UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de graduação em biotecnologia

Trabalho de Conclusão de Curso

Prospecção e análise *in silico* de enzimas celulases de espécies de fungos fitopatogênicos com potencial aplicação na produção de biocombustíveis

Rafael Cagliari

Pelotas, 2018

Rafael Cagliari

Prospecção e análise *in silico* de enzimas celulases de especies de fungos fitopatogênicos com potencial aplicação na produção de biocombustíveis

Trabalho Acadêmico apresentado ao Curso de Graduação em Biotecnologia da Universidade Federal de Pelotas como requisito parcial à obtenção do título de Bacharel em Biotecnologia

Orientador Acadêmico: Luciano da Silva Pinto
Orientador de Estágio: Frederico Schmitt Kremer

Pelotas, 2018

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

C131p Cagliari, Rafael

Prospecção e análise in silico de enzimas celulases de espécies de fungos fitopatogênicos com potencial aplicação na produção de biocombustíveis. / Rafael Cagliari ; Luciano da Silva Pinto, orientador ; Frederico Schmitt Kremer, coorientador. — Pelotas, 2018.

46 f. : il.

Trabalho de Conclusão de Curso (Bacharelado em Biotecnologia) — Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2018.

1. Bioetanol. 2. Docking. 3. Hemicelulase. 4. Modelagem de proteínas. I. Pinto, Luciano da Silva, orient. II. Kremer, Frederico Schmitt, coorient. III. Título.

CDD : 620.8

Elaborada por Ubirajara Buddin Cruz CRB: 10/901

Banca Examinadora:

Prof. Dr. Luciano da Silva Pinto (orientador)

Universidade Federal de Pelotas

Prof.ª Dra. Vanessa Galli
Universidade Federal de Pelotas

Doutorando Frederico Schmitt Kremer

Universidade Federal de Pelotas

Doutorando Rafael Danelon dos Santos Woloski

Universidade Federal de Pelotas

Agradecimentos

Aos meus pais e meus irmãos por todo o apoio e ajuda durante estes longos anos;

Aos amigos e colegas por todos os momentos juntos e superação de obstáculos;

Aos professores e demais pessoas que compartilharam seu conhecimento pela paciência e dedicação;

Enfim, a todas as pessoas que ajudaram a tornar esta incrível jornada e sua eventual conclusão realidade.

Resumo

A atual busca por fontes de energia limpas impulsionou o desenvolvimento dos chamados biocombustíveis de primeira geração a partir de açúcares e óleos de origem vegetal. Entretanto, evidências recentes colocam em dúvida sua capacidade de mitigar os problemas ambientais causados pelos atuais combustíveis fósseis, guiando o interesse de pesquisadores e das indústrias aos biocombustíveis de segunda geração, produzidos a partir da degradação da celulose e outros polissacarídeos contidos na parede celular das plantas. Este processo comumente utiliza enzimas celulases e hemicelulases na degradação destas que são consideradas algumas das moléculas de maior recalcitrância conhecidas atualmente, etapa esta responsável por aumentar os custos de produção, sendo a descoberta de novas enzimas de maior atividade catalítica uma das estratégias adotadas para a resolução deste problema. Fungos fitopatogênicos costumam apresentar celulases de alta eficiência e elevados padrões de secreção destas, sendo um recorrente alvo na busca por novas enzimas deste tipo. Sendo assim, este trabalho teve por objetivo desenvolver uma metodologia para a prospecção de enzimas degradadoras de material lignocelulósico através de buscas por sequências de aminoácidos em proteomas de fungos fitopatogênicos, modelagem de sua estrutura tridimensional e determinação de afinidade ao ligante através de ensaios de *docking* molecular. Foram identificadas ao menos dez enzimas com alto potencial catalítico, e possível aplicação biotecnológica para a produção de biocombustíveis de segunda geração.

Palavras-chave: bioetanol, docking, hemicelulases, modelagem de proteínas

Abstract

The current search for clean energy sources has boosted the development of first-generation biofuels from sugars and vegetable oils. However, recent evidence casts doubt on its ability to mitigate the environmental problems caused by current fossil fuels, leading researchers and industries to the second-generation biofuels produced from the degradation of cellulose and other polysaccharides contained in the cell wall of plants. This process uses cellulases and hemicellulases enzymes for the degradation of these that are considered some of the molecules of greater recalcitrance known, this stage is responsible for increasing the costs of production, being the discovery of new enzymes of higher catalytic action one of the strategies adopted for the resolution of this problem. Phytopathogenic fungi usually present cellulases of higher efficiency and high secretion rates, being a recurrent target in the search for new enzymes of this type. The aim of this work was to develop a methodology for the prospection of lignocellulosic material degrading enzymes by searching for amino acid sequences in proteomes of phytopathogenic fungi, modeling their three-dimensional structure and determining their binding affinity through molecular docking assays. At least ten enzymes with high catalytic potential and possible biotechnological application for the production of second generation biofuels were identified.

Keywords: bioetanol, docking, hemicellulases, protein modeling

Lista de Figuras

Figura 1: Fluxograma representando de maneira simplificada os processos empregados na conversão da biomassa lignocelulósica. 17

Figura 2: Representação da estrutura da parede celular vegetal 19

Figura 3: Representação gráfica dos três distintos tipos de celulases e seus mecanismos de ação sobre a degradação da celulose 20

Figura 4: Representação das etapas presentes nos processos de modelagem de proteínas comparativa e *ab initio* 25

Figura 5: Modelo de prospecção de celulases funcionais a partir de proteomas depositados no Uniprot. 20

Figura 6: Representação gráfica da estrutura tridimensional das enzimas modelo utilizadas para desenhar a *grid box* 32

Figura 7: Esquematização dos principais resultados encontrados. 38

Lista de Tabelas

Tabela 1: Lista de espécies de fungos fitopatogênicos com proteomas depositados no uniprot e seus respectivos Ids de organismo 34

Tabela 2: : Lista de sequências de aminoácidos selecionadas pelo método de alinhamento BLASTp contra banco de dados do PDB 36

Tabela 3: : Lista de sequências de aminoácidos selecionadas pelo método de alinhamento pela ferramenta HMMER contra um alinhamento múltiplo de celulases do Pfam. 37

Tabela 4: Resultados dos ensaios de *docking* molecular para endoglucanases prospectadas a partir de proteomas do Uniprot e mananase de *T. reesei* ao final. 38

Tabela 5: Resultados dos ensaios de *docking* molecular para mananases prospectadas a partir de proteomas do Uniprot e mananase de *T. reesei* ao final. 39

Lista de Abreviaturas e Siglas

DNA	Ácido Desoxirribonucleico
NMR	Espectroscopia de Ressonância Magnética Nuclear
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank

Sumário

Sumário **11**

1. Introdução **13**

2. Revisão Bibliográfica **15**

2.1. Biocombustíveis 15

2.2. Biocombustíveis de Segunda Geração 16

2.2.1. Celulose e outros constituintes da parede celular 17

2.2.2. Enzimas Celulases 19

2.2.3. Hemicelulases 21

2.3. Patógenos de plantas como fonte para a prospecção de enzimas lignocelulósicas com destaque para fungos 22

2.4. Ferramentas de bioinformática para a análise e prospecção de novas enzimas degradadoras de biomassa lignocelulósica 23

2.4.1. Modelagem de proteínas 24

2.4.2. *Docking* molecular 25

3. Objetivos **27**

3.1. Objetivos gerais 27

3.2. Objetivos específicos 27

4. Metodologia **28**

4.1. Identificação de fungos fitopatogênicos com proteomas depositados em bases de dados online 28

4.2. Prospecção de celulases 28

4.2.1. Identificação de celulases a partir de um banco de dados formado por proteínas depositadas no Protein Data Bank 29

4.2.2. Identificação de celulases a partir de alinhamento múltiplo através do Pfam 29

4.3. Modelagem de proteínas testes de estabilidade estrutural 30

4.4. *Docking* molecular 31

5. Resultados **33**

6. Discussão **39**

8. Referências **42**

1. Introdução

A demanda por fontes de energia limpas e renováveis se tornou uma constante em nossa sociedade, tendo em vista a inevitável escassez e consequências ambientais, originadas pelos combustíveis fósseis atualmente utilizados e dos quais a humanidade depende quase que completamente. Dentre as diversas fontes alternativas propostas destacam-se, os biocombustíveis de primeira geração, isto é, gerados a partir da fermentação de açúcares e óleos contidos em vegetais. Durante as duas últimas décadas a utilização destes combustíveis cresceu constantemente, até que dúvidas quanto a sua capacidade de reduzir a emissão de gases responsáveis pelo efeito estufa e disputas de espaço para plantio com alimentos voltaram a atenção mundial para novas alternativas (DOORNBOSCH; STEENBLIK, 2008; NAIK et al., 2010).

Neste contexto estão inseridos os biocombustíveis de segunda geração, gerados a partir de resíduos agrícolas por meio da fermentação da glicose retida na parede vegetal, uma estrutura de enorme abundância nas células vegetais formada principalmente pelo polissacarídeo celulose, formado por cadeias de glicose e a hemicelulose, constituída por longas cadeias de carboidratos associados à celulose. Apesar de seu crescente uso, o alto custo de sua produção é hoje a maior barreira impedindo sua plena inserção no mercado global. A etapa de quebra de celulose e hemicelulose é particularmente cara devido à altíssima recalcitrância destas moléculas, havendo demanda para o desenvolvimento de novas técnicas visando aumentar a eficiência deste processo (ROBAK; BALCEREK, 2018).

Uma estratégia muito utilizada é a prospecção de novas enzimas, isto é, a procura por celulases de maior capacidade funcional. Fungos fitopatogênicos desenvolveram durante sua história evolutiva verdadeiros arsenais enzimáticos para degradação da parede celular vegetal, sendo um possível alvo para a prospecção de novas celulases e outras enzimas degradadoras da lignocelulose. O método padrão consiste no isolamento do fungo e realização de testes enzimáticos, um processo custoso em tempo e recursos. Todavia, existem milhares de organismos cujo proteoma completo é conhecido e está depositado em bancos de dados online, sendo possível acessar todas suas proteínas

curadas de modo a viabilizar o desenvolvimento de novos métodos para a busca por enzimas a partir de ferramentas de bioinformática.

