portanto é a principal motivação deste trabalho. Assim, surgiu o projeto MENGOfams (<http://mengofams.lbi.iq.usp.br/>), o qual foi financiado pela FAPESP (2011/23296-7) e desenvolvido pelo grupo composto por João Carlos Setubal (coordenador), Cláudia Barros Monteiro-Vitorello (pesquisadora associada - ESALQ-USP) e Lucas Mitsuo Taniguti (à época, aluno de mestrado da ESALQ-USP), além da autora. O objetivo geral do projeto MENGOfams é produzir uma coleção de modelos computacionais de famílias de proteínas relacionadas a processos bioenergéticos em procariotos para propagar automaticamente termos do Gene Ontology. De acordo com o proposto, pretendeu-se criar uma coleção de HMMs curados e validados, e disponibilizá-los via web. A técnica escolhida para a construção das famílias tem como base os modelos de Markov com estados ocultos (HMMs).

Dentro do projeto MENGOfams, coube à autora desenvolver os seguintes objetivos:

- 1) Criação de cerca de 80 famílias de proteínas, resultando em cerca de 80 HMMs, por meio de um pipeline semi-automatizado (figura 3);
- 2) Desenvolvimento um protocolo de curadoria manual para ser usado em conjunto com o pipeline;
- 3) Validação dos HMMs criados;
- 4) Disponibilização dos HMMs com nota de corte através de uma página web pública, para que possam ser utilizados por qualquer pesquisador do mundo;
- 5) Criação de uma lista de proteínas cuja anotação deveria ser corrigida com base nas observações feitas durante a criação das famílias de proteínas relevantes em bioenergia de procariotos.

Nas páginas que se seguem, apresentamos a metodologia utilizada para alcançar os objetivos 1 e 2.

4.1 Pipeline para criação e curadoria de um HMM.

Figura 3. Pipeline do trabalho.

I. Proteínas de interesse

As famílias de proteínas são geradas a partir da seleção de uma proteína-semente, que irá guiar a busca por proteínas com a mesma função e que caracterizem uma família gênica. Nesta etapa, é necessário que a proteína-semente seja bem caracterizada, que tenha sido anotada manualmente, possua sua função descrita experimentalmente em publicação científica e que tenha a sequência completa disponível em bancos de dados públicos, como UniprotKB (APWEILER, *et al.*, 2004). Faz parte desta etapa a identificação de dados públicos relativos a cada proteína-semente. Entre esses dados podemos citar HMMs de PFAM (FINN *et al.*, 2014), HMMs do *website* PANTHER (MI, 2013), HMMs do projeto Phylofacts (KRISHNAMURTHY *et al.*, 2006), alinhamentos múltiplos do projeto HAMAP (PEDRUZZI, 2015), e HMMs de TIGRFams (HAFT *et al.*, 2013). Quando existirem, tais dados foram analisados para verificação de sua utilidade para o projeto ou

possível redundância com o que se pretende fazer. Na ausência de modelos computacionais já existentes, deve-se fazer um levantamento bibliográfico sobre a proteína-semente:

1. Contexto biológico que a proteína atua.
2. Em qual organismo a proteína foi identificada.
3. Componentes da proteína (domínios, peptídeo-sinal, regiões de ligação, regiões conservadas, propriedades químicas dos aminoácidos, tamanho da sequência).

Estas informações servirão de base para os próximos passos, por isso devem ser cuidadosamente analisadas. Outros recursos podem ser analisados a fim de se obter informações que não estejam presentes na literatura, como bancos de dados especializados e de vias metabólicas.

II. Busca via BLAST

De posse das informações sobre a proteína de interesse, esta é submetida a uma busca via BLAST. Neste protocolo, recomendamos a utilização do banco de dados UniprotKB para realização do BLAST. Os critérios citados referentes ao percentual de similaridade, cobertura e ao *e-value* visam coletar membros adicionais da família, mas minimizando a possibilidade de falsos positivos (sequências similares à sequência-semente, mas que não fazem parte da mesma família). Nesta etapa não estamos preocupados com falsos negativos (sequências de proteínas que fazem parte da família, mas que não são identificados pelos critérios de busca).

O resultado do BLAST deve ser analisado de forma acurada, uma vez que as proteínas selecionadas integrarão o modelo que deve melhor representar a família. Nesta etapa, deve-se analisar:

1. Porcentual de identidade.
2. Porcentual de cobertura.
3. Alinhamento em relação à sequência *query*.
4. Informações contidas no UniprotKB sobre a proteína apresentada no resultado (verificar se são informações compatíveis com as que foram anotadas para a sequência-semente).

5. Diversidade taxonômica: observar se as proteínas são em maior número de uma mesma espécie (o que caracteriza uma família restrita), ou se há outras espécies relacionadas.

Baixos valores de identidade não devem ser um problema, uma vez que há maior flexibilidade no modelo computacional por representar aquilo que é essencial para a manutenção da mesma função entre as proteínas que compõem uma mesma família. Já a diferença entre os tamanhos da proteína-semente e da proteína resultante do BLAST indica que a proteína resultante deverá ser descartada. Isto porque, provavelmente, houve o alinhamento de apenas uma região da proteína, que pode ter sido conservada ao longo da escala evolutiva e estar presente em diversos organismos que não necessariamente apresentam a mesma via. Através do alinhamento entre a *query* e a sequência recuperada via BLAST também se pode fazer tal verificação. É possível identificar quais as regiões são conservadas entre elas e qual o papel desta região na função proteica, lembrando que regiões cruciais para a função devem ser menos variáveis que o restante da proteína.

III. Alinhamento múltiplo de sequências

Ao final do passo anterior, são selecionadas as sequências que devem integrar o modelo computacional. Nesta etapa consultamos a literatura, para verificar informações relevantes que precisamos incorporar no processo. Um exemplo de informação relevante é a descrição de domínios que sejam necessários para a proteína-semente realizar sua função. Tais domínios precisam estar presentes e na mesma ordem em todos os membros da família. Se necessário, o alinhamento múltiplo é editado, principalmente para excluir sequências que se julguem serem falsos positivos, que podem ter sido recuperadas apesar dos critérios estritos de busca descritos no passo anterior. O alinhamento também é editado para excluir sequências redundantes e posições não informativas. Caso haja alguma discrepância neste alinhamento as proteínas são reanalisadas e, com base nisto, toma-se a decisão de manter ou não a proteína como integrante do modelo que será construído. Neste passo, é necessário que sejam observadas:

1. Sequências idênticas.
2. Presença de sequências que distorcem o alinhamento.
3. Existência de todos os domínios essenciais para a caracterização da proteína.

Sequências idênticas enviesam o alinhamento porque restringem o modelo, uma vez que o padrão apresentado por elas será mais representativo em relação à amostra. As sequências devem apresentar certa variação, respeitando os critérios estabelecidos nos passos iniciais, assim, um maior número de sequências poderá ser capturado pelo modelo. Sequências que distorcem o alinhamento são aquelas em que as regiões conservadas não são claramente identificadas. Tais sequências devem ser excluídas caso não haja evidência suficiente para mantê-las como integrantes do modelo. O alinhamento múltiplo é o *input* para gerar o modelo computacional, assim, conforme são necessárias as iterações, este passo é retomado. O alinhador que utilizamos foi o MUSCLE 3.7 (EDGAR, 2004), e os parâmetros são *default* e devem ser fornecidos a extensão do arquivo (- clw), o arquivo de entrada (- in "ArquivoMultifasta") e de saída (- out "ArquivoDeSaída").

IV. Árvore filogenética

O uso de árvores filogenéticas auxilia a interpretação de famílias que não estão bem resolvidas. Neste passo, fazemos uma etapa de curadoria para verificar se a família que pretendemos descrever é *simples* ou *complexa*. Uma família complexa é aquela que é constituída de várias subfamílias. A identificação da presença de subfamílias é uma atividade com a qual ainda não temos muita experiência, mas sabemos pela literatura que ocorre com alguma frequência. Quando tal situação ocorrer, em princípio iremos fragmentar a família em suas subfamílias constituintes, gerando diferentes HMMs para cada subfamília. O texto daqui por diante supõe que a família em estudo é simples, portanto não contendo subfamílias.

V. HMMs e busca de novos membros

As proteínas que serão reconhecidas pelo HMM dependerão muito da arquitetura do modelo. De acordo com as proteínas selecionadas para o alinhamento múltiplo e geração do modelo, as pontuações atribuídas serão diferentes. Assim, além dos parâmetros do próprio programa para construção do HMM, as proteínas selecionadas devem ser informativas o suficiente para capturar o que há de essencial na família. Após esta etapa, o HMM é gerado através do programa HMMbuild do pacote HMMER (EDDY, 1998). Nesta etapa, o arquivo de entrada é o próprio alinhamento múltiplo gerado no passo anterior e o arquivo de saída corresponde ao HMM com as probabilidades associadas ao aminoácido de cada posição da sequência.

Com o modelo em mãos, é feita a busca contra o banco do UniprotKB (Swiss-Prot + TrEMBL), através do HMMsearch do pacote HMMER (EDGAR, 2004) e o HMM é utilizado como *query*. Os parâmetros usados são *default* do programa, o arquivo de entrada é o próprio modelo computacional e os arquivos de saída correspondem a tabelas que mostram os resultados encontrados no Swiss-Prot e os resultados encontrados no TrEMBL com as pontuações obtidas pelas proteínas quando comparadas ao HMM. Nosso protocolo inclui a geração de um gráfico com a distribuição dos *nota de cortes* e uma planilha com os possíveis novos membros, que são aqueles com pontuação superior à nota de corte preliminar da família.

VI. Curadoria e definição da nota de corte

A construção das famílias está associada à produção de um modelo computacional, HMM, que por sua vez, possui um nota de corte, conforme já explicado no capítulo 1. Para cada família há uma nota de corte que permite classificar as proteínas que obtiverem pontuações igual ou superiores a tal valor naquela família, enquanto que as proteínas com pontuação abaixo deste valor podem pertencer ou não.

Para cada família construída, são necessárias iterações da busca com o HMM daquela etapa, com o objetivo de garantir que a nota de corte consiga separar bem os membros da família dos não-membros. A atividade curatorial existe durante todo o processo, mas neste passo é fundamental o olhar do curador para tomar decisões a respeito dos integrantes do modelo e membros das famílias. No resultado do HMMsearch, é necessário analisar:

1. A pontuação das proteínas que integraram o modelo: tais proteínas devem apresentar notas aproximadas. Caso alguma proteína receba uma nota muito abaixo das demais proteínas do modelo, esta deve ser reavaliada e a presença dos domínios selecionados no passo três deve ser confirmada.
2. A pontuação das proteínas que não integram o modelo, mas que são membros da família. Verificar se as proteínas que receberam notas acima da nota de corte são verdadeiros ou falsos positivos.

