

UNIVERSIDADE FEDERAL DE MINAS GERAIS  
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

DIEGO CÉSAR BATISTA MARIANO

**Uso de assinaturas estruturais para proposta de mutações em  
enzimas  $\beta$ -glicosidase usadas na produção de  
biocombustíveis**

Belo Horizonte  
2019

UNIVERSIDADE FEDERAL DE MINAS GERAIS  
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

DIEGO CÉSAR BATISTA MARIANO

**Uso de assinaturas estruturais para proposta de mutações em  
enzimas  $\beta$ -glicosidase usadas na produção de  
biocombustíveis**

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Doutor em Bioinformática.

Orientadora: Dra. Raquel Cardoso de Melo Minardi

Belo Horizonte  
2019

043

Mariano, Diego César Batista.

Uso de assinaturas estruturais para proposta de mutações em enzimas  $\beta$ -glicosidade usadas na produção de biocombustíveis [manuscrito] /Diego César Batista Mariano. – 2019.

98 f. : il. ; 29,5 cm.

Orientadora: Dra Raquel Cardoso de Melo Minardi.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática - Teses. 2. Biocombustíveis. 3. Mutação. I. Minardi, Raquel Cardoso de Melo. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha elaborada por Sônia M. S. Moraes CRB 6/1357 - ICB/UFMG



**"Uso de assinaturas estruturais para proposta de mutações em enzimas  $\beta$ -glicosidase usadas na produção de biocombustíveis"**

**Diego César Batista Mariano**

Tese aprovada pela banca examinadora constituída pelos Professores:

Profa Raquel Cardoso de Melo Minardi - Orientadora  
UFMG

Prof. Lucas Bleicher  
UFMG

Prof. José Miguel Ortega  
UFMG

Profa Cristiane Neri Nobre  
PUC Minas

Prof. Thiago de Souza Rodrigues  
CEFET

Belo Horizonte, 11 de março de 2019.

## Resumo

$\beta$ -glicosidases (EC 3.2.1.21) são enzimas chave no processo de produção de biocombustíveis de segunda geração. Elas agem em conjunto com endoglucanases e exoglucanases na conversão de biomassa em açúcares fermentáveis. Entretanto, a maior parte das  $\beta$ -glicosidases conhecidas são altamente inibidas por grandes concentrações de glicose. Assim, a busca por mutações que melhorem a atividade de  $\beta$ -glicosidases não tolerantes a inibição por glicose é de grande importância para a indústria. Nesta tese, é apresentada uma revisão sistemática da literatura para coletar informações sobre  $\beta$ -glicosidases glicose-tolerantes e construir uma base de dados, denominada BETAGDB. Além disso, são caracterizados resíduos importantes no bolsão catalítico para atividade e glicose-tolerância. Por fim, apresenta-se um método baseado na diferença de variação de assinaturas estruturais para propor mutações em enzimas, denominado SSV (*Structural Signature Variation*). SSV usa modelagem em grafos para criar uma assinatura estrutural identificadora de  $\beta$ -glicosidases glicose-tolerantes. O método SSV foi avaliado em três estudos de caso: (i) 27 mutações descritas na literatura foram manualmente classificadas como benéficas ou não. A seguir, a classificação foi reproduzida utilizando SSV. O método obteve uma acurácia de 0,74 e uma precisão de 0,89; (ii) 18 mutações benéficas foram propostas para a  $\beta$ -glicosidase não tolerante Bgl1B. Resultados experimentais de três mutações corroboram os resultados obtidos pelo segundo estudo de caso e demonstram que SSV é um método eficaz para proposta de mutações em  $\beta$ -glicosidases; e (iii) comparou-se SSV com SVM para verificar se a distância euclidiana, métrica usada por SSV para comparação de assinaturas, era eficaz. Comparou-se ainda com BioGPS, um método que usa *fingerprints* para proposta de mutações baseadas em estruturas tridimensionais. SSV obteve valores de precisão e especificidade superiores a SVM. Na comparação com BioGPS, SSV foi capaz de prever corretamente cinco em sete mutações validadas em bancada que inseriam atividade amidase em uma lipase. Os resultados obtidos nesta tese podem ajudar na produção de enzimas  $\beta$ -glicosidases mutantes capazes de aperfeiçoar a produção de biocombustíveis de segunda geração. O método SSV pode ser estendido a outras enzimas e pode ainda ser utilizado em conjunto com outras estratégias e ferramentas para propor mutações mais eficientes. SSV está disponível em: <<http://bioinfo.dcc.ufmg.br/ssv>>.

**Palavras-chave:** biocombustíveis,  $\beta$ -glicosidases, assinaturas estruturais, mutações, SSV

## Abstract

*$\beta$ -glucosidases (EC 3.2.1.21) are key enzymes in the second-generation biofuel production. They act synergically with endoglucanases and exoglucanases in the conversion of biomass to fermentable sugars. However, most known  $\beta$ -glucosidases are highly inhibited by high glucose concentrations. Hence, the search for mutations that improve the activity of non-tolerant  $\beta$ -glucosidases has great importance to the industry. In this thesis, I present a systematic review of the literature to collect information about glucose-tolerant  $\beta$ -glucosidases and to construct a database, called BETAGDB. In addition, important residues for the activity and glucose-tolerance were characterized in the catalytic pocket. Finally, I proposed a method based on the difference of variation of structural signatures to propose mutations in enzymes, called Structural Signature Variation (SSV). SSV uses graph modeling to create a structural signature that identifies glucose-tolerant  $\beta$ -glucosidases. The SSV method was evaluated in three case studies: (i) 27 mutations described in the literature were manually classified as beneficial or not. The classification was then reproduced using SSV. The method obtained an accuracy of 0.74 and a precision of 0.89; (ii) 18 beneficial mutations were proposed for the non-tolerant  $\beta$ -glucosidase Bgl1B. Experimental results of three mutations corroborate the outcomes obtained by the second case study and demonstrate that SSV is an effective method for the proposal of mutations in  $\beta$ -glucosidases; and (iii) SSV was compared with SVM to verify whether the Euclidean distance, metric used by SSV for comparison of signatures, was effective. It was also compared with BioGPS, a method that uses fingerprints to propose mutations based on three-dimensional structures. SSV obtained values of precision and specificity superior to SVM. In the comparison to BioGPS, SSV was able to correctly predict five in seven bench-validated mutations that inserted amidase activity into a lipase. The results obtained in this thesis may aid in the production of mutant  $\beta$ -glucosidase enzymes capable of enhancing the production of second-generation biofuels. The SSV method can be extended to other enzymes and can also be used together to other strategies and tools to propose more efficient mutations. SSV is available at <<http://bioinfo.dcc.ufmg.br/ssv>>.*

**Keywords:** biofuels,  $\beta$ -glucosidases, structural signatures, mutations, SSV

## Agradecimentos

Gostaria de agradecer às agências de fomento à pesquisa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); ao Programa de Pós-graduação em Bioinformática da UFMG; e à equipe do Laboratório de Bioinformática e Sistemas que me auxiliou nesta jornada.

Agradeço a minha família pelo apoio e incentivo aos estudos; a meus orientadores de iniciação científica, mestrado e doutorado, que auxiliaram na minha construção como pesquisador; agradeço à comunidade de desenvolvedores de *software* livre e de divulgação científica; e agradeço à Universidade Federal de Minas Gerais por fomentar minha formação intelectual, moral e ética.

“O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001”. Edital Biologia Computacional. Número de processo 23038.004007/2014-82.

# SUMÁRIO

<b>Resumo.....</b>	<b>i</b>
<b>Abstract.....</b>	<b>ii</b>
<b>Agradecimentos .....</b>	<b>iii</b>
<b>SUMÁRIO.....</b>	<b>iv</b>
<b>Lista de Figuras .....</b>	<b>vi</b>
<b>Lista de Tabelas.....</b>	<b>viii</b>
<b>Lista de abreviações .....</b>	<b>ix</b>
<b>Produção Acadêmica.....</b>	<b>x</b>
Artigos publicados diretamente relacionados a esta tese.....	x
Prêmios obtidos durante o período de doutorado.....	x
<b>1. Introdução .....</b>	<b>15</b>
<b>1.1 Enzimas <math>\beta</math>-glicosidase.....</b>	<b>19</b>
<b>1.2 Tolerância a glicose .....</b>	<b>22</b>
<b>1.3 Estimulação por glicose .....</b>	<b>25</b>
<b>1.4 Classificação de <math>\beta</math>-glicosidases baseada em tolerância e estimulação por glicose....</b>	<b>26</b>
<b>1.5 Mutações em <math>\beta</math>-glicosidases .....</b>	<b>29</b>
<b>1.6 Análises de bioinformática .....</b>	<b>32</b>
1.6.1 Sequências.....	32
1.6.2 Bioinformática estrutural.....	34
1.6.2.1 Modelagem comparativa.....	35
1.6.2.2 Ancoramento molecular.....	35
1.6.3 Assinaturas estruturais.....	36
1.6.4 Cutoff Scanning Matrix.....	37
1.6.5 Assinaturas de nível atômico (aCSM) .....	39
<b>1.7 Hipóteses.....</b>	<b>41</b>
<b>2. Objetivos .....</b>	<b>42</b>
<b>2.1 Objetivo geral.....</b>	<b>42</b>



2.2	Objetivos específicos.....	42
3.	Material e métodos.....	43
3.1	Coleta dos dados .....	43
3.2	<i>Pipeline</i> para modelagem comparativa.....	43
3.2.1	Base de dados BETAGDB.....	45
3.3	Ancoramento molecular de celobiose .....	45
3.4	Caracterização do bolsão catalítico .....	46
3.5	Construção das assinaturas estruturais.....	47
3.6	Estudos de caso .....	47
3.6.1	Avaliando mutações da literatura .....	47
3.6.2	Propondo mutações para uma $\beta$ -glicosidase não tolerante .....	48
3.6.3	Comparando SSV com outros métodos .....	49
4.	Resultados .....	51
4.1	Bolsão catalítico .....	52
4.2	Análise da interação com celobiose.....	55
4.3	Assinaturas estruturais em $\beta$ -glicosidases glicose-tolerantes .....	57
4.4	<i>Structural Signature Variation (SSV)</i> .....	58
4.4.1	Aplicação do SSV em $\beta$ -glicosidases .....	59
4.4.2	Estudo de caso 1: verificando mutações com SSV .....	60
4.4.3	Estudo de caso 2: propostas de mutações para a $\beta$ -glicosidase Bgl1B .....	64
4.4.4	Estudo de caso 3: comparação com outros métodos (SVM e BioGPS) .....	68
4.5	<i>Webtool SSV</i> .....	70
5.	Discussão .....	71
5.1	Importância dos resíduos do bolsão catalítico na tolerância a glicose .....	71
5.1.1	Sítio ativo .....	72
5.1.2	Canal do substrato .....	73
5.1.3	Mecanismo de glicose-tolerância.....	74
5.1.4	Papel dos resíduos do bolsão catalítico na atividade .....	77
5.2	Comparação do SSV com outros métodos de proposta de mutações.....	79
6.	Conclusões.....	84
7.	Perspectivas.....	86
8.	Referências bibliográficas .....	87
9.	Apêndices .....	98

## Lista de Figuras

Figura 1. Produção mundial de biocombustíveis. ....	15
Figura 2. Processo de decomposição da celulose. ....	17
Figura 3. Processo de sacarificação enzimática na produção de biocombustíveis de segunda geração. ....	18
Figura 4. Comparação estrutural entre $\beta$ -glicosidases GH1 e GH3. ....	21
Figura 5. Comparação entre bolsões catalíticos de $\beta$ -glicosidases GH1 e GH3. ....	22
Figura 6. Representação do efeito da concentração de glicose na atividade relativa de $\beta$ -glicosidases. ....	28
Figura 7. Dados estruturais de $\beta$ -glicosidases, aminoácidos relacionados com a tolerância a glicose e ao sítio ativo. ....	31
Figura 8. Alinhamento de sequências de duas proteínas. ....	34
Figura 9. Representação do <i>docking</i> entre substrato e proteína. ....	36
Figura 10. Construção da assinatura estrutural usando CSM. ....	37
Figura 11. Algoritmo para o cálculo do CSM. ....	38
Figura 12. Decomposição em valores singulares: a matriz A é decomposta nas matrizes U, S e $V^T$ . ....	39
Figura 13. Algoritmo de cálculo do aCSM. ....	40
Figura 14. <i>Pipeline</i> para modelagem de 18 sequências de $\beta$ -glicosidases glicose-tolerantes sem estruturas tridimensionais disponíveis. ....	43
Figura 15. Estruturas cristalográficas e modelos de $\beta$ -glucosidases. ....	44
Figura 16. Caixa utilizada para determinação da região do <i>docking</i> . ....	46
Figura 17. Metodologia para construção da assinatura de tolerância à glicose. ....	47
Figura 18. Passos para determinação do estudo de caso 1. ....	48
Figura 19. Metodologia para proposta de mutações para uma $\beta$ -glicosidase não-tolerante. ....	48
Figura 20. Região do bolsão catalítico de CaLB (PDB ID: 1TCA). ....	50
Figura 21. Alinhamento entre os resíduos de bolsões catalíticos das 21 $\beta$ -glicosidases GH1 glicose-tolerantes. ....	53
Figura 22. Resíduos presentes no bolsão catalítico de $\beta$ -glicosidases glicose-tolerantes. ....	54
Figura 23. <i>Docking</i> de celobiose no bolsão catalítico das $\beta$ -glicosidases glicose-tolerantes. ....	55
Figura 24. Contatos da celobiose com resíduos de $\beta$ -glicosidases. ....	56

Figura 25. Representação da assinatura de tolerância à glicose.....	59
Figura 26. Etapas do estudo de caso para validação do SSV usando 27 mutações coletadas na literatura para $\beta$ -glicosidases.....	61
Figura 27. Representação do estudo de caso 2.....	65
Figura 28. Atividade relativa em relação à concentração de glicose para Bgl1B e mutantes.....	66
Figura 29. Posicionamento da celobiose transferida e do resíduo H228 em Bgl1B.....	67
Figura 30. Comparação entre o número de contatos realizados pelos aminoácidos glicina e serina para mutação G246S de Bgl1B.....	68
Figura 31. Interface da <i>webtool</i> SSV.....	70
Figura 32. Mecanismo de ação catalítica das $\beta$ -glicosidases GH1.....	73
Figura 33. Comparação entre estruturas de Bgl1A e Bgl1B.....	76
Figura 34. Projeção da análise de componentes principais (PC1 e PC2) dos mutantes CaLB no modelo BioGPS-UPCA ( <i>score</i> global).....	82

## Lista de Tabelas

Tabela 1. Lista de $\beta$ -glicosidases tolerantes a inibição por glucose. ....	23
Tabela 2. Mutações coletadas na literatura e na base de dados UniProt. ....	29
Tabela 3. Experimento das oito mutações de CaLB usando SSV para comparação com BioGPS. ....	50
Tabela 4. Matriz com o percentual de identidade entre sequências determinado pelo algoritmo Needleman-Wunsch para alinhamento global. ....	51
Tabela 5. Predição do impacto de mutações a partir da pontuação da diferença da variação da assinatura de tolerância. ....	62
Tabela 6. Matriz de confusão dos resultados preditos comparados aos valores reais. Deve-se ressaltar que o resultado é considerado positivo se $\Delta\Delta\text{SSV} < 0$ (benéfica); e negativo se $\Delta\Delta\text{SSV} > 0$ (não benéfica). ....	63
Tabela 7. Valores de sensibilidade, especificidade, valor predito (+), valor predito (-) e acurácia para os testes de $\Delta\Delta\text{SSV}$ para as 27 mutações em $\beta$ -glicosidases. ....	64
Tabela 8. Mutações benéficas propostas para Bgl1B. ....	65
Tabela 9. Métricas usadas para comparar SSV com SVM. ....	69
Tabela 10. Mutantes de CaLB avaliados por SSV. ....	69
Tabela 11. Resíduos conservados e contatos com celobiose para os 22 resíduos correspondentes a subsequência consenso do bolsão catalítico das 21 glicose-tolerantes GH1. ....	72
Tabela 12. Importância dos resíduos do bolsão catalítico. ....	77

## Lista de abreviações

BETAGDB:  *$\beta$ -glucosidase data bank*

CaLB: *Candida antarctica lipase B*

CO<sub>2</sub>: dióxido de carbono

EP-PCR: *error-prone PCR*

GH1: Glicosídeo hidrolases (família) 1

GH3: Glicosídeo hidrolases (família) 3

k<sub>cat</sub>: constante catalítica

K<sub>i</sub>: constante de inibição

K<sub>M</sub>: constante de Michaelis

PDB: *Protein Data Bank*

pNPG: *4-nitrophenyl- $\beta$ -D-glucopyranoside*

SHF: *separate hydrolysis and fermentation*

SSF: *simultaneous saccharification and fermentation*

SSV: *structural signature variation*

SVM: *support vector machine*

## Produção Acadêmica

Durante o período de doutorado (2015-2019), o autor desta tese publicou 20 artigos em revistas científicas de alta relevância, sendo seis artigos como primeiro autor e oito publicados em revistas com estratos QUALIS A1/A2 nas áreas de avaliação: ciências biológicas I, ciência da computação ou interdisciplinar.

### *Artigos publicados diretamente relacionados a esta tese*

1. **Mariano, D.C.B.**; Leite, C.; Santos, L.H.S.; Marins, L.F.; Machado, K.S.; Werhli, A.V.; Lima, L.H.F.; De Melo-Minardi, R.C. *Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review*. GENETICS AND MOLECULAR RESEARCH, v. 16, p. 1-19, 2017.
2. **Mariano, D.C.B.**; SANTOS, L.H.S.; MACHADO, K.S. ; WERHLI, A.V. ; LIMA, L.H.F.; DE MELO-MINARDI, R.C. *A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV)*. *Int. J. Mol. Sci.* 2019.
3. Costa, L.S.C.\*; **Mariano, D.C.B.\***; Rocha, R.E.O.; Kraml, J.; Silveira, C.H.; Liedl, K.R.; de Melo-Minardi, R.C.; Lima, L.H.F. *Molecular Dynamics Gives New Insights into the Glucose Tolerance and Inhibition Mechanisms on  $\beta$ -Glucosidases*. *Molecules*, 24, 3215, 2019.

### *Prêmios obtidos durante o período de doutorado*

1. *ISCB Wikipedia Competition Winner* (2018);
2. **Best Poster Award: X-Meeting 2016** na categoria “*Proteins and Proteomics*”. Título do trabalho apresentado: “*Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production*”;
3. **Best Poster Award: X-Meeting 2016** na categoria “*Software Development and Databases*”. Título do trabalho apresentado: “*An approach for constructing a database of manually curated contacts in proteins*”;
4. **Best Paper Award: "Oral presentation", 2nd Brazilian Student Council Symposium**. Título do trabalho apresentado: “*A new method based on structural signatures to propose mutations for enzymes  $\beta$ -glucosidase used in biofuel production*”.

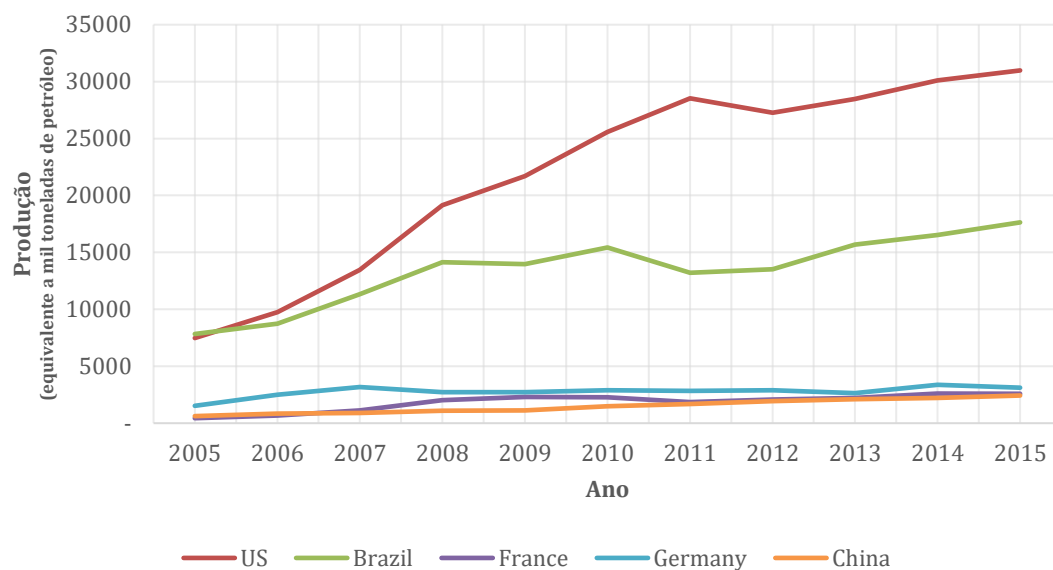
---

\* Estes autores contribuíram igualmente no artigo.

## 1. Introdução

Biocombustíveis são uma forma de energia limpa e renovável (CHOUDRI et al., 2017). Eles destacam-se como uma alternativa aos combustíveis fósseis, uma vez que seu processo de combustão libera na atmosfera CO<sub>2</sub> equivalente ao já fixado pelas plantas (HO; NGO; GUO, 2014). No caso dos combustíveis fósseis, moléculas de carbono, que já estavam removidas do ciclo do carbono, são extraídas e recolocadas no ciclo, o que agrava os efeitos do aquecimento global (BROWN; CALDEIRA, 2017).

Biocombustíveis são produzidos a partir de matérias agrícolas, como cana-de-açúcar, milho, algas, dentre outros tipos de matérias orgânicas (SOLOMON, 2010). Dentre os tipos de biocombustíveis mais produzidos destacam-se o biodiesel, o biogás e o bioetanol (VAZ, 2019). O Brasil tem se destacado como o segundo maior produtor de biocombustíveis do mundo, produzindo bioetanol a partir de cana-de-açúcar (Figura 1; RAMAN; GNANSOU-NOU, 2014).



**Figura 1. Produção mundial de biocombustíveis.**

Os cinco maiores produtores de biocombustíveis são, respectivamente: Estados Unidos, Brasil, França, Alemanha e China. Fonte: Adaptado de BP GLOBAL (2016).

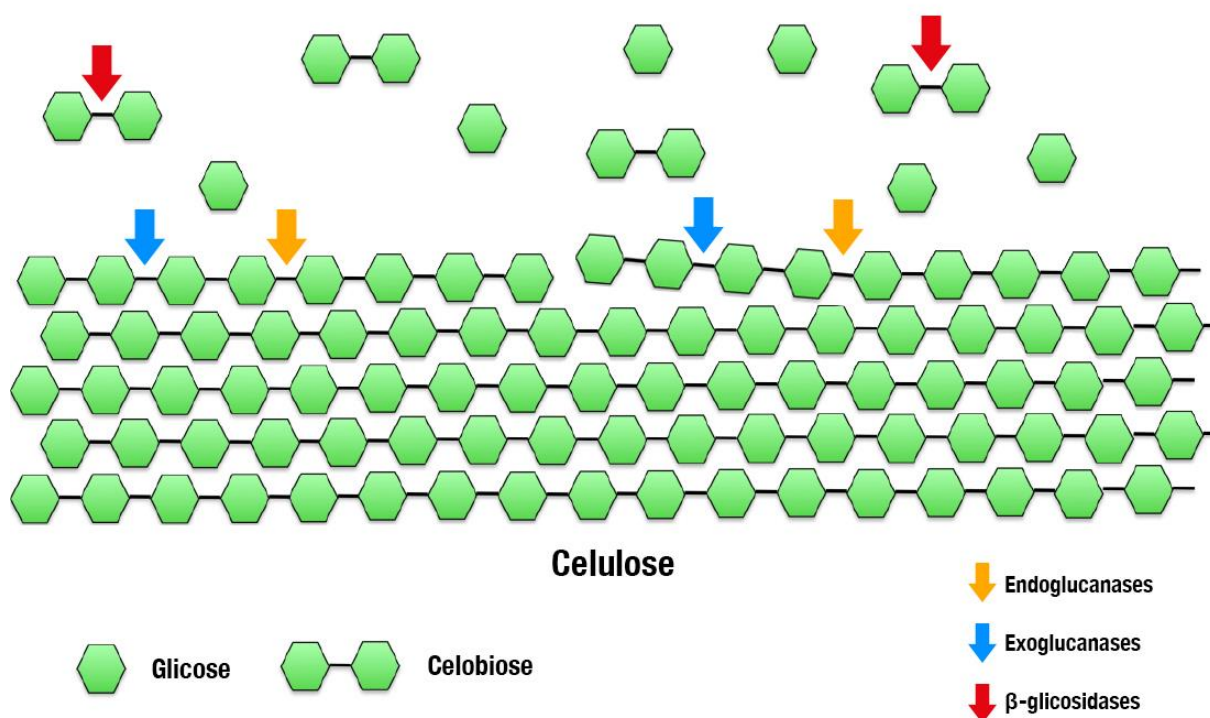
O processo de produção do bioetanol ocorre em várias etapas, destacando-se duas: sacarificação e fermentação. Na sacarificação, o açúcar é extraído da matéria-prima, enquanto na fermentação, o açúcar extraído é convertido em etanol.

Os processos de produção de biocombustíveis podem ser classificados de acordo com a forma a qual a matéria-prima para produção do biocombustível é obtida. Biocombustíveis de primeira geração são aqueles a qual, no processo de produção, o açúcar é extraído com facilidade a partir de alimentos, como beterraba, trigo, milho, cana-de-açúcar, dentre outros (MARTIN, 2010). O processo de produção de biocombustíveis de primeira geração pode entrar em conflito com a produção de alimentos, uma vez que parte das lavouras destinadas a produção de alimentos passaria a ser destinada a produção de combustíveis. Além disso, esse processo gera a produção de resíduos pouco aproveitados, como o bagaço da cana-de-açúcar. Uma solução para esses problemas pode estar na chamada segunda geração de biocombustíveis (ROBAK; BALCEREK, 2018).

Biocombustíveis de segunda geração são aqueles produzidos a partir de biomassa. Celulose é a maior fonte de biomassa da Terra, compondo em torno de 40 a 50% de todo peso da biomassa de plantas (UCHIMA et al., 2012; UCHIYAMA; YAOI; MIYAZAKI, 2015). Ela é composta por monômeros de glicose conectados por ligações  $\beta$ -1,4 glicosídicas (PERCIVAL ZHANG; HIMMEL; MIELENZ, 2006; RAMANI et al., 2015a). A glicose obtida pela degradação da celulose pode ser fermentada para produção de bioetanol, uma alternativa verde promissora e fonte de energia renovável para produção de combustíveis (TEUGJAS; VÄLJAMÄE, 2013).