Neste trabalho propõe-se a prospecção de celulases e enzimas relacionadas a partir de sequências de aminoácidos depositadas em bancos de dados online, passando por sua modelagem estrutural e averiguação da afinidade das enzimas pelo substrato através de *docking* molecular aplicando diversos parâmetros de qualidade e por fim comparando à atividade obtida para uma enzima de alto uso na indústria. Assim espera-se encontrar lignocelulases de alto valor industrial e biotecnológico para a produção de bioetanol de segunda geração.

2. Revisão Bibliográfica

2.1. Biocombustíveis

A atual situação de dependência da sociedade moderna por recursos fósseis para a produção, através de seu refinamento, de combustíveis como a gasolina e o óleo diesel (ambos derivados do petróleo) conflita diretamente com o fato de que estes não são recursos renováveis, nem sustentáveis. A queima destes combustíveis tem sido cada vez mais associada a mudanças climáticas no planeta por conta do fenômeno conhecido como efeito estufa, causado pelo abrupto aumento da concentração de gases como o dióxido de carbono (CO_2) na atmosfera. Sendo assim, a busca por fontes de energias renováveis e sustentáveis tornou-se umas das grande prioridades econômica e científica desta geração (NAIK et al., 2010).

Biocombustíveis surgem neste contexto como candidatos a tecnologia alternativa aos combustíveis fósseis, através da geração de bioetanol e biodiesel líquidos a partir de substratos de origem vegetal, utilizando principalmente milho e cana-de-açúcar como matéria-prima, os chamados biocombustíveis de primeira geração (HOLZMAN, 2008). A premissa de que o bioetanol e o biodiesel conferem energia limpa provém do fato de que o carbono liberado por estes combustíveis teria sido anteriormente fixado pelas plantas que lhe deram origem, assim mantendo a concentração de carbono na atmosfera inalterada (SIMS et al., 2010). O consumo destes combustíveis verdes vem aumentando expressivamente desde os anos 2000, com os Estados Unidos e o Brasil representando juntos mais da metade da produção de bioetanol mundial, constituindo cerca de 18% do consumo total de combustíveis no Brasil (HIRANI et al., 2018).

Apesar dos aparentes benefícios, os biocombustíveis de primeira geração encontram dificuldades em sua plena implementação, enfrentando obstáculos comerciais, uma vez que o ainda elevado custo de sua produção o torna pouco competitivo, sendo sua inserção no mercado possível apenas mediante subsídios governamentais (DOORNBOSCH; STEENBLIK, 2008). Os biocombustíveis também acabam conflituosamente esbarrando em barreiras ambientais, uma vez que a área utilizada para o plantio de sua matéria-prima

acaba por muitas vezes advindo do desflorestamento de regiões de alta densidade vegetal, detentoras de biomas complexos como a floresta tropical no Brasil (FARGIONE et al., 2008). O desmatamento dessas áreas ocasiona a liberação de mais carbono para a atmosfera através da decomposição ou queima da vegetação, além de prejudicar a longo prazo o sequestro de carbono promovido pela vegetação local. Uma solução apontada, o uso de plantações produtoras de alimentos já existentes para o cultivo de culturas para a geração de biocombustíveis, também contribui para o efeito estufa indiretamente pois a troca de culturas leva ao aumento dos preços para o alimento previamente cultivado, incentivando produtores a expandir suas lavouras a fim de produzir a cultura, fenômeno cunhado *fuel versus food*, ou combustível contra comida (SEARCHINGER et al., 2008). Embora tenham crescido muito nos últimos anos, é evidente que os biocombustíveis de primeira geração possuem diversos contrapontos ambientais e comerciais quanto à sua consolidação como fonte de energia limpa, levando a questionamentos inclusive quanto a sua aplicabilidade a um nível global e de longo prazo.

2.2. Biocombustíveis de Segunda Geração

Devido às limitações inerentes aos biocombustíveis de primeira geração surgiu o estímulo para o desenvolvimento dos biocombustíveis de segunda geração, isto é, a partir de biomassa lignocelulósica, encontrada em grande quantidade em plantas e disponível principalmente em resíduos que de outra forma seriam descartados, como o bagaço da cana-de-açúcar, desta forma não havendo disputa por recursos entre a plantação de combustíveis e alimentos (SIMS et al., 2010). O processo para a conversão da lignocelulose em etanol passa por três etapas: um pré-tratamento químico e/ou por calor ajuda a quebrar as fibras de celulose, aumentando a área de contato para a atuação enzimática de celulases diversas na segunda etapa, reduzindo os açúcares resultantes em glicose, que será fermentada por leveduras resultando na produção do etanol (Figura 1) (DIAS et al., 2011).

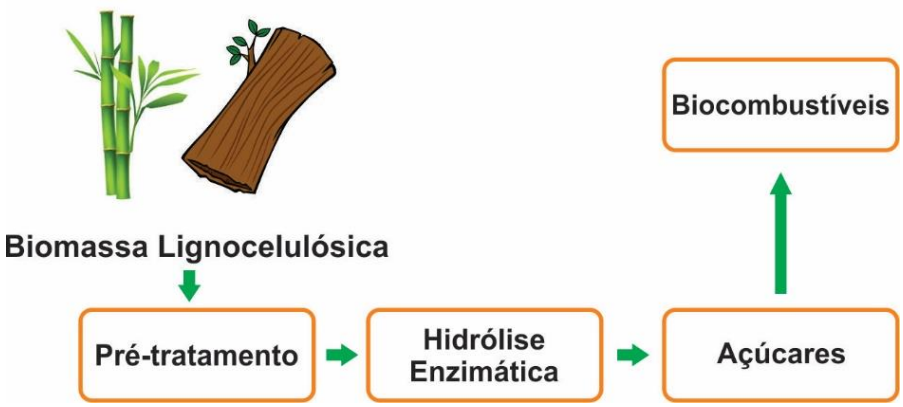

Figura 1: Fluxograma representando de maneira simplificada os processos empregados na conversão da biomassa lignocelulósica contida na parede celular vegetal em glicose para eventual produção de etanol através de fermentação. A biomassa passa por um pré-tratamento biofísico ou químico, desfazendo a estrutura altamente compactada adotada pelos polissacarídeos da parede para melhorar a ação de enzimas que irão então atuar na quebra das ligações glicosídicas gerando glicose.

Apesar de promissores, combustíveis de segunda geração ainda estão longe de ser economicamente viáveis, com as poucas usinas em atividade hoje representando apenas 0,1% do volume total de biocombustíveis produzidos no mundo. Entretanto existe uma enorme tendência de queda de preços para esta tecnologia, uma vez que se trata de uma técnica relativamente recente, com muito espaço para novas descobertas e ganho de experiência por pesquisadores e empresários. Uma vez estabelecidas as tecnologias de produção se espera uma rápida transição dos biocombustíveis de primeira geração para os de segunda geração. (SIMS et al., 2010).

Dentre os fatores travando o pleno desenvolvimento desta tecnologia podem ser citados a melhora das técnicas de pré-tratamento, melhor integração do processo, e principalmente o aprimoramento na eficácia das enzimas celulases utilizadas para a conversão de resíduos da celulose em glicose, configurando a etapa mais cara do processo de produção de bioetanol de segunda geração (ROBAK; BALCEREK, 2018).

2.2.1. Celulose e outros constituintes da parede celular

Presente em todas as células vegetais, a celulose é a principal constituinte da parede celular, estrutura que confere à planta rigidez e resistência ao ataque de patógenos, sendo assim considerada o material biológico mais abundante do

planeta, gerando grande interesse em seu uso como biocombustível (SOMERVILLE, 2004). A celulose é um polissacarídeo formado por ligações β-1-4 de monômeros de glicose, entretanto a biomassa lignocelulósica configura uma estrutura complexa. Além da celulose, outra fonte de carboidratos é a hemicelulose, definida como o conjunto dos carboidratos restantes presentes em associação com a celulose na parede celular, são polissacarídeos extremamente variados quanto a sua composição e propriedades biofísicas, geralmente constituídos por ligações β-1-4 e podem ser divididos em xilanas, xiloglicanas, mananos e glicomananos. Um terceiro componente denominado lignina formado por compostos fenólicos também está presente em grande quantidade na parede celular vegetal (SCHELLER; ULVSKOV, 2010). A celulose, hemicelulose e lignina representam respectivamente 20-50%, 15-35% e 10-30% do peso seco da parede celular de célula vegetal (PAULY; KEEGSTRA, 2008).

Os açúcares contidos na celulose e hemicelulose são fontes para a geração de biocombustíveis, porém para tanto é necessário superar a altíssima recalcitrância destas moléculas (WOLFENDEN; SNIDER, 2001). As ligações O-glicosídicas da celulose apresentam estabilidade de duas a quatro vezes superior às encontradas em ligações peptídicas e no ácido desoxirribonucleico (DNA). Além das forças covalentes, a degradação dessas estruturas também precisa superar sua alta compactação. A etapa de pré-tratamento é importantíssima na resolução deste problema, uma vez que quebra interações intermoleculares entre os constituintes da lignocelulose, diminuindo significativamente a sua compactação e expondo mais cadeias suscetíveis à hidrólise enzimática (PAYNE et al., 2015), conforme ilustrado na Figura 2.

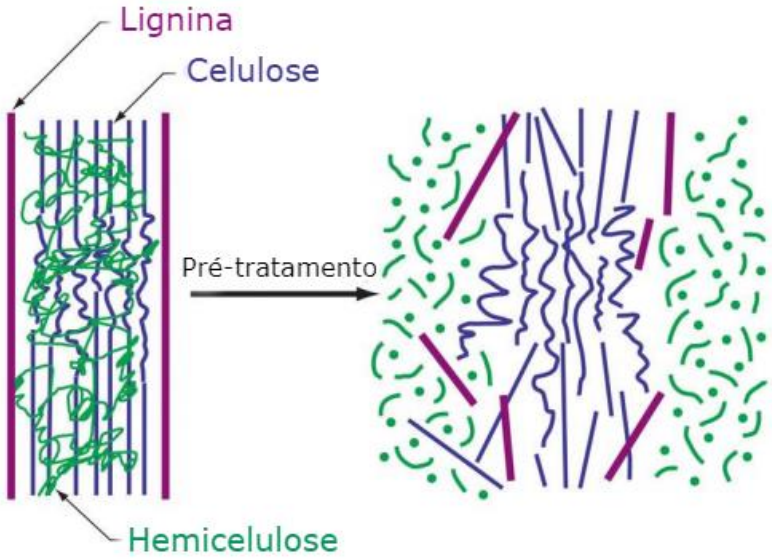

Figura 2: Representação da estrutura da parede celular vegetal ressaltando a conformação compactados e organizada de seus componentes em comparação com sua dispersão após a etapa de pré-tratamento utilizada durante a produção do bioetanol (adaptado de BRANDT et al., 2013).