3. A pontuação das proteínas que não integram o modelo, e que **não** são membros da família. Verificar se as proteínas que receberam notas abaixo da nota de corte são verdadeiros ou falsos negativos.

4. Distribuição taxonômica da família. É interessante que haja o maior número de diferentes táxons compondo uma família, assim esta será capaz de capturar membros mais distantemente relacionados, mas que mantém a função.

Se por um lado as famílias restritas são difíceis de trabalhar por causa do pequeno número de representantes, as famílias que são muito amplas tornam difícil a atribuição da nota de corte, já que este valor deverá ser capaz de não deixar nenhum membro de fora, ou seja, não deve contemplar falsos positivos, nem permitir falsos negativos. No caso das famílias restritas, isso ocorre pelo fato de que realmente não existe outra proteína com a mesma função em outros organismos (o que torna a família restrita por espécie), ou por tal função não ter sido estudada em outros organismos. No caso das famílias mais amplas, as proteínas que não possuem evidências para serem membros da família são eliminadas das análises, mas são utilizadas na definição da nota de corte como uma forma de limite, ou seja, a pontuação que esta sequência receber será um indicativo de que a nota de corte deve ser acima daquele valor.

5. Possíveis novos membros. Aquelas proteínas que possuem *score* maior que a pior pontuação de um integrante do modelo são ditas "novos membros". Estas proteínas são analisadas na tentativa de encontrar evidências que justifiquem que são membros da família. Muitas vezes não há literatura associada a estas sequências, no entanto, ao fazer o alinhamento, deve ser possível identificar as regiões conservadas conhecidas da proteína-semente nestes polipeptídios identificados pelo HMM. Assim, são tomadas as decisões de incluir ou não tais membros no modelo, tornando-os integrantes do HMM ou apenas supor que o *score* está capturando todos os membros para a família em estudo. Não incluí-las como integrantes pode ser mais indicado se todas ou a maior parte das sequências pertencem a uma mesma espécie, o que poderia ser um viés no modelo.

6. Definição da nota de corte. Uma vez definida a nota de corte, é necessário analisar as proteínas que apresentaram pontuação abaixo do valor necessário para pertencer à família. Esta análise é mais um passo para garantir que membros não sejam excluídos. Em alguns casos é fácil separar

os não-membros por causa da existência de um "degrau" nos *scores*, em que a diferença na pontuação é tal que dificilmente a proteína possui a mesma função. Em outros casos, a proteína pode cair numa zona cinza, assim chamada por se tratar de uma região “nebulosa” em relação à certeza de pertencer ou não à família. A existência da zona cinza, conforme citado anteriormente, permite que membros de uma família complexa, por exemplo, sejam acomodados dentro da família sem que a nota de corte se altere. Quanto melhor for o processo de curadoria do modelo, mais útil ele será, e menor será o risco de ter anotações inadequadas para novas sequências.

4.2 Comparação com a literatura.

Nosso modelo de trabalho assemelha-se ao modelo utilizado por outros grupos, como TIGRFAMs. Tanto estes quanto o projeto MENGOFams utilizam um pipeline que se baseia na utilização de alinhamentos múltiplos para geração das famílias representadas por HMMs. A diferença da nossa proposta reside no objetivo principal da construção das famílias que é a propagação dos novos termos da *Gene Ontology* gerados pelos curadores do MENGO. As famílias criadas são específicas para tais termos, podendo reunir mais de uma função em um mesmo modelo a depender do termo propagado.

Outra diferença é que alguns detalhamentos do processo de criação das HMMs não foram documentados por nenhum outro grupo. Estes detalhamentos referem-se a um protocolo para curadoria manual de HMMs, que explore aspectos que permitam a construção de profile-HMMs manualmente curados e que tenham uma nota de corte associada. Abaixo se tem um trecho retirado do próprio artigo do TIGRFAMs, onde são citadas as fases de criação de uma família:

“The TIGRFAMs database provides manually reviewed definitions of protein families with the following characteristics to make them useful for automated genome annotation, pathway reconstruction, creation of phylogenetic profiles and any number of other computationally driven studies. First, all annotation is traced to its original source without ever relying on transitive annotations from public sequence databases. Second, every protein is considered in the context of any related protein families that differ in function, if they exist, with examination of molecular phylogenetic trees. Third, only sequences that can be assigned with high confidence are selected as exemplars for the seed alignment. Multiple sequence alignments are examined for

misalignments, inconsistent domain architecture, altered active or binding sites, faulty gene models, long branches in phylogenetic trees that may suggest neofunctionalization, etc. Biocuration during model construction includes review of local synteny, metabolic context and phenotypic data. Fourth, each model is searched against several protein databases, including the CharProtDB collection of explicitly characterized proteins (1), sequences with known structures in PDB and NCBI's non-redundant protein sequence collection, which pulls annotation from multiple sources, including value-added databases such as UniProt and archives of sequences as originally submitted. Each completed HMM is intended to serve as a proxy for an expert curator, emulating expertise developed at the time of model construction but operating at BLAST-like speeds."

Neste parágrafo, retirado do artigo publicado por Haft e colaboradores em 2013, há uma descrição da metodologia aplicada por curadores do TIGRFAMs. Não há detalhes sobre o processo de criação dos modelos de forma a serem reproduzidos por outros pesquisadores. Não são citadas as ferramentas utilizadas, o que deve ser observado para a construção do alinhamento e os passos para definição da nota de corte. Além disto, não é mencionada a validação dos modelos criados.

5 Resultados

5.1 Conjunto de HMMs.

No total foram 74 modelos manualmente curados. Estes HMMs estão disponibilizadas no site: mengofams.lbi.iq.usp.br. Para cada uma destas famílias (apêndice C), além do HMM, há o alinhamento múltiplo e um gráfico contendo os *scores* apresentados pelos membros da família e os não-membros. Estas 74 famílias passaram por um processo de validação para garantir que a nota de corte fosse boa o suficiente para diminuir os problemas de falsos positivos e falsos negativos.

5.2 Validação.

O esquema de validação utilizado foi por meio de buscas com cada HMM gerado contra o banco NR do NCBI. Escolhemos o NR porque este banco apresenta sequências adicionais em relação ao UniprotKB. No passado, estes bancos eram sincronizados, mas hoje algumas sequências como: GI: 498188238, são encontradas apenas no NR.

Nessas buscas verificamos se as notas de corte atribuídas a cada HMM estão tendo um bom desempenho. Isto pode ser feito verificando possíveis falsos positivos (sequências com nota acima da nota de corte, mas que não pertencem à família, por avaliação manual) e falsos negativos (sequências com nota abaixo da nota de corte, mas que pertencem à família, por avaliação manual). Na Tabela 1 mostramos um resumo da validação realizada.

Tabela 1. Validação das notas de corte dos modelos. Aqui salientamos os casos em que os números de falsos positivos e/ou negativos forem considerados excessivamente altos.

MENGOfams	Total
Disponíveis no site	74
Validadas <u>com</u> revisão	10
Validadas <u>sem</u> revisão	64

Dentre as famílias que necessitaram de revisão, utilizaremos como exemplo a MENGOfam-009, *Extracellular solute-binding protein family 1*. Esta família propaga o termo *glucose binding*

(GO:0005536). Inicialmente (primeira iteração do passo VI do pipeline descrito no capítulo 4), a nota de corte desta família foi de 513.1.

O resultado da utilização do modelo para buscas no NR mostrou a presença de algumas proteínas com a mesma anotação e alinhamento completo em relação ao modelo, abaixo da nota de corte (figura 4). Assim, as proteínas que obtiveram nota menor que a nota de corte, porém se alinharam ao modelo, foram analisadas.

Significant Query Matches (20469)						Customize
Target	Description	Species	# significant hits	Bit Score	Hit Positions	E-value
> 493236739	ABC transporter substrate-binding protein	Paenibacillus vortex	1	650.3	=====	8.8e-192
> 573543679	extracellular solute-binding protein	Paenibacillus sp. FSL R5-808	1	649.9	=====	1.2e-191
> 494215275	ABC transporter substrate-binding protein	Paenibacillus lactis	1	646.8	=====	1.0e-190
> 511072830	hypothetical protein	Paenibacillus barengoltzii	1	627.1	=====	9.8e-185
> 57355864	extracellular solute-binding protein	Paenibacillus sp. FSL R7-269	1	622.5	=====	2.4e-183
> 516294966	ABC transporter substrate-binding protein	Paenibacillus sp. PAMC 26794	1	611.1	=====	7.3e-180
> 544839325	extracellular solute-binding protein	Paenibacillus alvei	1	601.4	=====	6.4e-177
> 544844915	extracellular solute-binding protein	Paenibacillus alvei	1	601.3	=====	6.7e-177
> 489612259	ABC transporter substrate-binding protein	Clostridium thermocellum	1	594.9	=====	5.8e-175
> 49609799	ABC transporter substrate-binding protein	Clostridium thermocellum	1	592.0	=====	4.4e-174
> 573578759	extracellular solute-binding protein	Paenibacillus sp. FSL R7-277	1	573.0	=====	2.7e-168
> 648229839	ABC transporter substrate-binding protein	Paenibacillus sp. OSY-SE	1	570.4	=====	1.6e-167
> 517589379	hypothetical protein	Paenibacillus terrigena	1	552.7	=====	3.8e-162
> 491696141	extracellular solute-binding protein family 1	Paenibacillus alvei	1	546.7	=====	2.5e-160
> 64768347	ABC transporter substrate-binding protein	Paenibacillus ehimensis	1	535.0	=====	9.2e-157
> 497954094	ABC transporter substrate-binding protein	Paenibacillus senegalensis	1	529.4	=====	4.4e-155
> 546656225	extracellular solute-binding protein family 1	Ruminococcus sp. CAG-108	1	513.1	=====	3.9e-150
> 498188238	extracellular solute-binding protein family 1	Paenibacillus elgii	1	511.5	=====	1.2e-149
> 517588084	hypothetical protein	Paenibacillus terrigena	1	467.5	=====	2.8e-136
> 522107039	hypothetical protein	Paenibacillus daejeonensis	1	437.8	=====	2.7e-127
> 647572121	ABC transporter substrate-binding protein	Paenibacillus ehimensis	1	390.7	=====	5.7e-113
> 491694301	ABC transporter substrate-binding protein	Paenibacillus alvei	1	379.9	=====	1.0e-109
> 544845718	extracellular solute-binding protein	Paenibacillus alvei	1	376.6	=====	1.1e-108
> 49372807	ABC transporter substrate-binding protein	Paenibacillus dendritiformis	1	376.4	=====	1.2e-108
> 544838586	extracellular solute-binding protein	Paenibacillus alvei	1	371.2	=====	4.5e-107
> 517587221	hypothetical protein	Paenibacillus terrigena	1	353.0	=====	1.6e-101
> 488122071	ABC transporter substrate-binding protein	Bacillus cereus	1	346.6	=====	1.4e-99
> 507062513	hypothetical protein	Bacillus cereus	1	346.2	=====	1.8e-99
> 498185145	ABC transporter substrate-binding protein	Paenibacillus elgii	1	345.1	=====	3.8e-99
> 507031903	hypothetical protein	Bacillus cereus	1	344.2	=====	7.2e-99
> 653620719	hypothetical protein	Clostridium cellobioparum	2	344.0	=====	8.4e-99
> 517612943	ABC transporter substrate-binding protein	Bacillus sp. 95MFCv12.1	1	343.5	=====	1.2e-98
> 647574300	ABC transporter substrate-binding protein	Paenibacillus ehimensis	1	341.7	=====	4.1e-98
> 488116280	ABC transporter substrate-binding protein	Bacillus cereus	1	339.7	=====	1.7e-97
> 490763390	ABC-type sugar transport system, periplasmic component	Clostridium termitidis	2	338.7	=====	3.3e-97