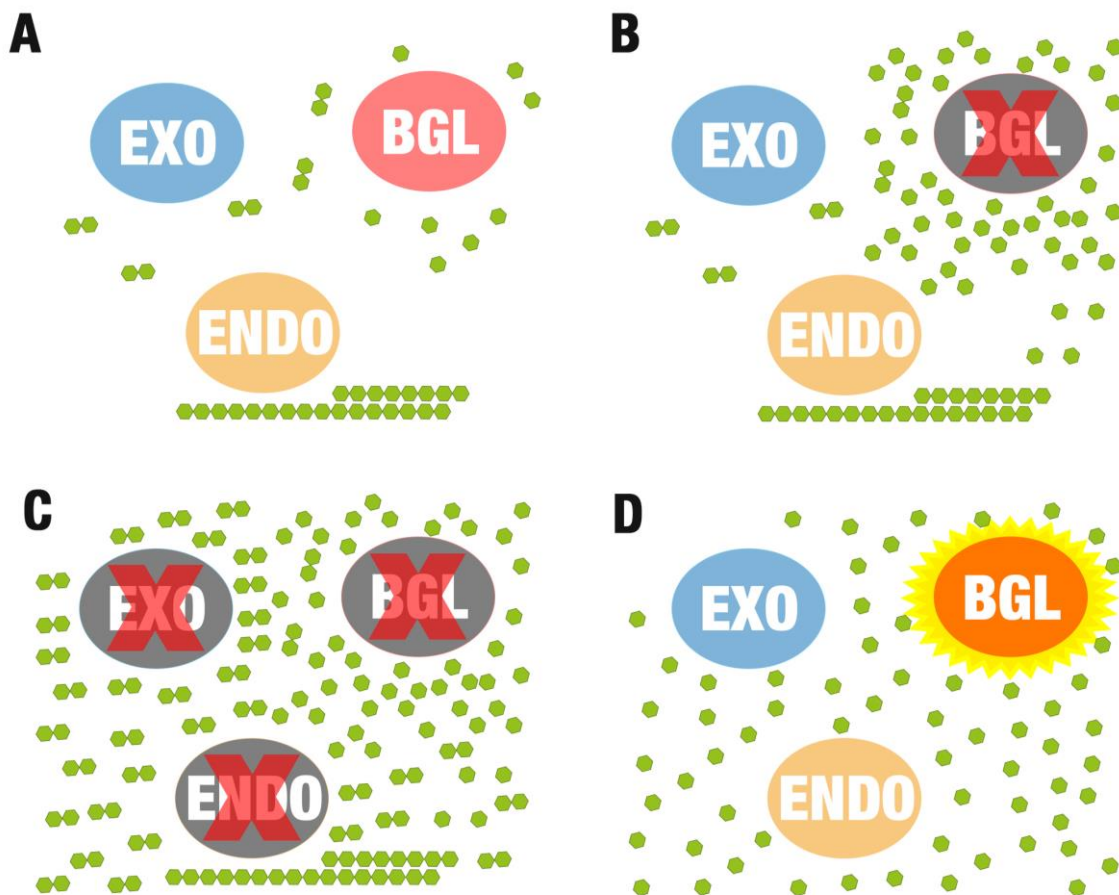
O bagaço da cana-de-açúcar é um exemplo de biomassa lignocelulósica. A lignocelulose é um componente presente na parede celular de plantas. É composta por lignina, hemicelulose e celulose, cuja degradação pode gerar uma grande quantidade de monossacarídeos fermentáveis (glicose) que podem ser utilizados na produção de etanol lignocelulósico. Estima-se que se 80% da glicose aprisionada na estrutura da parede celular do bagaço da cana-de-açúcar pudesse ser extraída, seria possível dobrar a produção brasileira de bioetanol sem a necessidade de aumentar as áreas destinadas ao cultivo de cana-de-açúcar (Prof. Dr. Luis Fernando Marins, comunicação pessoal). Entretanto, a produção de biocombustíveis de segunda geração ainda apresenta limitações. Fungos e bactérias possuem sistemas enzimáticos capazes de decompor com eficácia a parede celular de plantas. Um desses sistemas é composto por três celulasas: as endoglucanases, as exoglucanases e as  $\beta$ -glicosidases (Figura 2).





**Figura 2. Processo de decomposição da celulose.**  
 Fonte: Adaptado de CHANDEL & SILVA (2013).

$\beta$ -glicosidases agem sinergicamente com endoglucanases (E.C. 3.2.1.4) e exoglucanases (E.C. 3.2.1.91), conhecidas também como celobiohidrolases, em um sistema enzimático responsável pela bioconversão de celulose (KUMAR; SINGH; SINGH, 2008). Enquanto endoglucanases agem na cadeia da celulose produzindo oligossacarídeos de tamanhos variados, exoglucanases agem na produção principalmente de celobiose (Figura 3).  $\beta$ -glicosidases clivam as ligações  $\beta$ -1,4 glicosídicas ao reagir com uma molécula de água, produzindo dois monômeros de glicose (BÉGUIN; AUBERT, 1994).  $\beta$ -glicosidases têm o papel fundamental nesse sistema de remover a celobiose, que é um forte inibidor tanto das endoglucanases quanto das exoglucanases (CHAMOLI et al., 2016; KADAM; DEMAIN, 1989; MURPHY et al., 2013; WATANABE et al., 1992; ZHAO et al., 2013). Entretanto, a maior parte das  $\beta$ -glicosidases relatadas na literatura são inibidas pelo aumento da concentração do próprio produto (GUEGUEN et al., 1995; RAJASREE et al., 2013; TEUGJAS; VÄLJAMÄE, 2013; YANG et al., 2015b). Portanto, o interesse na descoberta de  $\beta$ -glicosidases termoestáveis e resistentes à inibição por glicose tem crescido nos últimos tempos. Além disso, a produção de tais enzimas pode ajudar na descoberta dos mecanismos de funcionamento de enzimas  $\beta$ -glicosidase eficientes e pode gerar melhorias no processo de sacarificação para produção de biocombustíveis (PEI et al., 2012). Sugere-se ainda que  $\beta$ -glicosidases melhoradas podem proporcionar um aumento da conversão de bagaço-de-cana em açúcar (CAO et al., 2015).



**Figura 3. Processo de sacarificação enzimática na produção de biocombustíveis de segunda geração.**

(A) Em geral, o processo ocorre pela ação de três enzimas: (i) endoglucanases (ENDO), (ii) exoglucanases ou celobiohidrolases (EXO), e  $\beta$ -glicosidases (BGL); (B)  $\beta$ -glicosidases podem sofrer inibição competitiva pela glicose, produto da quebra da celobiose; (C) a inibição de  $\beta$ -glicosidases proporciona um aumento de celobiose no meio, que é um inibidor de endoglucanases e exoglucanases; (D) portanto, a presença de  $\beta$ -glicosidases mais eficientes poderia melhorar o processo de sacarificação. Fonte: próprio autor.

Na produção industrial, essas enzimas costumam ser cultivadas em fungos. Entretanto, a produção em fungos requer a suplementação de enzimas  $\beta$ -glicosidase, que tendem a ser pouco secretadas por esses organismos. Novozyme® 188 é um coquetel comercial de enzimas celulósicas largamente utilizadas para complementar a ação das celulases. Novozyme® 188 ainda tem sido utilizado para comparar novas  $\beta$ -glicosidases e testá-las para conversão de celulose (CAO et al., 2015; UCHIMA et al., 2012; ZHAO et al., 2013). Por exemplo, o fungo *Trichoderma reesei*, um produtor de celulases em nível industrial, requer a suplementação desta enzima para melhorar a conversão de celulose. Entretanto, a  $\beta$ -glicosidase presente em Novozyme® 188 apresenta uma constante de inibição ( $K_i$ ) inferior a 0,1 M (CAO et al., 2015). Enzimas com uma constante de inibição superior a esse valor podem ser consideradas mais eficientes. Para que a produção industrial de biocombustíveis de segunda geração seja economicamente viável, ela deve ocorrer em nível industrial com um baixo custo tanto no

processo sacarificação quanto de fermentação. Em geral, suplementações encarecem o processo de produção. Ainda, para ser economicamente viável, a hidrólise de celulose deve ser conduzida em uma alta concentração de matéria seca, que inevitavelmente resulta em uma alta concentração de produtos de hidrólise, como a glicose e celobiose, o que faz da inibição por produtos o maior desafio do processo de engenharia de enzimas (TEUGJAS; VÄLJAMÄE, 2013). Logo, a descoberta de novas  $\beta$ -glicosidases tolerantes a inibição por glicose pode ser eficaz para suplementação em nível industrial e pode ainda reduzir custos do processo.

Recentemente,  $\beta$ -glicosidases com uma característica especial de alta resistência a inibição por glicose têm sido descritas na literatura (SALGADO et al., 2018). Essas  $\beta$ -glicosidases, denominadas como glicose-tolerantes, têm sido sugeridas como enzimas mais eficientes para produção de biocombustíveis de segunda geração. Além disso, o uso de técnicas de engenharia genética tem sido proposto para tornar  $\beta$ -glicosidases não tolerantes a inibição em  $\beta$ -glicosidases glicose-tolerantes. Um exemplo é a técnica de EP-PCR (*error-prone PCR*), que produz mutações aleatórias na estrutura da proteína (WILSON; KEEFE, 2001). Nessa técnica, uma DNA polimerase danificada insere mutações em posições aleatórias da sequência durante o processo de replicação do gene codificante da  $\beta$ -glicosidase. Na maior parte dos casos, essas mutações não geram qualquer impacto ou até mesmo pioram a atividade das  $\beta$ -glicosidases. Ocasionalmente alguma mutação benéfica pode ser descoberta, mas isso depende de extensos experimentos de bancada e validação manual de cada mutante produzido aleatoriamente. Logo, a descoberta de resíduos específicos para mutações sítio-dirigidas aparenta ser a estratégia mais eficiente para melhoria de enzimas  $\beta$ -glicosidase, e sua busca tem incentivado estudos com pesquisadores de diversas áreas. Tais problemas motivaram um projeto de pesquisa apresentado ao edital Biologia Computacional CAPES 051/2013. O projeto, denominado “Bioinformática Estrutural de Proteínas: modelos, algoritmos e aplicações biotecnológicas”, busca propor novas mutações em  $\beta$ -glicosidase que melhorem a produção de biocombustíveis. Esta tese faz parte desse projeto de pesquisa.

### 1.1 Enzimas $\beta$ -glicosidase

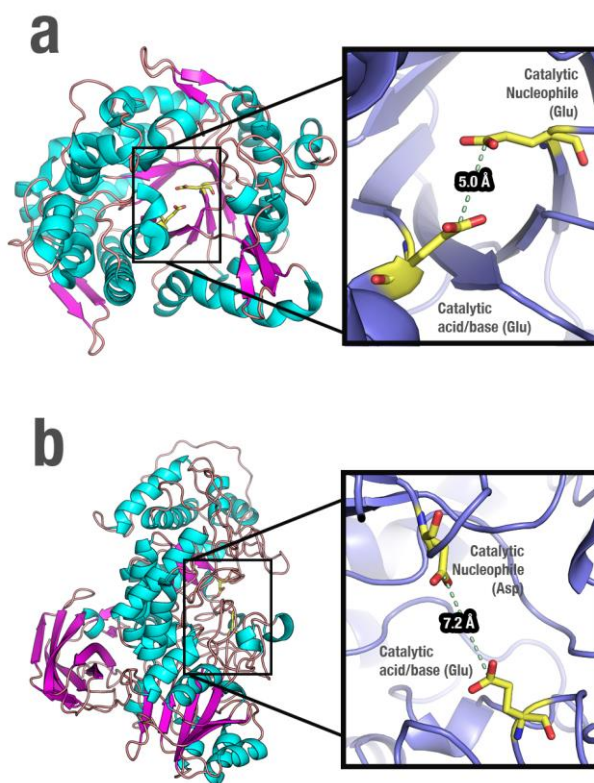
$\beta$ -glicosidase (E.C. 3.2.1.21) é uma classe de enzimas que hidrolisam as ligações glicosídicas de dissacarídeos, oligossacarídeos, aquil- e aril- $\beta$ -glicosídeos (CAIRNS; ESEN, 2010; LU et al., 2013). Elas são extraídas de diversos organismos, como por exemplo, animais (PENTZOLD et al., 2014; UCHIMA et al., 2011), fungos (SAHA; BOTHAST, 1996), plantas

(CAIRNS; ESEN, 2010; PENTZOLD et al., 2014), bactérias (CRESPIM et al., 2016) e até mesmo em metagenomas (UCHIYAMA; MIYAZAKI; YAOI, 2013; UCHIYAMA; YAOI; MIYAZAKI, 2015). Em animais, essas enzimas ajudam no metabolismo de glicosídeos, além de participar de diversas funções digestivas (CAIRNS; ESEN, 2010; MARANA et al., 2001). Em plantas, elas desempenham diversas funções, tais como defesa, liberação de compostos aromatizantes, catabolismo da parede celular e lignificação (CAIRNS; ESEN, 2010; PENTZOLD et al., 2014). Em bactérias e em fungos, elas são componentes essenciais para hidrólise da celulose (RAMANI et al., 2015b).  $\beta$ -glicosidases têm aplicações em diversas áreas das indústrias biotecnológicas, como melhoria do aroma de sucos e vinhos (SWANGKEAW et al., 2010), hidrólise de glicosídeos de isoflavonas de soja (LI et al., 2012; SINGHANIA et al., 2013), redução de toxicidade de ração animal (COTA et al., 2015; GOPALAN et al., 1992) e degradação de celulose para conversão de biomassa em produção de biocombustível (SINGHANIA et al., 2013; SØRENSEN et al., 2011).

Inicialmente, as  $\beta$ -glicosidases foram classificadas, baseado na especificidade por substrato, em três grupos: (i) aril-glicosidases; (ii) celobiasas; e (iii)  $\beta$ -glicosidases de ampla especificidade (RAMANI et al., 2015b, p. 1; SINGHANIA et al., 2013; YANG et al., 2015a). Entretanto, a limitação dessa classificação levou à necessidade de um novo método de classificação, a partir de então, baseado em similaridades de sequências (HENRISSAT, 1991). Assim,  $\beta$ -glicosidases passaram a ser classificadas com base nas famílias das glicosídeo hidrolases (GH) 1, 3, 5, 9, 30 e 116 (CAIRNS; ESEN, 2010; HENRISSAT; BAIROCH, 1993; JABBOUR; KLIPPEL; ANTRANIKIAN, 2012). Entretanto, a maior parte das  $\beta$ -glicosidases conhecidas pertencem às famílias 1 e 3 (CRESPIM et al., 2016; DEL POZO et al., 2012; SHIPKOWSKI; BRENCHLEY, 2005; SINGHANIA et al., 2013).

Enzimas  $\beta$ -glicosidase da família GH1 pertencem ao clã GH-A. Esse clã apresenta uma estrutura altamente conservada conhecida como  $(\alpha/\beta)_8$  TIM barril, composta por oito fitas-beta, que se enovelam em forma de barril, intercaladas com oito alfa-hélices (Figura 4a). Além disso,  $\beta$ -glicosidases GH1 apresentam dois glutamatos a uma distância de aproximadamente 5 Å localizados nas fitas-beta 4 e 7 do barril (CAIRNS; ESEN, 2010; CANTAREL et al., 2009; CRESPIM et al., 2016; HENRISSAT; DAVIES, 1997; JABBOUR; KLIPPEL; ANTRANIKIAN, 2012). Elas hidrolisam através de um mecanismo de retenção do carbono anomérico, usando os glutamatos como nucleófilo catalítico e ácido/base catalítico (Figura 4a). Em contrapartida, as  $\beta$ -glicosidases da família GH3 apresentam estruturas menos conservadas, em geral, utilizando um aspartato e um glutamato como nucleófilo catalítico e ácido/base catalítico (Figura 4b). Nos últimos anos,  $\beta$ -glicosidases da família GH1 têm chamado

a atenção devido à alta resistência a inibição por produto, o que pode ter aplicação para degradação de celulose (YANG et al., 2015b).



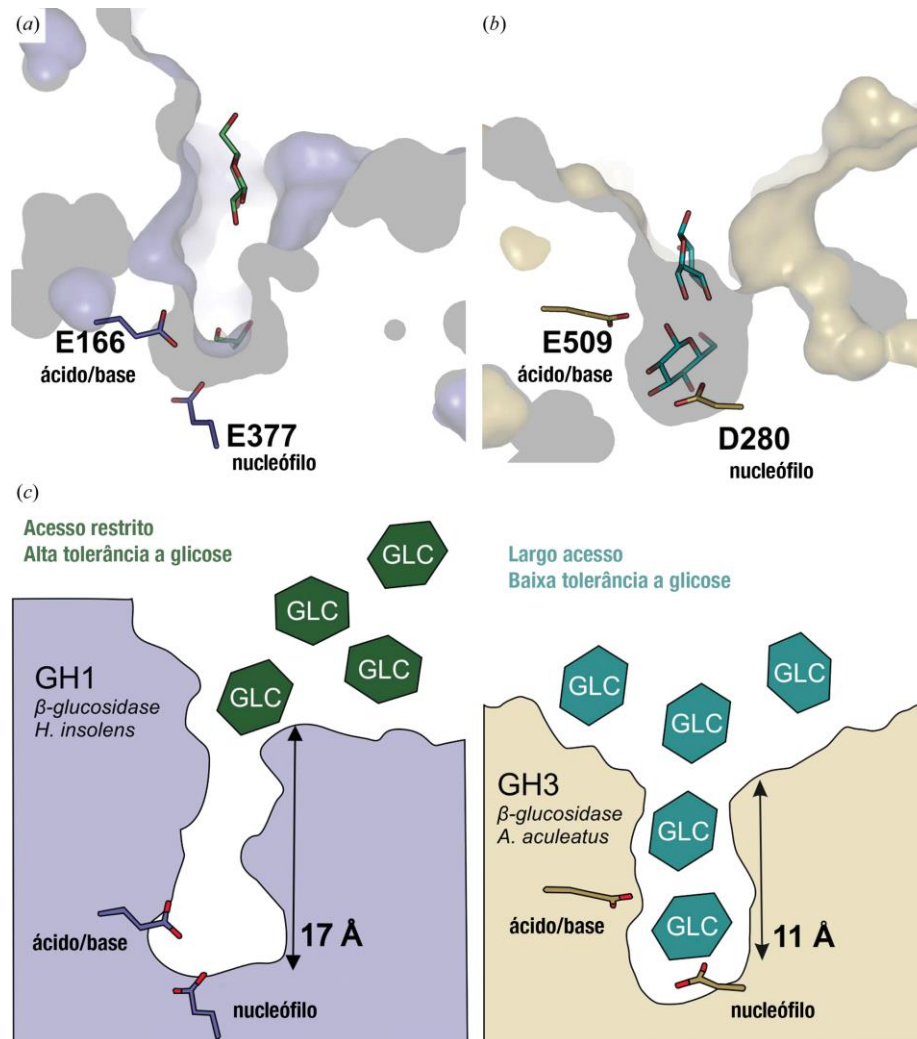
**Figura 4. Comparação estrutural entre  $\beta$ -glicosidases GH1 e GH3.**

(a)  $\beta$ -glicosidase GH1 de *Pyrococcus furiosus* (3APG). A estrutura apresenta um enovelamento  $(\alpha/\beta)_8$  TIM barril, com dois glutamatos como nucleófilo catalítico e ácido/base catalítico nas fitas-beta 4 e 7 presentes no barril; (b)  $\beta$ -glicosidase GH3 de *Mucor circinelloides* (modelado; *template*: 4I8D). A estrutura apresenta um aspartato como nucleófilo catalítico e um glutamato como ácido/base catalítico. Imagens geradas com PyMOL (<http://pymol.org>). Fonte: próprio autor.

A maior parte das  $\beta$ -glicosidases reportadas como glicose-tolerantes pertencem à família GH1. De fato,  $\beta$ -glicosidases da família GH1 têm sido reportadas entre 10 e 1000 vezes mais tolerantes que  $\beta$ -glicosidases GH3 (DE GIUSEPPE et al., 2014). Entretanto, detectou-se duas  $\beta$ -glicosidases GH3 com alta atividade em concentrações de glicose de 140 mM (HUANG et al., 2014) e 400 mM (RAMANI et al., 2015a).

COTA et al. (2015) destacaram o potencial para uso industrial de enzimas  $\beta$ -glicosidase GH1 devido a ampla especificidade de substratos e fraca inibição por glicose. DE GIUSEPPE et al. (2014) argumentam que as  $\beta$ -glicosidases GH1 têm um canal profundo e estreito na entrada do sítio ativo (Figura 5a-c). Tal canal pode estar relacionado com a maior tolerância a glicose das enzimas GH1. Em oposição, as  $\beta$ -glicosidases da família GH3 têm a concavidade que leva ao sítio ativo mais larga. Isso facilita a entrada do substrato, o que pode

levar a uma maior atividade catalítica. Entretanto, também facilita a retenção do produto, causando uma maior inibição, mesmo com baixas concentrações de glicose (Figura 5b-d).



**Figura 5. Comparação entre bolsões catalíticos de  $\beta$ -glicosidases GH1 e GH3.**

(a) Acesso de glicose ao sítio ativo da  $\beta$ -glicosidase glicose-tolerante GH1 de *H. insolens*. (b) Acesso de glicose ao sítio ativo da  $\beta$ -glicosidase GH3 de *A. aculeatus*. Comparação entre bolsões catalíticos de (c) GH1 e GH3: o canal das GH1 é mais profundo e estreito que o das GH3. Fonte: adaptado de DE GIUSEPPE et al. (2014).

CAO et al. (2015) sugeriram que uma  $\beta$ -glicosidase ideal poderia ser obtida melhorando a tolerância a glicose das GH3 ou melhorando a constante especificidade ( $k_{cat} / K_m$ ) das GH1. BREVES et al. (1997) reportaram duas  $\beta$ -glicosidases das famílias GH1 e GH3, respectivamente, presentes em um mesmo operon de *Thermoanaerobacter brockii*. Isso sugere que a diversidade de  $\beta$ -glicosidases de diferentes famílias pode ser de alguma forma benéfica para o organismo. Logo, isso pode ser uma evidência de que o uso de coquetéis suplementares com  $\beta$ -glicosidases de diferentes famílias pode melhorar a produção de biocombustíveis.

## 1.2 Tolerância a glicose

Glicose inibe a atividade de  $\beta$ -glicosidases competindo com o substrato (celobiose) na interação com o sítio ativo. Portanto,  $\beta$ -glicosidases tolerantes à glicose podem aumentar a produção e reduzir custos (YANG et al., 2015b). MELEIRO et al. (2015) relataram que a hidrólise enzimática produz concentrações de glicose que variam de 650 a 1000 mM. Assim, a  $\beta$ -glicosidase ideal deveria estar ativa mesmo em concentrações de glicose maiores que essas.

A constante de inibição ( $K_i$ ) para glicose tem sido reportada como o principal parâmetro utilizado para medir a inibição por produto. Quanto menor o valor da constante de inibição, conseqüentemente, maior a inibição. Uma análise da literatura demonstrou que a  $\beta$ -glicosidase de *Bacillus subtilis* (CHAMOLI et al., 2016) apresenta a mais alta constante de inibição por glicose (1,9 M) dentre as estruturas disponíveis em bancos de dados públicos (Tabela 1). Quando a constante de inibição para glicose não estava disponível, os artigos reportavam outro parâmetro de tolerância, como o  $IC_{50}$ . O  $IC_{50}$  mede a concentração de inibidor quando a velocidade da reação atinge 50% da velocidade máxima. Entretanto, em alguns trabalhos não se tem o  $IC_{50}$  calculado precisamente. Por exemplo, a  $\beta$ -glicosidase GH3 obtida no organismo *Mucor circinelloides* mantém 84% da sua atividade a uma concentração de glicose acima de 140 mM (HUANG et al., 2014).

**Tabela 1. Lista de  $\beta$ -glicosidases tolerantes a inibição por glicose.**

#	ORGANISMO	UNIPROT	PDB	GH	T <sup>A</sup>	PH	K <sub>i</sub> <sup>B</sup>	IC <sub>50</sub> <sup>C</sup>	K <sub>M</sub> <sup>D</sup>	FONTE
1	<i>Bacillus subtilis</i>	I3QIG4	ND*	1	60	6	1900	ND*	ND*	(CHAMOLI et al., 2016)
2	Metagenoma de <i>China South Sea</i>	D5KX75	ND*	1	40	6.5	ND*	1000	20.4	(Yang et al., 2015b)
3	Metagenoma de <i>Turpan Depression</i>	A0A0F7K KB7	ND*	1	50	5	ND*	3500	ND*	(CAO et al., 2015)
4	<i>Exiguobacterium antarcticum</i> B7	K0A8J9	5DT5	1	30	7	ND*	1000	6.18	(CRESPIM et al., 2016)
5	Metagenoma de <i>Kusaya gravy</i>	HV348683. 1 <sup>e</sup>	ND*	1	45	5-6.5	ND*	>750	0.36	(Uchiyama et al., 2015)
6	<i>Thermoanaerobacterium aotearoense</i>	A0A0H4N XH8	ND*	1	60	6	800	ND*	ND*	(Yang et al., 2015a)
7	<i>Talaromyces funiculosus</i> <sup>h</sup>	K4KB38	ND*	3	60	5	ND*	400	1.25	(RAMANI et al., 2015a)
8	<i>Humicola grisea</i> var. <i>thermoidea</i>	O93784	4MDO	1	40	6	ND*	>450	ND*	(Benoliel et al., 2010; de Giuseppe et al., 2014)
9	Metagenoma de solo	K4I4U1	ND*	1	50	6	ND*	>300	2.09	(LU et al., 2013)
10	<i>Thermoanaerobacterium thermosaccharo-</i>	D9TR57	ND*	1	70	6.4	600	ND*	7.9	(PEI et al., 2012)

	<i>lyticum</i>									
11	<i>Fervidobacterium islandicum</i>	G8YZD7	ND*	1	90	7	211	ND*	ND*	(Jabbour et al., 2012)
12	<i>Mucor circinelloides</i>	ND*	ND*	3	50	5	ND*	>140	4.55	(HUANG et al., 2014)
13	<i>Trichoderma reesei</i> <sup>g</sup>	O93785	ND*	1	30-40 <sup>f</sup>	5-7	ND*	650	2.48 <sup>f</sup>	(Guo et al., 2016)
14	<i>Thermotoga naphthophila</i>	D2C6W2	ND*	1	80	7	1200	ND*	7.76	(AKRAM et al., 2016)
15	<i>Caldicellulosiruptor bescii</i>	B9MNR1	ND*	1	85	6.8	113.8	ND*	90.8	(BAI et al., 2013)
16	<i>Neurospora crassa</i>	U9W8B8	ND*	1	40-45	5.5-6.5	ND*	950	0.21	(MELEIRO et al., 2015)
17	<i>Pyrococcus furiosus</i>	E7FHY4	3APG	1	> 80	ND*	207	ND*	ND*	(COTA et al., 2015)
18	<i>Thermotoga petrophila</i>	A5IL97	ND*	1	> 80	ND*	1100	ND*	ND*	(COTA et al., 2015)
19	<i>Acidilobus saccharovorans</i>	D9PZ08	4HA3	1	93	6	500	ND*	ND*	(GUMEROV et al., 2015)
20	Metagenoma <i>Hydrothermal Spring</i>	W8W3B8	ND*	1	90	6.5	150	ND*	ND*	(SCHRÖDER et al., 2014)
21	<i>Neotermes koshunensis</i>	Q8T0W7	3AHZ	1	50	5	ND*	>1000	ND*	(Uchima et al., 2011; de Giuseppe et al., 2014)
22	<i>Thermoanaerobacter brockii</i>	Q60026	ND*	1	ND*	ND*	200	ND*	ND*	(BREVES et al., 1997)
23	<i>Nasutitermes takasagoensis</i>	D0VYR9	ND*	1	65	5.5	ND*	>600	ND*	(Uchima et al., 2012)

<sup>a</sup> Temperatura ótima (em °C).