2.2.2. Enzimas Celulases

As enzimas secretadas por diversos organismos responsáveis pela degradação hidrolítica da celulose são denominadas celulases. O processo ocorre através de três etapas distintas realizadas por celulases específicas, sendo: i) a quebra das microfibrilas de celulose em unidades menores de polissacarídeos glicose através das endoglucanases; ii) a consequente conversão destas cadeias menores em celobiose (um dissacarídeo formado por duas glicoses unidas por ligação β-1-4) através das exoglucanases que age nas extremidades das fibras e finalmente iii) ocorre a hidrólise da celobiose em duas moléculas de glicose pelas β-glicosidades, conforme exemplificado na Figura 3. A variação na concentração e consequente participação de cada tipo de celulase no meio extracelular durante o processo de degradação da celulose é de fundamental importância para a eficiência do processo (HASUNUMA et al., 2013).

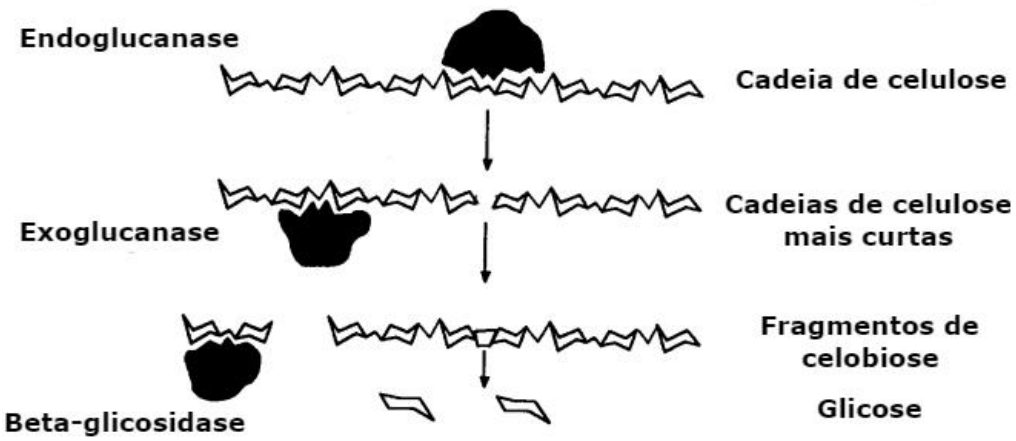

Figura 3: Representação gráfica dos três distintos tipos de celulases e seus mecanismos de ação sobre a degradação da celulose (adaptado de WRIGHT; WYMAN; GROHMANN, 1988).

As celulases são atualmente a terceira enzima mais comercializada no mundo, exercendo papel fundamental em diversos processos de importância econômica, como na indústria têxtil, desde o processamento do algodão até o envelhecimento do jeans, reciclagem de papel, manutenção de cor e brilho de tecidos através de sua adição ao sabão em pó dentre outros (MIETTINEN-OINONEN, Arja et al., 2004). Atualmente os coquetéis de enzimas mais utilizado na indústria são os dos fungos *Trichoderma reesei* e *Aspergillus niger*, por conta de sua resistência a condições industriais hostis, alta secreção e atividade enzimática e experiência por parte da indústria, uma vez que estes são os organismos utilizados a mais tempo em processos de quebra da lignocelulose (BISWAS; PERSAD; BISARIA, 2014; BISCHOF; RAMONI; SEIBOTH, 2016). Atualmente o principal método para a obtenção destas enzimas é através de sua secreção pelos organismos mencionados, entretanto a expressão de celulases e lignocelulases recombinantes é o foco de diversos estudos, havendo sido expressas em diversas espécies de fungos, bactérias e plantas. Mesmo sendo promissora, esta estratégia ainda não alcançou os níveis de produção de cepas industrialmente bem-sucedidas (GARVEY et al., 2013). Apesar de apresentar estabelecida viabilidade comercial para estes processos industriais, o patamar de eficiência ou de expressão requeridos para a degradação da lignocelulose visando a produção de bioetanol ainda não foram atingidos (WILSON, 2009).

Um estudo profundo acerca da viabilidade comercial do bioetanol de segunda geração nos EUA, remanescente da sua era de ouro no início da

década, apontou que a produção do combustível tinha custo aproximado de $0,47 por galão, valor entre duas e três vezes superior ao do etanol de primeira geração produzido a partir do milho, com grande possibilidade de diminuição de custos através da melhora das taxas de conversão enzimática (CURRENT STATE OF THE U S ETHANOL INDUSTRY, 2010). O Brasil encontra-se em melhor situação, estima-se que no ano de 2014 o litro do combustível celulósico custava cerca R$1,50, enquanto o valor do bioetanol de cana-de-açúcar beirava R$1,15. Apesar do sucesso da tecnologia de primeira geração no país, o aprimoramento da conversão da celulose é de grande interesse, tendo em vista a possibilidade de exportação e liderança técnica. Projeções estimam que até 2025 o bioetanol de segunda geração deve tornar-se mais atrativo que a tecnologia de biocombustíveis atualmente empregada, atingindo o preço de R$ 0,75 por litro caso os atuais avanços técnicos, principalmente em relação às enzimas, se mantenham (MILANEZ et al., 2015; FAPESP, 2017).

Diversos esforços vêm sido realizados visando o aperfeiçoamento da atividade destas moléculas, implicando diretamente em diminuição dos custos para os processos de degradação da celulose, uma vez que embora diversos organismos, especialmente fungos e bactérias, sejam capazes de degradar o substrato, poucos o fazem com alta eficiência. Sendo assim, diversas técnicas como mutagênese, engenharia genética e engenharia de proteínas vêm sendo empregadas na tentativa de aumentar a quantidade de enzimas secretados pelos microrganismos de interesse, além disso existem diversos estudos de prospecção com o objetivo de identificar e caracterizar novas enzimas de alta atividade em organismos encontrados no ambiente (WANG et al., 2012). Um fator limitante à atividade das celulases é a concentração de celobiose e glicose no meio, fato observado em diversos ensaios e que provavelmente está relacionado a um controle de sua atividade por excesso de produtos, uma vez que a conversão do açúcar em etanol restitui os níveis prévios de atividade. Acredita-se que esta inibição ocorre por ligação do produto ao sítio ativo das enzimas (ANDRIĆ et al., 2010).

2.2.3. Hemicelulases

Embora muito se discuta sobre a celulose, outro componente da parede celular vegetal, a hemicelulose também é um dos principais componentes desta estrutura. Além de estarem bastante presentes, os diversos polissacarídeos que compõem a hemicelulose são uma valiosa fonte de glicose para a produção de biocombustíveis (WANG et al., 2012). As hemicelulases, enzimas responsáveis pela degradação da hemicelulose também atuam sobre as ligações glicosídicas β-1-4 e costumam ser tão importantes quanto as celulases na degradação da parede celular, uma vez que estes carboidratos podem inibir a plena atividade da celulase, sendo utilizado um coquetel das duas enzimas em grande parte dos processos industriais (OBENG et al., 2017).

2.3. Patógenos de plantas como fonte para a prospecção de enzimas lignocelulósicas com destaque para fungos

Durante a evolução diversos organismos encontraram nas plantas condições ótimas para o seu desenvolvimento, utilizando-as como hospedeiras. Para tanto porém foi necessária a superação de uma barreira robusta: a parede celular e sua altíssima recalcitrância, levando assim ao desenvolvimento através das gerações de diferentes mecanismos de degradação do material lignocelulósico que evoluíram independentemente em distintos ramos da vida, como fungos, bactérias e protistas, originados provavelmente a partir de sapróbios, resultando em um verdadeiro arsenal de diferentes celulases e hemicelulases com distintos padrões de secreção que podem ser úteis à produção de biocombustíveis de segunda geração (HÉMATY; CHERK; SOMERVILLE, 2009).

Dentro deste contexto é importante o destaque para fungos fitopatogênicos, diversos organismos neste grupo demonstram em diferentes estudos padrões de expressão e secreção maiores em comparação aos de espécies de outros grupos taxonômicos. Sendo assim, fungos fitopatogênicos são ótimas fontes para a prospecção de enzimas hidrolisantes de material lignocelulósico com alta atividade catalítica (GIBSON et al., 2011).

2.4. Ferramentas de bioinformática para a análise e prospecção de novas enzimas degradadoras de biomassa lignocelulósica

Apesar de sua indiscutível importância para o aprimoramento dos processos de produção de biocombustíveis, a descoberta de novas enzimas de elevada eficiência catalítica segue sendo um desafio para pesquisadores e para a indústria, outras características desejáveis incluem alta termoestabilidade em temperaturas elevadas, resistência a faixas de pH extremas, conforme as presentes em processos industriais. A estratégia clássica para a descoberta de novas enzimas tem sido o isolamento e testes *in vitro* de atividade, e embora seja eficaz, demanda muito tempo e recursos (TIWARI et al., 2018).

Novos métodos de análise e prospecção de novas enzimas celulases envolvem técnicas de sequenciamento de nova geração, gerando imensas fontes de dados de sequências de genomas, filtrando lignocelulases a partir da subsequente anotação dos genes identificados (DUAN; FENG, 2010). A análise se inicia a partir de um alvo, sendo este um único organismo ou mais comumente a totalidade dos microrganismos presentes em determinada região, o chamado metagenoma (ROUMPEKA et al., 2017). Alvos utilizados para a prospecção de celulases costumam obedecer a um único critério: serem organismos/ambientes associados a alta degradação da biomassa vegetal. Exemplos de alvos em estudos prévios incluem uma biblioteca de cDNA para do *gribble,* um artrópode marinho que costuma se alimentar de madeira, causando danos à estrutura de cais, de onde foram isoladas celulases com alta resistência a salinidade (KERN et al., 2013); o metagenoma de um rúmen bovino (HESS et al., 2011) e o metagenoma de uma amostra de solo da floresta amazônica (BERGMANN et al., 2014).