Figura 4. Resultado da busca no NR (*Immsearch*) utilizando a família MENGOfam-009 como isca. Abaixo da nota de corte (destacada em verde) existem seqüências se alinham completamente ao HMM e estas foram analisadas.

Inicialmente, analisamos as seguintes proteínas: GI: 507062513 – hypothetical protein [Bacillus cereus], GI: 517596770 – ABC transporter substrate-binding protein [Bacillus sp.], GI: 498188238 – extracellular solute binding protein family 1 [Paenibacillus elgii]. Encontramos evidências de que essas proteínas pertencem à família, tais como: presença do domínio SBP - Sugar-binding protein (PF01547) e a presença de peptídeo sinal de aproximadamente 26 aminoácidos no início da sequência nas três proteínas (figura 5) .

Figura 5. Resultado da predição da região de peptídeo sinal feita pelo SignalP 4.1 (PETERSEN *et al.*, 2011) para as três proteínas seleccionadas para compor o modelo.

A partir destas análises, estas proteínas recuperadas do NR foram incluídas no modelo. Após a atualização da família, o modelo foi submetido novamente à validação. Este processo se repetiu três vezes, onde outras proteínas foram incluídas porque foram consideradas falsos negativos. Os passos citados anteriormente foram repetidos para as proteínas incluídas. Ao final da última iteração, definimos a nova nota de corte da família como sendo 452.1. Proteínas (representadas por pontos azuis no gráfico) presentes apenas na coluna *Model members*, foram incluídas no modelo durante a validação e estão presentes apenas no NR – GenBank. Outras proteínas podem ser vistas apenas na coluna *Uniprot db* e não na coluna *Model members*, ou seja, as proteínas pertencem à família, porém não integram o modelo (figura 6). Abaixo da nota, as proteínas se alinham em partes ao HMM, não contemplando os domínios importantes para a função da proteína, como pode ser verificado através do alinhamento da proteína-semente com uma proteína abaixo da nota de corte (figura 7).

Figura 6. Gráfico com a distribuição de *scores* das proteínas pertencentes à família MENOfam-009 (acima da nota de corte 452,1) e não pertencentes (abaixo da nota de corte) após a validação. Este gráfico foi gerado utilizando-se o modelo contra o banco de dados do UniprotKB.

corresponde à proteína descrita para o microrganismo *Bacteroides thetaiotaomicron*, cuja sequência completa possui 617 aminoácidos.

Os termos MANGO associados a este produto gênico são: *neopullulanase activity* (GO:0031216) e *plasma membrane* (GO:0005886). O primeiro termo corresponde à função molecular de catálise da hidrólise de pululano a panose. Pululano é um polissacarídeo composto por unidades de maltotriose e relacionado ao metabolismo de amido. O segundo termo corresponde ao componente celular, membrana plasmática (D'ELIA e SALYERS, 1996).

A proteína-semente passou por um filtro inicial, no qual procuramos em outros recursos como TIGRfams, HAMAP e PANTHER se havia alguma família já estabelecida. Foi encontrado um registro no PANTHER: PTHR10357. Como o nosso objetivo é a propagação de um termo específico através de uma família manualmente curada (diferentemente do PANTHER, conforme citado em seções anteriores), seguimos o estudo da proteína-semente fazendo um levantamento sobre as informações mais relevantes para a recuperação de membros que constituam a família proteica. Este levantamento inclui a caracterização da proteína em termos de sequência: a SusA possui 617 aminoácidos; função: possui atividade catalítica de hidrólise de ligações do amido em di- ou tri-sacarídeos no processo de degradação de amido; sítio ativo: posições Asp-331, Glu-360 e Asp-440 da cadeia de aminoácidos; regiões de ligação a cálcio: 138, 143, 144, 164 e 166 e uma região peptídeo sinal entre os aminoácidos das posições 1-22 (figura 8). A presença do peptídeo sinal, neste caso, é fundamental já que a proteína é destinada à membrana, ou seja, proteínas similares sem peptídeo sinal não são incorporadas a família. Esta proteína-semente possui três domínios principais: PF00128 -Alpha-amylase PF10438 - Cyc-maltodext_C PF09087 - Cyc-maltodext_N.

10	20	30	40	50
MKRNLFIIL	LLLLPGLHQV	FATSTIKKVA	PTFWWAGMKN	PELQILLYGD
60	70	80	90	100
RISSADVLS	ADNITLQEVV	KQENPNYLVL	YDLLSKAAPQ	NFDIILKQ GK
110	120	130	140	150
KQTKIPYELK	QRRPNASAVE	GFDSSDVLYL	IMPDRFANGN	PSNDIIPGML
160	170	180	190	200
EGNVDRNEPF	ARHGGDLKGI	ENHLDYIADL	GVTSIWLNP I	QENDMKEGSY
210	220	230	240	250
HGYAITDYYQ	VDRRFGSNEE	FRKLTQEANA	KGLKVVMDMI	FNHCGSDNYL
260	270	280	290	300
FKDMPSKDF	NFEGNYVQTS	FKTATQMDPY	ASDYEEKIAI	DGWFTLTMPD
310	320	330	340	350
FNQQRNRHVAT	YLIQSSIWI	EYAGINGIRQ	DTHPYADFDM	MARWCKAVNE
360	370	380	390	400
EYPKENIVGE	TWLGNNVLIS	YWQKDSRLAY	PKNSNLP TVM	DFPLMEEMNK
410	420	430	440	450
AFDEETTEWN	GGLFRLYEYL	SQDIVYSHPM	SLLTFLDNHD	TSRFYRSEAD
460	470	480	490	500
TKNLDYKQA	LTFLLTTRGI	PQIYYGTEIL	MAADKANGDG	LLRCDFPGGW
510	520	530	540	550
PNDTKNCFDA	ANRTPQQNEA	FSFMQKLLQW	RKGNEVIAKG	QLKHFAPNKG
560	570	580	590	600
VYVYERKYGD	KSVVVFNLGN	DREQTIDLVP	YQEILPASSA	FDLLTEKKVE
610				
LRNELTLPSR	EIYLLSF			

Figura 8. Representação das posições importantes para a proteína-semente. A região de peptídeo-sinal: quadro lilás; sítios de ligação ao cálcio: amarelo; sítios ativos: rosa, verde e azul.

Outra caracterização corresponde ao levantamento de informações que corroborem a função descrita. *Bacteroides thetaiotaomicron* pode utilizar três formas de amido, sendo o pululano uma delas. Dentre as principais enzimas de degradação do amido, tem-se a *neopullulanase* (EC 3.2.1.135). Estas proteínas são similares à alfa-amilases (EC 3.2.1.1) encontradas em *Clostridium thermocellum*, no entanto, diferem no mecanismo de ação, pois quebram ligações em diferentes substratos.

Após a coleta de informações, foram feitas as buscas através do *blastp* utilizando esta proteína-semente como *query* no banco de dados do UniprotKB. A partir do resultado do *blastp*, foram selecionadas quatro proteínas, do mesmo gênero que a nossa *query*, que atendiam aos critérios estabelecidos. Selecionamos apenas estas sequências porque partimos de um modelo mais restrito, com maior confiabilidade de que as sequências são verdadeiramente pertencentes à família, para então afrouxar os parâmetros de acordo com o conhecimento adquirido com as

iterações. Estas proteínas foram utilizadas na geração de um primeiro alinhamento múltiplo, bastante conservado devido a alta similaridade entre as sequências. A partir do alinhamento chegamos ao HMM, o qual foi utilizado como *query* no banco do UniprotKB (TrEMBL + Swiss-Prot). Nesta primeira versão, a maior pontuação foi 1631.4 e a menor nota de uma proteína integrante do modelo foi 1478.8; no entanto esta família estava restrita a um único gênero.

Na segunda iteração fomos menos conservadores na escolha das sequências e incluímos dois registros de proteínas identificadas em amostras de metagenoma de estômago. Na análise do alinhamento múltiplo, foram identificados blocos conservados em regiões importantes da sequência proteica, como regiões de ligação ao cálcio e regiões de sítio ativo, na ponta N-terminal. A menor nota de uma sequência neste modelo foi de 1225.3, mas algumas proteínas abaixo desta nota mereceram ser analisadas devido a evidências de apresentarem a mesma função que representa a família em estudo como: a presença de uma região de peptídeo-sinal, os mesmos domínios apresentados pela proteína-semente (confirmados através de buscas no PFAM), os quais foram confirmados em análises do alinhamento múltiplo.

Na terceira iteração incluímos estes registros (UniprotKB ids:W4PIN3; F8WZF8) que ficaram abaixo da nossa nota de corte e montamos um novo alinhamento múltiplo para ser analisado (figura 9). Este modelo apresentou como menor pontuação o valor 1182.7, que foi adotado como nova nota de corte. No entanto, analisando as proteínas abaixo desta pontuação, fizemos uma nova iteração incluindo uma sequência (UniprotKB id: W4UQW7). Esta proteína, na quarta iteração apresentou pontuação de 1029.6. A sequência seguinte na lista de hits teve pontuação de 988.5.