<sup>b</sup> Constante de inibição (K<sub>i</sub>) para glicose (em mM).

<sup>c</sup> IC<sub>50</sub> (em mM).

<sup>d</sup> Afinidade (K<sub>m</sub>) para celobiose (em mM).

<sup>e</sup> Sequência obtida no GenBank (gi|347872408|dbj|HV348683.1).

<sup>f</sup> Mutante L167W/P172L aumenta a temperatura ótima para 55°C e reduz afinidade por celobiose para 0,23 mM.

<sup>g</sup> *Trichoderma reesei* (teleomorfo *Hypocrea jecorina*);

<sup>h</sup> *Talaromyces funiculosus* (anamorfo *Penicillium funiculosum*).

\* ND: não determinado.

FANG et al. (2010) suspeitaram que a tolerância a inibição por glicose estava relacionada com a diferença geométrica ao redor do *loop C*, uma região que está próxima a entrada para o canal que leva ao sítio ativo. Eles sugerem que parte da cavidade deveria desempenhar um importante papel na ligação do substrato. Como destacado anteriormente, DE GIUSEPPE et al. (2014) argumentaram que β-glicosidases da família GH1 tolerantes a inibição por glicose têm um canal profundo e estreito que limita o acesso da glicose ao sítio ativo. Assim, a



forma do canal do substrato (bolsão catalítico) é responsável pela redução do acesso da glicose e, conseqüentemente, a resistência à inibição por glicose em  $\beta$ -glicosidases GH1. Os autores ainda detectaram que na  $\beta$ -glicosidase de *Humicola grisea* var. *thermoidea* (PDB: 4MDO), os aminoácidos W168 e L173 contribuem para redução do efeito inibitório, devido à aplicação de restrições nos subsítios +2, que conseqüentemente limita o acesso da glicose ao subsítio -1. Entretanto, YANG et al. (2015b) utilizaram mutações para verificar que a afinidade por glicose por resíduos na entrada e no meio do canal pode ser mais importante para o mecanismo de tolerância a glicose do que a profundidade e estreiteza do canal.

### 1.3 Estimulação por glicose

Além da tolerância a glicose, um controverso efeito estimulatório de glicose para  $\beta$ -glicosidases foi reportado em diversos trabalhos (AKRAM et al., 2016; CRESPIM et al., 2016; GUO; AMANO; NOZAKI, 2016; LU et al., 2013; MELEIRO et al., 2015; PEI et al., 2012; RAMANI et al., 2015a; SOUZA et al., 2014; UCHIYAMA; YAOI; MIYAZAKI, 2015; YANG et al., 2015a, 2015b; ZHAO et al., 2013). Esse efeito é descrito unicamente para  $\beta$ -glicosidases da família GH1 (DE GIUSEPPE et al., 2014; YANG et al., 2015b). Entretanto, nem todas as  $\beta$ -glicosidases glicose-tolerantes apresentam esse efeito estimulatório (MELEIRO et al., 2015).

Estimulação por glicose consiste num aumento da atividade de  $\beta$ -glicosidases em determinadas concentrações de glicose. Por exemplo, SOUZA et al. (2014) relataram que na presença de 50 mM de glicose a atividade da  $\beta$ -glicosidase de *Humicola insolens* RP86 era estimulada em até 1,8 vezes. CAO et al. (2015) argumentam que o efeito estimulatório por glicose ocorre devido a um efeito alostérico pela ligação da glicose a um sítio secundário ou devido a transglicosilação, reação em que o intermediário é atacado por um aceptor glicosídico formando um novo glicosídeo, como por exemplo celotriose, celotetraose ou celopentose (FRUTUOSO, 2011). Além disso, argumentaram que o resíduo V174 pode estar relacionado com tal efeito. GUO, AMANO e NOZAKI (2016) sugerem que durante o processo de sacarificação, enquanto a concentração de glicose aumenta, a inibição por substrato é gradualmente reduzida. Tal hipótese suporta a ideia de que o efeito estimulatório não ocorre devido a presença de certas quantidades de glicose, mas devido à redução de concentração de celobiose no ambiente. O que sugere que algumas  $\beta$ -glicosidases também podem ser inibidas por celobiose. Eles ainda sugeriram que os resíduos W168, L173 e F348 podem ser os responsáveis pela tolerância e estimulação por glicose, restringido o volume e largura da entrada do local ativo e

prendendo uma molécula de glicose no subsítio +2 através de interações hidrofóbicas. Entretanto, deve-se ressaltar que a base para a tolerância e estimulação não está totalmente clara (YANG et al., 2015b).

#### 1.4 Classificação de $\beta$ -glicosidases baseada em tolerância e estimulação por glicose

Muitos métodos tradicionais de classificação de enzimas não consideram a importância das características de tolerância e estimulação por glicose em  $\beta$ -glicosidases, como por exemplo a classificação baseada na especificidade da enzima ou a classificação em famílias propostas pelo CAZy (*Carbohydrate-Active enZymes*; CANTAREL et al., 2009). Por esse motivo, novos métodos de classificação têm sido propostos recentemente.

CAO et al. (2015) propuseram que, em relação à inibição por glicose,  $\beta$ -glicosidases poderiam ser divididas em três grupos:

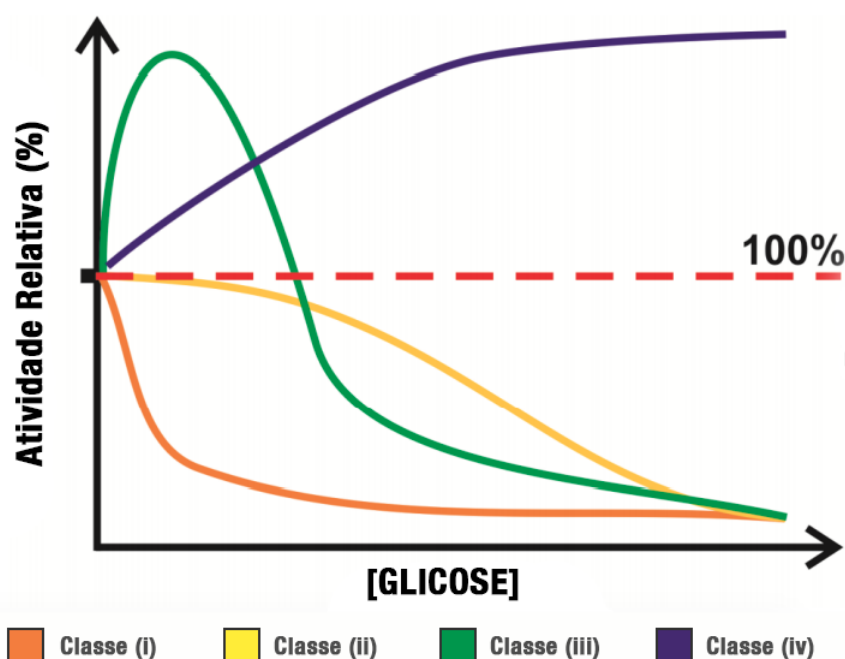
- (i)  *$\beta$ -glicosidases que são fortemente inibidas por glicose*: em  $K_i$  menores que 0,1 M (TEUGJAS; VÄLJAMÄE, 2013);
- (ii)  *$\beta$ -glicosidases que são tolerantes a baixa concentração de glicose, mas são inibidas em altas concentrações*: como por exemplo as  $\beta$ -glicosidases de *Aspergillus oryzae* ( $K_i$  de 1,36 M; UCHIMA et al., 2011); de *Candida petala* ( $K_i$  de 1,36 M; SAHA; BOTHAST, 1996) e de uma bactéria não cultivada ( $K_i$  de 4,28 M; LI et al., 2012);
- (iii)  *$\beta$ -glicosidases que são estimuladas por baixa concentração de glicose e são inibidas por alta concentração de glicose*: como por exemplo, a  $\beta$ -glicosidase do metagenoma de *Turpan Depression* que apresenta sua atividade aumentada em até quatro vezes entre concentrações de 0,2-0,6 M de glicose.

Recentemente, SALGADO et al. (2018) estenderam a classificação proposta por CAO et al. (2015) e propuseram uma nova classificação, agora levando em consideração quatro categorias de  $\beta$ -glicosidases (Figura 6):

- (i)  *$\beta$ -glicosidases fortemente inibidas por baixas concentrações de glicose*: a maior parte das  $\beta$ -glicosidases são classificadas nessa categoria. Apresenta enzimas cuja constante de inibição por glicose seja menor que 0,1 M, como por exemplo,  $\beta$ -glicosidases da família GH3 (CHAUVE et al., 2010; DECKER; VISSER; SCHREIER, 2001; HARNPICHARNCHAI et al., 2009; MELEIRO et al., 2014; YOON; KIM; CHA, 2008), algumas da família GH1 (FANG et al.,

2010; PÉREZ-PONS; REBORDOSA; QUEROL, 1995; WIERZBICKA-WOŚ et al., 2013), e diversas  $\beta$ -glicosidases ainda não categorizadas nos grupos da plataforma CAZy (AŠIĆ et al., 2015; BENOLIEL et al., 2010; BOU-DABBOUS et al., 2017; MALLEK-FAKHFAKH; BELGHITH, 2016);

- (ii)  *$\beta$ -glicosidases tolerantes à glicose*: classe composta por enzimas com valores de constante de inibição para glicose maiores que 0,1 M.  $\beta$ -glicosidases glicose-tolerantes mostram uma atividade catalítica entre 40-65° C, e pH ótimo variando entre 4,5-6,5. Esse grupo é composto majoritariamente por  $\beta$ -glicosidases da família GH1, com exceção das  $\beta$ -glucosidases GH3 de *Mucor circinelloides* (HUANG et al., 2014) e de *Talaromyces funiculosus*, anamorfo de *Penicillium funiculosum* (RAMANI et al., 2015a);
- (iii)  *$\beta$ -glicosidases estimuladas por baixas concentrações de glicose, mas inibidas por altas concentrações*: enzimas desta categoria podem apresentar um aumento na atividade, dado por um fator de estimulação (FE), que pode variar de 1,1 a 4,5 vezes. Paradoxalmente, a melhoria na atividade ocorre devido ao aumento da concentração de glicose no meio, em geral, variando entre 20 a 500 mM (SALGADO et al., 2018). Entretanto, essas enzimas ainda são inibidas em altas concentrações de glicose;
- (iv)  *$\beta$ -glicosidases não inibidas por altas concentrações de glicose*: enzimas desta categoria apresentam atividade catalítica na presença de glicose maior do que na ausência. Por exemplo, as  $\beta$ -glicosidases de *Bacillus halodurans* (XU et al., 2011), de um metagenoma microbial (UCHIYAMA; MIYAZAKI; YAOI, 2013) e de um metagenoma de solo (MATSUZAWA; YAOI, 2017) foram estimuladas a uma concentração de 1 M de glicose; a  $\beta$ -glicosidase de *Jeotgali-bacillus malaysiensis* (LIEW et al., 2018) e uma outra extraída de um solo agrícola (BIVER et al., 2014) foram estimuladas a uma concentração 2,5 M de glicose; por fim, a  $\beta$ -glicosidase de *Anoxybacillus* sp. DT3-1 foi estimulada a 5 M e reteve atividade relativa alta a até 15 M de glicose (CHAN et al., 2016).



**Figura 6. Representação do efeito da concentração de glicose na atividade relativa de β-glicosidases.**

Linha laranja: β-glicosidases fortemente inibidas por glicose (classe i); linha amarela: β-glicosidases tolerantes a glicose (classe ii); linha verde: β-glicosidases estimuladas por baixas concentrações de glicose, mas inibidas por altas concentrações (classe iii); e linha roxa: β-glicosidases não inibidas por altas concentrações de glicose (classe iv). Fonte: adaptado de SALGADO et al. (2018).

TEUGJAS & VÄLJAMÄE (2013) sugerem que os parâmetros mais importantes para avaliar enzimas β-glicosidase são a constante de especificidade para celobiose, dada pela razão entre a constante catalítica e constante de Michaelis ( $k_{cat}/K_M$ ), e constante de inibição por glicose ( $K_i$ ). Assim, β-glicosidases eficientes devem apresentar para celobiose valores mais altos de  $k_{cat}$  e mais baixos de  $K_M$ . Além disso, enzimas que apresentam altas temperaturas ótimas podem ser importantes, uma vez que, nesse mesmo estudo, o aumento de temperatura ambiente reduziu a inibição por glicose e aumentou a eficiência catalítica. Apesar de que também relataram que um aumento na tolerância por glicose tende a gerar uma redução na eficiência catalítica, ou seja, redução nos valores de constante de especificidade. Muitas β-glicosidases são inibidas pelo substrato (celobiose), o que ocorre possivelmente devido a uma reação de transglicosilação que compete com hidrólise (TEUGJAS; VÄLJAMÄE, 2013; WOODWARD; WISEMAN, 1982).

De fato, nem todos os fatores acima apresentados são levados em consideração nas classificações propostas por CAO et al. (2015) ou por SALGADO et al. (2018). Pode-se então inferir que o processo de seleção de enzimas para a hidrólise de celulose é complexo e depende de muitos fatores. Logo, a escolha da enzima β-glicosidase ideal para uma aplicação indus-

trial depende de um profundo estudo das características das enzimas obtidas e das condições às quais elas serão empregadas. Além disso, a glicose inibe a atividade enzimática das  $\beta$ -glicosidases competindo com o substrato para ligar no sítio ativo, logo a inibição aparenta ser inevitável (YANG et al., 2015b). Como retratado anteriormente, CAO et al. (2015) sugerem que a  $\beta$ -glicosidase ideal poderia ser obtida melhorando a tolerância das GH3 ou melhorando o constante especificidade das GH1. Assim, a produção de enzimas mutantes por meio de engenharia genética aparenta ser uma boa estratégia para melhoria do processo de produção de biocombustíveis de segunda geração.

### 1.5 Mutações em $\beta$ -glicosidases

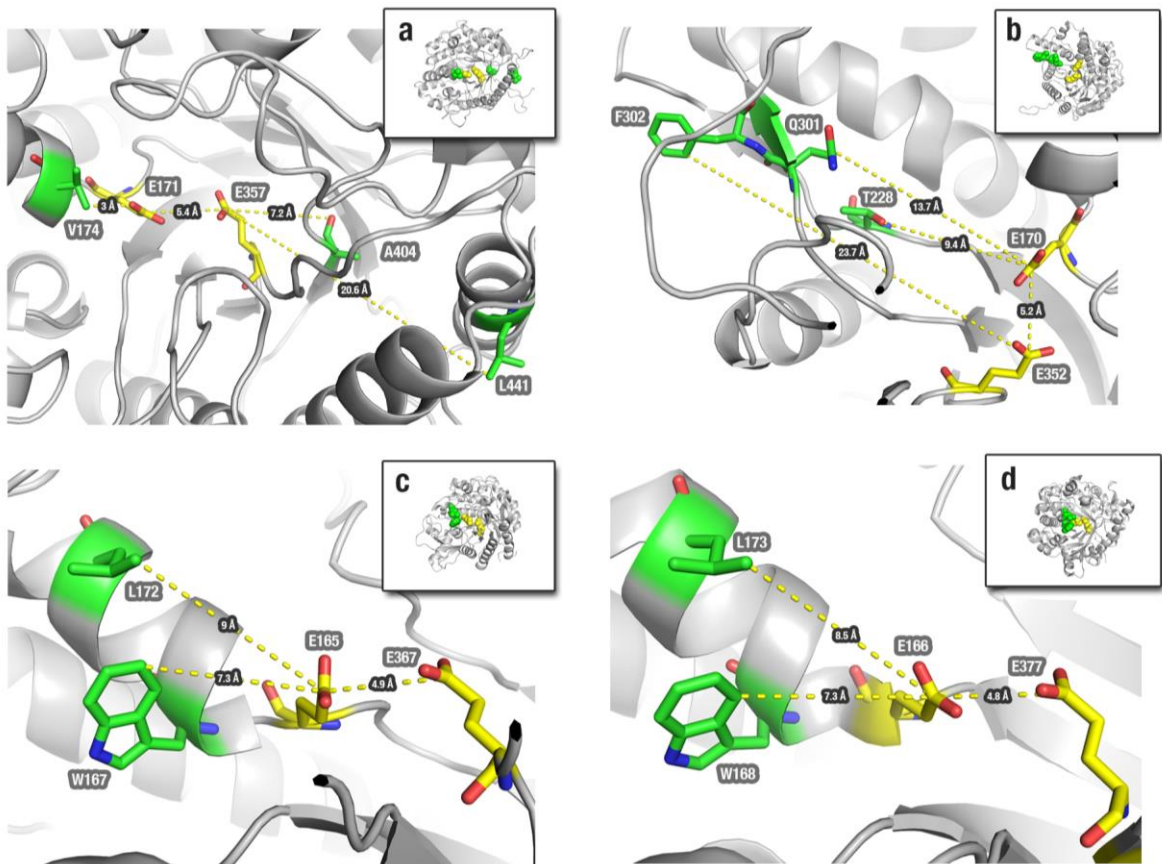
Nos últimos tempos, algumas mutações têm sido retratadas na literatura por promover um efeito negativo para a atividade de  $\beta$ -glicosidases (aqui denominadas como mutações não benéficas). Entretanto, muitas outras mutações têm sido descritas como importantes para tornar  $\beta$ -glicosidases mais eficientes para degradação de celulose (aqui denominadas como mutações benéficas). Diversos estudos têm relatado mutações pontuais, aquelas que ocorrem em um único aminoácido, ou um conjunto de mutações que trazem benefícios para a atividade da enzima, como aumento da resistência a tolerância, da temperatura ótima, da constante de inibição por glicose, da eficiência catalítica, da atividade e da constante de especificidade (Tabela 2).

**Tabela 2. Mutações coletadas na literatura e na base de dados UniProt.**

<i>id</i>	<i>Mutação</i>	<i>Efeito</i>	<i>Classificação</i>	<i>Fonte</i>
1	H228T	Responsável por prender a glicose no centro do canal que leva ao sítio ativo, levando assim à resistência a tolerância.	Benéfica	(YANG et al., 2015b)
2	V174C / A404V / L441F	Leva a um aumento da temperatura ótima de 50°C para 60°C, reduz pH ótimo de 6 para 5,5 e aumenta a meia vida de 1h para entre 2 a 20h.	Benéfica	(CAO et al., 2015)
3	H184F	Gera um aumento na constante de inibição por glicose.	Benéfica	(LIU et al., 2011)
4	P172L	Aumento na eficiência catalítica ( $V_{max}/K_m$ ).	Benéfica	(LEE et al., 2012)
5	P172L / F250A	Aumento na eficiência catalítica ( $V_{max}/K_m$ ).	Benéfica	(LEE et al., 2012)
6	L167W	Aumento da temperatura ótima para 50°C e aumento da tolerância a glicose.	Benéfica	(LEE et al., 2012)
7	L167W / P172L	Aumento da atividade em duas vezes.	Benéfica	(GUO; AMANO; NOZAKI, 2016)
8	L167W /	Aumento da atividade em 1,3x.	Benéfica	(GUO; AMANO;

	P172L / P338F			NOZAKI, 2016)
9	V168Y	Redução na atividade específica.	Não benéfica	(BERRIN et al., 2003)
10	F225S	Redução na atividade específica.	Não benéfica	(BERRIN et al., 2003)
11	Y308F	Redução na atividade específica.	Não benéfica	(BERRIN et al., 2003)
12	Y308A	Redução na atividade específica.	Não benéfica	(BERRIN et al., 2003)
13	I207V	Aumento na constante de especificidade ( $K_{cat}/K_m$ ) de nove a 24 vezes, dependendo do substrato.	Benéfica	(SANSENYA; MANEESAN; CAIRNS, 2012)
14	N218H	Redução da $K_m$ pela metade.	Benéfica	(CHUENCHOR et al., 2008)
15	N273V	Aumento da $K_m$ em cinco vezes	Não benéfica	(CHUENCHOR et al., 2008)
16	F252I	Redução da afinidade para o substrato.	Não benéfica	(ZOUHAR et al., 2001)
17	F252W	Redução da afinidade para o substrato.	Não benéfica	(ZOUHAR et al., 2001)
18	F252Y	Redução da afinidade para o substrato.	Não benéfica	(ZOUHAR et al., 2001)
19	M284N	Redução do $K_{cat}/K_m$ de sete a 30 vezes, dependendo do substrato.	Não benéfica	(SANSENYA; MANEESAN; CAIRNS, 2012)
20	H276M	Redução do $K_{cat}/K_m$ de dois a seis vezes, dependendo do substrato.	Não benéfica	(SANSENYA et al., 2011)
21	V173C	Reduz a afinidade por celobiose pela metade.	Não benéfica	(TSUKADA et al., 2008)
22	M177L	Pequena redução na afinidade para celobiose.	Não benéfica	(TSUKADA et al., 2008)
23	D229N	Redução na afinidade para celobiose em 17 vezes.	Não benéfica	(TSUKADA et al., 2008)
24	H231D	Redução na afinidade para celobiose em três vezes.	Não benéfica	(TSUKADA et al., 2008)
25	E96K	Melhoria da termoestabilidade.	Benéfica	(SANZ-APARICIO et al., 1998)
26	N223G	Redução da transglicosilação, da tolerância a glicose e da atividade de 4,62 para 0,62.	Não benéfica	(MATSUZAWA et al., 2016)
27	N223Q	Redução da transglicosilação, da tolerância a glicose e da atividade de 4,62 para 0,5.	Não benéfica	(MATSUZAWA et al., 2016)

Mutações em aminoácidos na entrada e no meio do canal que leva ao sítio ativo em uma  $\beta$ -glicosidase da família GH1 foram utilizadas para caracterizar sítios em que a glicose tem uma maior preferência de ligação em relação ao sítio ativo (Figura 7b; YANG et al., 2015b). Através das mutações H228T e N301Q/V302F, os autores desse estudo aumentaram a tolerância a glicose de uma  $\beta$ -glicosidase considerada não tolerante, denominada Bgl1B. Além disso, através de mutações, é possível revelar a importância de resíduos para atividade catalítica de  $\beta$ -glicosidases. Por exemplo, a importância do resíduo 184 para tolerância a glicose foi retratada através da mutação H184F (LIU et al., 2011).



**Figura 7. Dados estruturais de  $\beta$ -glicosidases, aminoácidos relacionados com a tolerância a glicose e ao sítio ativo.**

(a)  $\beta$ -glicosidase do metagenoma de *Turpan Depression*. Em amarelo os dois resíduos de aminoácidos do sítio ativo (E357 e E171). Em verde, resíduos da proteína selvagem. As mutações V174C/A404V/L441F aumentaram a atividade catalítica, entretanto a tolerância a glicose reduziu de 3,5M a 3M; (b)  $\beta$ -glicosidase do metagenoma de *China South Sea*. Em amarelo o sítio ativo (E170 e E352). Em verde, os aminoácidos mutados (H228T, N301Q e V302F) na entrada e no meio do canal que leva ao sítio ativo; (c)  $\beta$ -glicosidase de *Trichoderma reesei*. Em amarelo, os dois glutamatos catalíticos (E165 e E367). Em verde, as mutações L167W e P172L; (d)  $\beta$ -glicosidase de *Humicola grisea*. Ela apresenta um estreito e profundo canal que leva ao sítio ativo. Em amarelo, os resíduos ácido/base catalítico e o nucleófilo. E166 e E377 estão a uma distância de  $\sim 5$  Å. Em verde, os resíduos de aminoácidos W168 e L173 responsáveis por contribuir com a redução da inibição. Imagens geradas pelo PyMOL (<http://pymol.org>). Fonte: próprio autor.

Além da tolerância à glicose, estabilidade térmica e de pH são importantes características para melhoria da degradação de biomassa por  $\beta$ -glicosidases. Mutações sítio-dirigidas na

entrada do sítio ativo (L167W e P172L) foram utilizadas para melhorar tais características de uma  $\beta$ -glicosidase extraída de *Trichoderma reesei* (Figura 7c) (GUO; AMANO; NOZAKI, 2016).

CAO et al. (2015) realizaram mutações aleatórias em uma  $\beta$ -glicosidase obtida em uma biblioteca de metagenoma. Eles detectaram três importantes aminoácidos, V173C, A404V e L441F, e usando mutações sítio-dirigidas construíram um mutante termoestável e tolerante a inibição por glicose, chamado M3. Esse mutante aumentou a conversão do bagaço de cana entre 14 e 35% (Figura 7a).

## 1.6 Análises de bioinformática

A principal estratégia adotada para melhoria de enzimas  $\beta$ -glicosidase tem consistido no uso de mutações aleatórias, como por exemplo, por meio da técnica de EP-PCR. Entretanto, a probabilidade de uma mutação benéfica ser alcançada por essa técnica é baixa. Por exemplo, uma enzima  $\beta$ -glicosidase pertencente à família GH1 possui aproximadamente 400 resíduos. Se cada resíduo pode ser mutado em outros 19 aminoácidos é possível que se tenha um total de aproximadamente  $2,58 \times 10^{520}$  combinações possíveis ( $20^{400}$ ).

Nesse contexto, a bioinformática surge como um elemento transformador. Utilizando-se técnicas da bioinformática tradicional e estrutural é possível estabelecer as bases estruturais para melhor compreender o fenômeno da glicose-tolerância em  $\beta$ -glicosidases, e assim propor mutações para teste em bancada. Neste tópico serão abordados conceitos relacionados a análises de bioinformática.

### 1.6.1 Sequências

A evolução das tecnologias de sequenciamento de próxima geração (NGS - *Next-Generation Sequencing*) tem proporcionado um expressivo aumento na quantidade de genomas sequenciados, tanto de fungos, bactérias e até mesmo de metagenomas, principais fontes de  $\beta$ -glicosidases. Por exemplo, uma busca em janeiro de 2017 na base de dados de sequências de proteínas UniProt (<http://uniprot.org>) pelo *E.C. number* “3.2.1.21” (identificador das  $\beta$ -glicosidases) retornou 10.153 sequências. A mesma busca repetida em janeiro de 2018 retornou 17.013 sequências. Uma busca por estruturas tridimensionais no PDB (<http://www.rcsb.org>) retornou apenas 153 estruturas tridimensionais em janeiro de 2017, e 176 em janeiro de 2018, o que demonstra um grande contraste entre o número de sequências e estruturas tridimensionais disponíveis. Assim, apesar da grande quantidade de sequências de



$\beta$ -glicosidases disponíveis, pouco se sabe sobre a maioria delas. De fato, muitas dessas sequências só foram obtidas devido ao aumento de sequenciamentos de genoma completos, principalmente de bactérias, ocasionado pela recente redução dos custos (MARIANO et al., 2016).