Devido à impossibilidade da utilização de qualquer método tradicional para a prospecção de lignocelulases para o alvo “fungos fitopatogênicos” em função de sua abrangência, foi adotada uma estratégia alternativa, recorrendo a bancos de dados online de proteomas para a obtenção de possíveis enzimas de organismos previamente sequenciados e anotados. O passo seguinte à descoberta dos genes é a sua expressão proteica e caracterização funcional e estrutural por técnicas diversas. Entretanto, o número de sequências com potencial biotecnológico na geração de combustíveis é elevado, sendo pouco

viável a produção de sua totalidade. Neste contexto, diversas técnicas de bioinformática podem auxiliar no processo de filtragem funcional destas enzimas.

Novas técnicas de análise de proteínas *in silico* como modelagem, *docking*, dinâmica molecular, design de proteínas, e análise de bancos de dados online vêm renovando a maneira como buscamos por novas enzimas de interesse industrial (SELVAM et al., 2017). Embora a qualidade final destas análises varie muito e o resultado para enzimas individuais não substitua ensaios *in vitro* para a plena caracterização da atividade enzimática, a bioinformática pode ser uma ferramenta muito útil para uma pré-seleção de enzimas com maior potencial biotecnológico, uma vez que dispomos de uma enorme quantidade de genomas sequenciados com seus genes anotados e estruturas tridimensionais de enzimas disponíveis (MANZONI et al., 2018). A correta filtração destes dados e utilização de ferramentas adequadas para medir sua atividade como o *docking* molecular podem gerar excelentes candidatos a enzimas de alta atividade (KRISHNARAJ et al., 2017), que deverão ser posteriormente testadas em ensaios comprovando sua atividade, assim tornando o processo de prospecção e caracterização muito mais inteligente e prático.

2.4.1. Modelagem de proteínas

Estruturas tridimensionais de proteínas são fundamentais para a realização de diversos estudos, estes modelos nos permitem não apenas estudar sua conformação e sítios como também possibilita a execução de importantes técnicas *in silico* como *docking* e dinâmica molecular, com inúmeras aplicações em diversas áreas das ciências da vida. Um grave problema enfrentado por pesquisadores que utilizam estruturas 3D em seus estudos é a baixíssima disponibilidade destes modelos, uma vez que técnicas capazes de gerar tais dados como cristalografia de raios-x demandam muito tempo e recursos (ZHANG, 2009).

A modelagem de proteínas surge como uma alternativa a esta escassez de estruturas bem caracterizadas. Uma vez que até o ano de 2009 existiam cerca de apenas 50.000 proteínas com modelos 3D disponíveis, em comparação com os mais de 7 milhões de sequências de nucleotídeos disponíveis devido a sua fácil obtenção, especialmente após o advento das tecnologias de

sequenciamento de DNA, levando à maior utilização desta técnica que visa determinar a estrutura de proteínas cuja sequência primária é conhecida através do reconhecimento de regiões homólogas em proteínas de estrutura terciária ou quaternária conhecidas, chamada modelo (SCHWEDE, 2013; ZHANG, 2009).

Duas principais estratégias são empregadas na montagem destas moléculas: modelagem comparativa, onde se emprega um modelo evolutivamente próximo e *threading*, quando a sequência é alinhada procurando padrões de enovelamento parecidos sem a necessidade de uma origem evolutiva próxima. Também é possível realizar modelagens *ab initio* para proteínas sem modelos adequados conhecidos, visando diminuir a energia livre global conforme visto na Figura 4 (LAM et al., 2017).

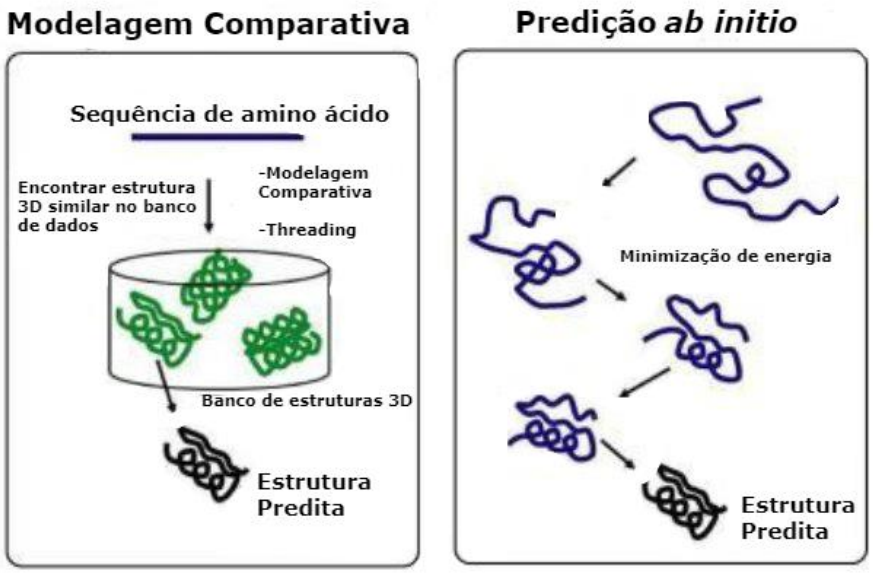

Figura 4: Representação das etapas presentes nos processos de modelagem de proteínas comparativa e *ab initio*. Adaptado de (NARA INSTITUTE OF TECHNOLOGY, 2018) .

.

2.4.2. *Docking* molecular

A técnica de *docking* molecular é uma importante ferramenta de bioinformática muito utilizada pela indústria farmacêutica para a descoberta de novos fármacos, sendo também útil ao estudo de diversas outras áreas que precisem determinar interações biológicas proteína-ligante ou proteína-proteína (LIONTA et al., 2014). Esta ferramenta baseia-se em simulações a partir de modelos tridimensionais para aferir o grau de afinidade entre duas moléculas,

executando testes para as possíveis interações bem como conformações que podem ser adotadas. O *docking* possui assim dois objetivos principais: caracterizar estruturalmente a interação do complexo formado bem como quantificar sua energia livre, determinando a afinidade da ligação entre as partes envolvidas.

Mesmo para o cálculo da interação entre uma proteína e um ligante simples, algoritmos que utilizem a estrutura inteira da proteína levam a um número de conformações e interações possíveis muito alto, podendo gerar resultados pouco específicos além de demandar um elevado custo computacional (GRINTER; ZOU, 2014). Sendo assim, é importante conhecer o sítio ativo da proteína através de modelos cuja estrutura tridimensional ou a de proteínas semelhantes em complexo com o ligante tenha sido determinada por técnicas como cristalografia de raios-x ou espectroscopia de ressonância magnética nuclear (NMR), possibilitando limitar o cálculo do *docking* apenas à área de interesse (KITCHEN et al., 2004).

No estudo da degradação da celulose e polissacarídeos associados uma maior afinidade do sítio ativo da enzima pelo substrato está diretamente relacionada a uma maior atividade catalítica da molécula quando submetida a testes *in vitro* (KRISHNARAJ et al., 2017), sendo assim a utilização do *docking* molecular para fins de prospecção de enzimas celulase é uma estratégia eficiente.

3. Objetivos

3.1. Objetivos gerais

O trabalho tem como objetivo a busca por novas enzimas celulases de alta eficiência, cuja sequência de aminoácidos tenha sido depositada em bancos de dados online, buscando modela-las a partir de modelos já caracterizados por técnicas de cristalografia de raios-x ou NMR a fim de gerar modelos 3D de alta estabilidade. Estas estruturas terão como objetivo performar ensaios *in silico* de atividade através de *docking* molecular, revelando a afinidade destas celulases por seus respectivos substratos.

3.2. Objetivos específicos

1. Catalogar espécies de fungos fitopatogêncios cujos proteomas estejam disponíveis em bancos de dado.

2. Encontrar sequências que apresentem alta similaridade para com celulases nestes organismos.

3. Modelar as sequências encontradas, utilizando os modelos tridimensionais gerados para ensaios de *docking* molecular.

4. Comparar as enzimas celulases de maior afinidade pelo substrato com a de *Trichoderma reesei*.

5. Identificar pelo menos 1 enzima para posterior expressão em sistema nativo ou heterologo.

4. Metodologia

4.1. Identificação de fungos fitopatogênicos com proteomas depositados em bases de dados online

Para o primeiro passo deste trabalho foi prospectada uma base de dados contendo todas as espécies taxonomicamente classificadas como pertencendo ao reino fungi, utilizando o catálogo desenvolvido pela Comissão Internacional da Taxonomia dos Fungos (http://www.fungaltaxonomy.org/index.php/download_file/view/132/107/).

Em seguida foi realizada uma busca no Uniprot (UNIPROT, 2018), uma coleção de banco de dados para sequências de proteínas e anotação detalhada de genomas, com atualmente cerca de 60 milhões de sequências depositadas (THE UNIPROT CONSORTIUM, 2017). Foi feita uma pesquisa por proteomas de fungos utilizando o termo "*taxonomy:"Fungi [4751]"* ". Por fim foi utilizado um script simples em linguagem de programação Python 2.7 (PYTHON, 2018) para extrair os nomes das espécies presentes nas duas listas, sobrepondo-as para gerar um documento final contendo o nome de todas as espécies de fungos fitopatogêncos cujos proteomas se encontram disponíveis.

4.2. Prospecção de celulases

Após a identificação dos organismos fitopatogênicos se fez necessária a identificação de possíveis enzimas celulases dentre as milhares de proteínas anotadas para cada proteoma. Para isso foram adotadas duas estratégias experimentais distintas. A primeira abordagem consistiu na procura destas proteínas contra um banco de dados formado por modelos depositados no *Research Collaboratory for Structural Bioinformatics Protein Data Bank* (RCSB PDB) (PDB, 2018), enquanto a segunda utilizou um alinhamento múltiplo do Pfam (PFAM, 2018).

4.2.1. Identificação de celulases a partir de um banco de dados formado por proteínas depositadas no Protein Data Bank

O PDB é um formato de arquivo utilizado para guardar a um nível atômico as informações relativas à estrutura tridimensional de macromoléculas biológicas como DNA, ácido ribonucleico (RNA) e proteínas obtidas através de técnicas de determinação de estrutura, estando todos estes arquivos depositados no RCSB PDB (ROSE et al., 2017). Visando a otimização de futuras etapas de modelagem foi montado um banco de dados somente com a sequência de aminoácidos de proteínas resultantes da busca pelo termo "*cellulase*", restrito apenas a organismos eucariotos e com resolução atômica igual ou inferior a 2.0 angström.