Figura 9. Alinhamento múltiplo da terceira iteração do modelo para a família.

A análise do quarto alinhamento revelou a alta conservação dos blocos principais para manutenção da função da proteína. Além disto, apesar da restrição inicial ao gênero *Bacteroides* (figura 10), o modelo gerado conseguiu capturar sequências dos gêneros *Dysgonomonas* e *Parabacteroides*, além de sequências provenientes de amostras de metagenomas. Assim, ao final das análises deste modelo, definimos a nota de corte desta família: 1069.2 (figura 11).

Figura 10. Gráfico com a distribuição de táxons das proteínas pertencentes à família.

Figura 11. Gráfico com a distribuição de *scores* das proteínas pertencentes à família (acima da nota de corte 1069.2) e não pertencentes (abaixo da nota de corte).

Endoglucanase – MENOOfam-001

Outro exemplo, mas agora de uma família mais complexa é o estudo da família Endoglucanase (F, N, Z). Esta família é formada por três endoglucanases, F, N e Z, diferentemente da maior parte das outras famílias, que é constituída por apenas uma proteína-semente. O termo MENO a ser propagado é: *cellulase activity* (GO:0008810).

Endoglucanases são enzimas envolvidas no processo de degradação de matéria orgânica, principalmente de celulose, resultando em biocombustíveis (YENNAMALLI *et al.*, 2013). Esta conversão é mediada por celulasas, enzimas específicas que compõem o celulosomo (figura 12), estrutura complexa presente em alguns procariotos, responsável pela degradação de celulose. Dentre as proteínas associadas ao celulosomo, nem todas possuem atividade de celulase, algumas possuem função de ligação e outras atuam como “estabilizadoras”, garantindo a manutenção estrutural do complexo proteico (FONTES e GILBERT, 2010).

Figura 12. Representação do complexo enzimático celulosomo. Adaptado de Fontes e Gilbert, 2010.

Ao iniciarmos o processo de construção da família, observamos que nos resultados de *blastp* da endoglucanase F apareciam endoglucanase N e Z, idem para as outras, indicando alta similaridade entre as mesmas. Assim, investigamos a função de cada uma destas proteínas e verificamos que as três possuem atividade catalítica na quebra de ligações específicas no processo de degradação de celulose. Além disto, apresentam similaridade entre as sequências, contendo os mesmos domínios (CBM3, GH9 e doquerina) e uma região de peptídeo sinal

aproximadamente entre as posições 1-38. Assim, ao invés de continuarmos com as três famílias tratadas separadamente, optamos por testar o *pipeline* utilizando as três sequências como *query*.

Para montar o alinhamento múltiplo, utilizamos sequências das três proteínas semente e as sequências provenientes do resultado da busca através do *blastp* para cada uma separadamente. Uma vez que as famílias já existiam separadamente, fizemos a seleção de algumas sequências que apareceram como pertencentes a cada uma das famílias. Nesta etapa, seguimos a recomendação do protocolo sobre os critérios para serem selecionadas. Assim, escolhemos àquelas com os domínios necessários para executar a função de celulase, a região de peptídeo-sinal e priorizamos a diversidade de espécies.

Conseguimos reunir as três proteínas-semente e outras proteínas com a mesma função em um único alinhamento. A primeira iteração do modelo teve como nota de corte estabelecida 1119.6 (figura 13). Logo abaixo desta nota, algumas proteínas possuíam pontuações muito próximas (1118.6, por exemplo) e evidência de pertencerem à família, como os mesmos domínios CBM3 e GH9.

Figura 13. Gráficos de distribuição taxonômica (esquerda) e gráfico com a distribuição de *scores* (direita) das proteínas pertencentes à família (acima da nota de corte 1119.6) e não pertencentes (abaixo da nota de corte).

Analisamos as proteínas abaixo da nota de corte. Concluimos que tais proteínas não contemplavam todos os critérios para pertencerem à família. Em alguns casos diferiam na composição dos domínios, seja pela classificação diferente (GH5 ao invés de GH9, por exemplo) ou pela falta do domínio doquerina. Sem este domínio, a enzima não consegue se ligar ao celulosomo, dado que esta ocorre por meio da interação entre o domínio doquerina da enzima e o domínio coesina do complexo. Consequentemente, a função proteica esperada para esta família estará comprometida.

O mecanismo de degradação através de celulosomo até então é descrito como exclusivo de Clostridia, por isto, a restrição taxonômica a este gênero. Devido aos estudos prévios, com as proteínas-semente separadas, conseguimos chegar à nota de corte com uma única iteração.

5.4 Proteínas candidatas a serem reanotadas.

Sabe-se que há muitos erros de anotação ou anotações incompletas nos bancos de dados de sequências. Ao utilizar HMMs para encontrar novos membros para cada família, nos deparamos com proteínas cuja anotação deve ser modificada. A Tabela 2 mostra alguns exemplos, como no caso da MENOfam011. Esta família está descrita como Endo-1,3-beta-xylanase C e os membros desta família possuem um domínio GH10 (*glycoside hydrolase* 10), que pode ter originado a anotação incompleta de GI:48960934.

Outro exemplo é o caso da MENOfam064. Esta família corresponde ao Elemento P do celulosomo, e conforme descrito na literatura, os membros desta família possuem um domínio doquerina. Isto sugere que GI: 489613942 foi anotada apenas pela presença deste domínio. Assim, a anotação da MENOfam064 seria mais precisa.

Idealmente a equipe do NCBI deveria ser alertada para que modifiquem essas anotações. Infelizmente, pelo alto volume de dados que essa equipe tem que gerir, acreditamos que não há mecanismos para fazer essas correções.

Tabela 2. Proteínas candidatas a serem reanotadas com base na avaliação a partir das notas de corte atribuídas a cada família.

Família	Proposta de reanotação	Anotação original
MENGOfam001	Endoglucanase (F, G, N, Z)	GI:270288699 hypothetical protein
MENGOfam002	Extracellular solute-binding protein family 1	GI:575089985 hypothetical protein
MENGOfam004	Endoglucanase D	GI:40672 unnamed protein product
MENGOfam005	Sugar ABC transporter	GI:317389231 hypothetical protein
MENGOfam011	Endo-1,3-beta-xylanase C	GI:48960934 glycoside hydrolase
MENGOfam015	Endo-1,4-beta-xylanase Z	GI:506405016 sugar-binding protein
MENGOfam045	Cellulose 1,4-beta-cellobiosidase	GI:545459126 chitinase
MENGOfam064	Cellulosomal element P	GI:489613942 dockerin GI: 490768495 spore coat assembly protein
MENGOfam072	Endo-1,4-beta-xylanase Y	GI:640310871 glycoside hydrolase
MENGOfam077	Endoglucanase A	GI:575088163 cellulase
MENGOfam079	Endoglucanase Y	GI: 489607585 cellulose 1,4-beta-cellobiosidase
MENGOfam082	Arabinoxylan-specific endo- β -1,4-xylanase	GI:489609288 carbohydrate-binding protein

5.5 Aplicação dos modelos em dados de metagenômica.

Fizemos testes com nossos HMMs em dados de metagenômica na tentativa de identificar possíveis membros das famílias estudadas nestas amostras e classificar tais sequências através de uma metodologia confiável. Utilizamos os seguintes conjuntos de dados públicos: compostagem do Zoológico de São Paulo (MARTINS *et al.*, 2013), biorreator de degradação de benzeno (IMG ID Submission: 1537), reator de biocombustíveis (LI *et al.*, 2009), mini-reator de compostagem

(ALLGAIER *et al.*, 2010), sedimento de lago (KALYUZHNYA *et al.*, 2008), solo da floresta tropical (DeANGELIS *et al.*, 2010), intestino de cupim (WARNECKE *et al.*, 2007), biorreator de biomassa de álamo (*Populus nigra*) (van der LELIE *et al.*, 2012).

HMMs para famílias de proteínas capturam todas as posições dos membros da família. Isto faz com que a utilização da nota de corte privilegie os hits que tenham sequências completas ou próximas de completas, em relação ao comprimento modelado pelo HMM. Ao se comparar nossos modelos com os dados públicos de metagenômica, quase sempre verificamos que os genes preditos em metagenomas são parciais, provavelmente devido à baixa cobertura do sequenciamento em relação à diversidade existente na amostra. No caso dos dados de metagenômica que utilizamos a maioria dos genes é representada por fragmentos de cerca de 180-200 aminoácidos (figura 14). Na figura 15 apresentamos o alinhamento entre o modelo da MENGOfam-001, cujo tamanho em posições é de 728, e um dos *hits* encontrado. O alinhamento ocorreu entre as posições 1-167 do *hit* com as posições 249-421 do modelo. Outro *hit* alinhou-se das posições 2-167 às posições 99-264 do modelo e o terceiro *hit* mostrado na figura alinhou-se das posições 1-126 às posições 372-493 do modelo. Assim, mesmo que existam proteínas que pertençam às famílias para as quais geramos HMMs curados, tais fragmentos nunca alcançarão a nota de corte.

Este problema possivelmente existiu em nossos testes pelo fato dos conjuntos de dados utilizados serem proveniente de amostras com alta diversidade microbiana, como ocorre com amostras de solo (LOMBARD *et al.*, 2011). Podemos mencionar que no estudo de caso de *neopullulanase* visto na Seção 5.3 encontramos proteínas pertencentes à família que provieram de metagenomas, as quais puderam ser classificadas como pertencentes à família. Isto pode ter ocorrido porque as sequências que alcançaram a nota de corte exigida para a família mantiveram os domínios essenciais para função. Apesar de serem provenientes de projetos de metagenomas, o tamanho dos fragmentos é muito próximo ao tamanho da proteína-semente (617 aminoácidos).