A bioinformática tradicional tem proporcionado uma grande evolução na biologia molecular, principalmente nas análises de dados ômicos, como sequências de DNA, RNA e proteína. Novos algoritmos e ferramentas têm proporcionado uma melhor compreensão das sequências biológicas, o que poderia levar a novas descobertas científicas. Uma correta anotação da função de proteínas e correlação com sua estrutura é a base para compreensão de seu papel, o que pode ser proveitoso para construção de proteínas mais eficientes para a indústria baseado na comparação com outras proteínas.

Proteínas podem ser comparadas por meio de alinhamentos de sequências. Tais alinhamentos são utilizados para detecção de parentes evolutivos (proteínas homólogas) de uma sequência por meio de buscas em um grupo de sequências selecionadas ou nos grandes bancos de dados públicos, como UniProt, Genbank ou PDB. Alinhamentos de sequências também são úteis para detecção de mutações (NELSON et al., 2014).

Mutações podem se originar de erros no processo de replicação genética ou de danos nas fitas do DNA, que podem levar a alterações na estrutura de proteínas e que, quando passadas aos descendentes, podem ser danosas ou letais. Ocasionalmente uma mutação pode equipar melhor um organismo para sobreviver melhor em um dado ambiente. Foram mutações genéticas ocasionais, combinadas com a seleção natural, que resultaram nessa enorme quantidade de enzimas com variadas características (NELSON et al., 2014). Como exemplo, pode-se citar as  $\beta$ -glicosidases de bactérias termofílicas, que são capazes de desempenhar sua atividade catalítica em altas temperaturas, o que tem grande aplicação na indústria (AKRAM et al., 2016; BAI et al., 2013; BREVES et al., 1997; COTA et al., 2015).

Nos últimos anos, diversos métodos, algoritmos e programas de computador têm aperfeiçoado os processos de alinhamento. O alinhamento de sequências é um processo de busca eletrônica ao qual é feito o deslizamento de uma sequência sobre outra ou mais sequências até que seja encontrada um trecho com boa correspondência. A qualidade de um alinhamento é medida pelo total de posições onde resíduos de aminoácidos nas duas sequências sejam idênticos (identidade). Programas de alinhamento usam um método baseado em matrizes para selecionar um alinhamento de melhor pontuação que maximize os resíduos de aminoácidos idênticos enquanto minimiza a introdução de lacunas (intervalos ou *gaps*). Lacunas são inseridas quando sequências apresentam trechos idênticos conectados por sequências diferentes.

Para permitir que trechos correspondentes sejam alinhados ao mesmo tempo, a ferramenta de alinhamento introduz lacunas em uma das sequências para registrar os segmentos correspondentes (Tabela 8). Os programas incluem penalidades na pontuação global do alinhamento para cada lacuna introduzida para evitar alinhamentos sem informações significativas (NELSON et al., 2014).

```

Escherichia coli TGNRTIAVYDLGGGTFDISIIEIDEVDGEKTFEVLATNGDTHLGGEDFDSRLIHYL
Bacillus subtilis DEDQTILLYDLGGGTFDVSILELGDG TFEVRS TAGDNRLGGDDFDQVIIDHL
                                     └───┬───┘
                                     Intervalo

```

**Figura 8. Alinhamento de sequências de duas proteínas.**

Trecho de sequência da proteína Hsp70 de *E. coli* e *B. subtilis*. Resíduos de aminoácidos idênticos estão destacados. O intervalo (*gap*) introduzido na sequência de Hsp70 de *B. subtilis* melhora o alinhamento. Resíduos de aminoácidos idênticos estão sombreados. Fonte: (NELSON et al., 2014).

Alinhamentos podem ser globais, quando consideram que as sequências devem ser alinhadas em toda sua extensão, ou locais, quando partes das sequências podem ser alinhadas. Para alinhamento global, destaca-se o algoritmo de Needleman-Wunsch, adotado, por exemplo, pelo programa ggsearch36 (PEARSON, 2016). Dentre outros programas que realizam alinhamento de sequência, destacam-se a ferramenta BLAST (JOHNSON et al., 2008) para alinhamentos locais e buscas em grandes bases de dados, e as ferramentas ClustalW2 e Clustal Omega para alinhamentos globais e múltiplos (SIEVERS et al., 2011).

### 1.6.2 Bioinformática estrutural

Bioinformática estrutural consiste num campo da bioinformática que prevê o uso de estruturas tridimensionais para compreender as interações que ocorrem em macromoléculas ou outros tipos de moléculas, como proteínas e suas interações com substrato e produto. Semelhanças entre estruturas tridimensionais podem ser úteis para revelar relações evolutivas nas quais a homologia de sequências foi apagada pelo tempo (NELSON et al., 2014).

Estruturas tridimensionais de proteínas são cruciais para inferir sobre o mecanismo de glicose-tolerância em  $\beta$ -glicosidases. Entretanto, poucas estruturas de  $\beta$ -glicosidases glicose-tolerantes têm sido disponibilizadas em bancos de dados públicos, como o *Protein Data Bank* (PDB). Assim, modelagem comparativa tem sido a técnica mais utilizada para se obter estruturas tridimensionais a partir de sequências (BAI et al., 2013; CHAMOLI et al., 2016; CRESPIM et al., 2016; JABBOUR; KLIPPEL; ANTRANIKIAN, 2012; RAMANI et al., 2015a; YANG et al., 2015a, 2015b).

### **1.6.2.1 Modelagem comparativa**

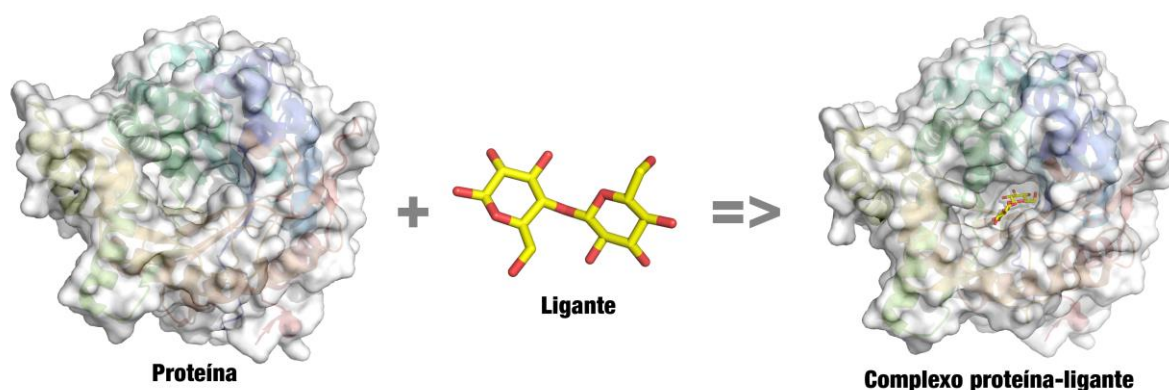
Modelagem comparativa consiste na obtenção da estrutura tridimensional de uma proteína, em geral não resolvida experimentalmente, a partir de sua sequência e de uma estrutura tridimensional referência, denominada *template* ou estrutura molde. A escolha do *template* é realizada a partir de alinhamentos de sequências entre a proteína estudada e o banco de dados PDB. Convencionou-se que a sequência molde deve ter uma identidade maior que 25% com a proteína estudada (BITAR; FRANCO, 2014).

A modelagem comparativa se difere da modelagem *ab initio* (*de novo*) por depender de uma estrutura referência. Entretanto, essa estratégia tende a apresentar estruturas modelo mais próximas da estrutura real. Um dos programas que realiza modelagem comparativa é o MODELLER (MARTÍ-RENOM et al., 2000; SALI; BLUNDELL, 1993; WEBB; SALI, 2014). MODELLER utiliza a técnica de restrição espacial, na qual a distância entre átomos de resíduos equivalentes ao *template* é utilizada por heurísticas na construção de modelos. Para avaliação e seleção dos melhores modelos, MODELLER apresenta o potencial estatístico DOPE (*Discrete Optimized Protein Energy*). Além disso, o número de resíduos em posições favoráveis pode ser utilizado nessa avaliação. O gráfico de Ramachandran permite a visualização de ângulos de torção  $\Psi$  (*psi*) e  $\Phi$  (*phi*) de resíduos de uma proteína. O ângulo de torção *phi* define a rotação em torno da ligação do carbono alfa ao nitrogênio de um resíduo, enquanto *psi* define a rotação em torno da ligação carbono alfa a outro carbono do mesmo resíduo. Devido às propriedades das cadeias laterais dos resíduos, nem todos os ângulos *phi* e *psi* são permitidos. Assim, a presença de ângulos não favoráveis pode ser utilizada para eliminar modelos com alta probabilidade de erros. Dentre os programas que geram o gráfico de Ramachandran pode-se citar o RAMPAGE (LOVELL et al., 2003).

### **1.6.2.2 Ancoramento molecular**

Ancoramento molecular (também conhecido como atracamento molecular, docagem molecular ou *docking*) é um método da bioinformática estrutural que busca simular as interações entre diferentes moléculas, prevendo orientações mais favoráveis para uma das moléculas em relação a outra (Figura 9). Programas de *docking* possuem estratégias distintas, que em geral, utilizam campos de força, que são representações que definem as interações que ocorrem nas moléculas. Ancoramento molecular tem sido utilizado para simular as interações entre proteína e substrato (BAI et al., 2013; CHAMOLI et al., 2016; YANG et al., 2015b). Por exemplo,

para entender as ligações de celobiose e glicose em  $\beta$ -glicosidases (KHAIRUDIN; MAZLAN, 2013).



**Figura 9. Representação do *docking* entre substrato e proteína.**

Nesta exemplificação, celobiose foi utilizada como ligante e a proteína  $\beta$ -glicosidase (PDB ID: 1BGA) como receptor. Figura gerada pelo PyMOL (<http://pymol.org>). Fonte: próprio autor.

*Docking* tem sido utilizado, por exemplo, para detectar diferenças entre  $\beta$ -glicosidases mutantes. Em um estudo recente, diversos mutantes de uma  $\beta$ -glicosidase obtida em um metagenoma marinho demonstraram que a glicose prefere se ligar a um sítio secundário ao invés de se ligar ao sítio ativo (YANG et al., 2015b). Tais estudos podem ajudar a elucidar o mecanismo de resistência a inibição das GH1. Em um outro estudo com uma  $\beta$ -glicosidase de *Bacillus subtilis*, CHAMOLI et al. (2016) reportaram que a docagem molecular da celobiose mostrou uma maior afinidade do que outros substratos, o que não foi consistente com os dados cinéticos. Isso demonstra que ainda há dificuldades em lidar com ancoramento dos açúcares em  $\beta$ -glicosidases. *Docking* tende a trabalhar com moléculas estáticas, o que não condiz o estado natural da molécula. Estudos mais precisos poderiam ser obtidos com simulações de dinâmica molecular, entretanto a realização de dinâmica molecular requer altos custos computacionais.

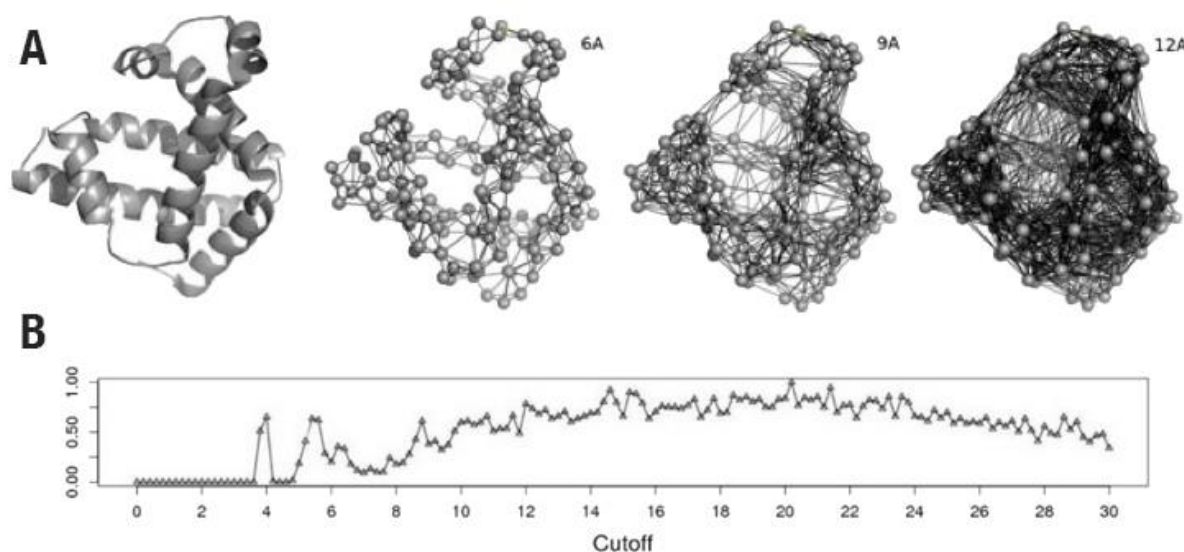
### 1.6.3 Assinaturas estruturais

Assinaturas estruturais, também conhecidas como *fingerprints*, são representações de conjuntos de características que descrevem semelhanças entre estruturas tridimensionais, visando definir, agrupar ou discriminar um conjunto de dados biológicos (DE MELO-MINARDI, 2008; PIRES, 2012). Assinaturas estruturais de proteínas podem ser construídas a partir de modelagem computacional em grafos. Um grafo consiste de um conjunto de vértices interligados ou não por um conjunto de arestas (PIRES, 2012).

Assinaturas baseadas em grafos têm sido utilizadas para classificação e anotação automática de proteínas (PIRES et al., 2011; PIRES, 2012), predição das interações proteína-ligante (PIRES et al., 2013), predição do efeito de mutações na estabilidade da proteína (PIRES; ASCHER; BLUNDELL, 2014), predição do impacto de mutações na afinidade entre proteína e ligante (PIRES; BLUNDELL; ASCHER, 2016) e predição do impacto de mutações na afinidade entre anticorpos e antígenos (PIRES; ASCHER, 2016).

### 1.6.4 Cutoff Scanning Matrix

*Cutoff Scanning Matrix* (CSM) é um método para construção de assinaturas estruturais baseada em grafos (PIRES et al., 2011). O método CSM permite a construção de matrizes numéricas que, por meio de técnicas de aprendizado de máquina, podem ser utilizadas para classificar proteínas com alta acurácia. Segundo o algoritmo do CSM, cada centroide de um resíduo é representado por um vértice e a ligação entre resíduos até uma certa distância (*cutoff*) é representada por arestas (Figura 10a). A contagem de pares de resíduos é realizada em variadas distâncias (*steps*) e, por fim, uma matriz com a quantidade de pares de resíduos para cada distância representa a assinatura da proteína (Figura 10b).



**Figura 10. Construção da assinatura estrutural usando CSM.**

(A) Representação da modelagem em grafos de uma proteína. *Cutoff* de 6 a 12 Å (com *steps* de 3 Å). (B) Matriz de assinatura da proteína. Nesse exemplo, avalia-se um *cutoff* variando de 0 a 30 Å, com valor de *step* de 0,2 Å. Fonte: Adaptado de PIRES (2012).

A função que gera a assinatura CSM recebe um conjunto de proteínas, que serão representadas com uma linha da matriz final, uma distância mínima e máxima para *cutoff* e a distância de variação (*step*). A seguir, o programa calcula a distância euclidiana entre todos os resíduos de cada proteína, levando em consideração a posição espacial do carbono alfa como

ponto representativo para um resíduo (centroide). Por fim, CSM calcula, para cada distância *cutoff*, quantos pares de resíduos existem. Por exemplo, se o programa receber como distância mínima o valor zero, distância máxima o valor dez e como distância de variação o valor dois, CSM inicialmente irá calcular quantos carbonos alfa estão a uma distância maior que 0 e menor ou igual a 2 Å, a seguir entre 2-4 Å, 4-6 Å, 6-8 Å, e por fim entre 8-10 Å. Cada um desses *cutoff* é representado por uma coluna na matriz final (Figura 11).

---

### Algoritmo. Cálculo da Cutoff Scanning Matrix

---

```

1:  função geraCSM (conjuntoProteínas, CSM, distânciaMIN, distânciaMAX, distânciaSTEP)
2:  para cada proteína i ∈ (conjuntoProteínas) faça
3:    j = 0
4:    Calcula a distância entre todos os pares de carbonos-alfa (cα)
5:    para dist = distânciaMIN; até distânciaMAX; variando distânciaSTEP faça
6:      CSM[i][j] = frequência de pares de cα dentro da distância dist
7:      j++
8:  retorna CSM

```

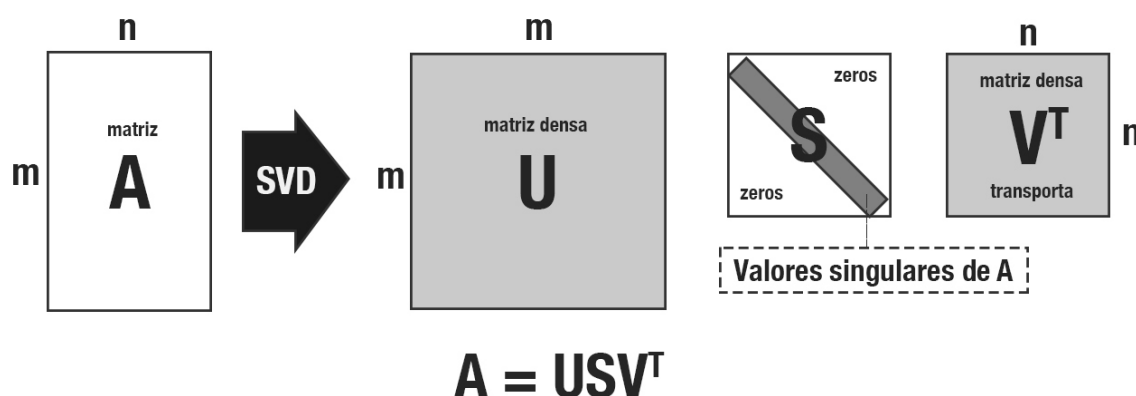
---

**Figura 11. Algoritmo para o cálculo do CSM.**

Fonte: Adaptado de PIRES et al. (2011).

O método CSM foi utilizado para classificação estrutural de proteínas levando em consideração classes, enovelamento, superfamília e família, e obteve uma precisão de 0,927, 0,868, 0,871, 0,888, respectivamente (PIRES et al., 2011). Para isso, foi utilizada a base de dados Full-SCOP e o algoritmo de KNN (*k-nearest neighbors*). Os autores ainda propuseram que o uso de técnicas de redução de ruídos melhoraria a precisão da classificação. Aplicando a técnica de decomposição por valores singulares obtiveram uma precisão de 0,954, 0,922, 0,926, 0,938, respectivamente.

A redução dimensional e de ruídos apresentada no método CSM utiliza a técnica de decomposição em valores singulares, também conhecida como SVD (*singular value decomposition*). SVD é uma técnica de álgebra linear que permite a redução dimensional de uma matriz A, de tamanho m x n, em três matrizes: U, S e V (Figura 12). Onde U é uma matriz de tamanho m x m, V é uma matriz de tamanho n x n, e S é uma matriz cuja diagonal armazena os valores singulares da matriz A. U e V são matrizes ortonormais, *i.e.* quando multiplicadas por sua matriz transposta obtém-se a matriz de identidade (PIRES, 2012).



**Figura 12.** Decomposição em valores singulares: a matriz  $A$  é decompostas nas matrizes  $U$ ,  $S$  e  $V^T$ .

Fonte: próprio autor.

O uso da decomposição em valores singulares permite que dados pouco representativos (ruídos) sejam removidos das matrizes de assinatura, reduzindo o custo computacional para processamento das matrizes e melhorando resultados de algoritmos de classificação. Ao aplicar uma técnica de redução dimensional para reduzir o tamanho da matriz, entretanto, o número de dimensões representativas pode ser maior do que dois, o que dificulta a visualização dos dados. O método de MARCOLINO, COUTO e SANTOS (2010) pode ser utilizado para visualizar os dados multidimensionais em um ambiente bidimensional, entretanto os cálculos para comparação entre assinaturas deve ser realizado usando o vetor com todas as dimensões representativas.

### 1.6.5 Assinaturas de nível atômico (aCSM)

*Atomic Cutoff Scanning Matrix* (aCSM) é um método de assinatura baseado em grafos que leva em consideração o nível atômico para modelagem (PIRES et al., 2013). Nessa abordagem cada átomo é representado por um nó do grafo e os pares de átomos, que atendam as distâncias de corte (*cutoff*), como arestas. O aCSM foi desenvolvido especialmente para modelagem computacional da estrutura de bolsões de proteínas. Três modelos de assinatura foram propostos para aCSM:

- **aCSM:** gera um valor por *cutoff* correspondente ao total de pares de átomos na faixa de distância analisada;
- **aCSM-HP (*Hydrophobic patche*):** gera três valores por *cutoff*. O primeiro remete ao total de pares de átomos na faixa de distância analisada entre dois átomos polares; o segundo indica quantos desses pares de átomos são compostos por dois átomos hidro-

fóbicos; e o terceiro indica quantos pares de átomos são compostos por um átomo hidrofóbico e um outro átomo polar;

- **aCSM-ALL:** considera oito categorias para propriedades farmacofóricas de átomos propostas pelo programa PMapper (pH 7). As oito categorias são: hidrofóbico, positivo, negativo,ceptor, doador, aromático, enxofre e neutro. Entretanto, nem todas as combinações entre categorias são realizadas, e por isso, aCSM gera apenas 36 combinações por *cutoff*.

O algoritmo de aCSM (Figura 13) é bastante similar ao algoritmo do CSM, apresentando apenas algumas diferenças. Ao invés de utilizar uma única posição espacial para representar um resíduo (CSM utiliza o carbono alfa), aCSM calcula a distância entre todos os átomos. Além disso, aCSM define uma classe de acordo com os grupos farmacofóricos dos resíduos de aminoácidos, que poderão ser utilizados na construção da matriz de acordo com a versão da assinatura utilizada. Por exemplo, para o cálculo da assinatura de uma proteína, utilizando-se como parâmetros de entrada a distância mínima de 0 Å, máxima de 30 Å e variação de 0,2 Å seria gerado um vetor com 151 colunas (aCSM), 453 colunas (aCSM-HP) e 5436 colunas (aCSM-ALL).

---

### Algoritmo. Cálculo da Atomic Cutoff Scanning Matrix

---

```

1:  função aCSM (conjuntoProteínas, classeDeÁtomo, distânciaMIN, distânciaMAX, distânciaSTEP)
2:  para cada proteína i ∈ (conjuntoProteínas) faça
3:    j = 0
4:    distMatriz = calculaDistânciaDosParesDeÁtomos (proteína)
5:    para dist = distânciaMIN; até distânciaMAX; variando distânciaSTEP faça
6:      para cada classe ∈ (classeDeÁtomo) faça
7:        aCSM[i][j] = retornaFrequência (distMatriz, dist, classe)
8:      j++
9:  retorna aCSM

```

---

**Figura 13. Algoritmo de cálculo do aCSM.**

Fonte: Adaptado de PIRES et al. (2013).

$$V_{TAM} = Q_{tipos} \left[ \frac{(D_{MAX} - D_{MIN})}{D_{STEP}} + 1 \right] \quad (1)$$

A equação (1) apresenta a fórmula para cálculo do tamanho do vetor de assinatura, onde  $Q_{tipos}$  indica a quantidade de tipos de pares avaliado (um para aCSM, três para aCSM-HP e 36 para aCSM-ALL);  $D_{MAX}$  a distância máxima;  $D_{MIN}$  a distância mínima; e  $D_{STEP}$  a distância de variação avaliada por *cutoff*.



Por levar em consideração as informações farmacofóricas dos átomos, aCSM-ALL é indicado para determinar assinaturas de bolsões em proteínas e prever as interações com ligantes. Por essa razão, ele poderia ser utilizado para determinar a assinatura dos bolsões catalíticos de  $\beta$ -glicosidases.

### 1.7 Hipóteses

Após uma análise sistemática da literatura (MARIANO et al., 2017a), detectou-se estudos referentes a enzimas  $\beta$ -glicosidase tolerantes a inibição por glicose com dados disponíveis para análises de bioinformática (MARIANO et al., 2017b; ver apêndices). Dada a importância das enzimas  $\beta$ -glicosidase e a necessidade de melhoria dessas enzimas para a indústria nacional de biocombustíveis de segunda geração, as seguintes hipóteses foram levantadas:

- (i) Resíduos importantes para a característica de tolerância a inibição por glicose em  $\beta$ -glicosidases possivelmente se localizam na concavidade que leva ao sítio ativo;
- (ii) logo, toda essa região, aqui denominada bolsão catalítico, seus resíduos e propriedades farmacofóricas devem apresentar padrões conservados em  $\beta$ -glicosidases tolerantes a inibição que poderiam ser utilizados para classificá-las;
- (iii) assim, assinaturas estruturais poderiam ser utilizadas para identificar  $\beta$ -glicosidases glicose-tolerantes e propor mutações que melhorassem a atividade de  $\beta$ -glicosidases não tolerantes.

Com base nessas hipóteses, propõe-se um método para verificação da variação das assinaturas de enzimas  $\beta$ -glicosidase. Esse método pode ser utilizado para simulação de mutações pontuais, além da classificação dessas mutações como benéficas ou não benéficas para melhoria da atividade de  $\beta$ -glicosidases.

## 2. Objetivos

### 2.1 Objetivo geral

Utilizar técnicas de bioinformática, como assinaturas estruturais, para propor mutações sítio-dirigidas em enzimas  $\beta$ -glicosidase que proporcionem uma melhoria no processo de produção de biocombustíveis de segunda geração.

### 2.2 Objetivos específicos

- Encontrar estruturas de  $\beta$ -glicosidases resistentes à inibição por glicose;
- Construir uma base de dados para armazená-las;
- Detectar resíduos importantes para tolerância a inibição por glicose;
- Detectar uma assinatura estrutural que caracterize enzimas resistentes à inibição por glicose;
- Desenvolver um modelo computacional para sugestão de mutações em  $\beta$ -glicosidases e possivelmente estender esse modelo para outras enzimas.