Em seguida foi realizado um alinhamento local de regiões homólogas da sequência peptídica dos proteomas contra o banco de dados criado por meio da ferramenta BLASTp (WHEELER; BHAGWAT, 2007), gerando um arquivo com diversos alinhamentos que foi então filtrado por escore (um parâmetro que descreve a chance de dado alinhamento ocorrer por acaso) inferior a 1e-100 e homologia da sequência superior a 65%, retornando apenas as proteínas de maior similaridade e possibilidade de modelagem.

4.2.2. Identificação de celulases a partir de alinhamento múltiplo através do Pfam

Visando efetuar uma busca mais ampla (uma vez que a busca por celulases no RCSB PDB não retorna todas as enzimas celulases depositadas no site por conta da falta de uma completa padronização e indexação das submissões) optou-se por realizar uma segunda busca utilizando um alinhamento múltiplo para celulases. Para isto utilizou-se o Pfam, um banco de famílias proteicas curadas (FINN et al., 2014). O alinhamento múltiplo gerado foi então processado e alinhado contra os proteomas pela ferramenta HMMER (FINN; CLEMENTS; EDDY, 2011), cujos resultados foram filtrados por escore menor que 1e-50 e similaridade da sequência superior a 60%. Um esquema simplificado desta etapa pode ser observado na Figura 5.

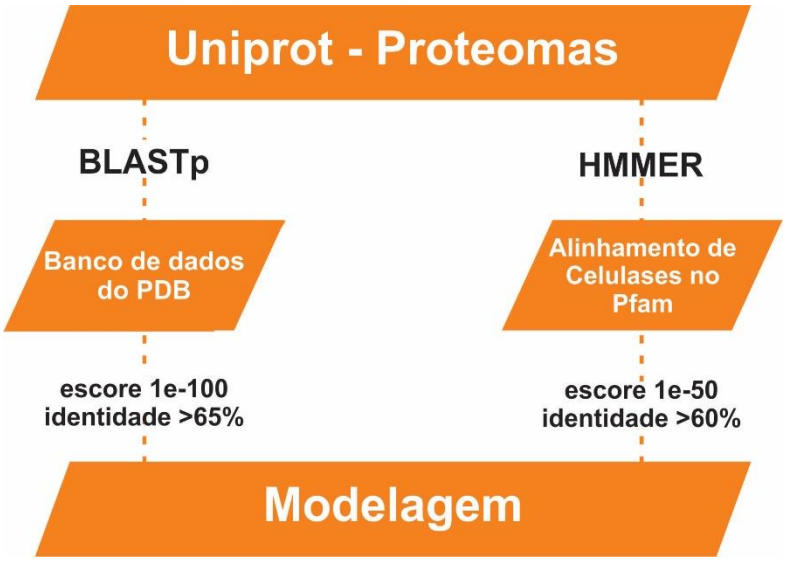

Figura 5: Modelo de prospecção de celulases funcionais a partir de proteomas depositados no Uniprot. Foram elaboradas duas estratégias distintas visando maximizar o número de enzimas encontradas, através de algoritmos de alinhamento contra um banco de dados de enzimas do PDB e contra um alinhamento múltiplo de celulases do Pfam, sendo os resultados filtrados por escore e identidade.

4.3. Modelagem de proteínas testes de estabilidade estrutural

As sequências de aminoácidos selecionadas foram submetidas à ferramenta SignalP para a identificação e excisão de regiões correspondentes ao peptídeo sinal (PETERSEN et al., 2011). As possíveis celulases tiveram então sua estrutura tridimensional modelada através de duas ferramentas: Swiss-Model e RaptorX, ambos são servidores automáticos online de modelagem através de uma proteína modelo previamente caracterizada e depositada, foram escolhidos por conta de sua praticidade e rapidez, uma vez que se fez necessária a predição de estrutura para diversas enzimas (PENG; XU, 2011; SCHWEDE et al., 2003).

Para a avaliação de qualidade e estabilidade conformacional das estruturas geradas foram utilizadas pontuações de QMEAN, automaticamente calculadas para proteínas geradas pelo Swiss-model e 3D1D da ferramenta Verify3D para os modelos gerados pelo Swiss-model e RaptorX. Ambos cálculos visam determinar a repulsão entre as cadeias laterais dos aminoácidos, indicando assim a probabilidade de determinadas conformações ocorrerem (EISENBERG; LÜTHY; BOWIE, 1997; SOMERVILLE, 2004). A pontuação mínima considerada para o controle de qualidade do QMEAN foi de -4.0,

enquanto o parâmetro mínimo utilizado para o parâmetro 3D1D da ferramenta Verify3D foi de 0.2 para que a posição da cadeia lateral do aminoácido fosse estável, sendo necessário que no mínimo de 70% dos resíduos possua este valor. Outro controle de qualidade consistiu em excluir proteínas cujo valor de identidade com a sequência modelo fosse inferior a 60%.

4.4. *Docking* molecular

Foram gerados diversos modelos tridimensionais para duas enzimas degradadoras de material lignocelulósico a partir das ferramentas de modelagem de proteínas: endoglucanases e mananases. A primeira etapa para a realização do *docking* molecular foi a escolha dos ligantes. As endoglucanases clivam as cadeias de celulose indiscriminadamente, sendo assim possível utilizar polímeros de glicose com diferentes números de ligações β-1-4. Optou-se pela celotetraose, carboidrato comporto de quatro glicoses, pois um ligante menor não iria representar adequadamente as interações presentes entre a enzima e uma molécula de celulose, enquanto uma cadeia com número maior de glicoses elevaria em muito o tempo de processamento requerido pelas simulações. Para a mananase se utilizou uma molécula de manano. As estruturas 3D dos ligantes foram obtidas através do banco de dados de moléculas Pubchem (THE PUBCHEM PROJECT, 2018), PubChemCID: 439626 e 870.

Para a localização do sítio ativo da proteína foram utilizados os arquivos PDB da estrutura de uma enzima endoglucanase de *Fusarium oxysporum* em complexo com celobiose, provavelmente atuando na inibição de sua atividade, entretanto sabe-se que esta ação ocorre por meio da interação do inibidor diretamente com o sítio funcional da enzima, podendo ser utilizado para fins de simulação de *docking (*KRISHNARAJ et al., 2017*)*. Para a mananase foi utilizada uma estrutura de *Amphibacillus xylanus* em complexo com a manose, carboidrato cuja cadeia de ligações β-1-4 formam o manano (PDB ID: 2OVW e 5YLI). Por fim também foram realizados ensaios de *docking* molecular utilizando uma endoglucanase e mananase da espécie *Trichoderma reesei* (PDB ID: 1EG1 e 1QNR) a fim de comparar a afinidade pelo substrato das novas enzimas prospectadas com as deste organismo atualmente utilizado no mercado.

O *docking* foi executado pela ferramenta AutoDock Vina por conta de sua rapidez e precisão comparada a outros software do tipo (TROTT; OLSON, 2010). As simulações foram restritas a regiões específicas das enzimas previamente identificadas como pertencendo ao sítio ativo, de forma a economizar o tempo total de computação. Ambas as *grid box* (caixa que delimita a área e propriedades da simulação) foram desenhadas com tamanho X = 80, Y = 80 e Z = 80; enquanto as coordenadas espaciais para as endoglucanases foram X = 50, Y = 70 e Z = 35 e para as mananases X = -20, Y = -10 e Z = 40, conforme ilustrado na Figura 6. Para a representação visual da interação entre as proteínas e os ligantes bem como o desenho da *grid box* foi utilizada ferramenta gráfica PyMOL (YUAN; CHAN; HU, 2017).

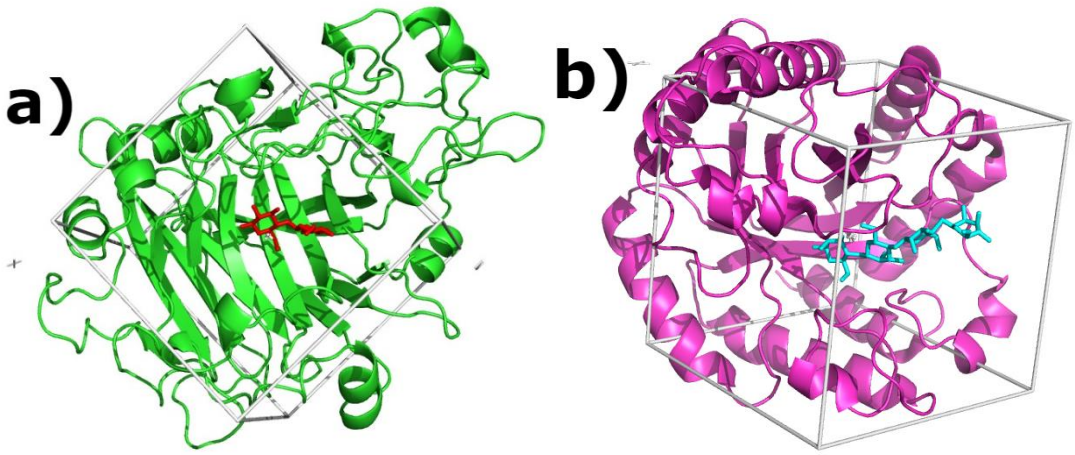

Figura 6: Representação gráfica da estrutura tridimensional das enzimas modelo utilizadas para desenhar a *grid box*. a) endoglucanase ligada a uma molécula de celobiose b) mananase em complexo com quatro moléculas de manose.

5. Resultados

A busca por proteomas de fungos no Uniprot retornou 1061 resultados, que após cruzados com os nomes dos organismos do catálogo da Comissão Internacional da Taxonomia dos Fungos gerou uma lista com o nome de 52 espécies de fungos fitopatogênicos cujos proteomas estão disponíveis, citadas na Tabela 1:

Tabela 1: Lista de espécies de fungos fitopatogênicos com proteomas depositados no uniprot e seus respectivos Ids de organismo. Espécies com mais de um ID possuem mais de um genoma de referência.