Query: mengofam-001-03 [M=728]
Scores for complete sequences (score includes all domains):

--- full sequence ---			--- best 1 domain ---			#dom-			
E-value	score	bias	E-value	score	bias	exp	N	Sequence	Description
2.5e-215	720.6	12.4	3e-215	720.4	12.4	1.0	1	GBAN_2580090	Cellulose binding domain./Glycosyl hydrolase fa
8.4e-61	209.8	4.1	9.3e-61	209.6	4.1	1.0	1	GBAN_3256720	Glycosyl hydrolase family 9.
1.7e-60	208.8	2.4	1.9e-60	208.6	2.4	1.0	1	GBAN_3977050	Glycosyl hydrolase family 9.
3.9e-50	174.5	1.4	4.5e-50	174.3	1.4	1.0	1	GBAN_1087170	Glycosyl hydrolase family 9.
2.6e-09	39.5	0.0	2.8e-09	39.4	0.0	1.0	1	GBAN_893120	
4.2e-09	38.9	0.3	2.4e-08	36.4	0.3	1.9	1	GBAN_3635130	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
1.1e-08	37.5	0.1	1.2e-08	37.4	0.1	1.0	1	GBAN_876250	Glycosyl hydrolase family 9.
1.5e-08	37.0	0.0	1.9e-08	36.7	0.0	1.2	1	GBAN_2175030	Glycosyl hydrolase family 9.
5.8e-08	35.1	3.7	2.9e-07	32.8	3.7	1.8	1	GBAN_543930	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
1.7e-06	30.2	3.8	0.015	17.3	0.4	2.2	2	GBAN_3635120	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
3.3e-05	26.0	0.5	3.9e-05	25.8	0.5	1.1	1	GBAN_3250510	Glycosyl hydrolase family 9./N-terminal ig-like
0.00071	21.6	1.2	0.00087	21.3	1.2	1.1	1	GBAN_3688120	Glycosyl hydrolase family 9./N-terminal ig-like
0.00077	21.5	0.1	0.00094	21.2	0.1	1.0	1	GBAN_2941280	Glycosyl hydrolase family 9.
0.00081	21.4	0.0	0.00093	21.2	0.0	1.1	1	GBAN_3812460	Glycosyl hydrolase family 9.
0.0023	19.9	0.1	0.0026	19.7	0.1	1.1	1	GBAN_3485190	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
0.0028	19.6	0.2	0.0031	19.5	0.2	1.2	1	GBAN_2419630	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
0.0056	18.6	0.1	0.0056	18.6	0.1	1.1	1	GBAN_559430	Glycosyl hydrolase family 9.
----- inclusion threshold -----									
0.018	17.0	0.1	0.025	16.5	0.1	1.1	1	GBAN_2987770	Glycosyl hydrolase family 9.
0.02	16.8	0.0	0.023	16.6	0.0	1.0	1	GBAN_1734890	
0.027	16.3	1.2	0.031	16.2	1.2	1.0	1	GBAN_385810	Glycosyl hydrolase family 9.
0.032	16.1	0.1	0.035	16.0	0.1	1.0	1	GBAN_1399940	
0.048	15.6	0.0	0.052	15.4	0.0	1.0	1	GBAN_3161070	
0.064	15.1	0.5	0.068	15.0	0.5	1.0	1	GBAN_3104300	
0.15	13.9	0.0	0.16	13.8	0.0	1.0	1	GBAN_3265350	
0.17	13.7	0.0	0.19	13.6	0.0	1.0	1	GBAN_3033140	Glycosyl hydrolase family 9.
0.29	13.0	0.1	0.33	12.8	0.1	1.0	1	GBAN_2580300	
0.61	11.9	0.0	0.67	11.8	0.0	1.0	1	GBAN_2109680	
0.71	11.7	0.1	0.73	11.6	0.1	1.0	1	GBAN_3106530	
0.75	11.6	0.1	0.94	11.3	0.1	1.1	1	GBAN_1940490	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
1.1	11.1	0.3	1.3	10.8	0.3	1.1	1	GBAN_3989700	cellulose 1,4-beta-cellobiosidase (EC 3.2.1.91)
1.1	11.0	0.0	1.3	10.8	0.0	1.0	1	GBAN_962990	Glycosyl hydrolase family 9.
1.4	10.7	0.0	21	6.8	0.0	1.9	2	GBAN_17870	Uncharacterized protein conserved in bacteria
2.3	10.0	0.0	2.9	9.7	0.0	1.1	1	GBAN_1605730	Predicted membrane-associated HD superfamily hy
2.9	9.7	0.1	5.8	8.7	0.0	1.5	2	GBAN_2442960	Phage QLRG family, putative DNA packaging.

Figura 14. Resultado do *hmmsearch*, onde o mini-reator de compostagem foi dado como banco de dados de seqüências e o modelo MENGOfam-001 foi utilizado como isca. O maior *score* alcançado por uma proteína do metagenoma do mini-reator foi 720.6, enquanto que o necessário para pertencer à família é 1119.6.


```

>> GBAN_3256720 Glycosyl hydrolase family 9.
#   score bias c-Value i-Value hmmfrom hmm to   alifrom ali to   envfrom env to   acc
---
1 ! 209.6 4.1 1.2e-64 9.3e-61 249 421 .. 1 167 [] 1 167 [] 0.95

Alignments for each domain:
== domain 1 score: 209.6 bits; conditional E-value: 1.2e-64
mengofam-001-03 249 wlylatgdetyldkaesyvekwekekqtiltsykwahcwwdvsyGaalllakitnkelykeaieknlwywtvlgngervkytPkgLawldswGs 342
GBAN_3256720 1 WLVRATGEQIYLQKAESLYA----QHVTGASLRWTHNWDDKIYGASILLAQLTGKSIYTTNAERFLNYWTVGDNGAKITTTTGGGLAWLQQWGS 89
*****999.....66778899*****

mengofam-001-03 343 lryattaafasvyadwegvsrekaeykdfakkqidyalGss..grsyvvGfGenpPkhPhrtahsswadslevPdyhr 421
lrya+ +a+la++yad r+ + y+dfak+q+dyaG++ grsyvvGfG+npP +Phhr ah+swa+s+++P r
GBAN_3256720 90 LRYAANTALLAFIYADRV---RDYNGRYHDFAKRQVDYALGNnpLGRSYVVVGfGNNPPVNPPhrGAHGSWANSIQEPRNRR 167
*****74...677889*****95569*****8766 PP

>> GBAN_3977050 Glycosyl hydrolase family 9.
#   score bias c-Value i-Value hmmfrom hmm to   alifrom ali to   envfrom env to   acc
---
1 ! 208.6 2.4 2.5e-64 1.9e-60 99 264 .. 2 167 .. 1 168 [] 0.95

Alignments for each domain:
== domain 1 score: 208.6 bits; conditional E-value: 2.5e-64
mengofam-001-03 99 lawayeeekdaleksGqldyldeikwatdyfikchPepnvyyqvGdgldhsvvpaevldlktdrpsykvdlslPGstvtvaetaaalaas 192
law e ++ +e s q +y+ld+++w d+++khPepnv+y qvG gldh++w+p e++++ +rp++k+d++nPGs++++e+aaalaaa+
GBAN_3977050 2 LAWGGIEFRKGYEESAQWNYLLDAVRWGDDWLMKAHPPEPNVYGVQVGRGDLDAFAGPPELMKM- -PRPAFKIDAQNPGLAGEAAAAALAAA 93
g*****

mengofam-001-03 193 ivfkdtPeyaetclkhakelfefadktrs..dagytaaegfydshsgfydelswaavwlylatgdetyldkae 264
i+f+d dP+y+++cl+hak+lf+fad+r +++ +++++y+s sg+ del+waa+wl atg++ yl+++e
GBAN_3977050 94 ILFRDVPDYSAKCLEHAKQLDFADTYRGtyNDAIPDSRAYYTSFSGYEDELAWAAAWLAWATGEKRYLEQSE 167
*****9622344456789*****987 PP

>> GBAN_1087170 Glycosyl hydrolase family 9.
#   score bias c-Value i-Value hmmfrom hmm to   alifrom ali to   envfrom env to   acc
---
1 ! 174.3 1.4 5.9e-54 4.5e-50 372 493 .. 1 126 [. 1 134 [. 0.94

Alignments for each domain:
== domain 1 score: 174.3 bits; conditional E-value: 5.9e-54
mengofam-001-03 372 dfakkqidyalGssgr..syvvGfGenpPkhPhrtahsswadslevPdyhrhvlvGalvGGPdk.ddsykddindyvlnevacdynaGfvGal 462
df +qi+yalG+++r syvvGfG+npP++Phrtah+sw+ds+++P+ rhvlyGalvGGP++ +d+y+d+ +dy++neva+dynaGf al
GBAN_1087170 1 DFGVRQINVALGDNPrnsSYVVGfGNNPPRNPPhrtAHGSWTDSITSPADNRHVLYGALVGGPGSANDAYTDSRQDYIANEVATDYNAGFSSAL 94
7889*****g*****8538*****

mengofam-001-03 463 aknyekygGePipnfkaieekte.effveaai 493
a + +ygG P+ nf ee e+fvea+
GBAN_1087170 95 AMLVAEYGGTPLANFPPEEPPDGPpEMFVEAQD 126
*****99998878****9975 PP

```

Figura 15. Alinhamento resultado do *hmmsearch*, onde o mini-reator de compostagem foi dado como banco de dados de sequências e o modelo MENGOfam-001 foi utilizado como isca. É possível visualizar que apenas fragmentos, regiões muito curtas, são alinhados ao HMM.

6 Conclusão

As principais contribuições deste trabalho foram:

- Geração de 74 modelos HMM, constituindo a base do recurso MENGINEfams.
- Estabelecimento de um protocolo de curadoria manual para famílias de proteínas mais detalhado do que o que se encontra na literatura.
- Contribuição para o estabelecimento de um website amigável, em que os HMMs gerados e outros resultados são disponibilizados para a comunidade científica.
- Identificação de 13 sequências no GenBank que poderiam ter suas anotações revisadas com base neste trabalho.
- Apresentação de resultados preliminares na conferência internacional Biocuration 2013 (<http://www.ebi.ac.uk/biocuration2013/>) realizada entre 7 a 10 de abril de 2013 em Cambridge, Inglaterra. O resumo apresentado na conferência está no Apêndice A.

Os principais passos que poderiam ser tomados para o prosseguimento deste trabalho são:

- À medida que novas anotações de proteínas-semente utilizando termos MENGINE sejam geradas pela comunidade científica, criar os respectivos HMMs, desta forma aumentando o número de famílias abarcadas por MENGINEfams.
- Integrar MENGINEfams a uma plataforma de famílias já existente (tal como TIGRfams), para que nossos modelos ganhem mais visibilidade.
- Identificar outras áreas em que haja necessidade de HMMs curadas para famílias específicas, e aplicar o protocolo desenvolvido neste trabalho para proteínas-semente dessas áreas.