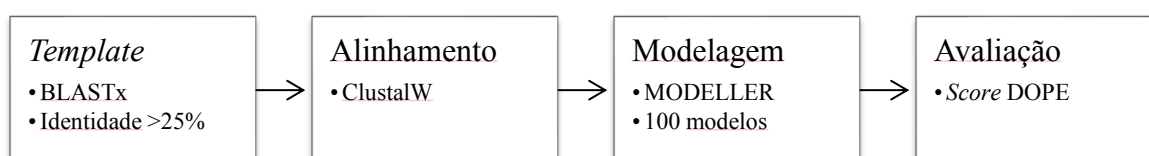
### 3. Material e métodos

#### 3.1 Coleta dos dados

Sequências de 23  $\beta$ -glicosidases glicose-tolerantes foram coletadas nas bases de dados UniProt (<http://www.uniprot.org>) e GenBank (<http://www.ncbi.nlm.nih.gov/genbank>). Estruturas tridimensionais de cinco proteínas foram obtidas na base de dados *Protein Data Bank* (PDB; <http://pdb.org>; BERMAN et al., 2000). Foi realizada a comparação entre as sequências coletadas utilizando-se o algoritmo de Needleman-Wunsch para alinhamento global com o *software* ggsearch36 (PEARSON, 2016).

#### 3.2 Pipeline para modelagem comparativa

A fim de realizar a modelagem de sequências sem estruturas tridimensionais disponíveis foi construído um *pipeline* automático para modelagem comparativa (Figura 14). Um *pipeline* automático constitui num conjunto de *software* que são sistematicamente executados por um *script*. A construção de um *pipeline* automático permite que modelagens comparativa sejam executadas de maneira rápida, sistemática e consistente com uma metodologia única.

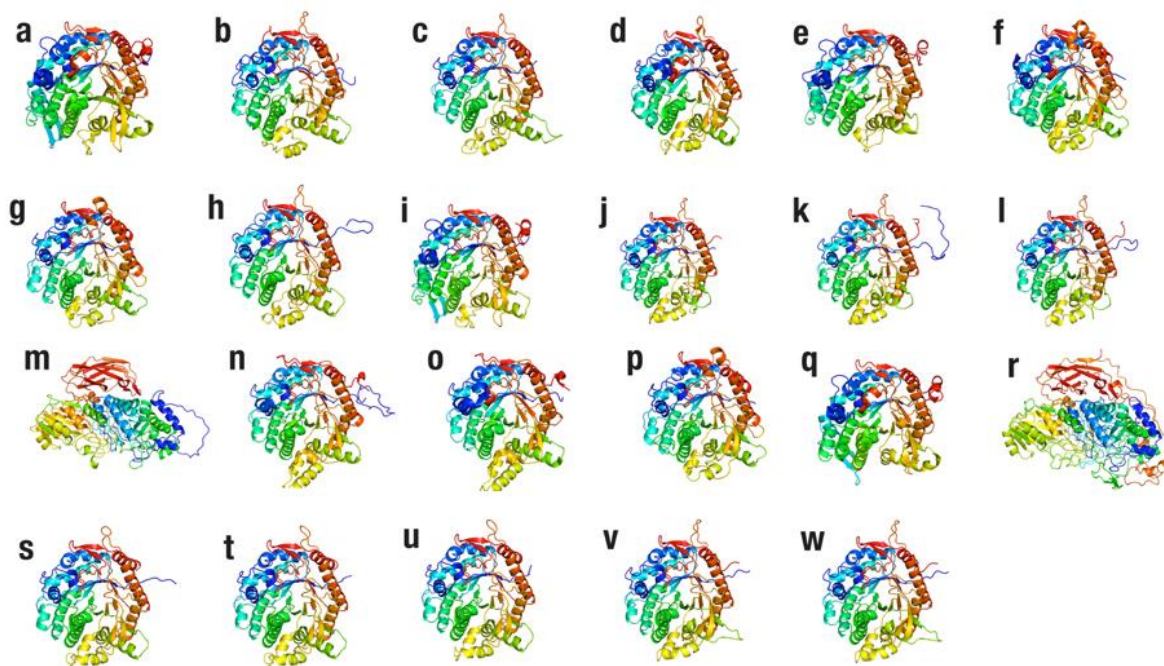


**Figura 14. Pipeline para modelagem de 18 sequências de  $\beta$ -glicosidases glicose-tolerantes sem estruturas tridimensionais disponíveis.**

Quatro etapas são propostas: (i) definição do *template* com BLASTx na base de dados PDB; (ii) alinhamento de sequências com ClustalW; (iii) modelagem comparativa com MODELLER; e (iv) avaliação e definição do melhor modelo usando a pontuação DOPE. Fonte: próprio autor.

A metodologia foi baseada no *pipeline* de BITAR & FRANCO (2014) e foi dividida em quatro etapas: (i) definição do *template*; (ii) alinhamento de sequências; (iii) modelagem comparativa; e (iv) avaliação e definição do melhor modelo (Figura 14). A estrutura modelo é escolhida com base nos resultados de maior identidade do BLASTx usando como base de dados de busca o PDB (Figura 15). Se não houver estruturas referência com identidade maior

que 25% no PDB não é possível realizar a modelagem. A cobertura também é um fator importante. A maior parte das  $\beta$ -glicosidases coletadas pertencem a família GH1, que possui uma estrutura tridimensional altamente conservada. O alinhamento entre molde e sequência estudada foi realizado com clustalw v2.1. Para cada sequência foram construídos 100 modelos usando MODELLER (usando parâmetros padrão). O melhor modelo foi selecionado baseado no menor valor da função DOPE (SHEN; SALI, 2006).



**Figura 15. Estruturas cristalográficas e modelos de  $\beta$ -glicosidases.**

(a) *A. saccharovorans* (PDB ID: 4HA3); (b) *B. subtilis* (template: 4IPL); (c) *C. bescii* (template: 3AHX); (d) *E. antarcticum* (PDB ID: 5DT5); (e) *F. islandicum* (template: 4HA3); (f) *H. grisea* (PDB ID: 4MDO); (g) *T. reesei* (template: 3AHY); (h) Metagenome China South Sea (template: 3AHX); (i) Metagenome hydrothermal spring (template: 4HA3); (j) Metagenome Kusaya gravy (template: 1OD0); (k) Metagenome soil (template: 1OD0); (l) Metagenome Turpan Depression (template: 1OD0); (m) *M. circinelloides* (GH3; template: 4I8D); (n) *N. takasagoensis* (template: 3AHZ); (o) *N. koshunensis* (PDB ID: 3AHZ); (p) *N. crassa* (template: 4MDO); (q) *P. furiosus* (PDB ID: 3APG); (r) *P. funiculosum* (GH3; template: 4D0J); (s) *T. brockii* (template: 5DT5); (t) *T. aotearoense* (template: 5DT5); (u) *T. thermosaccharolyticum* (template: 5DT5); (v) *T. naphthophila* (template: 1OD0); e (w) *T. petrophila* (template: 1OD0). Somente a cadeia A foi exibida. Imagens geradas com PyMOL (<http://pymol.org>). Fonte: próprio autor.

O pipeline foi utilizado para modelar as 18  $\beta$ -glicosidases sem estrutura tridimensional disponível. Ao todo 1800 modelos foram construídos. Selecionou-se um representante para cada  $\beta$ -glicosidase baseado no menor *score* DOPE. A quantidade de resíduos em posições desfavoráveis para cada modelo também foi verificada (dados não disponíveis), mas constatou-se poucas variações entre diferentes modelos de uma mesma proteína.

### 3.2.1 Base de dados BETAGDB

Foi construída uma base de dados, denominada BETAGDB, e um website para disponibilizar as estruturas tridimensionais coletadas ou modeladas e visualizá-las usando 3Dmol (REGO; KOES, 2015). Assim, os 18 modelos de  $\beta$ -glicosidases, além das cinco estruturas disponíveis no PDB passaram a constituir a primeira base de dados de  $\beta$ -glicosidases glicose-tolerantes (Figura 15). BETAGDB está disponível em: <<http://bioinfo.dcc.ufmg.br/betagdb>>.

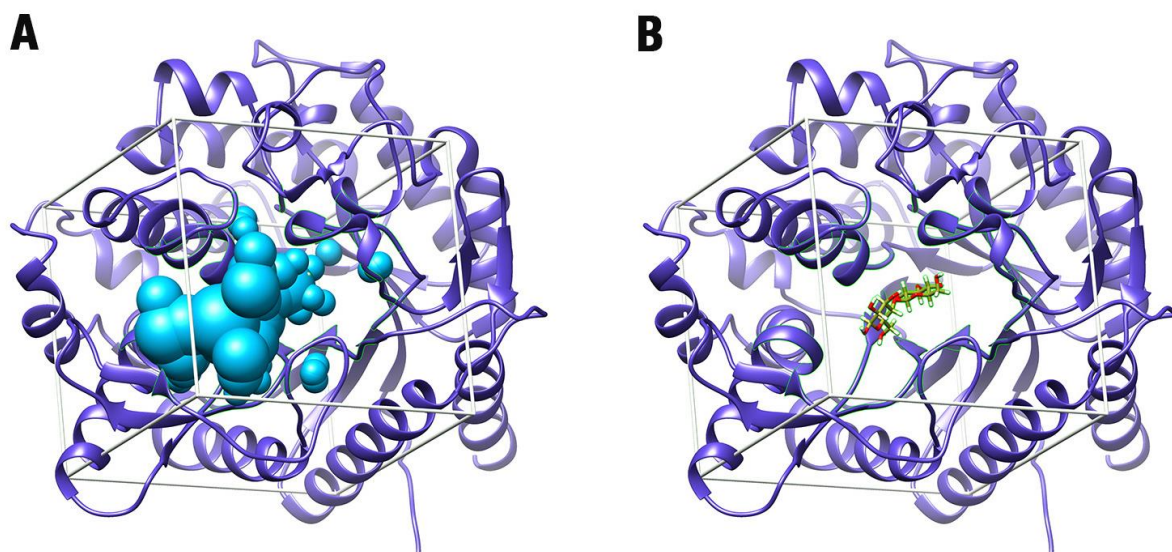
### 3.3 Ancoramento molecular de celobiose

Ancoramento molecular foi utilizado para verificar quais resíduos das  $\beta$ -glicosidases glicose-tolerantes pertencentes a família GH1 estavam diretamente relacionadas com a ligação do substrato. Além disso, foi utilizado para determinar o número de contatos realizados por cada resíduo com a celobiose e assim determinar a importância do resíduo.

O *software* DOCK v6.7 foi usado para ancoramento molecular (ALLEN et al., 2015). Receptor e ligante foram preparados antes do *docking*. A estrutura cristalográfica de uma  $\beta$ -glicosidase ligada a celobiose (PDB ID: 3VIK; JENG et al., 2012) foi escolhida como referência. As outras estruturas de glicose-tolerantes foram sobrepostas a ela usando o *software* Chimera (PETTERSEN et al., 2004). Para estruturas extraídas do PDB, íons e cofatores foram removidos. Hidrogênios foram adicionados e os resíduos de receptores padrão foram atribuídos com AMBER ff14sb *atomic partial charges* (MAIER et al., 2015), enquanto para celobiose foi utilizado AM1-BCC *charges* (JAKALIAN; JACK; BAYLY, 2002). Todos os sistemas receptor-celobiose foram submetidos a minimização (*500 steepest descent minimization steps*) usando *Chimera's Minimize Structure module* (PETTERSEN et al., 2004). Esse passo de minimização permite que os novos átomos de hidrogênios adicionados no receptor e no ligante se ajustem em posições razoáveis, enquanto também transportam a proteína a um ponto de energia potencial inferior ao anterior.

Receptor e celobiose foram separados para que preparações específicas do DOCK pudessem ser executadas. O programa sphgen do DOCK (DESJARLAIS et al., 1988) foi usado para gerar esferas no sítio de ligação (Figura 16A). As esferas que estavam dentro de um raio de 8 Å da posição da celobiose cristalográfica (PDB ID: 3VIK) foram mantidas para o acoplamento (Figura 16B). A caixa ao redor das esferas, além de uma margem de 5 Å em todas as direções, foi usada para restringir o espaço do receptor para cálculos da grade de energia. Celobiose foi tratada como flexível em todos os resultados de *docking*, de acordo com o protocolo de *docking* flexível padrão (ALLEN et al., 2015). Um máximo de 300 conformações

de ligantes foram obtidas e agrupadas (corte de RMSD de 2 Å) para todos os sistemas receptor-celbiose a fim de reduzir redundâncias e permitir uma melhor análise do *docking*.



**Figura 16. Caixa utilizada para determinação da região do *docking*.**

(A) Esferas do sítio de ligação; (B) Caixa foi definida com base nos resíduos a uma distância de 8 Å da posição cristalográfica da celbiose. Uma margem de 5 Å em todas as direções foi usada para restringir o espaço do receptor para cálculos da grade de energia. Fonte: próprio autor.

### 3.4 Caracterização do bolsão catalítico

Para determinação do bolsão catalítico utilizou-se o método de PIRES et al. (2013). Nesse método é proposto que, para uma determinada estrutura tridimensional de proteína em complexo com ligante, é possível extrair uma assinatura estrutural identificadora do bolsão (*pocket*) analisando-se os resíduos a uma distância de 4 a 8 Å do ligante, sendo 6 Å a distância ideal. Optou-se por estender essa distância para 6,5 Å, uma vez que na  $\beta$ -glicosidase de *H. insolens*, o triptofano, descrito como importante para o processo de resistência a inibição, estava a uma distância superior a 6 Å e inferior a 6,5 Å.

Uma vez que a determinação dos resíduos do bolsão requer uma proteína em complexo com um ligante, utilizou-se a  $\beta$ -glicosidase do cupim *Neotermes koshunensis* em complexo com celbiose (PDB ID: 3VIK) como referência. Foram selecionados 24 resíduos em um raio de 6,5 Å da celbiose. A seguir, foi realizado alinhamento estrutural entre as 21  $\beta$ -glicosidase da família GH1 e 3VIK para coletar os resíduos correspondentes do bolsão catalítico. Para essa etapa foi utilizado o *software* MultiProt (SHATSKY; NUSSINOV; WOLFSON, 2004). As coordenadas dos átomos pertencentes a resíduos do bolsão catalítico foram extraídas e salvas no formato PDB. Linhas correspondentes a águas foram removidas. Por fim, as sequências dos novos arquivos PDBs foram extraídas e salvas no formato FASTA. Por não corres-

ponderem à sequência numérica das estruturas originais, a partir daqui serão referidas como subsequências.

Para analisar a conservação das sequências de aminoácidos, realizou-se alinhamento de sequências utilizando Clustal Omega (GOUJON et al., 2010; SIEVERS et al., 2011).

### 3.5 Construção das assinaturas estruturais

A construção da assinatura de tolerância a glicose iniciou-se pela extração dos resíduos do bolsão catalítico (ver resultados). Os resíduos foram obtidos com base no alinhamento estrutural com a  $\beta$ -glicosidase de *N. koshunensis* (3VIK), cujos resíduos do bolsão catalítico a uma distância de 6,5 Å da celobiose foram extraídos conforme descrito no capítulo anterior. A construção da assinatura foi realizada usando aCSM-ALL, valor de distância mínimo de 0 Å, máximo de 10 Å e variação de 0,1 Å (o método usado para definição desses valores é descrito nos estudos de caso). Por fim, os valores de  $\Delta$ SSV e  $\Delta\Delta$ SSV foram calculados usando distância euclidiana simples entre vetores.

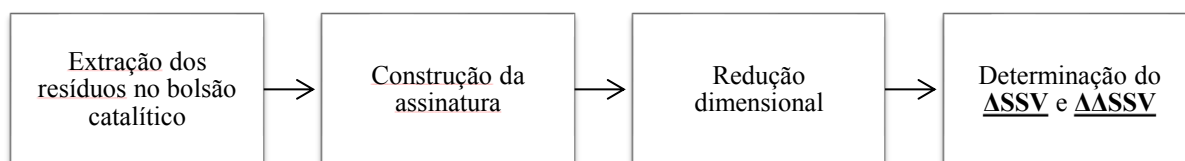


Figura 17. Metodologia para construção da assinatura de tolerância à glicose. Fonte: próprio autor.

### 3.6 Estudos de caso

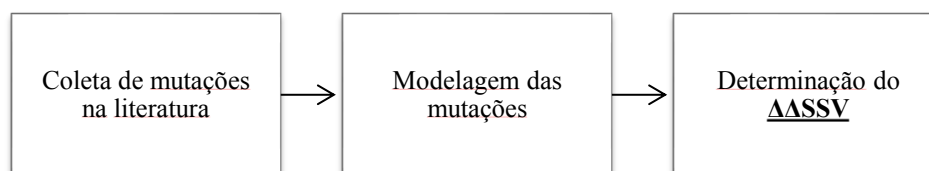
Foram propostos três estudos de caso para avaliar o método de assinatura de tolerância à glicose. O primeiro estudo de caso compara a eficácia do método para avaliar o impacto de mutações. O segundo estudo de caso propõe mutações para uma  $\beta$ -glicosidase não tolerante, e os resultados são comparados com dados experimentais. No terceiro estudo de caso, é feita uma comparação entre o método proposto e outros métodos.

#### 3.6.1 Avaliando mutações da literatura

O primeiro estudo de caso foi dividido em algumas etapas. Inicialmente, 27 mutações em  $\beta$ -glicosidases foram coletadas a partir de buscas na base de dados UniProt (Tabela 2). Cada mutação foi avaliada individualmente e classificada como benéfica (melhoram a atividade da enzima) ou não benéfica (pioram a atividade enzima) com base nos efeitos sugeridos nos artigos. A seguir as mutações foram modeladas utilizando o *script* de mutações pontuais do *soft-*

*tware* MODELLER. Os bolsões catalíticos foram coletados usando *in-house scripts* e a assinatura construída usando aCSM-ALL. Para a  $\beta$ -glicosidase extraída do metagenoma de *Turpan Depression* (mutação V174C/A404V/L441F; (CAO et al., 2015) foi escolhida a segunda assinatura mais próxima, uma vez que ela é glicose-tolerante e faz parte de BETAGDB.

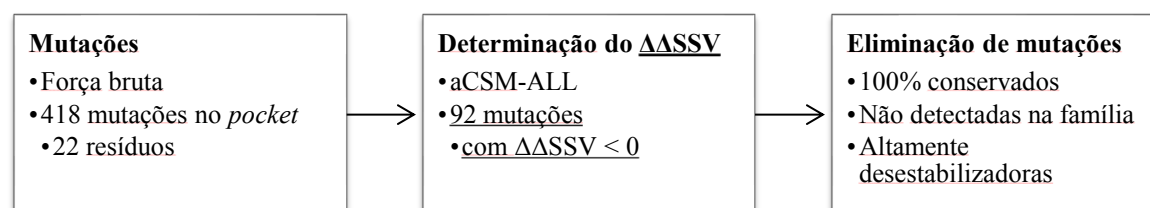
Foi definida a  $\beta$ -glicosidase glicose-tolerante de BETAGDB com assinatura mais próxima da  $\beta$ -glicosidase testada. Os melhores valores dos parâmetros foram definidos com base em testes recursivos variando as distâncias máxima de *cutoff* de 1 a 30 Å e a distância de variação de 0,1 a 0,9 Å. O resultado com maior acurácia apresentou distância máxima de 10 Å e distância de variação de 0,1 Å. Por fim, foi avaliado se a pontuação de  $\Delta\Delta$ SSV correspondia com o valor esperado para determinada mutação.



**Figura 18.** Passos para determinação do estudo de caso 1. Fonte: próprio autor.

### 3.6.2 Propondo mutações para uma $\beta$ -glicosidase não tolerante

No segundo estudo de caso, os 22 resíduos do bolsão catalítico foram mutados por força bruta, ou seja, todas as mutações pontuais possíveis foram testadas. As mutações foram modeladas *in silico* usando o *software* MODELLER. A seguir, as assinaturas das selvagens e das mutantes foram construídas usando aCSM-ALL (distância mínima de 0 Å, máxima de 10 Å e distância de variação de 0,1 Å). Foi determinada a melhor referência calculando o valor de  $\Delta$ SSV. Em seguida, foi avaliado o  $\Delta\Delta$ SSV para mutantes e selvagens. Eliminou-se mutações com  $\Delta\Delta$ SSV não benéfico, ou seja, aquelas que apresentam pontuações maiores que zero. A seguir foram avaliados: conservação nas tolerantes, conservação na família e estabilidade da estrutura (Figura 19).



**Figura 19.** Metodologia para proposta de mutações para uma  $\beta$ -glicosidase não-tolerante. Fonte: próprio autor.

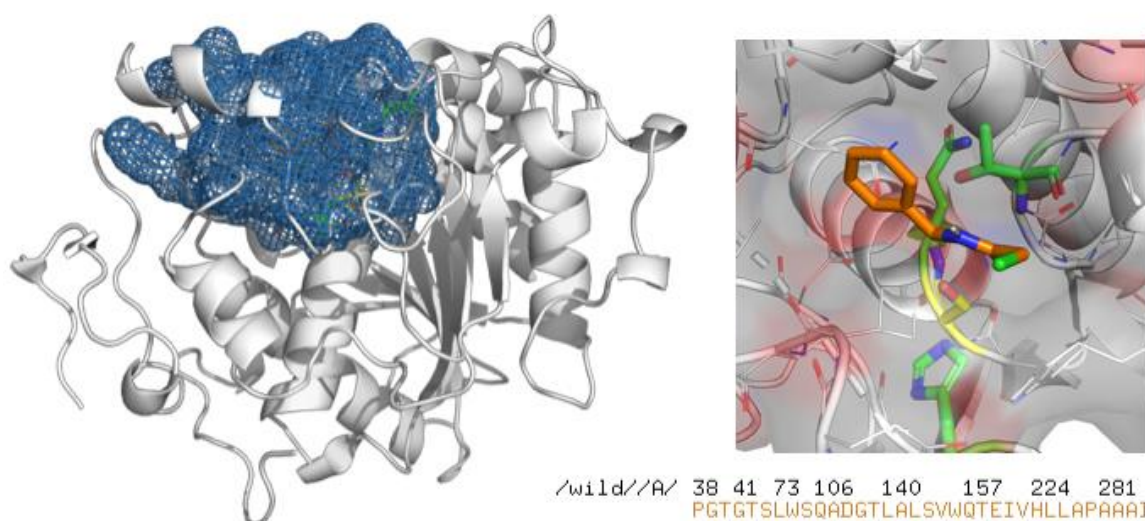


Foram construídos 418 mutantes por força bruta. Das 418 mutações, 92 apresentavam um valor de  $\Delta\Delta\text{SSV}$  menor que zero, e foram classificadas para próxima etapa. A seguir eliminou-se 12 mutações que ocorriam em resíduos 100% conservados nas  $\beta$ -glicosidases GH1 de BETAGDB. A ferramenta SIFT (<http://sift.jcvi.org/>) foi utilizada para analisar resíduos permitidos em determinada posição baseado em alinhamentos de toda a família GH1 (NG; HENIKOFF, 2003). Das 80 mutações, restaram 24. Também avaliou-se o impacto das mutações na estabilidade da proteína usando mCSM (PIRES; ASCHER; BLUNDELL, 2014). O *software* mCSM prediz a diferença de variação de energia livre ( $\Delta\Delta G$ ). Eliminou-se mutações altamente desestabilizadoras seguindo as definições descritas no trabalho de PIRES, ASCHER e BLUNDELL (2014), ou seja, mutações com  $\Delta\Delta G$  menor que -2 Kcal/mol.

### 3.6.3 Comparando SSV com outros métodos

O uso de distância euclidiana para classificação no método SSV foi comparado com SVM. Para essa avaliação SSV foi utilizado usando *in-house scripts* na linguagem Python e a SVM foi executada usando a ferramenta WEKA (FRANK; HALL; WITTEN, 2016). Para esse teste foram utilizados dados do primeiro estudo de caso (avaliando mutações da literatura). Quatro experimentos foram realizados, sendo um deles usando SSV e outros três usando SVM com entradas diferentes: (i) SSV; (ii) SVM usando assinaturas de proteínas selvagens; (iii) SVM usando assinaturas de proteínas mutantes; (iv) SVM usando assinaturas de proteínas selvagens e mutantes. As seguintes métricas foram avaliadas: precisão, acurácia, especificidade, sensibilidade e *F1-score* (SILVA; LEIJOTO; NOBRE, 2017).

A seguir foi realizado um teste comparativo com a ferramenta BioGPS. Nessa comparação, foi utilizada a estrutura cristalográfica de CaLB obtida no PDB (ID: 1TCA). Moléculas de água e ligantes foram removidos. Os mutantes M1-M8 foram modelados usando a ferramenta de mutagêneses do PyMOL (<http://pymol.org>). Detectou-se os resíduos do bolsão catalítico da estrutura selvagem e mutantes usando AutoDock Vina (TROTT; OLSON, 2010). A molécula *N-benzyl-2-chloroacetamide*, o mesmo ligante utilizado para determinar a atividade de amidase em CaLB (FERRARIO et al., 2014b), foi obtida na base de dados Zinc (IRWIN et al., 2012). Foram usados os parâmetros “exhaustiveness = 50” e uma caixa de 15x15x15 Å cujo centro foi definido baseado na posição do último átomo da serina catalítica (resíduo S105; átomo OG). Utilizou-se a primeira conformação obtida pelo *docking* e coletou-se todos os resíduos a uma distância de 6.5 Å de qualquer átomo do ligante (Figura 20). O ligante foi removido e a estrutura salva no formato PDB.



**Figura 20. Região do bolsão catalítico de CaLB (PDB ID: 1TCA).**

Resíduos a 6.5 Å do ligante docado em CaLB (região azul); em amarelo a serina catalítica (S105); em laranja o ligante *N-benzyl-2-chloroacetamide*; em verde o ácido/base catalítico (H224), o *oxyanion* 1º termo (Q106) e 2º termo (T40). Subsequência do bolsão catalítico: “PGTGTSLSWSQADGTLALSVMQTEIVHLLAPAAAI”. Imagem gerada com PyMOL (<http://pymol.org>). Fonte: próprio autor.

Os testes foram realizados com a ferramenta *web* do SSV usando o bolsão catalítico da proteína selvagem, os oito mutantes e a base de dados de *templates* (Tabela 3). Para esse passo, o arquivo PDB do mutante M3 foi comprimido no formato zip. O mutante M3 foi utilizado na base de dados de *templates* por apresentar o maior fator de melhoria descrito na literatura para CaLB.