Espécie	ID de organismo do Uniprot
Alternaria alternata	5599
Armillaria ostoyae	22192
Ashbya gossypii	284811
Aureobasidium pullulans	1043002
Bipolaris oryzae	930090
Bipolaris victoriae	12879
Blumeria graminis	1268274 e 546991
Botryotinia fuckeliana	1290391, 999810 e 425009
Ceratocystis fimbriata	1035309 e 88771
Cercospora beticola	12433
Chaetomium globosum	306901
Choanephora cucurbitarum	101091
Claviceps purpurea	1111077
Cochliobolus heterostrophus	701091
Cochliobolus sativus	665912
Colletotrichum gloeosporioides	1237896 e 1213859
Colletotrichum graminicola	645133
Corynespora cassiicola	17125
Diaporthe helianthi	158607
Didymella rabiei	5454
Dothistroma septosporum	675120
Fusarium oxysporum	426428, 1229664, 1229665, 1089458, 660025, 327505 e 5507
Gaeumannomyces graminis	644352

Espécie	ID de organismo do Uniprot
Geotrichum candidum	1173061
Grifola frondosa	5627
Leptosphaeria maculans	985895
Macrophomina phaseolina	1126212 e 1527523
Magnaporthe oryzae	242507, 1580607, 271257 e 441999
Melampsora larici-populina	747676
Moniliophthora perniciosa	554373
Moniliophthora roreri	1381753 e 221103
Mucor ambiguus	91626
Mucor circinelloides	1220926 e 747725
Neonectria ditíssima	78410
Ophiostoma piceae	1262450
Phaeosphaeria nodorum	321614
Pseudocercospora fijiensis	383855
Pseudocercospora musae	113226
Pyrenophora tritici-repentis	426418
Rhizoctonia solani	45151
Rhynchosporium secalis	38038
Schizophyllum commune	578458
Sclerotinia sclerotiorum	665079
Sporisorium scitamineum	49012
Taphrina deformans	1097556
Thielaviopsis punctulata	72032
Tilletia controversa	13291
Tilletia indica	43049
Ustilago maydis	237631
Uncinula necator	52586
Verticillium dahliae	498257
Zymoseptoria tritici	498257

O banco de dados construído a partir das enzimas celulases depositadas no RCSB PDB teve um total de 60 enzimas, o alinhamento por BLASTp nos parâmetros previamente citados retornou 20 sequências. O alinhamento múltiplo do Pfam para um modelo de celulase utilizou como base 15602 sequências de

aminoácidos, resultando em 38 proteínas que foram alinhadas contra os proteomas pela ferramenta.

A ferramenta online de montagem RaptorX gerou estruturas de baixa qualidade quando analisada por verificadores de estabilidade proteica, além disso as proteínas estavam visivelmente comprometidas estruturalmente (resultados não mostrados). Todavia, as estruturas tridimensionais montadas pela ferramenta Swiss-model tiveram excelentes resultados de qualidade, gerando enzimas endoglucana e mananases, das quais grande parte esteve apta a ser utilizada nos ensaios de *docking* molecular (Tabelas 2 e 3). Além disso foi possível verificar a acurácia das duas distintas técnicas de prospecção abordadas. Todas as sequências buscadas contra o banco de dados construído a partir do RCSB PDB estiveram aptas ao docking. Apenas 8 das 38 enzimas prospectadas através do cruzamento de dados do Pfam (além das já encontradas pelo primeiro método) apresentaram boa estrutura, estas, entretanto de celulases e hemicelulases diversas por conta da altíssima abrangência de enzimas classificadas sob o termo celulase pelo Pfam.

Tabela 2: Lista de sequências de aminoácidos selecionadas pelo método de alinhamento BLASTp contra banco de dados do PDB cujos modelos tridimensionais gerados pelo Swiss-model passaram pelos seguintes controles de qualidade: QMEAN ≥ -4,0; Identidade com a estrutura modelo ≥ 60% e parâmetro 3D1D da ferramenta VERIFY3D ≥ 70%, estando aptos aos ensaios de *Docking* molecular.

ID UNIPROT	QMEAN	IDENTIDADE MODELO	VERIFY3D 3D1D	TIPO	DOCKING
E5A2B2	**-1,32**	**64,04%**	**100%**	**Endoglucanase**	**APTO**
K2RE11	**-0,83**	**69,84%**	**94%**	**Endoglucanase**	**APTO**
G2XV26	**-0,97**	**78,83%**	**100%**	**Endoglucanase**	**APTO**
A0A0C4DHQ3	**0,06**	**99,51%**	**98%**	**Endoglucanase**	**APTO**
N4UKU3	**-1,16**	**65,89%**	**94%**	**Endoglucanase**	**APTO**
N1RUE4	**-1,27**	**65,89%**	**94%**	**Endoglucanase**	**APTO**
N1S229	**0,11**	**72,04%**	**100%**	**Endoglucanase**	**APTO**
X0BJP8	**-1,23**	**65,55%**	**94%**	**Endoglucanase**	**APTO**
X0J1P8	**-0,09**	**70,89%**	**100%**	**Endoglucanase**	**APTO**
X0MNI9	**0,12**	**72,04%**	**100%**	**Endoglucanase**	**APTO**
W9PEE9	**0,08**	**99,27%**	**98%**	**Endoglucanase**	**APTO**
W9K2M9	**0,19**	**72,51%**	**100%**	**Endoglucanase**	**APTO**
W9JU27	**-1,23**	**65,55%**	**94%**	**Endoglucanase**	**APTO**
W9K9E4	**0,22**	**71,36%**	**100%**	**Endoglucanase**	**APTO**
W6ZSU6	**-0,75**	**66,17%**	**100%**	**Endoglucanase**	**APTO**
A0A177D8G6	**-1,29**	**66,34%**	**100%**	**Endoglucanase**	**APTO**
A0A2P5HZN4	**-1,46**	**64,04**	**99%**	**Endoglucanase**	**APTO**
A0A1E1MUR1	**-1,78**	**65,33%**	**95%**	**Endoglucanase**	**APTO**
A0A1E1LX78	**-0,62**	**69,09%**	**99%**	**Endoglucanase**	**APTO**
A0A2J8EMS2	**0,21**	**71,43%**	**100%**	**Endoglucanase**	**APTO**

Tabela 3: Lista de sequências de aminoácidos selecionadas pelo método de alinhamento pela ferramenta HMMER contra um alinhamento múltiplo de celulases do Pfam, cujos modelos tridimensionais gerados pelo Swiss-model passaram pelos seguintes controles de qualidade: QMEAN ≥ -4,0; Identidade com a estrutura modelo ≥ 60% e parâmetro 3D1D da ferramenta VERIFY3D ≥ 70%, estando aptos aos ensaios de *Docking* molecular.

ID UNIPROT	QMEAN	IDENTIDADE MODELO	VERIFY3D 3D1D	TIPO	DOCKING
G4N1K6	-6,33	18,61%	-	Endoglicoceramidase	NÃO APTO
A0A0B7FSB3	-3,88	17,01%	-	Exoglucanase	NÃO APTO
M5C2A7	-3,96	18,24%	-	Galactosidase	NÃO APTO
A0A074RQ48	-4,27	19,86%	-	Galactosidase	NÃO APTO
A0A0B7FQC5	-3,62	19,05%	-	Galactosidase	NÃO APTO
W9K8P3	-5,82	20,51%	-	Glicosidase	NÃO APTO
X0K505	-5,61	20,19%	-	Glicosidase	NÃO APTO
A0A074RU01	-9,87	25,95%	-	-	NÃO APTO
X0HHG5	-4,52	22,93%	-	-	NÃO APTO
A0A1C7MFD6	-6,13	24,86%	-	Exoglucanase	NÃO APTO
A0A0B7FFE5	-4,15	23,45%	-	Endocelulase	NÃO APTO
F9X9U3	-4,94	22,15%	-	Glicosidase	NÃO APTO
A0A1X7RSL7	-3,82	20,33%	-	Exoglucanase	NÃO APTO
A0A074YMU1	-5,21	22,80%	-	Exoglucanase	NÃO APTO
V2X3X3	-2,11	47,31%	-	Mananase	NÃO APTO
A0A2G5I3V5	-4,78	22,15%	-	Glicosidase	NÃO APTO
W9ZQ23	-4,8	20,52%	-	Exoglucanase	NÃO APTO
X0K4T1	-3,62	19,05	-	Galactosidase	NÃO APTO
X8JKS9	-2,66	44,09%	-	Mananase	NÃO APTO
A0A284RUZ3	-1,95	20,52%	-	Mananase	NÃO APTO
G2YMG1	-5,20	23,13%	-	-	NÃO APTO
Q2H1Y9	-0,04	79,29%	94%	Endoglucanase	APTO
K2RHF8	-6,15	22,15%	-	Glicosidase	NÃO APTO
G4NHU4	-5,06	22,15%	-	Exoglucanase	NÃO APTO
A0A284RI75	-3,84	16,20%	-	Exoglucanase	NÃO APTO
A0A177DAP8	-0,89	60,19%	100%	Endoglucanase	APTO

ID UNIPROT	QMEAN	IDENTIDADE MODELO	VERIFY3D 3D1D	TIPO	DOCKING
G4NI55	-1,14	67,29%	98%	Mananase	APTO
A0A0B7FHC3	-2,26	27,54%	-	Endoglucanase	NÃO APTO
A0A0B7FT22	-1,21	45,24%	-	Mananase	NÃO APTO
A0A2J8C779	-1,05	66,84%	95%	Mananase	APTO
V2WCT7	-4,32	21,14%	-	-	NÃO APTO
A0A177DSV3	-1,05	70,53%	100%	Mananase	NÃO APTO
A0A0W0G2R0	-3,8	38,94%	-	Mananase	NÃO APTO
Q2GMU6	-0,87	71,43%	98%	Mananase	APTO
N1PMD5	-1,49	65,53%	98%	Mananase	APTO
A0A177E1M2	-0,67	63,25%	100%	Mananase	APTO

Os resultados individuais referentes à energia livre para cada complexo formado durante os ensaios de *docking* molecular, incluindo os das enzimas de *Trichoderma reesei* podem ser visualizados na Tabela 4 (endoglucanases) e Tabela 5 (mananases). Pelo menos duas glucanases pertencentes à espécie *Fusarium oxysporum* possuem afinidade pela celibiose semelhante à apresentada pela mesma enzima de *T. reesei* e cinco demonstram afinidade muito próxima. Duas mananases pertencentes às espécies *Aureobasidium pullulans* e *Alternaria alternata* apresentaram afinidade próxima à da hemiceluase de *T. reesei*. Uma esquematização dos principais resultados encontrados pode ser observada na Figura 7.

Tabela 4: Resultados dos ensaios de *docking* molecular para endoglucanases prospectadas a partir de proteomas do Uniprot e endoglucanase de *T. reesei* ao final.