7 Referências bibliográficas

- Allgaier, M. et al. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. **PLoS One**, v. 5, n. 1, p. e8812, 2010. ISSN 1932-6203. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/20098679> >.
- Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res**, v. 25, n. 17, p. 3389-402, Sep 1997. ISSN 0305-1048. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/9254694> >.
- Anton, B. P. et al. Objective: biochemical function. **Front Genet**, v. 5, p. 210, 2014. ISSN 1664-8021. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/25071837> >.
- Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. **Nucleic Acids Res**, v. 32, n. Database issue, p. D115-9, Jan 2004. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/14681372> >.
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nat Genet**, v. 25, n. 1, p. 25-9, May 2000. ISSN 1061-4036. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/10802651> >.
- Balakrishnan, R. et al. A guide to best practices for Gene Ontology (GO) manual annotation. **Database (Oxford)**, v. 2013, p. bat054, 2013. ISSN 1758-0463. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/23842463> >.
- Bateman, A.; Haft, D. H. HMM-based databases in InterPro. **Brief Bioinform**, v. 3, n. 3, p. 236-45, Sep 2002. ISSN 1467-5463. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/12230032> >.
- Brenner, S. E. Errors in genome annotation. **Trends Genet**, v. 15, n. 4, p. 132-3, Apr 1999a. ISSN 0168-9525. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/10203816> >.
- Cristianini, N., Hahn M.W. Introduction to computational genomics: a case studies approach. **Cambridge University Press**, 2006.

- DeAngelis K.M., Gladden J.M., Allgaier M., D'haeseleer P., Fortney J.L., et al. Strategies for enhancing the effectiveness of metagenomic-based enzyme discovery in lignocellulolytic microbial communities. **BioEnergy Research**, 2010, 3: 146–158.
- D'Elia, J. N.; Salyers, A. A. Contribution of a neopullulanase, a pullulanase, and an alpha-glucosidase to growth of *Bacteroides thetaiotaomicron* on starch. **J Bacteriol**, v. 178, n. 24, p. 7173-9, Dec 1996. ISSN 0021-9193. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/8955399>>.
- Eddy, S. R. Profile hidden Markov models. **Bioinformatics**, v. 14, n. 9, p. 755-63, 1998. ISSN 1367-4803. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/9918945>>.
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Res**, v. 32, n. 5, p. 1792-7, 2004. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/15034147>>.
- Engelhardt, B. E. et al. Protein molecular function prediction by Bayesian phylogenomics. **PLoS Comput Biol**, v. 1, n. 5, p. e45, Oct 2005. ISSN 1553-7358. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/16217548>>.
- Engelhardt, B. E. et al. Genome-scale phylogenetic function annotation of large and diverse protein families. **Genome Res**, v. 21, n. 11, p. 1969-80, Nov 2011. ISSN 1549-5469. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21784873>>.
- Finn, R. D. et al. Pfam: the protein families database. **Nucleic Acids Res**, v. 42, n. Database issue, p. D222-30, Jan 2014. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24288371>>.
- Finn, R. D.; Clements, J.; Eddy, S. R. HMMER web server: interactive sequence similarity searching. **Nucleic Acids Res**, v. 39, n. Web Server issue, p. W29-37, Jul 2011. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21593126>>.
- Fontes, C. M.; Gilbert, H. J. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. **Annu Rev Biochem**, v. 79, p. 655-81,

- 2010a. ISSN 1545-4509. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/20373916> >.
- Gaudet, P. et al. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. **Brief Bioinform**, v. 12, n. 5, p. 449-62, Sep 2011. ISSN 1477-4054. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/21873635> >.
- Gieg, L. M.; Duncan, K. E.; Suflita, J. M. Bioenergy production via microbial conversion of residual oil to natural gas. **Appl Environ Microbiol**, v. 74, n. 10, p. 3022-9, May 2008a. ISSN 1098-5336. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/18378655> >.
- Haft, D. H. et al. TIGRFAMs: a protein family resource for the functional identification of proteins. **Nucleic Acids Res**, v. 29, n. 1, p. 41-3, Jan 2001. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/11125044> >.
- Haft, D. H. et al. TIGRFAMs and Genome Properties in 2013. **Nucleic Acids Res**, v. 41, n. Database issue, p. D387-95, Jan 2013. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/23197656> >.
- Holm, L., Heger, A. Automated sequence-based approaches for identifying domain families. Protein Families: Relating Protein Sequence, Structure, and Function. Hoboken: **Wiley**, 2013. chap. 1, p. 1 – 22.
- Kalia, V. C.; Purohit, H. J. Microbial diversity and genomics in aid of bioenergy. **J Ind Microbiol Biotechnol**, v. 35, n. 5, p. 403-19, May 2008a. ISSN 1367-5435. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/18193465> >.
- Kalyuzhnaya, M. G. et al. High-resolution metagenomics targets specific functional types in complex microbial communities. **Nat Biotechnol**, v. 26, n. 9, p. 1029-34, Sep 2008a. ISSN 1546-1696. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/18711340> >.
- Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. **Annu Rev Genet**, v. 39, p. 309-38, 2005a. ISSN 0066-4197. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/16285863> >.

- Krishnamurthy, N. et al. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. **Genome Biol**, v. 7, n. 9, p. R83, 2006. ISSN 1465-6914. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/16973001> >.
- Krogh, A. An introduction to hidden markov models for biological sequences. Computational methods in molecular biology. **Elsevier**, 1998. chap. 4, p. 45 – 63.
- Li, L. L. et al. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. **Biotechnol Biofuels**, v. 2, p. 10, 2009. ISSN 1754-6834. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/19450243> >.
- Lima, T. et al. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. **Nucleic Acids Res**, v. 37, n. Database issue, p. D471-8, Jan 2009. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/18849571> >.
- Lombard, N. et al. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. **FEMS Microbiol Ecol**, v. 78, n. 1, p. 31-49, Oct 2011. ISSN 1574-6941. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/21631545> >.
- Ma, J. et al. MRAlign: protein homology detection through alignment of Markov random fields. **PLoS Comput Biol**, v. 10, n. 3, p. e1003500, Mar 2014a. ISSN 1553-7358. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/24675572> >.
- Martins, L. F. et al. Metagenomic analysis of a tropical composting operation at the são paulo zoo park reveals diversity of biomass degradation functions and organisms. **PLoS One**, v. 8, n. 4, p. e61928, 2013. ISSN 1932-6203. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/23637931> >.
- Mi, H.; Muruganujan, A.; Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. **Nucleic Acids Res**, v. 41, n. Database issue, p. D377-86, Jan 2013. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/23193289> >.
- Nelson, D. L., Cox, M. M. Lehninger princípios de bioquímica. **Omega**, 2009.

- Orengo, C., Bateman, A. Introduction. Protein Families: Relating Protein Sequence, Structure, and Function. Hoboken: **Wiley**, 2013. Introduction, p. 1 – 4
- Orengo, C. A.; Thornton, J. M. Protein families and their evolution-a structural perspective. **Annu Rev Biochem**, v. 74, p. 867-900, 2005. ISSN 0066-4154. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/15954844> >.
- Pedruzzi, I. et al. HAMAP in 2015: updates to the protein family classification and annotation system. **Nucleic Acids Res**, v. 43, n. Database issue, p. D1064-70, Jan 2015. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/25348399>>.
- Petersen, T. N. et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. **Nat Methods**, v. 8, n. 10, p. 785-6, 2011a. ISSN 1548-7105. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/21959131> >.
- Punta, M. et al. The Pfam protein families database. **Nucleic Acids Res**, v. 40, n. Database issue, p. D290-301, Jan 2012. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/22127870>>.
- Purwantini, E. et al. Genetic resources for methane production from biomass described with the Gene Ontology. **Front Microbiol**, v. 5, p. 634, 2014a. ISSN 1664-302X. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/25520705> >.
- Remmert, M. et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. **Nat Methods**, v. 9, n. 2, p. 173-5, Feb 2012a. ISSN 1548-7105. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/22198341> >.
- Sjölander, K. Phylogenomic databases and orthology prediction. Protein Families: Relating Protein Sequence, Structure, and Function. Hoboken: **Wiley**, 2013. chap. 5, p. 99 – 125
- Stein, L. Genome annotation: from sequence to biology. **Nat Rev Genet**, v. 2, n. 7, p. 493-503, Jul 2001a. ISSN 1471-0056. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/11433356>>.

- Thomas, P. D. et al. PANTHER: a library of protein families and subfamilies indexed by function. **Genome Res**, v. 13, n. 9, p. 2129-41, Sep 2003. ISSN 1088-9051. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/12952881> >.
- Torto-Alalibo, T. et al. Genetic resources for advanced biofuel production described with the Gene Ontology. **Front Microbiol**, v. 5, p. 528, 2014a. ISSN 1664-302X. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/25346727> >.
- van der Lelie, D. et al. The metagenome of an anaerobic microbial community decomposing poplar wood chips. **PLoS One**, v. 7, n. 5, p. e36740, 2012a. ISSN 1932-6203. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/22629327> >.
- Warnecke, F. et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. **Nature**, v. 450, n. 7169, p. 560-5, Nov 2007. ISSN 1476-4687. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/18033299> >.
- Wu, M.; Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. **Genome Biol**, v. 9, n. 10, p. R151, 2008a. ISSN 1465-6914. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18851752> >.
- Yennamalli, R. M. et al. Endoglucanases: insights into thermostability for biofuel applications. **Biotechnol Biofuels**, v. 6, n. 1, p. 136, 2013a. ISSN 1754-6834. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/24070146> >.
- Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. **Nucleic Acids Res**, v. 40, n. Web Server issue, p. W445-51, Jul 2012. ISSN 1362-4962. Disponível em: < <http://www.ncbi.nlm.nih.gov/pubmed/22645317> >.
-

Apêndice A – Resumo apresentado na Biocuration 2013.