**Tabela 3. Experimento das oito mutações de CaLB usando SSV para comparação com BioGPS.**

Mutante	Mutação	FM	Classificação esperada	Link
M1	G39A/W104F/L278A	6.3	Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVC69F173">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVC69F173</a>
M2	G39A/T103G/L278A	3.8	Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVCF1AB06">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVCF1AB06</a>
M3	G39A/T103G/W104F/L278A	11.2	Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV84D8A0C">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV84D8A0C</a>
M4	G39A	2.8	Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV273B233">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV273B233</a>
M5	G39A/L278A	3.3	Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV2911964">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV2911964</a>
M6	I189A	0.4	Não benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVE6997DF">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSVE6997DF</a>
M7	T40A	0.4	Não benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV6D921F4">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV6D921F4</a>
M8	T103G	1.1	Neutra/Benéfica	<a href="http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV703D402">http://bioinfo.dcc.ufmg.br/ssv/project/id/SSV703D402</a>

## 4. Resultados

Coletou-se 23 estruturas de  $\beta$ -glicosidases descritas na literatura como resistentes a inibição por glicose (Tabela 1). Dessas, 21 pertenciam à família GH1 e duas à família GH3. Apenas cinco delas apresentavam estrutura tridimensional resolvida experimentalmente. As outras foram modeladas por comparação (dados foram disponibilizados na base de dados BETAGDB, disponível em <http://bioinfo.dcc.ufmg.br/betagdb>). Uma análise comparativa das estruturas de  $\beta$ -glicosidases tolerantes a inibição coletadas demonstra que apenas  $\beta$ -glicosidases da mesma família (GH1 ou GH3) possuem similaridades em suas sequências (Tabela 4).

**Tabela 4. Matriz com o percentual de identidade entre sequências determinado pelo algoritmo Needleman-Wunsch para alinhamento global.**

	<i>B. subtilis</i>	<i>C. bescii</i>	<i>T. naphthophila</i>	<i>T. petrophila</i>	<i>T. brockii</i>	<i>T. aotearoense</i>	Metagenome KG	<i>T. thermosaccharolyticum</i>	Metagenome TD	Metagenome S	<i>E. antarcticum</i> B7	<i>N. koshunensis</i>	<i>N. crassa</i>	<i>T. reesei</i>	<i>N. takasagoensis</i>	<i>H. grisea</i>	Metagenome CSS	Metagenome HS	<i>A. saccharovorans</i>	<i>P. furiosus</i>	<i>F. islandicum</i>	<i>T. funiculosus</i>	<i>M. circinelloides</i>
<i>B. subtilis</i>	100	35	34	34	35	33	32	31	31	30	29	28	30	30	26	30	27	27	23	24	23	0	0
<i>C. bescii</i>	35	100	49	49	49	48	44	49	44	42	42	37	38	38	37	38	38	29	27	26	28	0	0
<i>T. naphthophila</i>	34	49	100	100	53	52	45	50	51	48	44	40	40	40	39	39	43	31	29	28	28	0	0
<i>T. petrophila</i>	34	49	100	100	53	52	45	50	51	48	44	40	40	40	39	39	43	31	29	28	28	0	0
<i>T. brockii</i>	35	49	53	53	100	79	47	65	46	48	55	38	42	40	35	42	43	30	27	30	25	0	0
<i>T. aotearoense</i>	32	48	52	52	79	100	45	62	44	45	54	38	40	39	37	39	44	29	27	27	24	0	0
Metagenome KG	32	43	45	45	47	45	100	47	40	38	42	35	37	37	33	37	40	27	26	27	26	0	0
<i>T. thermosaccharolyticum</i>	31	49	51	51	65	62	47	100	44	45	53	38	40	40	36	38	45	30	30	28	25	0	0
Metagenome TD	31	43	51	51	45	44	39	44	100	49	41	37	39	40	36	39	40	26	28	27	24	0	0
Metagenome S	30	40	45	45	46	42	37	43	48	100	41	34	37	38	32	37	40	27	28	26	24	0	0
<i>E. antarcticum</i> B7	29	42	43	43	55	54	42	53	41	42	100	33	35	36	32	35	43	27	25	28	26	0	0
<i>N. koshunensis</i>	27	35	38	38	36	36	33	36	36	33	31	100	36	37	69	36	33	28	26	24	24	0	0
<i>N. crassa</i>	29	38	40	40	42	39	38	39	40	39	35	38	100	74	41	90	39	28	27	27	26	0	0
<i>T. reesei</i>	30	37	39	39	41	39	38	40	41	40	36	39	74	100	41	73	38	28	29	25	25	0	0
<i>N. takasagoensis</i>	25	35	38	38	34	35	31	34	35	33	32	69	39	39	100	38	31	25	24	23	24	0	0
<i>H. grisea</i>	29	38	39	39	41	39	38	38	39	39	35	39	90	73	40	100	37	29	27	26	26	0	0
Metagenome CSS	27	38	43	43	44	44	40	45	40	41	43	34	39	38	31	37	100	27	28	28	27	0	0
Metagenome HS	26	28	30	30	29	28	26	29	27	27	27	29	27	27	25	28	26	100	52	49	40	0	0
<i>A. saccharovorans</i>	23	26	28	28	27	26	25	29	27	28	25	25	27	28	24	26	27	51	100	53	44	0	0
<i>P. furiosus</i>	23	25	28	28	30	26	26	28	26	27	27	25	27	24	24	26	27	49	53	100	46	0	0
<i>F. islandicum</i>	22	28	28	28	24	24	26	24	25	24	26	26	26	25	24	26	26	41	45	47	100	0	0
<i>T. funiculosus</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	38
<i>M. circinelloides</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	100

Sequências *subject* estão nas linhas e a *query* nas colunas. Devido ao tamanho da cobertura os valores presentes nas colunas e linhas podem apresentar diferenças. Legenda: *Metagenome Kusaya gravy* (KG), *Metagenome Tur-*

*pan Depression (TD), Metagenome soil (S), Metagenome China South Sea (CSS), Metagenome hydrothermal spring (HS)*. Fonte: próprio autor.

Apesar de algumas sequências apresentarem alta identidade, como no caso das  $\beta$ -glicosidases obtidas nas bactérias hipertermofílicas *T. naphthophila* e *T. petrophila*, muitas enzimas definidas na literatura como glicose-tolerantes possuíam baixa identidade entre si. Isso gerou um indício inicial de que pode haver diferentes conjuntos de resíduos responsáveis pela tolerância a inibição por glicose.

Baseado nas informações coletadas dos trabalhos de YANG et al., (2015b) e DE GIUSEPPE et al. (2014) inferiu-se que os resíduos presentes no canal que leva aos glutamatos catalíticos podem ser vitais para atividade de  $\beta$ -glicosidases glicose-tolerantes. Alguns desses aminoácidos, como uma leucina e um triptofano presentes no meio desse canal, têm sido apontados como responsáveis pela redução da inibição por glicose em  $\beta$ -glicosidases glicose-tolerantes (DE GIUSEPPE et al., 2014). Entretanto, outros aminoácidos também podem estar correlacionados com a característica da resistência a inibição. A fim de compreender melhor resíduos importantes para característica da glicose-tolerância, realizou-se um estudo da conservação de aminoácidos nas regiões próximas aos resíduos catalíticos nas 21  $\beta$ -glicosidases da família GH1 de BETAGDB (por terem a estrutura mais conservada, o que facilita a proposta de mutações).

#### **4.1 Bolsão catalítico**

Determinou-se todos os resíduos a uma distância de 6,5 Å do ligante utilizando-se como referência a estrutura da  $\beta$ -glicosidase de *N. kosshunensis* em complexo com celobiose (3VIK). A seguir, as posições desses resíduos foram extrapoladas para as 21 estruturas GH1 de BETAGDB (denominadas neste manuscrito como  $\beta$ -glicosidase glicose-tolerantes) por meio de alinhamento estrutural e os átomos foram extraídos para arquivos PDB. Além disso, as sequências foram extraídas a partir dos novos arquivos PDB dos bolsões catalíticos (para detalhes, consulte a seção “materiais e métodos”). A fim de estabelecer os mais importantes resíduos do bolsão catalítico, conservações foram analisadas por meio de alinhamento múltiplo das subsequências (Figura 21).

Bacillus_subtilis	HYNENHA--ATNSYYMM-EWSAY
Fervidobacterium_islandicum	HFNEHVLFSTANWYVR-EWEWF
Acidilobus_saccharovorans	HWNEVLFANSANYYTQ-EWEWF
Pyrococcus_furiosus	HWNEVQFAFAINYYSL-EWEWF
Metagenome_hydrothermal_spring	HWNEVIFATSSNYYSN-EWEWF
Metagenome_Kusaya_gravy	HWNEQCLHTCGENIYNY-EWEWF
Thermotoga_petrophila	HWNEWVVHNNGMNYYSH-EWEWF
Thermotoga_naphthophila	HWNEWVVHNNGMNYYSH-EWEWF
Metagenome_Turpan_Depression	HWNEWVGLNIERNWYTA-EWEWF
Metagenome_soil	HWNELCIMNLSANYYFA-EWEWF
Metagenome_China_South_Sea	HWNEFCLHNFTANFYTQ-EWEWF
Caldicellulosiruptor_bescii	HWNEYCLHNLTSNYYTA-EWEWF
Exiguobacterium_antarcticum_B7	HWNEWCLHNLAANFYSN-EWEWF
Thermoanaerobacterium_thermosaccharolyticum	HWNEWCLHNLNANYYTA-EWEWF
Thermoanaerobacterium_aotearoense	HWNEWCLHNLNANYYS-EWEWF
Thermoanaerobacter_brockii	HWNEWCLHNLNANYYTS-EWEWF
Neurospora_crassa	HWNEWCLFNGDKNHYTNWEWEWF
Humicola_grisea_var_thermoidea	HWNEWCLFNGDKNHYTNWEWEWF
Hypocrea_jecorina_Trichoderma_reesei	HWNEWCLFNGDRNHYTNWEWEWF
Nasutitermes_takasagoensis	HWNERDTMDNAFNFYTNWEWEWF
Neotermes_koshunensis	HWNELTDMNINYNFYTLWEWEWF
	*:** . * ** :

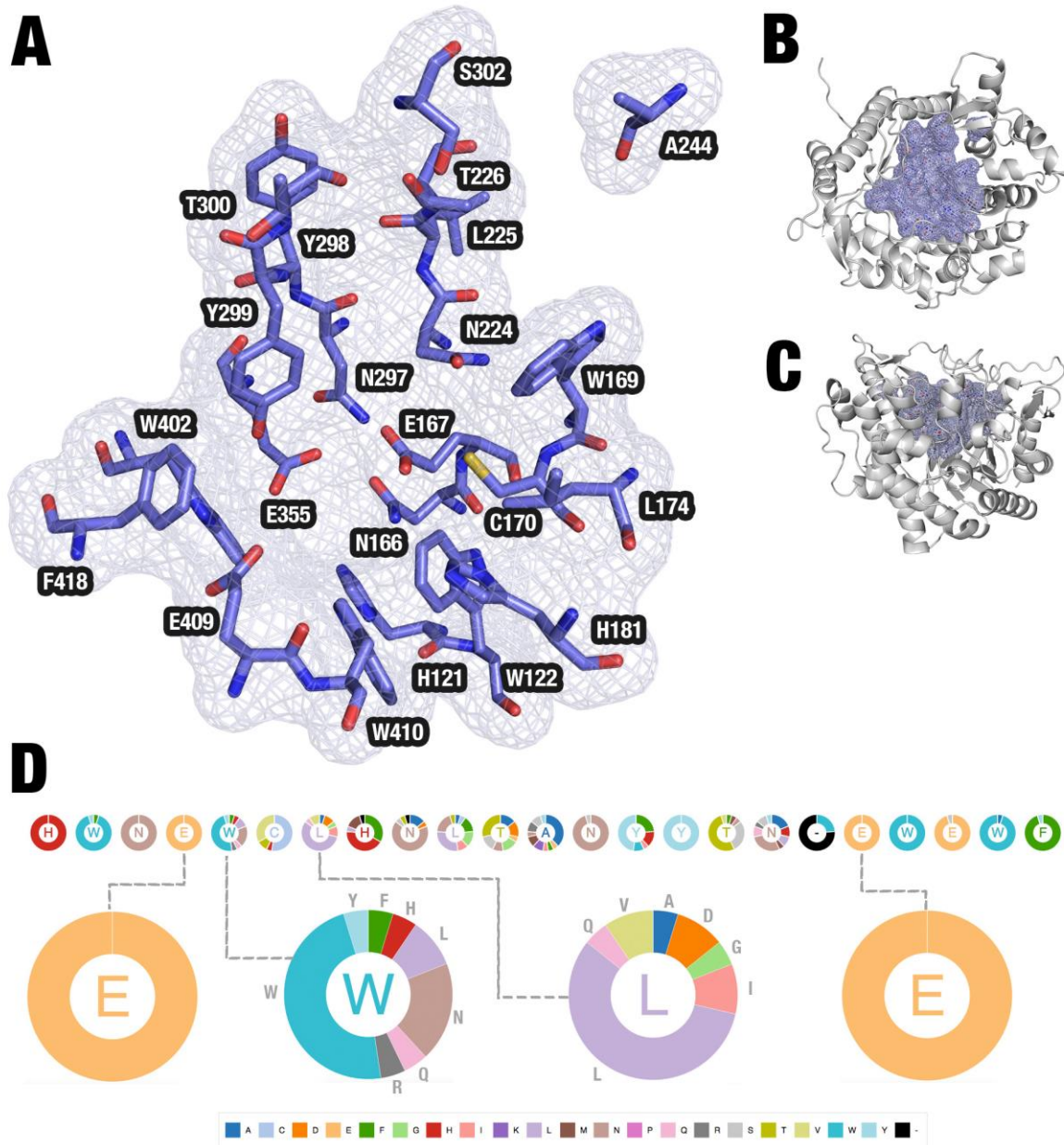
**Figura 21. Alinhamento entre os resíduos de bolsões catalíticos das 21  $\beta$ -glicosidases GH1 glicose-tolerantes.**

Aminoácidos correspondem na  $\beta$ -glicosidase de *T. brockii* aos resíduos: H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410, e F418.  $\beta$ -glicosidases da família GH3 não foram incluídas no alinhamento. Alinhamento gerado com Clustal Omega. Fonte: (MARIANO et al., 2017b).

Detectou-se uma subsequência consenso “HWNEWCLHNLNANYYTNEWEWF” formada por 22 resíduos mais conservados nos bolsões catalíticos das  $\beta$ -glicosidases GH1 de BETAGDB (Figura 22). Esses aminoácidos correspondem na  $\beta$ -glicosidase de *T. brockii* aos resíduos: H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410 e F418 (Figura 22a). O bolsão catalítico da  $\beta$ -glicosidase de *T. brockii* foi o que apresentou maior similaridade com a subsequência consenso (21 dos 22 resíduos). Por essa razão, a partir de agora, será utilizado como referência para identificação dos resíduos.

O alinhamento múltiplo dos resíduos do bolsão catalítico revelou ainda que apenas seis aminoácidos eram conservados em todas as estruturas: H121, N166, E167, Y299, E355, e W402. Os resíduos E167 e E355 são ácido/base catalítico e nucleófilo, respectivamente. Os outros resíduos conservados estão localizados ao redor dos resíduos catalíticos. Isso sugere que podem ser essenciais para o reconhecimento e acomodação do substrato no sítio ativo. Os resíduos W122, N297, E409, W410 e F418 são conservados em mais de 90% das sequências. Enquanto isso, os resíduos W169, C170, L174, N224, A244, Y299 e T300 são conservados na maior parte das estruturas. A baixa conservação dos resíduos L225, T226 e S302 (que na

subseqüência consenso é substituído por uma asparagina) sugere que eles tenham uma menor importância. Entretanto mais estudos são necessários para validar essa hipótese. Um triptofano foi detectado como um 23° resíduo no bolsão catalítico, entretanto aparece apenas em cinco estruturas (Figura 21; Figura 22D).

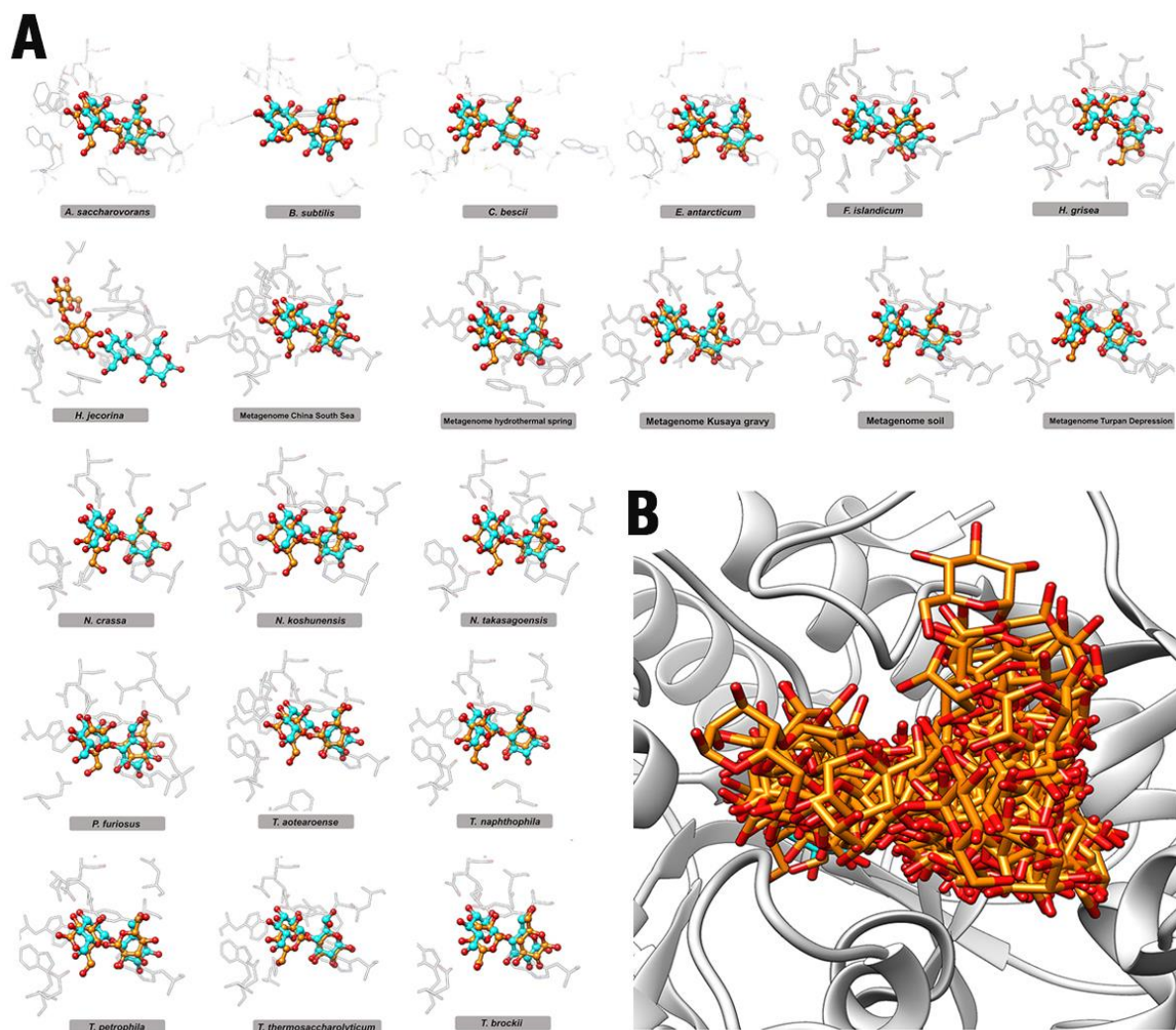


**Figura 22. Resíduos presentes no bolsão catalítico de  $\beta$ -glicosidases glicose-tolerantes.**

(A) Resíduos a 6,5 Å do ligante na  $\beta$ -glicosidase de *T. brockii*. (B) visão superior do bolsão catalítico no canto direito acima; (C) no canto direito abaixo, visão lateral. (D) Resíduos correspondentes a H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410 e F418. A conservação é demonstrada a partir de um gráfico de pizza. Foram destacados os resíduos catalíticos, E167 e E355, e os resíduos W169 e L173, descritos na literatura como responsáveis pela tolerância a glicose. Imagens geradas com PyMOL (<http://pymol.org>) e D3.js (<http://d3js.org>). Fonte: próprio autor (MARIANO et al., 2017b).

## 4.2 Análise da interação com celobiose

A fim de melhor compreender a importância dos resíduos do bolsão catalítico, realizou-se *docking* de celobiose em todas as estruturas GH1 de BETAGDB. Inicialmente foram realizadas análises dos resultados de *docking* com menor RMSD (Figura 23A). Entretanto, percebeu-se que uma análise estática dos resultados de *docking* poderia gerar a perda de informações relevantes, como por exemplo, resíduos que interagem com a celobiose no caminho até o sítio ativo. Por essa razão, optou-se por utilizar todas as conformações obtidas para calcular o número de contatos entre resíduos e substratos (Figura 23B).

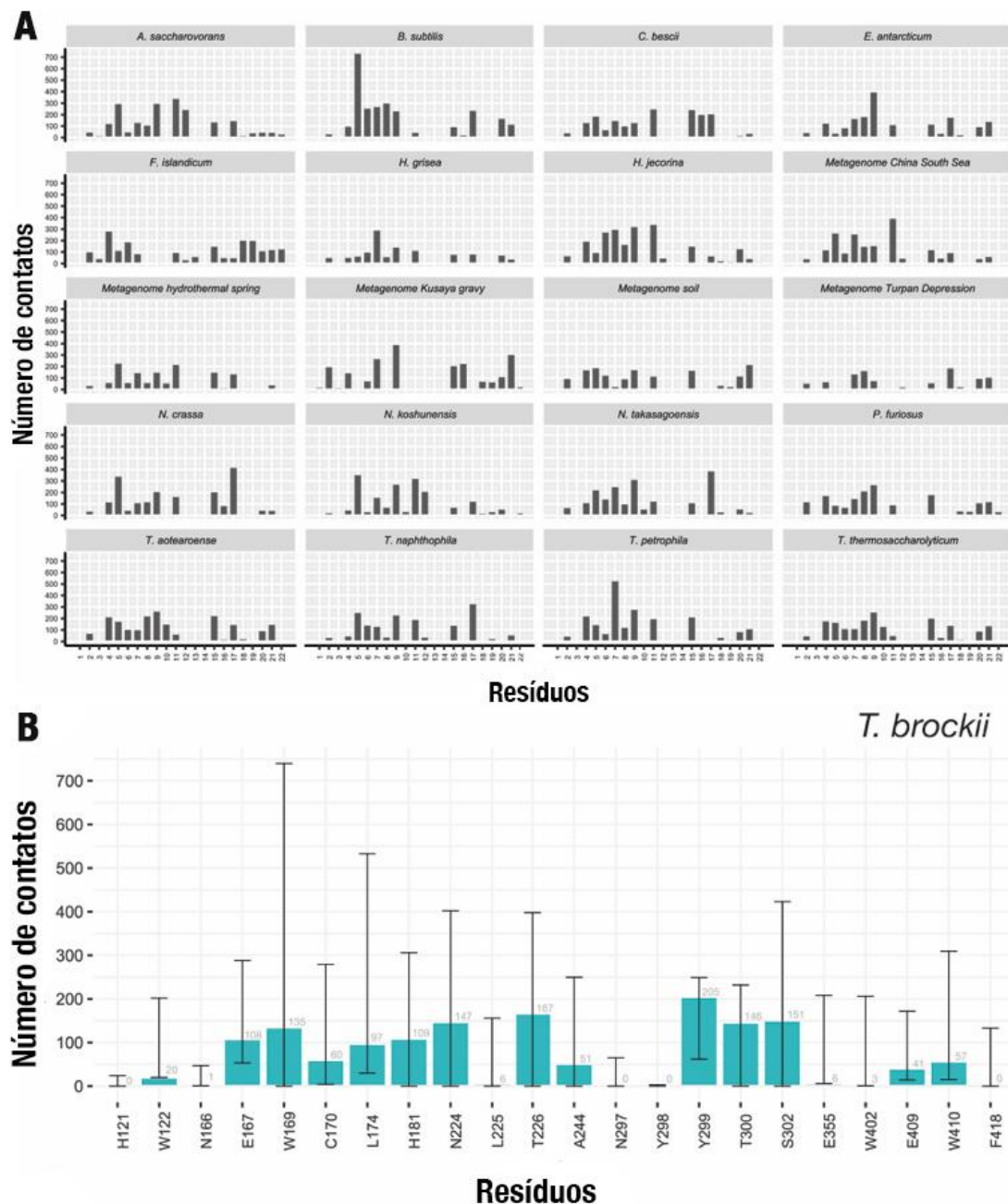


**Figura 23. Docking de celobiose no bolsão catalítico das  $\beta$ -glicosidases glicose-tolerantes.**

(A) Posição cristalográfica extraída de 3VIK é representada na cor azul. Na cor laranja é representada a posição de *docking* com menor RMSD. Em cinza, os resíduos que realizam contatos. (B) Representação de todas as conformações detectadas para celobiose no bolsão catalítico de uma  $\beta$ -glicosidase. Celobiose representada na cor laranja. Fonte: próprio autor (MARIANO et al., 2017b).

Foi calculado o número de contatos entre todos resíduos dos receptores e da celobiose a partir das conformações agrupadas obtidas no *docking*. Um contato pode ser definido como

um tipo de interação direta: polar, apolar, favorável ou não-favorável (incluindo conflitos). Dentre os tipos de contatos, pode-se citar: (i) ligações de hidrogênio, quando átomos acceptor e doador de diferentes aminoácidos encontram-se em uma distância e angulação que favoreçam uma interação; (ii) ligação iônica, ligação que ocorre entre aminoácidos polares positivos e negativos; (iii) empilhamento aromático, interação que ocorre entre aminoácidos que possuem anel aromático; e (iv) interação hidrofóbica, que ocorre entre compostos apolares (SILVA et al., 2019).



**Figura 24. Contatos da celbiose com resíduos de  $\beta$ -glicosidases.**

(A) Número de contatos com celbiose para cada resíduo das  $\beta$ -glicosidases de BETAGDB. Número de contatos são listados no eixo Y, enquanto a posição dos resíduos correspondentes é listada no eixo X. Uma tabela com resíduos correspondentes a cada posição pode ser encontrada nos apêndices. (B) Número de contatos com a ce-



lobiose para cada resíduo de *T. brockii*. A linha no centro indica os valores mínimo e máximo de contatos realizados por resíduos na mesma posição nas outras 20 glicose-tolerantes. Fonte: próprio autor (MARIANO et al., 2017b).

Os resíduos correspondentes a W122, N166, E167, C170, L174, Y299, E355, W402, E409 e W410 realizaram contatos com a celobiose em todas as estruturas analisadas (Figura 24). Além desses, os resíduos correspondentes a H181 e N224 efetuaram contatos na maioria das estruturas, não presentes apenas na  $\beta$ -glicosidase de *B. subtilis*. Alguns resíduos realizaram diversos contatos, exceto quando substituídos por outros aminoácidos, como por exemplo, W169 quando substituído por uma glutamina ou T226 quando substituído por uma glicina. O resíduo apresentado na 17<sup>a</sup> posição da subsequência consenso do bolsão catalítico, que na  $\beta$ -glicosidase de *T. brockii* é S302, apresenta muitos contatos quando é uma serina, mas quase nenhum quando é uma alanina. Quando o resíduo nessa posição é uma asparagina, o mais conservado aminoácido na subsequência consenso para essa posição, em algumas  $\beta$ -glicosidases ele apresenta muitos contatos, enquanto em outras nenhum, como por exemplo nas  $\beta$ -glicosidases de *N. crassa* e *T. reesei*, respectivamente.