Uniprot ID	Espécie	Energia livre (kcal/mol)
A0A0C4DHQ3	*Fusarium oxysporum* f. sp.	-8,6
W9PEE9	*Fusarium oxysporum* f. sp. pisi	-8,6
A0A1E1LX78	*Rhynchosporium secalis*	-7,8
K2RE11	*Macrophomina phaseolina* (strain MS6)	-7,7
G2XV26	*Botryotinia fuckeliana* (strain T4)	-7,6
A0A177DAP8	*Alternaria alternata*	-7,4
A0A2P5HZN4	*Diaporthe helianthi*	-7,3
N1RUE4	*Fusarium oxysporum* f. sp. Cubense	-3,8
A0A1E1MUR1	*Rhynchosporium secalis*	-1,9
W9K9E4	*Fusarium oxysporum* Fo47	0,0
W9K2M9	*Fusarium oxysporum* Fo47	0,0
X0J1P8	*Fusarium oxysporum* f. sp. Conglutinans	0,0

Uniprot ID	Espécie	Energia livre (kcal/mol)
N1S229	*Fusarium oxysporum* f. sp. Cubense	0,0
N1RUE4	*Fusarium oxysporum* f. sp. Cubense	0,0
Q2H1Y9	*Chaetomium globosum*	0,0
E5A2B2	*Leptosphaeria maculans*	0,0
W6ZSU6	*Bipolaris oryzae* ATCC 44560	0,0
A0A177D8G6	*Alternaria alternata*	0,0
Q2GU45	*Chaetomium globosum*	0,0
-	***Trichoderma reesei***	**-8,6**

Tabela 5: Resultados dos ensaios de *docking* molecular para mananases prospectadas a partir de proteomas do Uniprot e mananase de *T. reesei* ao final.

Uniprot ID	Espécie	Energia livre (kcal/mol)
A0A074XP68	*Aureobasidium pullulans*	-7,4
A0A177E1M2	*Alternaria alternata*	-7,4
Q2GMU6	*Chaetomium globosum*	-5,3
A0A177DSV3	*Macrophomina phaseolina* (strain MS6)	-5,2
A0A2J8C779	*Alternaria alternata*	-4,8
G4NI55	*Magnaporthe oryzae*	-4,4
N1PMD5	*Dothistroma septosporum*	-4,1
-	***Trichoderma reesei***	**-7,5**

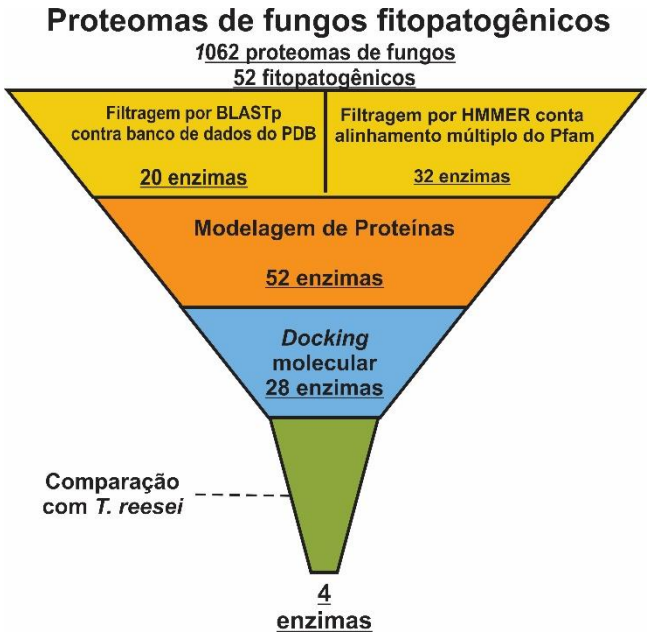

Figura 7: Esquematização dos principais resultados encontrados, desde a filtragem dos proteoma até a descoberta de duas enzimas endoglucanases e mananases com afinidade pelo substrato semelhante à de enzimas atualmente utilizadas pela indústria.

6. Discussão

As duas abordagens distintas empregadas com o objetivo de prospectar celulases e hemicelulases de fungos fitopatogênicos foram bem-sucedidas, apesar dos resultados distintos gerados por cada método, ambos foram essenciais para a exploração dos proteomas e sua filtragem.

Primeiramente a busca por enzimas a partir de um banco de celulases coletadas no RCSB PDB garantiu resultados mais sólidos, com taxa de 100% das enzimas encontradas estando aptas ao *docking*. Resultado esperado, uma vez que estes modelos tridimensionais foram os mesmos utilizados subsequentemente pelo Swiss-model para a etapa de modelagem. Apesar disto, a utilização do Pfam foi imprescindível para o resultado final, abrindo um amplo leque de oportunidades de busca, uma vez que o seu banco de mais de 15 mil sequências de proteínas classificadas como pertencentes à família das celulases. Isto permitiu o retorno de resultados mais diversos, como todos os três tipos de celulases e diversas hemicelulases, ainda que em geral menos específicos e com menor taxa de aproveitamento para o *docking*. Fica claro, entretanto que a maior limitação para estas abordagens, especialmente utilizando o Pfam é a falta de proteínas como exoglucanases, betaglicosidases e diversas hemicelulases devidamente caracterizadas por métodos biofísicos e depositadas em formato PDB em plataformas online.

A escolha de fungos fitopatogênicos como alvo para a procura por novas celulases e hemicelulases de alta atividade mostrou-se até então excelente, foram encontradas sete endoglucanases com afinidade ao substrato semelhante ou igual a do fungo *Trichoderma reesei*, amplamente utilizado na indústria (BISCHOF; RAMONI; SEIBOTH, 2016). Além disso foram identificados ao menos duas mananases com afinidade semelhante às deste mesmo organismo. Por outro lado, algumas enzimas apresentaram energia de ligação igual a zero ao substrato, longe de ser um resultado esperado. Isto provavelmente se deu por alguma falha no *docking* da proteína ou no modelo montado pela ferramenta Swiss-model.

O principal resultado deste trabalho é a lista de enzimas gerada, cabendo a estas sua expressão em sistema nativo ou heterologo para estudos futuros

quanto a sua atividade *in vitro* e potencial como insumo biotecnológico. A endoglucanase de maior afinidade foi encontrada para o organismo *Fusarium oxysporum*, uma espécie de fungo cuja produção de celulases ainda é pouco explorada. Um estudo utilizando cinco cepas distintas de *F. oxysporum* extraídas a partir de plantas infectadas registrou a atividade enzimática máxima das celulases secretadas por estes organismos entre as temperaturas de 60-65ºC, porém ainda funcional entre a faixa de 4 e 50ºC. Estas enzimas são igualmente bem adaptadas a uma ampla faixa de pH, atingindo sua taxa de conversão máxima entre 4,5 e 5,5, mas também mantendo sua atividade até 9,5 (YUAN et al., 2012), características estas desejáveis em enzimas utilizadas em processos industriais.

Quanto ao restante das espécies, a literatura acerca da atividade de celulases para fins industriais é extremamente escassa ou inexistente, havendo o trabalho de prospecção sinalizado pelo menos cinco novas espécies com potencial utilização biotecnológica de celulases, que não apenas possuem alta afinidade pelo substrato, mas também tendência a uma expressiva produção destas enzimas por conta de sua natureza fitopatogênica.

Dentre as duas mananases encontradas e os seus respectivos organismos de origem, a espécie *Aureobasidium pullulans* já é conhecida como uma ótima fonte para a enzima. Uma triagem com diversos fungos encontrou em cepas de *A. pullulans* as melhores produtoras de mananases (CHI et al., 2009). Este é um excelente resultado, uma vez que temos a comprovação da elevada afinidade da enzima pelo manano *in silico* e de sua alta secreção *in vitro*, isto além de caracterizar uma possível fonte de enzimas lignocelulósicas, também ajuda a validar a hipótese de fungos fitopatogênicos como alvo para a descoberta destas enzimas. A literatura para mananases de *Alternaria alternata* é praticamente inexistente, configurando uma fonte completamente nova de possíveis enzimas hemicelulases.

Ademais, a utilização deste novo método não está limitada somente ao conjunto dos fungos fitopatogênicos, podendo ser aplicada a outros grupos abrangentes onde a utilização de técnicas de análise de metagenoma não é possível, como organismos extremófilos, decompositores e até fungos em geral, sendo este também um resultado importante e promissor do trabalho.

7. Conclusão

A metodologia empregada foi bem-sucedida, sendo capaz de a partir de 1000 proteomas contendo milhares de proteínas filtrar, modelar e realizar ensaios de *docking* para gerar modelos tridimensionais para ao menos dez enzimas com afinidade ao substrato próxima ou igual a de enzimas secretadas pela *T. reesei*, organismo de estabelecido uso comercial. Concebendo não apenas modelos de celulases e hemicelulases de alto potencial industrial, mas também um protocolo para a prospecção de enzimas degradadoras de material lignocelulósico a partir de fontes gigantes de dados.

Este modelo, porém, ainda pode e deverá ser melhorado. Perspectivas futuras incluem modificações no protocolo utilizado para o *docking* molecular, a fim de garantir que todas as proteínas se liguem corretamente ao substrato. Apesar de a literatura indicar uma forte tendência de altos padrões de expressão nativa e estabilidade estrutural para estas enzimas, é imprescindível a elaboração de um modelo para a submissão das celulases a testes de termoestabilidade por dinâmica molecular e a realização de testes *in vitro* utilizando as enzimas encontradas pelo estudo. Podendo assim validar suas capacidades funcionais como produtos biotecnológicos de alta eficiência para a produção de bioetanol de segunda geração. Outra perspectiva inclui o estabelecimento da melhor estratégia para a obtenção destas proteínas em alta quantidade, verificando seus padrões de expressão nativa e recombinante, definindo o melhor organismo para este fim.

8. Referências

ANDRIĆ, P. et al. Reactor design for minimizing product inhibition during enzymatic lignocellulose hydrolysis: I. Significance and mechanism of cellobiose and glucose inhibition on cellulolytic enzymes. **Biotechnology Advances**, v. 28, n. 3, p. 308–324, 2010.

BERGMANN, J. C. et al. Discovery of two novel β-glucosidases from an Amazon soil metagenomic library. **FEMS Microbiology Letters**, v. 351, n. 2, p. 147–155, 2014.