An HMM-based manual curation protocol for microbial protein families related to bionergy

Fernanda O. R. Rego¹, Lucas M. Taniguti², Claudia B. Monteiro-Vitorello², João C. Setubal¹

¹Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil

²Department of Genetics, College of Agriculture “Luiz de Queiroz”, University of São Paulo, São Paulo, Brazil

Hidden Markov Models (HMMs) are essential tools for automated annotation of protein sequences. For many years new protein family resources based on HMMs have been made available to the scientific community (e.g. TIGRfams). Much effort has also been devoted to the automated generation of protein family HMMs (e.g Panther). However, manually curated protein family HMMs remain the gold standard for use in genome annotation. Here we present a protocol for manual curation of protein family HMMs applied to microbial protein families of bioenergy interest. As input we assume a query sequence coding for a protein whose function has been experimentally demonstrated, linked to a publication, and manually annotated with Gene Ontology terms. We then apply the following steps: (1) definition of additional family member selection criteria: sequence size, minimum %-identity threshold to query sequence, minimum alignment coverage, presence of signal peptide or trans-membrane domain, and presence of conserved regions and domains; (2) search of additional family members on UniprotKB with BLAST. We use UniprotKB (TrEMBL + SwissProt) because its entries have richer records than those from GenBank; (3) multiple alignment and HMM building; (4) analysis of results and iteration of the process, with preliminary HMMs used for further searches; (5) definition of the minimum score threshold for the final HMM, or depending on family complexity, the establishment of subfamilies. Literature and web resource scans are also part of the process, so that we can link our results to existing related resources and assign comments and specific features to families or individual family members. We will present a case study of Glycoside hydrolase family 9 using as seeds proteins of the anaerobic species of *Clostridium*.

Apêndice B – Criação e manutenção do site.

Lucas Taniguti, aluno de mestrado da profa. Claudia Vitorello foi colaborador deste projeto, tendo sido responsável pela criação do banco de dados e de uma plataforma web para a construção e anotação de famílias de maneira automática.

A descrição que se segue foi baseada na dissertação de mestrado do aluno Lucas, defendida em 18 de novembro de 2014 no programa de pós-graduação em genética da ESALQ-USP. A intenção deste apêndice é apenas de dar ao leitor uma visão mais completa do recurso Web que foi montado no projeto MENGOfams. Embora a aluna tenha contribuído para o design do recurso Web, o crédito principal desta parte do trabalho é do aluno Lucas.

Foi utilizado o SQLite3, um dos vários programas que permitem administrar bancos de dados relacionais. Os dados do banco residem em apenas uma tabela com os seguintes atributos, brevemente descritos:

1. id: Identificador único de uma família que possui três campos. O primeiro é o prefixo ‘mengofam’ seguido por outro campo com três dígitos indicando o número da família (001-999) e por último um campo com dois dígitos indicando a versão da anotação (01-99).
2. title: Texto curto que deve descrever a função ou características da família criada.
3. description: Texto com propriedades da família como: domínios requeridos, presença da sequência de peptídeo sinal, mecanismo de atuação, sítios conservados, classificação enzimática. Cada um desses atributos são presença optativa a cada família.
4. referenceProtein: Identificador Uniprot da proteína utilizada como referência para a criação da família. Esse campo indica a proteína utilizada pela equipe MENGO para a atribuição do termo GO relacionado a bioenergia microbiana.
5. matches: Campo com os dados de score que cada proteína do modelo recebeu pelo modelo HMM. Utilizado para criar uma lista e um gráfico com pontuações de corte para a introdução de novos membros àquela família.
6. pieSpecies: Após a geração do modelo a classificação taxonômica de cada membro é obtida e armazenada para a construção de um gráfico tipo pizza representando a distribuição dos membros que pertencem àquela família.

7. author: Identifica o autor da anotação da família em questão.

8. status: Classifica a família como preliminar (preliminary), quando ainda está nas etapas mais iniciais, sem muitas descrições e/ou com incertezas sobre os grupos taxonômicos representados; processo em revisão (under review); e revisada (reviewed), em que todas as informações descritivas estão devidamente referenciadas.

Com o intuito de tornar a anotação de famílias e construção dos modelos computacionais menos dependentes de interação direta com linha de comando elaboramos um serviço web que disponibiliza interface gráfica (figura B1).

Figura B1. Página inicial da interface gráfica criada para construção de famílias dentro do contexto do projeto MENO Fams.

A utilização da plataforma para a construção dos modelos HMM inicia com a inserção dos arquivos no formato fasta de todas as sequências selecionadas como membros de uma família na caixa de diálogo apropriada (Figura B2).

Create a new family

Status	<input style="width: 90%;" type="text"/>	→	<div style="display: inline-block; vertical-align: middle;"> <div style="display: flex; align-items: center;"> <div style="font-size: 3em; margin-right: 5px;">{</div> <div style="text-align: left;"> Preliminary Under Review Reviewed </div> </div> </div>
Reference Protein	<input style="width: 90%;" type="text"/>	→	List of all seed-proteins
Family ID	<input style="width: 90%;" type="text"/>	→	MENGOfam id pattern
Description	Family brief description		
Seed sequences	Fasta sequence of all seed proteins for the HMM profile		
<div style="border: 1px solid black; padding: 5px 20px; display: inline-block;">Submit!</div>			

Figura B2. Representação esquemática das informações visualizadas no formulário de submissão da plataforma MENGOFams.

Após o preenchimento completo do formulário de submissão, todo o pipeline descrito a seguir é realizado automaticamente, ou seja:

1) alinhamento múltiplo: com as sequências em um arquivo multifasta a plataforma gera automaticamente o alinhamento múltiplo utilizando o programa MUSCLE (<https://www.ebi.ac.uk/Tools/msa/muscle/>) (Figura B3).

MUSCLE (3.8) multiple sequence alignment

```

tr|F0SY74|F0SY74_SYNGF      MKRFFQKLTVFILIAALSLSVYPQNVFSLTAFSYSFESGQIDGWSGSAILQENVGSCTLQ
tr|E6WTD1|E6WTD1_PSEUU      MSRTTL-----LAGLCLGLMALLP-----
tr|C4ID52|C4ID52_CLOBU      MKKGKVLMTVAA-----IILIVGMVISYKA-----
tr|R0AFG5|R0AFG5_CLOBU      MKKGKVLMTVAA-----IILIVGMVISYKA-----
tr|Q182N3|Q182N3_CLOD6      MFGGINIKRIGIIM-----FFLSISFIGFYLV-----
tr|D5Q297|D5Q297_CLODI      MMM-----FFLGIVLIVFYFV-----
tr|C6XKU0|C6XKU0_HIRBI      -MKQI-----WWGLLGAVFSVCT-----
tr|R7ZM94|R7ZM94_9BACT      -----MFFYTGGGLFAQQS-----
tr|M2AMX7|M2AMX7_9PLAN      MFDRDFRLP-----AWSPVLILAATCVLSVNCSTASAQSSDSEQVLFGP
tr|F2AY34|F2AY34_RHOBT      MFVRTTDVGLRPL-----RWLPAVMLAAIGIVGSSLSASAQTALGEKVLFG
tr|K2PFP0|K2PFP0_9THEM      MRNFFK-----VFTLVLVVISVLL-----
tr|H9UEE9|H9UEE9_FERPD      MVERHFRYVLIC-----ILFLVMLLISSTQ-----
tr|K9UT57|K9UT57_FERGO      -MRFRKNFAVLM-----LIVLISTLFLSTQ-----
tr|D4PEB3|D4PEB3_FERNB      -----
tr|E8QX85|E8QX85_ISOPI      MSA-----IWILTCCLVGVDP-----
tr|H1XTT3|H1XTT3_9BACT      -MKFIYR-----IYLFLFAVIFAM-----
tr|E8N637|E8N637_ANATU      MKMNKTQIPSFSP-----LWMVLIILTLFGC-----
tr|F3ZYT5|F3ZYT5_MAHAS      MKKSII-----LYIIVLSLIFMTA-----
tr|B5YCY0|B5YCY0_DICT6      -----
tr|B5YAS2|B5YAS2_DICT6      -MRLK-----ILLLLVLLIFLAG-----

```

Figura B3. Exemplo de alinhamento múltiplo gerado automaticamente na plataforma MENGOFams.

2) definição do modelo HMM: os modelos HMM gerados ficam disponíveis na página para download (Figura B4).

Family Entry
 ::General information::

Multiple sequence alignment | Score distribution | **Hidden Markov Model**

Download

Family ID:	mengofam-013-01
Family name:	Endoglucanase C
Family status:	Preliminary
Author:	Monteiro-Vitorello

HMMER3/b [3.0 | March 2010] NAME mengofam-013-01 LENG

0.69996HMM	A	C	D	E	F	Q	H
3.72511	2.99094	3.29059	3.06770	2.94668	2.86867	2.86852	
0.77269	0.00000		1	3.03822	4.57374	4.14239	3.7421
4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	2.77519	
2.81387	2.59001	4.93784	3.68480	2		2.60618	4.422
4.88788	2.73230	2.33486	4.16021	3.26863	3.58964	3.59133	
2.89001	2.37867	2.77519	2.98518	4.58477	3.61503	0.07	
2.48618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72	
3.52849	2.67894	2.33573	2.73470	3.59689	3.07986	3.84644	
3.61593		0.08275	4.00393	2.79400	0.61958	0.77255	0.69
3.46323	2.40538	3.72539	3.29273	2.67763	2.66277	4.24652	
2.40239	3.79685	4.11852	3.89572	3.21955	3.05235	2.85790	
0.61958	0.77255	0.48576	0.95510	8	2.65433	4.25738	4.4
2.69365	4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	

Figura B4. HMMs gerados ficam disponíveis publicamente para download.

3) criação dos gráficos (taxonomia): A plataforma disponibiliza uma visualização gráfica representando a porcentagem de cada grupo taxonômico que fazem parte da MENGOfam (Figura B5).

Figura B5. Representação esquemática da distribuição taxonômica dos organismos que são considerados membros de uma determinada família.

4) utilização do modelo para captura de novos membros nos bancos de dados do Uniprot e criação do gráfico da distribuição de scores. A representação gráfica permite ao anotador avaliar o score e a necessidade de alteração no modelo HMM para ampliar a abrangência de determinada família (Figura B6).

Figura B6. Representação gráfica para visualização da nota de corte e dos membros de uma determinada família. Em Model members estão as proteínas que fizeram parte do modelo, e em Uniprot db estão todas as sequências que foram capturadas pelo modelo diretamente nos bancos de dados do Uniprot (SwissProt + TrEMBL).

5) tabela com a lista de novos membros e outras proteínas com score fora da linha de corte que podem ser analisados para o update do modelo (com links para o Uniprot) (Figura B7).