Os resíduos correspondentes a L225, A244, N297, Y298, T300 e F418 fizeram poucos ou nenhum contato (Figura 24). Os resíduos na posição correspondente de N297 e F418 realizaram diversos contatos quando substituídos por uma serina e uma tirosina, respectivamente.

### 4.3 Assinaturas estruturais em $\beta$ -glicosidases glicose-tolerantes

A análise de conservação de resíduos e de contatos com celobiose demonstrou que os padrões internos dos bolsões catalíticos têm um nível de complexidade que transcende a capacidade de análise manual humana. Portanto, métodos computacionais que permitam uma abstração do problema e análise em larga escala são bem-vindos.

Assinaturas estruturais poderiam ser utilizadas para descrever características singulares e grupais de enzimas  $\beta$ -glicosidase. Desta forma, criou-se a hipótese de que  $\beta$ -glicosidases que apresentam a característica de alta tolerância a inibição por glicose possuem a assinatura do bolsão catalítico parecida com a de outras  $\beta$ -glicosidases que possuem a mesma característica. Como explicado anteriormente, uma assinatura pode ser vista como uma identificação única capaz de representar uma proteína ou parte de uma proteína. Logo, é possível inferir que proteínas com características similares possuem assinaturas parecidas. Assim, também é possível comparar  $\beta$ -glicosidases, ou até mesmo propor mutações que melhorem a atividade

catalítica, utilizando as assinaturas extraídas das  $\beta$ -glicosidases presentes na base de dados BETAGDB.

#### 4.4 *Structural Signature Variation (SSV)*

Para comparar diferenças e similaridades entre proteínas usando assinaturas estruturais, propõe-se aqui um método denominado *Structural Signature Variation (SSV)*, ou na tradução “Variação de Assinatura Estrutural”. A variação da assinatura estrutural pode ser estabelecida pela distância euclidiana entre vetores de assinaturas de duas proteínas, conforme descrito na equação da variação da assinatura de tolerância (2):

$$\Delta SSV = SSV_{P1} - SSV_{P2} \quad (2)$$

onde  $SSV_{P1}$  e  $SSV_{P2}$  representam vetores de assinatura de duas proteínas e  $\Delta SSV$  a distância euclidiana entre os dois vetores.

$SSV$  pode ainda ser utilizado para transferir características benéficas de uma proteína para outra e inferir o impacto de mutações por meio do cálculo da diferença de variação de assinatura estrutural ( $\Delta\Delta SSV$ ). Para isso são necessárias três estruturas: (i) proteína selvagem; (ii) proteína mutante; e (iii) proteína modelo (com características benéficas que deseja-se transmitir para o mutante).

Para o cálculo do  $\Delta\Delta SSV$  é necessário inicialmente determinar  $\Delta SSV$  entre a proteína selvagem e a proteína modelo, e a seguir, determinar o  $\Delta SSV$  entre a proteína mutante e a proteína modelo. Por fim, calcula-se a diferença entre os dois valores de  $\Delta SSV$ , conforme descrito na equação da diferença da variação das assinaturas estruturais entre uma proteína mutante (Mt) e uma selvagem (Wt) (3):

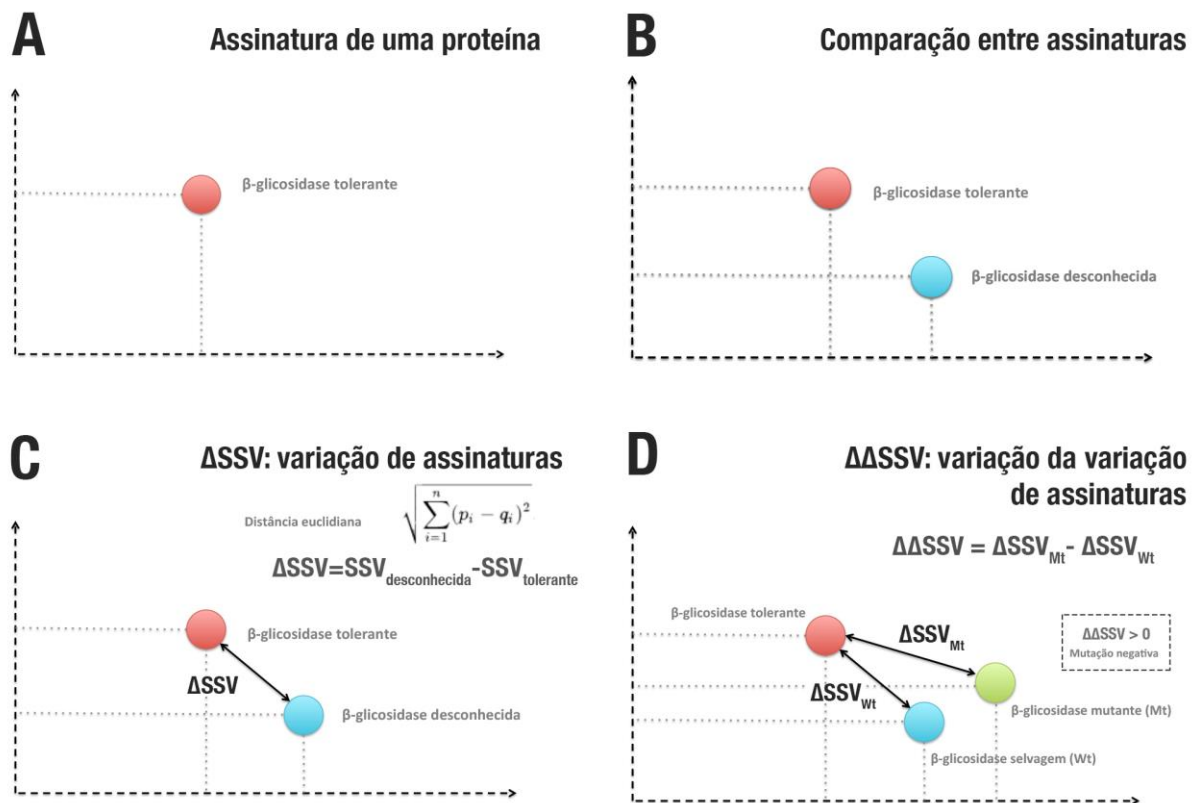
$$\Delta\Delta SSV = \Delta SSV_{Mt} - \Delta SSV_{Wt} \quad (3)$$

Espera-se que assinaturas mais próximas apresentem menores valores de distância euclidiana. Logo, se uma mutação inseriu padrões de distância cumulativa entre átomos (considerando também suas propriedades farmacofóricas) mais próximas da proteína modelo, então pode-se inferir que essa proteína mutante poderá apresentar características funcionais mais parecidas com as da estrutura modelo. Dessa forma, valores de  $\Delta\Delta SSV$  negativos indicam que o mutante possui uma assinatura mais próxima do modelo. Logo, aquela seria uma mutação

benéfica. Conseqüentemente, valores de  $\Delta\Delta\text{SSV}$  positivo indicam que o selvagem possui uma assinatura mais próxima do modelo. Logo, a mutação não seria benéfica. Para melhor ilustrar isso, será apresentado a implementação do SSV em enzimas  $\beta$ -glicosidase e, a seguir, uma comparação com outros métodos.

#### 4.4.1 Aplicação do SSV em $\beta$ -glicosidases

Como exemplo, pode-se representar a assinatura de uma  $\beta$ -glicosidase glicose-tolerante como um ponto em um plano bidimensional (Figura 25A). É possível comparar  $\beta$ -glicosidases mensurando a variação dos vetores de assinatura. Dada uma segunda  $\beta$ -glicosidase cuja tolerância não seja descrita na literatura (Figura 25B), denomina-se variação das assinaturas o valor obtido pela distância euclidiana entre dois vetores de assinaturas (Figura 25C). A variação de assinatura ( $\Delta\text{SSV}$ ) pode ser utilizada para definir qual  $\beta$ -glicosidase glicose-tolerante tem uma assinatura mais próxima de uma  $\beta$ -glicosidase de tolerância desconhecida.



**Figura 25. Representação da assinatura de tolerância à glicose.**

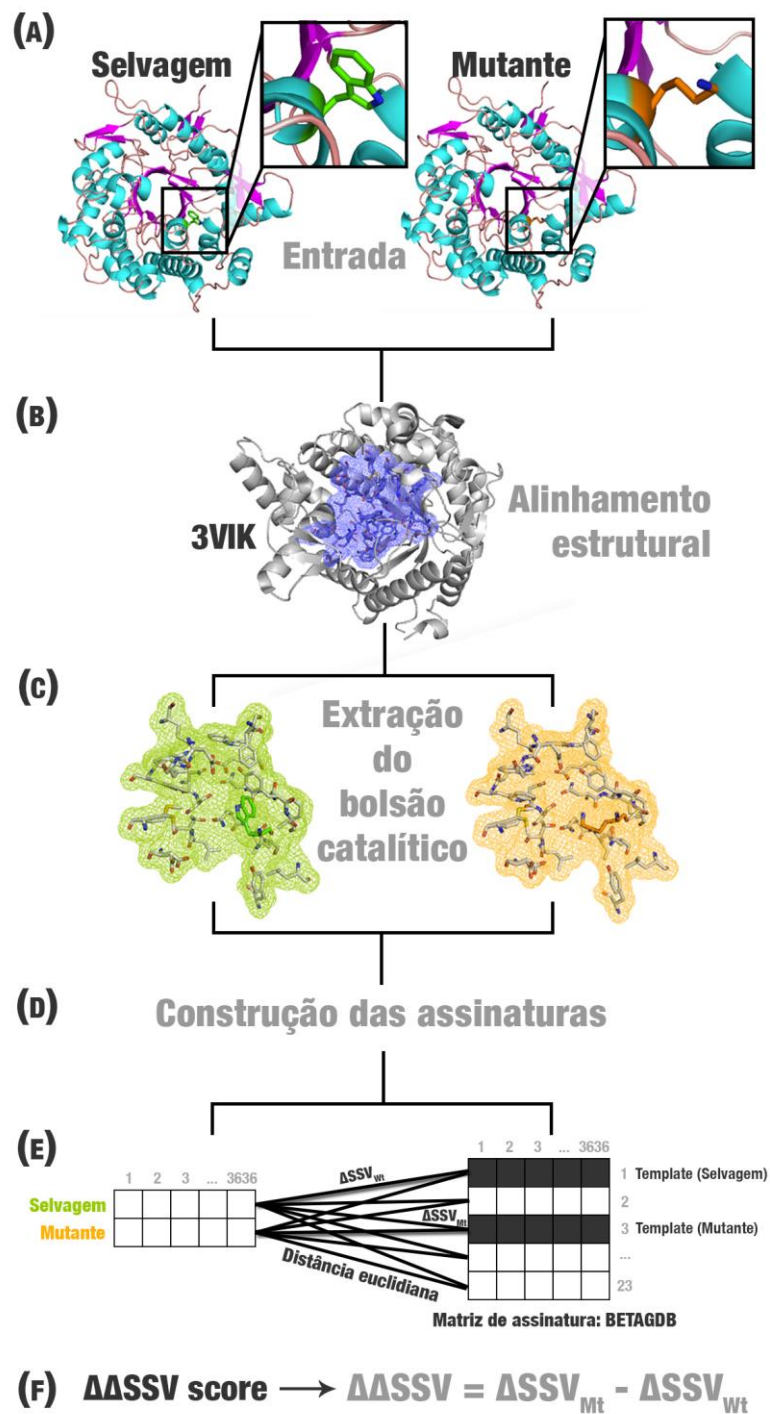
(A) Representação do vetor de assinatura de uma  $\beta$ -glicosidase tolerante e de uma segunda  $\beta$ -glicosidase (B), cuja tolerância é desconhecida, no espaço bidimensional. É importante ressaltar que um vetor de assinatura pode ter  $n$  dimensões, logo a figura retratada é apenas ilustrativa. A distância euclidiana entre os dois pontos (C) define a variação entre assinaturas e pode ser utilizada para definir se  $\beta$ -glicosidases possuem características similares. Além disso, a diferença da variação de assinatura (D) pode ser utilizada para medir se uma mutação aproxima ou distancia o ponto representativo de uma  $\beta$ -glicosidase de tolerância desconhecida do ponto de uma  $\beta$ -glicosidase tolerante - algoritmo de *k-nearest neighbors* (KNN) com  $k = 1$ . Fonte: próprio autor (MARIANO et al., 2019).

Além disso, é possível comparar as assinaturas de  $\beta$ -glicosidases selvagens e mutantes por meio da diferença da variação de assinatura ( $\Delta\Delta$ SSV). Por exemplo, dada uma terceira  $\beta$ -glicosidase, obtida por uma mutação experimental ou determinada computacionalmente, a variação da distância de assinatura dela para a tolerante pode ser comparada com a variação da distância da selvagem também para a tolerante (Figura 25D). A diferença da variação de tolerância pode ser utilizada para determinar se uma dada mutação é benéfica ou não para a enzima. Se o valor de  $\Delta\Delta$ SSV for maior que zero, como no exemplo da Figura 25D, a assinatura da mutada estará mais distante da assinatura da tolerante, logo a mutação tende a não ser benéfica. Caso o  $\Delta\Delta$ SSV seja menor do que zero, a mutação aproximaria as assinaturas, o que indicaria que ela tende a ser benéfica.

#### ***4.4.2 Estudo de caso 1: verificando mutações com SSV***

Para verificar a eficácia de SSV para avaliar a proposta de mutações em enzimas  $\beta$ -glicosidase realizou-se três estudos de caso. O primeiro estudo de caso foi dividido nas seguintes etapas:

- (i) 27 mutações em  $\beta$ -glicosidases foram coletadas a partir de buscas na base de dados UniProt (Tabela 2), classificadas em benéfica ou não benéfica, e modeladas por comparação quando estruturas tridimensionais obtidas por métodos experimentais não estavam disponíveis (Figura 26A);
- (ii) foi realizado alinhamento estrutural com a estrutura de 3VIK, cujos resíduos do bolsão catalítico foram anteriormente definidos (Figura 26B);
- (iii) bolsões catalíticos foram coletados (Figura 26C);
- (iv) assinaturas estruturais foram construídas para proteínas selvagens, mutantes e para todas as estruturas de BETAGDB (Figura 26D);
- (v) para cada proteína selvagem e seu mutante foi definida a  $\beta$ -glicosidase glicose-tolerante de BETAGDB com assinatura mais próxima (Figura 26E);
- (vi) por fim, foi avaliado se a pontuação de  $\Delta\Delta$ SSV correspondia com o valor esperado para determinada mutação (Figura 26F).



**Figura 26.** Etapas do estudo de caso para validação do SSV usando 27 mutações coletadas na literatura para  $\beta$ -glicosidases.

(A) Proteínas selvagem e mutante; (B) definição do bolsão catalítico com base no alinhamento estrutural com a proteína 3VIK (resíduos a uma distância de 6,5 Å do ligante); (C) Extração do bolsão catalítico; (D) Construção de assinaturas usando aCSM-ALL; (E) Definição da estrutura modelo; e (F) cálculo da diferença da variação de assinatura estrutural. Fonte: próprio autor (MARIANO et al., 2019).

O método  $\Delta\Delta\text{SSV}$  acertou 20 das 27 mutações analisadas (Tabela 5). A mutação H184F foi um dos erros apresentados pelo método. O resíduo H184 corresponde a H181 em *T. Brockii*. Analisando o alinhamento múltiplo das  $\beta$ -glicosidases glicose-tolerantes (Figura

21), percebe-se que na posição correspondente a H181, o resíduo mais conservado é uma histidina. Entretanto percebe-se ainda que uma fenilalanina é o segundo resíduo mais conservado nessa posição. É possível inferir que durante a determinação da  $\beta$ -glicosidase referência para cálculo do  $\Delta$ SSV, o método deve ter escolhido como referência uma  $\beta$ -glicosidase com uma histidina na mesma posição. É importante destacar também que a mutação H184F gera um aumento na constante de inibição para glicose, entretanto, como discutido nos capítulos anteriores, não há relato do impacto dessa mutação na tolerância a inibição por glicose quando o substrato era celobiose (LIU et al., 2011).

**Tabela 5. Predição do impacto de mutações a partir da pontuação da diferença da variação da assinatura de tolerância.**

<i>id</i>	<i>Mutação</i>	<i>Fonte</i>	$\Delta$ SSV esperado	$\Delta$ SSV score	<i>Predição</i>
1	H228T	(YANG et al., 2015b)	$\Delta$ SSV < 0	-231,09	✓
2	V174C/A404V/L441F	(CAO et al., 2015)	$\Delta$ SSV < 0	-254,94	✓
3	H184F	(LIU et al., 2011)	$\Delta$ SSV < 0	102,59	✗
4	P172L	(LEE et al., 2012)	$\Delta$ SSV < 0	-5,18	✓
5	P172L/F250A	(LEE et al., 2012)	$\Delta$ SSV < 0	-5,18	✓
6	L167W	(LEE et al., 2012)	$\Delta$ SSV < 0	-635,90	✓
7	L167W/P172L	(GUO; AMANO; NOZAKI, 2016)	$\Delta$ SSV < 0	-649,96	✓
8	L167W/P172L/P338F	(GUO; AMANO; NOZAKI, 2016)	$\Delta$ SSV < 0	-649,96	✓
9	V168Y	(BERRIN et al., 2003)	$\Delta$ SSV > 0	321,87	✓
10	F225S	(BERRIN et al., 2003)	$\Delta$ SSV > 0	-375,63	✗
11	Y308F	(BERRIN et al., 2003)	$\Delta$ SSV > 0	32,74	✓
12	Y308A	(BERRIN et al., 2003)	$\Delta$ SSV > 0	-111,33	✗
13	I207V	(SANSENYA; MANEESAN; CAIRNS, 2012)	$\Delta$ SSV < 0	-71,13	✓
14	N218H	(CHUENCHOR et al., 2008)	$\Delta$ SSV < 0	-243,87	✓
15	N273V	(CHUENCHOR et al., 2008)	$\Delta$ SSV > 0	-59,65	✗

16	F252I	(ZOUHAR et al., 2001)	$\Delta\Delta\text{SSV} > 0$	77,46	✓
17	F252W	(ZOUHAR et al., 2001)	$\Delta\Delta\text{SSV} > 0$	131,62	✓
18	F252Y	(ZOUHAR et al., 2001)	$\Delta\Delta\text{SSV} > 0$	34,27	✓
19	M284N	(SANSENYA; MANEESAN; CAIRNS, 2012)	$\Delta\Delta\text{SSV} > 0$	-152,13	✗
20	H276M	(SANSENYA et al., 2011)	$\Delta\Delta\text{SSV} > 0$	-515,50	✗
21	V173C	(TSUKADA et al., 2008)	$\Delta\Delta\text{SSV} > 0$	5,66	✓
22	M177L	(TSUKADA et al., 2008)	$\Delta\Delta\text{SSV} > 0$	22,93	✓
23	D229N	(TSUKADA et al., 2008)	$\Delta\Delta\text{SSV} > 0$	13,65	✓
24	H231D	(TSUKADA et al., 2008)	$\Delta\Delta\text{SSV} > 0$	-58,61	✗
25	E96K	(SANZ-APARICIO et al., 1998)	$\Delta\Delta\text{SSV} < 0$	-30,69	✓
26	N223G	(MATSUZAWA et al., 2016)	$\Delta\Delta\text{SSV} > 0$	50,47	✓
27	N223Q	(MATSUZAWA et al., 2016)	$\Delta\Delta\text{SSV} > 0$	266,27	✓

Das 18 mutações não benéficas, que deveriam apresentar um  $\Delta\Delta\text{SSV}$  maior que zero, 12 foram preditas corretamente. Das nove mutações benéficas, que deveriam apresentar um  $\Delta\Delta\text{SSV}$  menor que zero, oito foram preditas corretamente (Tabela 6).

**Tabela 6. Matriz de confusão dos resultados preditos comparados aos valores reais. Deve-se ressaltar que o resultado é considerado positivo se  $\Delta\Delta\text{SSV} < 0$  (benéfica); e negativo se  $\Delta\Delta\text{SSV} > 0$  (não benéfica).**

		Predito	
		Não benéfica	Benéfica
Real	Não benéfica	12	6
	Benéfica	1	8

Para as 27 mutações propostas, o método de  $\Delta\Delta\text{SSV}$  obteve valores de sensibilidade (*recall*/revocação), especificidade, valor predito positivo (precisão), valor predito negativo, acurácia, precisão e *F1-score* (média harmônica) de 0,57, 0,92, 0,89, 0,67, 0,74, 0,89 e 0,70, respectivamente (Tabela 7).

**Tabela 7. Valores de sensibilidade, especificidade, valor predito (+), valor predito (-) e acurácia para os testes de  $\Delta\Delta$ SSV para as 27 mutações em  $\beta$ -glicosidases.**

<i>Teste</i>	<i>Fórmula</i>	<i>Resultado</i>
<i>Sensibilidade</i>	$VP / (VP+FN)$	0,57
<i>Especificidade</i>	$VN / (FP+VN)$	0,92
<i>Valor predito (+)</i>	$VP / (VP+FP)$	0,89
<i>Valor predito (-)</i>	$VN / (FN+VN)$	0,67
<i>Acurácia</i>	$(VP+VN) / N$	0,74
<i>Precisão</i>	$VP / (VP+FP)$	0,89
<i>F1-score</i>	$2 \times (PxS) / (P+S)$	0,70

\* Valores: VP = 8; VN = 12; FP = 1; FN = 6; e N=27. VP: verdadeiros positivos; FN: falsos negativos; VN: verdadeiros negativos; FP: falsos positivos; P: precisão; S: sensibilidade; N: total de elementos.

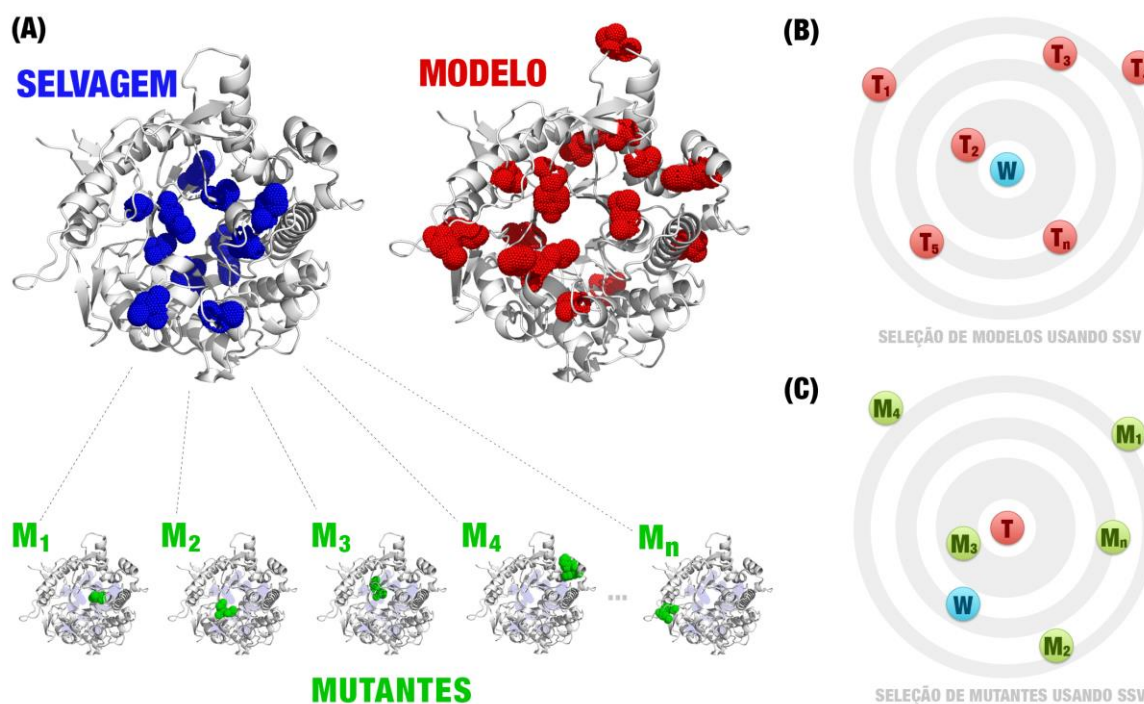
#### **4.4.3 Estudo de caso 2: propostas de mutações para a $\beta$ -glicosidase Bgl1B**

Propôs-se anteriormente que os resíduos-chave para o processo de conversão de celobiose a glicose com alta eficiência estão localizados na concavidade que leva ao sítio ativo. Ao todo 22 resíduos foram caracterizados no bolsão catalítico das  $\beta$ -glicosidases glicose-tolerantes. Entretanto, apesar de algumas similaridades, essas  $\beta$ -glicosidases apresentam diferentes resíduos em certas posições desse bolsão.

No segundo estudo de caso, propôs-se mutações no bolsão catalítico da  $\beta$ -glicosidase Bgl1B, extraída de um metagenoma marinho (FANG et al., 2009). Mutações experimentais nessa  $\beta$ -glicosidase foram comparada com uma outra  $\beta$ -glicosidase glicose-tolerante Bgl1A (FANG et al., 2010) no trabalho de YANG et al. (2015b). Esse trabalho servirá como base para validação das mutações aqui propostas. Bgl1B apresenta um IC50 de 50 mM, enquanto Bgl1A de 1000 mM.

Para o segundo estudo de caso, cada um dos resíduos do bolsão catalítico foi mutado para um dos outros 19 possíveis aminoácidos. Ao todo foram construídos 418 mutantes (Figura 27A). Para proteína selvagem e mutantes foram definidos modelos (*templates*) obtidos em BETAGDB usando SSV (Figura 27B). Por fim, calculou-se a diferença da variação de assinaturas estruturais. O mutante que apresentasse assinatura mais próxima ao modelo do que a proteína selvagem provavelmente apresentaria características mais similares ao modelo, como por exemplo poderia adquirir características de glicose-tolerantes (Figura 27C).





**Figura 27. Representação do estudo de caso 2.**

(A) 418 mutações pontuais foram propostas para Bgl1B; (B) o melhor modelo (*template*) foi selecionado na base BETAGDB usando SSV; (C) a mutação que apresenta uma assinatura mais próxima ao *template* (T) do que a estrutura selvagem (W) é definida como a provável mutação mais eficiente. Fonte: próprio autor (MARIANO et al., 2019).