BISCHOF, R. H.; RAMONI, J.; SEIBOTH, B. Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei. **Microbial Cell Factories**, v. 15, n. 1, p. 106, 2016.

BISWAS, R.; PERSAD, A.; BISARIA, V. S. Production of Cellulolytic Enzymes. In: BISARIA, V. S.; KONDO, A. (Eds.). **Bioprocessing of Renewable Resources to Commodity Bioproducts**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014. p. 105–132.

BRANDT, A. et al. Deconstruction of lignocellulosic biomass with ionic liquids. **Green Chemistry**, v. 15, n. 3, p. 550, 2013.

CHI, Z. et al. Bioproducts from Aureobasidium pullulans, a biotechnologically important yeast. **Applied Microbiology and Biotechnology**, v. 82, n. 5, p. 793–804, 2009.

Current State of the US Ethanol Industry. p. 48, 2010.

DIAS, M. O. S. et al. Second generation ethanol in Brazil: Can it compete with electricity production? **Bioresource Technology**, v. 102, n. 19, p. 8964–8971, 2011.

DOORNBOSCH, R.; STEENBLIK, R. Biofuels: Is the cure worse than the disease? p. 38, 2008.

DUAN, C.-J.; FENG, J.-X. Mining metagenomes for novel cellulase genes. **Biotechnology Letters**, v. 32, n. 12, p. 1765–1775, 2010.

EISENBERG, D.; LÜTHY, R.; BOWIE, J. U. [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. In: **Methods in Enzymology**. Macromolecular Crystallography Part B: Academic Press, 1997. v. 277p. 396–404.

FAPESP, 2017. Disponível em: <http://agencia.fapesp.br/etanol-de-segunda-geracao-podera-ser-economicamente-viavel-a-partir-de-2025/26272/>. Acesso em: 8 dez. 2018.

FARGIONE, J. et al. Land Clearing and the Biofuel Carbon Debt. **Science**, v. 319, n. 5867, p. 1235–1238, 2008.

FINN, R. D. et al. Pfam: the protein families database. **Nucleic Acids Research**, v. 42, n. Database issue, p. D222-230, 2014.

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

GARVEY, M. et al. Cellulases for biomass degradation: comparing recombinant cellulase expression platforms. **Trends in Biotechnology**, v. 31, n. 10, p. 581–593, 2013.

GIBSON, D. M. et al. Plant pathogens as a source of diverse enzymes for lignocellulose digestion. **Current Opinion in Microbiology**, v. 14, n. 3, p. 264–270, 2011.

GRINTER, S.; ZOU, X. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. **Molecules**, v. 19, n. 7, p. 10150–10176, 2014.

HASUNUMA, T. et al. A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. **Bioresource Technology**, v. 135, p. 513–522, 2013.

HÉMATY, K.; CHERK, C.; SOMERVILLE, S. Host-pathogen warfare at the plant cell wall. **Current Opinion in Plant Biology**, v. 12, n. 4, p. 406–413, 2009.

HESS, M. et al. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. **Science**, v. 331, n. 6016, p. 463–467, 2011.

HIRANI, A. H. et al. A Review on First- and Second-Generation Biofuel Productions. In: KUMAR, A.; OGITA, S.; YAU, Y.-Y. (Eds.). **Biofuels: Greenhouse Gas Mitigation and Global Warming**. New Delhi: Springer India, 2018. p. 141–154.

HOLZMAN, D. C. The Carbon Footprint of Biofuels: Can We Shrink It Down to Size in Time? **Environmental Health Perspectives**, v. 116, n. 6, p. A246–A252, 2008.

KERN, M. et al. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. **Proceedings of the National Academy of Sciences**, v. 110, n. 25, p. 10189–10194, 2013.

KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nature Reviews Drug Discovery**, v. 3, n. 11, p. 935–949, 2004.

KRISHNARAJ, R. N. et al. Bioprospecting of Thermostable Cellulolytic Enzymes through Modeling and Virtual Screening Method. **Canadian Journal of Biotechnology**, v. 1, n. 1, p. 19–25, 2017.

LAM, S. D. et al. An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. **Acta Crystallographica. Section D, Structural Biology**, v. 73, n. Pt 8, p. 628–640, 2017.

LIONTA, E. et al. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. **Current Topics in Medicinal Chemistry**, v. 14, n. 16, p. 1923–1938, 2014.

MANZONI, C. et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. **Briefings in Bioinformatics**, v. 19, n. 2, p. 286–302, 2018.

MILANEZ, A. Y. et al. De promessa a realidade: como o etanol celulósico pode revolucionar a indústria da cana-de-açúcar -– uma avaliação do potencial competitivo e sugestões de política pública. p. 58.

NAIK, S. N. et al. Production of first and second generation biofuels : A comprehensive review. v. 14, p. 578–597, 2010.

Nara Institute of Technology. Disponível em: <http://isw3.naist.jp/IS/Bio-Info-Unit/gogroup/study/study-en.html>. Acesso em: 8 dez. 2018.

OBENG, E. M. et al. Lignocellulases: a review of emerging and developing enzymes, systems, and practices. **Bioresources and Bioprocessing**, v. 4, n. 1, 2017. Disponível em: <http://bioresourcesbioprocessing.springeropen.com/articles/10.1186/s40643-017-0146-8>. Acesso em: 17 nov. 2018.

PAULY, M.; KEEGSTRA, K. Cell-wall carbohydrates and their modification as a resource for biofuels. **The Plant Journal: For Cell and Molecular Biology**, v. 54, n. 4, p. 559–568, 2008.

PAYNE, C. M. et al. Fungal Cellulases. **Chemical Reviews**, v. 115, n. 3, p. 1308–1448, 2015.

PDB, 2018. Disponível em: <https://www.rcsb.org/>. Acesso em: 19 nov. 2018.

PENG, J.; XU, J. RaptorX: exploiting structure information for protein alignment by statistical inference. **Proteins**, v. 79, n. Suppl 10, p. 161–171, 2011.

PETERSEN, T. N. et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. **Nature Methods**, v. 8, n. 10, p. 785–786, 2011.

Pfam, 2018. Disponível em: <https://pfam.xfam.org/>. Acesso em: 19 nov. 2018.

Python. Disponível em: <https://www.python.org/>. Acesso em: 19 nov. 2018.

ROBAK, K.; BALCEREK, M. Review of Second Generation Bioethanol Production from Residual Biomass. **Food Technology and Biotechnology**, v. 56, n. 2, p. 174–187, 2018.

ROSE, P. W. et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. **Nucleic Acids Research**, v. 45, n. Database issue, p. D271–D281, 2017.

ROUMPEKA, D. D. et al. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. **Frontiers in Genetics**, v. 8, 2017. Disponível em: <http://journal.frontiersin.org/article/10.3389/fgene.2017.00023/full>. Acesso em: 9 dez. 2018.

SCHELLER, H. V.; ULVSKOV, P. Hemicelluloses. **Annual Review of Plant Biology**, v. 61, n. 1, p. 263–289, 2010.

SCHWEDE, T. et al. SWISS-MODEL: an automated protein homology-modeling server. **Nucleic Acids Research**, v. 31, n. 13, p. 3381–3385, 2003.

SCHWEDE, T. Protein Modeling: What Happened to the “Protein Structure Gap”? **Structure**, v. 21, n. 9, p. 1531–1540, 2013.

SEARCHINGER, T. et al. Use of U.S. Croplands for Biofuels Increases Greenhouse Gases Through Emissions from Land-Use Change. **Science**, v. 319, n. 5867, p. 1238–1240, 2008.

SELVAM, K. et al. Cellulase enzyme: Homology modeling, binding site identification and molecular docking. **Journal of Molecular Structure**, v. 1150, p. 61–67, 2017.

SIMS, R. E. H. et al. An overview of second generation biofuel technologies. **Bioresource Technology**, v. 101, n. 6, p. 1570–1580, 2010.

SOMERVILLE, C. Toward a Systems Approach to Understanding Plant Cell Walls. **Science**, v. 306, n. 5705, p. 2206–2211, 2004.

The PubChem Project, 2018. Disponível em: <https://pubchem.ncbi.nlm.nih.gov/>. Acesso em: 20 nov. 2018.

THE UNIPROT CONSORTIUM. UniProt: the universal protein knowledgebase. **Nucleic Acids Research**, v. 45, n. D1, p. D158–D169, 2017.

TIWARI, R. et al. Bioprospecting of functional cellulases from metagenome for second generation biofuel production: a review. **Critical Reviews in Microbiology**, v. 44, n. 2, p. 244–257, 2018.

TROTT, O.; OLSON, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. **Journal of computational chemistry**, v. 31, n. 2, p. 455–461, 2010.

UniProt. Disponível em: <https://www.uniprot.org/>. Acesso em: 19 nov. 2018.

WANG, M. et al. Cellulolytic Enzyme Production and Enzymatic Hydrolysis for Second-Generation Bioethanol Production. In: BAI, F.-W. et al. (Eds.). **Biotechnology in China III: Biofuels and Bioenergy**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 128p. 1–24.

WHEELER, D.; BHAGWAT, M. **BLAST QuickStart**. Humana Press, 2007. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK1734/>. Acesso em: 19 nov. 2018.

WILSON, D. B. Cellulases and biofuels. **Current Opinion in Biotechnology**, v. 20, n. 3, p. 295–299, 2009.

WOLFENDEN, R.; SNIDER, M. J. The Depth of Chemical Time and the Power of Enzymes as Catalysts. **Accounts of Chemical Research**, v. 34, n. 12, p. 938–945, 2001.

WRIGHT, J. D.; WYMAN, C. E.; GROHMANN, K. Simultaneous saccharification and fermentation of lignocellulose: Process evaluation. **Applied Biochemistry and Biotechnology**, v. 18, n. 1, p. 75–90, 1988.

YUAN, L. et al. Screening and Identification of Cellulase-Producing Strain of Fusarium Oxysporum. **Procedia Environmental Sciences**, 2011 International Conference of Environmental Science and Engineering. v. 12, 2011 International Conference of Environmental Science and Engineering, p. 1213–1219, 2012.

YUAN, S.; CHAN, H. C. S.; HU, Z. Using PyMOL as a platform for computational drug design. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 7, n. 2, p. e1298, 2017.

ZHANG, Y. Protein structure prediction: when is it useful? **Current Opinion in Structural Biology**, v. 19, n. 2, p. 145–155, 2009.