Uniprot Access.	Score
1. R7ZM94_9BACT *seed	581.4
2. Q9P867_PIREQ	553.4
3. F2AY34_RH0BT *seed	520.8
4. K5E3V5_RH0BT	518.5
5. Q7UGF4_RH0BA	518.3
6. M2AMX7_9PLAN *seed	516.3
7. L7CK55_RH0BT	515.5
8. H1XTT3_9BACT *seed	505.0
9. H9UEE9_FERPD *seed	503.6
10. K9UT57_FERGO *seed	503.3
11. B5YAS2_DICT6 *seed	500.5
12. E8N637_ANATU *seed	500.1
13. B8DYX6_DICTD	491.2
14. E8QX85_ISOPI *seed	491.0
15. A7HNC0_FERNB	489.8
16. F3ZYT5_MAHA5 *seed	486.4
17. D4PEB3_FERNB *seed	485.3
18. K2PFP0_9THEM *seed	484.1
19. B1LAS2_THESQ	479.5
20. Q9X273_THEMA	478.5
21. B5YCY0_DICT6 *seed	478.5
22. B8DZM2_DICTD	474.3
23. B5YCX9_DICT6	457.8
24. B8DZM1_DICTD	456.5
25. G9IS52_9BACT	448.0
26. Q182N3_CLOD6 *seed	446.8
27. C9XMR5_CLODC	444.7

Figura B7. Lista contendo os identificadores e score de cada proteína utilizada ou não no modelo, que foram capturadas no banco de dados do Uniprot. Todos os identificadores apresentam links para o Uniprot.

Uma vez analisado o modelo e a tabela com os membros, o usuário poderá ainda reavaliar a família e fazer um update caso seja necessário. Neste update todas as informações já inseridas são mantidas na página de forma que o usuário poderá alterar somente os campos que desejar. A interface pretende facilitar a construção das famílias de maneira amigável tornando o trabalho do anotador mais rápido e eficiente.

Apêndice C – Tabela com as famílias que foram estudadas e estão disponíveis na página web. Em destaque (verde) as famílias que foram estudadas apenas pela autora do trabalho.

Identificador	Nome	Status	Autor
mengofam-001-03	Endoglucanase (F, N, Z)	Reviewed	Monteiro-Vitorello
mengofam-002-01	Extracellular solute-binding protein family 1	Reviewed	Orpinelli
mengofam-003-02	Dockerin type 1	Reviewed	Orpinelli
mengofam-004-01	Endoglucanase D	Reviewed	Monteiro-Vitorello
mengofam-005-05	Sugar ABC transporter	Reviewed	Orpinelli
mengofam-006-04	Acetyl-CoA decarbonylase/synthase complex subunit gamma 1	Reviewed	Orpinelli
mengofam-007-05	Starch-utilization system protein F	Reviewed	Monteiro-Vitorello
mengofam-008-01	ABC-type sugar transport system periplasmic component-like protein	Reviewed	Orpinelli
mengofam-009-05	Extracellular solute-binding protein family 1	Reviewed	Orpinelli
mengofam-011-07	Endo-1,3-beta-xylanase C	Reviewed	Monteiro-Vitorello
mengofam-012-06	Cellulose 1,4-beta-cellobiosidase	Reviewed	Monteiro-Vitorello
mengofam-013-01	Acetyl-CoA decarbonylase/synthase complex subunit delta 1	Reviewed	Orpinelli
mengofam-014-04	Endoglucanase E	Reviewed	Monteiro-Vitorello
mengofam-015-04	Endo-1,4-beta-xylanase Z	Reviewed	Monteiro-Vitorello
mengofam-016-03	Methyl-coenzyme M reductase I subunit alpha	Reviewed	Orpinelli
mengofam-017-01	Methyl-coenzyme M reductase I subunit beta	Reviewed	Orpinelli
mengofam-018-01	Methyl-coenzyme M reductase I subunit gamma	Reviewed	Orpinelli
mengofam-019-01	Butyrate--acetoacetate CoA-transferase subunit B	Reviewed	Orpinelli
mengofam-020-02	Aldehyde-alcohol dehydrogenase	Reviewed	Monteiro-Vitorello

mengofam-021-05	Butyrate--acetoacetate CoA-transferase subunit A	Reviewed	Orpinelli
mengofam-022-06	Glucan 1,4 - alpha-glucosidase SusB	Reviewed	Monteiro-Vitorello
mengofam-023-01	Phosphate butyryltransferase	Reviewed	Orpinelli
mengofam-024-03	ACDS epsilon 1	Reviewed	Orpinelli
mengofam-025-01	Phosphate acetyltransferase	Reviewed	Orpinelli
mengofam-026-03	Acetate kinase	Reviewed	Orpinelli
mengofam-027-03	Endoglucanase J	Reviewed	Monteiro-Vitorello
mengofam-028-01	ACDS complex subunit beta 1	Reviewed	Orpinelli
mengofam-029-05	Scaffoldin dockerin binding protein A	Reviewed	Orpinelli
mengofam-031-04	NADH-dependent butanol dehydrogenase A	Reviewed	Orpinelli
mengofam-032-05	Endoglucanase C	Reviewed	Monteiro-Vitorello
mengofam-033-02	NADH-dependent butanol dehydrogenase B	Reviewed	Orpinelli
mengofam-034-01	Scaffolding protein	Reviewed	Orpinelli
mengofam-035-02	Cell surface glycoprotein 1	Reviewed	Orpinelli
mengofam-036-02	L-sulfolactate dehydrogenase	Reviewed	Monteiro-Vitorello
mengofam-037-02	Formate dehydrogenase subunit beta	Reviewed	Monteiro-Vitorello
mengofam-038-01	Butyrate kinase 1	Reviewed	Orpinelli
mengofam-039-03	Endoglucanase G	Reviewed	Monteiro-Vitorello
mengofam-040-02	Putative substrate binding protein	Reviewed	Orpinelli
mengofam-041-01	Carbohydrate-binding protein (ctaA)	Reviewed	Orpinelli
mengofam-042-03	Endoglucanase H	Reviewed	Monteiro-Vitorello
mengofam-043-02	CoB--CoM heterodisulfide reductase iron-sulfur subunit A	Reviewed	Monteiro-Vitorello

mengofam-044-09	Lactaldehyde dehydrogenase	Reviewed	Monteiro-Vitorello
mengofam-045-04	Cellulose 1,4-beta-cellobiosidase	Reviewed	Monteiro-Vitorello
mengofam-046-05	Endo-1,4-beta-xylanase D	Reviewed	Monteiro-Vitorello
mengofam-047-03	Xyloglucan-specific endo-beta-1,4-glucanase activity	Reviewed	Monteiro-Vitorello
mengofam-048-01	ScaC cellulosomal scaffoldin protein	Reviewed	Orpinelli
mengofam-049-01	SusR - regulatory protein	Reviewed	Orpinelli
mengofam-050-04	Neopullulanase SusA	Reviewed	Monteiro-Vitorello
mengofam-051-05	Starch-utilization system protein C	Reviewed	Orpinelli
mengofam-052-02	Starch-utilization system protein D	Reviewed	Monteiro-Vitorello
mengofam-053-05	Alpha-amylase SusG	Reviewed	Monteiro-Vitorello
mengofam-054-02	3-hydroxybutyryl-CoA dehydrogenase	Reviewed	Orpinelli
mengofam-055-04	MalR - regulator protein	Reviewed	Orpinelli
mengofam-056-02	Endoglucanase T	Reviewed	Monteiro-Vitorello
mengofam-057-02	ScaA cellulosomal scaffoldin protein	Reviewed	Orpinelli
mengofam-059-06	F420-dependent methylenetetrahydromethanopterin dehydrogenase	Reviewed	Orpinelli
mengofam-060-01	Acetyl-CoA decarbonylase/synthase complex subunit delta 2	Reviewed	Orpinelli
mengofam-061-01	Acetyl-CoA decarbonylase/synthase complex subunit epsilon 2	Reviewed	Orpinelli
mengofam-062-02	ACDS complex subunit alpha 2	Reviewed	Orpinelli
mengofam-063-04	Cellulose 1,4-beta-cellobiosidase (celO)	Reviewed	Monteiro-Vitorello
mengofam-064-03	Cellulosomal element P (cseP)	Reviewed	Orpinelli
mengofam-065-01	Acetyl-CoA decarbonylase/synthase complex subunit gamma 2	Reviewed	Orpinelli
mengofam-066-01	Acetyl-CoA decarbonylase/synthase complex subunit beta 2	Reviewed	Orpinelli

mengofam-067-01	Basic membrane lipoprotein	Reviewed	Orpinelli
mengofam-068-03	Monomethylamine methyltransferase MtmB1	Reviewed	Monteiro-Vitorello
mengofam-069-02	Monomethylamine methyltransferase	Reviewed	Monteiro-Vitorello
mengofam-071-03	Methylated-thiol--corrinoid protein MtsB	Reviewed	Monteiro-Vitorello
mengofam-072-03	Endo-1,4-beta-xylanase Y	Reviewed	Monteiro-Vitorello
mengofam-073-01	Cellulosomal-scaffolding protein A	Reviewed	Orpinelli
mengofam-074-04	Aldehyde/alcohol dehydrogenase	Reviewed	Monteiro-Vitorello
mengofam-077-02	Endoglucanase A	Reviewed	Monteiro-Vitorello
mengofam-078-02	Endo-1,4-glucanase	Reviewed	Orpinelli
mengofam-079-05	Endoglucanase Y	Reviewed	Monteiro-Vitorello
mengofam-081-03	HepII	Reviewed	Orpinelli
mengofam-082-06	arabinoxylan-specific endo- β -1,4-xylanase	Reviewed	Orpinelli
mengofam-089-02	Pectin acetyltransferase	Reviewed	Orpinelli
mengofam-092-02	Fe-hydrogenase beta subunit	Reviewed	Orpinelli
mengofam-093-02	Fe-hydrogenase alpha subunit	Reviewed	Orpinelli
mengofam-098-06	Methenyltetrahydromethanopterin cyclohydrolase	Reviewed	Orpinelli
mengofam-100-04	3-hydroxybutyryl-CoA dehydratase	Reviewed	Orpinelli
mengofam-102-06	F420-non-reducing hydrogenase iron-sulfur subunit D	Reviewed	Monteiro-Vitorello
mengofam-103-06	F420-non-reducing hydrogenase subunit G	Reviewed	Monteiro-Vitorello
mengofam-107-06	Coenzyme F420-dependent sulfite reductase	Reviewed	Orpinelli
mengofam-109-02	Hydrogenase small beta subunit	Reviewed	Orpinelli
mengofam-110-04	Hydrogenase large alpha subunit	Reviewed	Orpinelli

mengofam-112-01	Diacetylchitobiose deacetylase	Reviewed	Orpinelli
mengofam-122-01	Phosphate acetyltransferase	Reviewed	Orpinelli