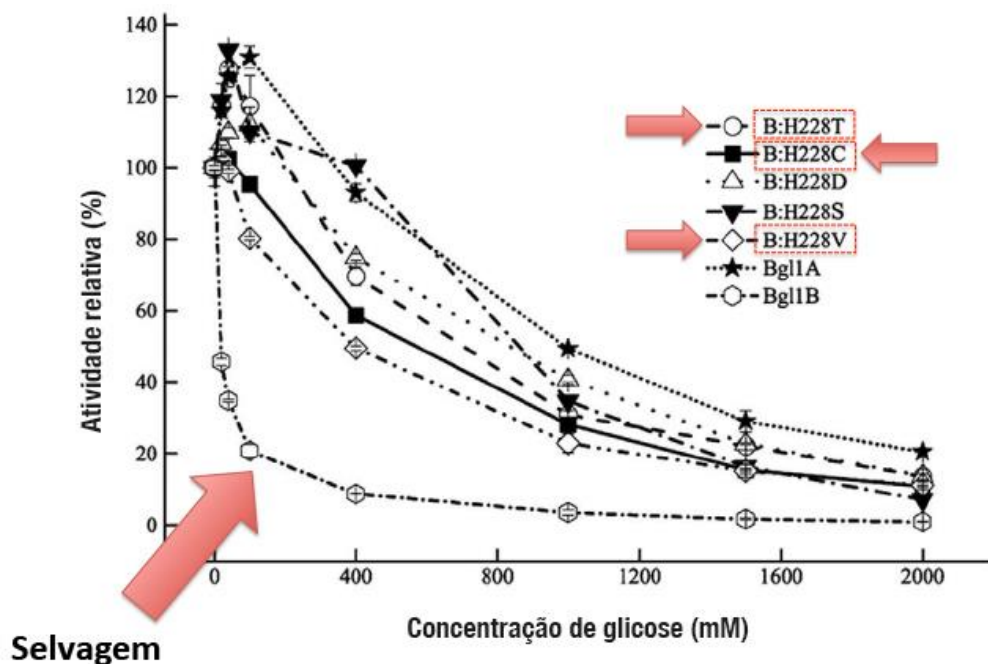
Após o cálculo do  $\Delta\Delta\text{SSV}$  e uma etapa de filtragem (ver materiais e métodos), um total de 18 mutações em cinco diferentes resíduos foram propostas para Bgl1B (Tabela 8).

**Tabela 8. Mutações benéficas propostas para Bgl1B.**

<i>F172</i>	<i>V227</i>	<i>H228</i>	<i>G246</i>	<i>T299</i>
F172C	V227M	H228A	G246S	T299S
F172I	V227R	<b>H228C</b>	G246T	
F172K		H228M		
F172P		H228N		
F172V		H228P		
		H228Q		
		<b>H228T</b>		
		<b>H228V</b>		

As mutações H228C, H228T e H228V possuem dados experimentais que comprovam que são eficazes para melhoria da atividade relativa em relação à concentração de glicose no meio.

O resíduo H228 foi o que mais apresentou possíveis mutações benéficas: oito no total. De fato, três dessas oito mutações apresentam dados experimentais disponíveis na literatura: H228T, H228C e H228V (Figura 28). Segundo os experimentos de YANG et al. (2015b), as três mutações mantiveram taxas de atividade relativa mais altas que Bgl1B mesmo em altas concentração de glicose. Isso demonstra que SSV é capaz de detectar mutações benéficas.

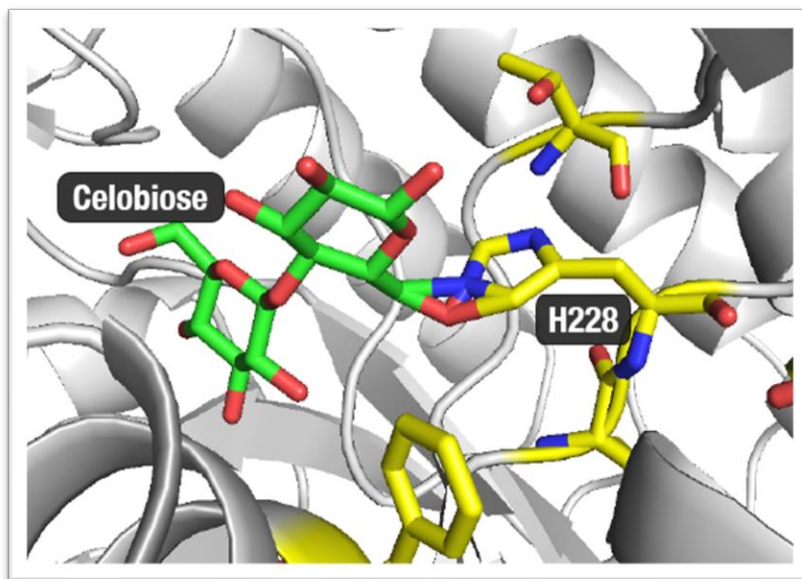


**Figura 28. Atividade relativa em relação à concentração de glicose para Bgl1B e mutantes.** Seta vermelha maior indica a atividade de Bgl1B. As outras três setas menores indicam as três mutações detectadas por SSV. Fonte: Adaptado de YANG et al. (2015b).

Para melhor compreender a importância das mutações propostas para os cinco resíduos de aminoácidos, realizou-se alinhamento estrutural entre Bgl1B e as  $\beta$ -glicosidases em complexo com celobiose (3VIK) e com glicose (3WH6) usando-se a ferramenta PyMOL. Após o alinhamento, as moléculas de glicose e celobiose foram transferidas separadamente para Bgl1B. Para esta etapa não foi utilizado nenhum campo de força ou realizada minimização para detectar o melhor posicionamento dos ligantes. Optou-se por realizar a simples transferência dos ligantes para detectar possíveis conflitos entre resíduos sugeridos para mutação e a posição em que os ligantes são encontrados naturalmente em outras  $\beta$ -glicosidases.

Dos cinco resíduos aos quais são propostas mutações dois chamaram a atenção: H228 e G246. Analisando a estrutura de Bgl1B com celobiose ligada, percebe-se que a posição do resíduo H228 entra em conflito com a posição da celobiose (Figura 29). Isso sugere que a presença de uma histidina força o substrato a adotar outra posição de ligação que provavelmente acarreta em uma menor afinidade. Histidina é um resíduo polar e carregado positivamente. As

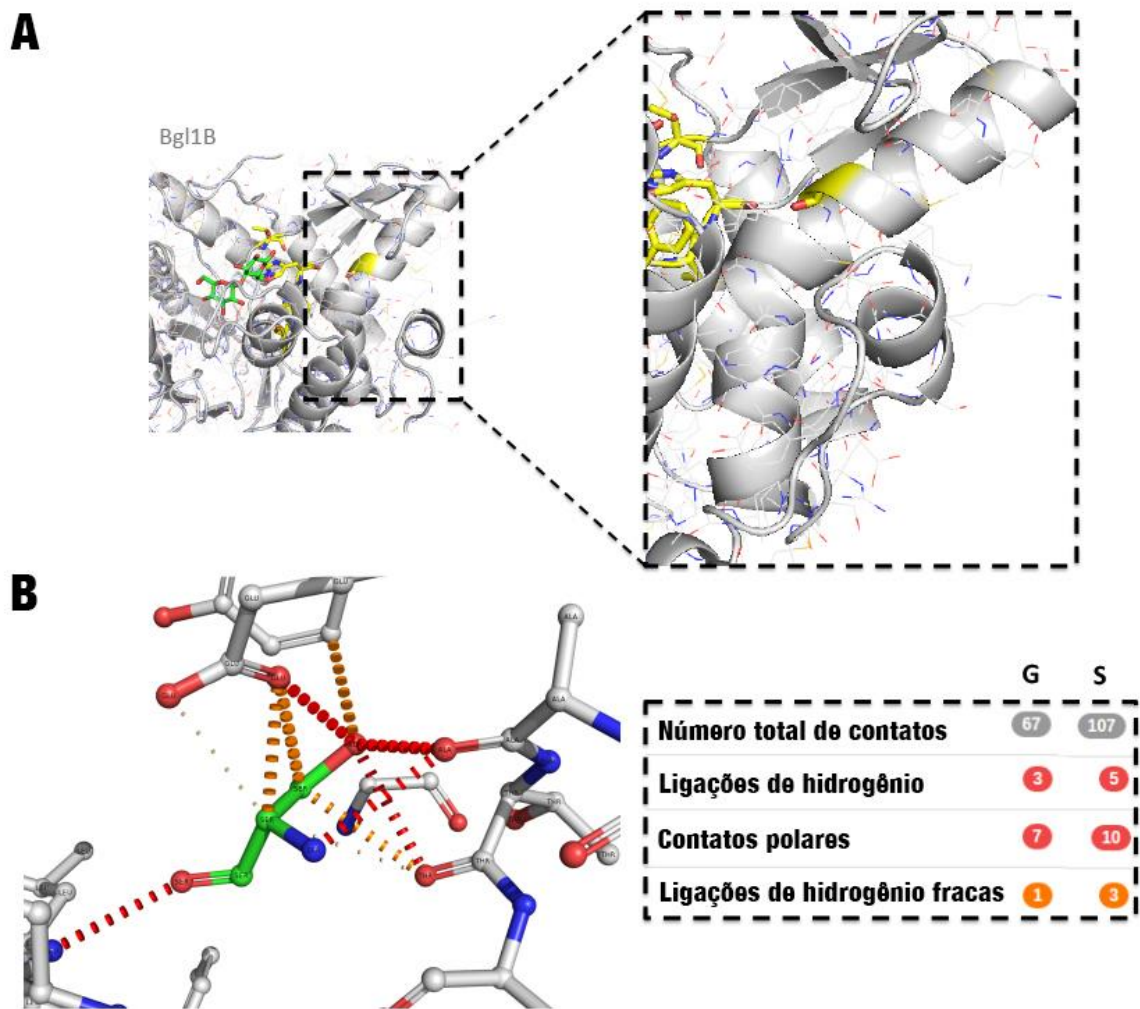
mutações propostas por SSV sugerem que resíduos de menor volume apolares ou polares neutros aumentariam a resistência à inibição por glicose de Bgl1B.



**Figura 29. Posicionamento da celobiose transferida e do resíduo H228 em Bgl1B.**

Em verde, a celobiose, em amarelo os resíduos aos quais sugere-se mutações. Imagem gerada com PyMOL. Fonte: próprio autor.

O resíduo G246 está localizado em uma alfa-hélice que pertence a estrutura do  $\beta$ -barril (Figura 30A). Glicose é um resíduo capaz de atingir ângulos diédricos *phi* e *psi* que nenhum outro resíduo consegue atingir, o que permite que ele tenha uma maior mobilidade. Além disso, é um resíduo de pequena cadeia lateral. A mutação desse resíduo por outro, como por exemplo por uma serina, poderia aumentar a quantidade de contatos realizados entre resíduos próximos na região (Figura 30B). O aumento na quantidade de contatos tende a aumentar a estabilidade das ligações da região. Por ser uma alfa-hélice localizada no bolsão catalítico, um aumento da estabilidade da região poderia reduzir a mobilidade e abertura da entrada ao sítio, o que poderia restringir o acesso do inibidor ao sítio e aumentar a tolerância a glicose.



**Figura 30.** Comparação entre o número de contatos realizados pelos aminoácidos glicina e serina para mutação G246S de Bgl1B.

(A) Localização da G246 em Bgl1B. (B) Quando o resíduo 246 é uma glicina, ele realiza 67 contatos na região próxima. Quando é substituído por uma serina, o número de contatos aumenta para 107. Imagem gerada com Arpeggio (JUBB et al., 2017). Fonte: próprio autor.

#### 4.4.4 Estudo de caso 3: comparação com outros métodos (SVM e BioGPS)

SSV foi comparado ao método *Support Vector Machine* (SVM) implementado pela ferramenta WEKA (FRANK; HALL; WITTEN, 2016). SVM é um algoritmo de aprendizado supervisionado usado para classificação e análise de regressão. A comparação com SVM foi realizada para verificar se a distância euclidiana, métrica simples utilizada por SSV, apresentaria bons resultados quando comparada com um conjunto de métodos mais robustos, como o encontrado no SVM.

Quatro experimentos foram realizados: (i) SSV; (ii) SVM usando assinaturas de proteínas selvagens; (iii) SVM usando assinaturas de proteínas mutantes; (iv) SVM usando assinaturas de proteínas selvagens e mutantes. Observou-se que SSV apresentou precisão e especificidade superior aos outros métodos, obtendo os valores de 0,89 e 0,92, respectivamente

(Tabela 9). SSV previu corretamente mais mutações benéficas do que SVM em todas as variações testadas.

**Tabela 9. Métricas usadas para comparar SSV com SVM.**

Métrica	SSV	SVM <sub>S</sub>	SVM <sub>M</sub>	SVM <sub>SM</sub>
Precisão	<b>0,89</b>	0,64	0,36	0,36
Acurácia	<b>0,74</b>	0,81	0,74	0,74
Especificidade	<b>0,92</b>	0,79	0,70	0,70
Sensibilidade	<b>0,57</b>	0,88	1,00	1,00
F1-score	<b>0,70</b>	0,74	0,53	0,53

SVM<sub>S</sub>: usando apenas as assinaturas de selvagens; SVM<sub>M</sub>: usando apenas as assinaturas de mutantes; SVM<sub>SM</sub>: usando apenas as assinaturas de selvagens e mutantes. Para execução do SVM foram utilizados os parâmetros padrão.

A seguir, comparou-se SSV com a ferramenta BioGPS (FERRARIO et al., 2014a). Oito mutações na lipase B de *Candida antarctica* (CaLB) foram usadas na comparação. CaLB é uma lipase pertencente à superfamília das serina-hidrolases, cuja inserção de atividade de amidase tem muitas aplicações industriais (BRAIUCA et al., 2006; FERRARIO et al., 2014a, 2014b). BioGPS classificou as oito mutações de CaLB baseado no fator de melhoria (FM), que é dado pela razão da atividade de amidase de CaLB mutante pela selvagem. Assim, a mutação é benéfica para  $FM > 1$ , não benéfica para  $FM < 1$  e neutra para  $FM = 0$ .

SSV previu corretamente cinco dos sete testes realizados (Tabela 10).

**Tabela 10.** Mutantes de CaLB avaliados por SSV.

Mutante	Mutação	FM	Classificação	$\Delta\Delta SSV$	Status
M1	G39A/W104F/L278A	6,3	Benéfica	-841	✓
M2	G39A/T103G/L278A	3,8	Benéfica	-121	✓
M3*	G39A/T103G/W104F/L278A	11,2	Benéfica	-841	✓
M4	G39A	2,8	Benéfica	150	
M5	G39A/L278A	3,3	Benéfica	-121	✓
M6	I189A	0,4	Não benéfica	-94	
M7	T40A	0,4	Não benéfica	40	✓
M8	T103G	1,1	Neutra/Benéfica	0	✓

\* M3 foi utilizado para treinamento e não deve ser levado em consideração na avaliação de predição.  
Fonte: (BRAIUCA et al., 2006; FERRARIO et al., 2014a, 2014b).

#### 4.5 Webtool SSV

Para facilitar o uso do método SSV, desenvolveu-se um website com interface amigável (Figura 31A). No website é possível executar o método SSV pela Internet (Figura 31B) ou baixar o código fonte do SSV para execução em larga escala (requer Python e Perl instalados). Para execução da ferramenta online deve-se usar como entrada: (i) nome do projeto (opcional), (ii) e-mail (opcional), (iii) descrição da mutação (opcional), (iv) arquivo PDB da proteína selvagem (obrigatório), (v) arquivo PDB da proteína mutante (obrigatório), e (vi) base de dados de *templates* no formato zip (obrigatório). A base de dados do primeiro estudo de caso (27 mutações em  $\beta$ -glicosidases) foi disponibilizada publicamente para testes (Figura 31C). Na página de resultados é possível visualizar a descrição do projeto criado (identificador, mutação, e-mail e data de criação), classificação da mutação com base nos valores de  $\Delta\Delta$ SSV, além de uma visualização interativa da estrutura selvagem e mutante (Figura 31D). SSV está disponível no endereço: <<http://bioinfo.dcc.ufmg.br/ssv>>.

**A** Página inicial do SSV webtool. O cabeçalho contém o logo 'SSV' e links para 'About', 'Documentation', 'Run online', 'Download' e 'Help'. O conteúdo principal apresenta o texto: 'Build new enzymes using SSV. SSV (Structural Signature Variation) is a method to propose mutations for enzymes used in industrial applications. SSV uses structural signatures to detect patterns, which can improve the activity of enzymes. A real and important application for SSV is the second-generation biofuel production.' Há um botão 'Run now!' destacado por uma seta vermelha.

**B** Tela de execução online ('Run online'). Possui campos para 'Project name', 'E-mail', 'Description of mutation', 'Wild protein PDB file', 'Mutant protein PDB file' e 'Templates database'. Um botão 'Submit' está visível.

**C** Exibição da base de dados de amostras ('Download sample database'). Uma tabela com as seguintes colunas: 'ID', 'Protein', 'File Name', 'Mutation', 'SSV score', 'SSV score'. A tabela contém 27 linhas de dados.

**D** Página de resultados para o projeto '1TCA (M1)'. O ID é 'SSVC69F173', a mutação é 'G39A/W104F/L278A', o e-mail é 'diego@dcc.ufmg.br' e a data é '12-13-2018'.  
 **$\Delta\Delta$ SSV score:** -841.98277892128  
**Classification:** Beneficial  
 **$\Delta$ SSVwt:** 841.98277892128  
 **$\Delta$ SSVmt:** 0

Abaixo dos dados, há duas visualizações de estruturas de proteínas: 'Wild' e 'Mutant', mostrando a estrutura em fitas e bastões dentro de um contorno de superfície branca.

**Figura 31. Interface da webtool SSV.**

(A) Página inicial. Para executar a ferramenta, inicialmente clique em um dos botões apontados pela seta. (B) Seção de execução da ferramenta online. (C) A base de dados utilizada no estudo de caso 1 (mutantes, selvagens e *templates*) pode ser baixada na página inicial. (D) Página de resultados. Fonte: próprio autor.

## 5. Discussão

Neste estudo, propôs-se estratégias usando bioinformática para melhor compreender as bases estruturais relacionadas aos mecanismos de tolerância a inibição por glicose em enzimas  $\beta$ -glicosidase utilizadas na produção de biocombustíveis de segunda geração e, assim, propor mutações que melhorassem a produção. Realizou-se uma busca na literatura através de uma revisão sistemática (ver apêndices) e criou-se uma base de dados com estruturas de  $\beta$ -glicosidases glicose-tolerantes (BETAGDB). Além disso, desenvolveu-se um método (SSV) para a proposta de mutações em enzimas baseadas na diferença de variação de assinaturas estruturais (ver apêndices).

### 5.1 Importância dos resíduos do bolsão catalítico na tolerância a glicose

As estruturas de  $\beta$ -glicosidases coletadas foram usadas para caracterizar resíduos conservados no bolsão catalítico. A conservação de resíduos é a primeira indicação de padrões conservados que podem ser utilizados para propor mutações em  $\beta$ -glicosidases não tolerantes. Utilizando alinhamento múltiplo de sequências, foi possível estabelecer uma subsequência consenso para o bolsão catalítico de  $\beta$ -glicosidases glicose-tolerantes e, assim, tentar estabelecer os aminoácidos necessários no bolsão catalítico para  $\beta$ -glicosidases com alta eficiência na produção de biocombustíveis.

A subsequência consenso “HWNEWCLHNLTYNYTNEWWEWF”, formada por 22 resíduos mais conservados das glicose-tolerantes, apesar de não aparecer por completo em nenhuma estrutura, poderia indicar os principais resíduos de aminoácidos para o efeito da glicose-tolerância (Figura 22).

Onze resíduos aparecem em mais de 90% das sequências e devem ter um papel fundamental para atividade enzimática (Tabela 11). Os outros resíduos podem estar correlacionados com os variantes valores de tolerância a glicose e sua importância pode ser verificada através dos estudos de interação com o ligante. Os resíduos correspondentes a W122, N166, E167, C170, L174, Y299, W402, E409 e W410 realizaram diversos contatos com o ligante e aparentam ser essenciais para o reconhecimento e acomodação do substrato (Tabela 11).

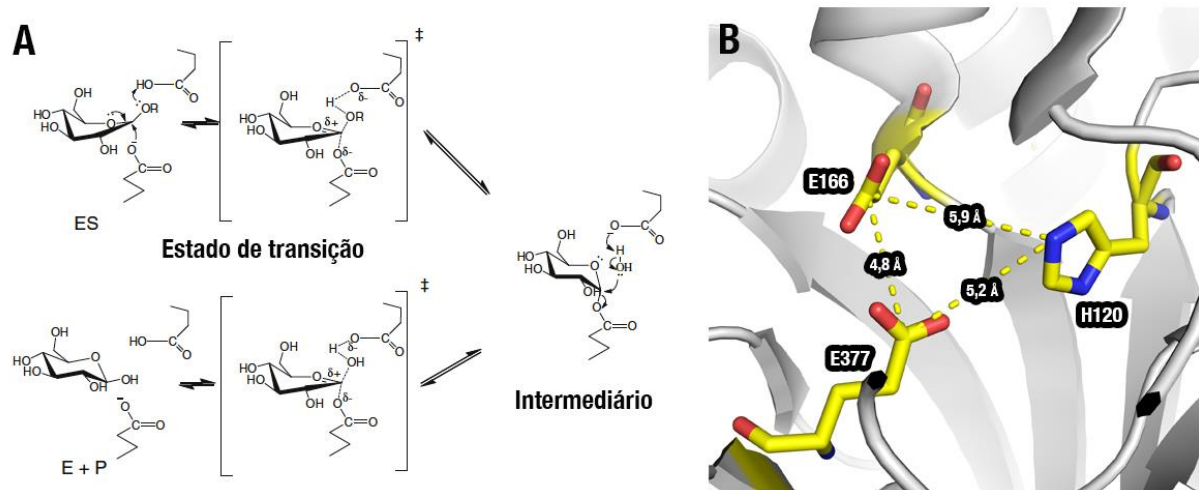
**Tabela 11. Resíduos conservados e contatos com celobiose para os 22 resíduos correspondentes a subsequência consenso do bolsão catalítico das 21 glicose-tolerantes GH1.**

#	Resíduos ( <i>T. brockii</i> )	Conservação dos resíduos		Contatos com celobiose	
		> 90% conservado	100% conservado	Resíduo correspondente faz contato em todas as estruturas	Não realiza contatos quando substituído (referência <i>T. brockii</i> )
1	H121	x	x		
2	W122	x		x	
3	N166	x	x	x	
4	E167	x	x	x	
5	W169				x
6	C170			x	
7	L174			x	
8	H181				x
9	N224				x
10	L225				
11	T226				x
12	A244				
13	N297	x			
14	Y298				
15	Y299	x	x	x	
16	T300				
17	S302				x
18	E355	x	x	x	
19	W402	x	x	x	
20	E409	x		x	
21	W410	x		x	
22	F418	x			

### 5.1.1 Sítio ativo

O resíduo E167, ácido/base catalítico, também realiza diversos contatos com o ligante. Entretanto, E355, nucleófilo catalítico, realiza poucos contatos. H121 é outro resíduo altamente conservado, mas que também realiza poucos contatos com o ligante. O resíduo H121 provavelmente está envolvido na ligação do substrato a enzima ou na estabilização do estado de transição, participando da ação catalítica junto a E167 e E355 (BARRETT et al., 1995; SANZ-APARICIO et al., 1998). Além disso, a distância da histidina para os glutamatos pode sugerir que ela contribui com um próton no mecanismo de ação catalítico das  $\beta$ -glicosidasas GH1 (Figura 32).





**Figura 32. Mecanismo de ação catalítica das  $\beta$ -glicosidases GH1.**

(A) Mecanismo de retenção do carbono anomérico. Um dos grupos glicosídicos se liga, formando o estado de transição. Inicialmente ocorre a glicosilação: o ácido catalítico doa um próton para o grupo que irá sair, enquanto o nucleófilo catalítico ataca no lado oposto. A seguir ocorre um passo de desglicosilação, em que a base catalítica (mesmo carboxilato do ácido catalítico) extrai um próton de uma molécula de água e ataca o carbono anomérico (CAIRNS; ESEN, 2010). (B) Distância entre os resíduos H120, E166 e E377 da  $\beta$ -glicosidase de *H. Grisea* (equivalentes a H121, E167 e E355 de *T. Brockii*). A proximidade pode sugerir que a histidina contribuiu com o próton para um dos glutamatos. Foi estabelecido que mutações em resíduos em posições análogas podem levar a total inibição da enzima (WITHERS et al., 1992). Fonte: Adaptado de KETUDAT CAIRNS; ESEN (2010).

### 5.1.2 Canal do substrato

Como discutido anteriormente,  $\beta$ -glicosidases glicose-tolerantes apresentam um profundo e estreito canal que limita o acesso da glicose ao sítio ativo. A forma do canal e os resíduos que o compõem podem estar envolvidos na redução do acesso da glicose, e consequentemente, na característica da glicose-tolerância (DE GIUSEPPE et al., 2014). Na  $\beta$ -glicosidase de *H. grisea*, os resíduos W168 e L173 (W169 e L174 na  $\beta$ -glicosidase de *T. Brockii*) foram descritos como importantes para liberação da glicose. Esses resíduos aparecem na maioria das glicose-tolerantes. Tal conservação reforça a importância desses aminoácidos para a regulação da entrada e saída de produto e substrato, demonstrada na literatura (DE GIUSEPPE et al., 2014; GUO; AMANO; NOZAKI, 2016). Além disso, o resíduo C170 aparece em muitas sequências de glicose-tolerantes, o que sugere que a inserção de uma cisteína nesta posição pode ser benéfica. Em um recente estudo, mutações aleatórias em uma  $\beta$ -glicosidase detectaram três mudanças benéficas: V174C, A404V e L441F, onde V174 corresponde a C170 em *T. Brockii* (Figura 7a). Essas mutações permitiram a construção de uma  $\beta$ -glicosidase mutante glicose-tolerante e termoestável, que aumentou a conversão do bagaço da cana entre 14 e 35% (CAO et al., 2015).

Os resultados demonstrados ainda sugerem que o aparecimento de uma cisteína (C170) nessa posição parece estar correlacionado com o aparecimento de um triptofano

(W169) e uma leucina (L174) em uma região próxima. Além disso, de acordo com os resultados do *docking*, quando os três resíduos, triptofano, cisteína e leucina, aparecem juntos, eles realizam menos contatos do que quando na mesma posição aparecem outros resíduos. Isso sugere que a presença de resíduos que realizam poucos contatos com a celobiose no meio do canal que leva ao sítio ativo pode ser benéfico para tolerância a inibição por glicose.

As informações coletadas na revisão sistemática indicaram que mutações ocorridas nas regiões próximas ao canal que leva ao sítio ativo (bolsão catalítico) provocam maior impacto na tolerância a inibição por glicose em  $\beta$ -glicosidases GH1. As análises de conservação de sequência indicaram resíduos essenciais para enzimas  $\beta$ -glicosidase, entretanto nem todas as  $\beta$ -glicosidases glicose-tolerantes coletadas apresentavam resíduos considerados importantes para tolerância pela literatura. A exemplo pode-se destacar o triptofano (W169) e a leucina (L174) localizados no centro do bolsão catalítico. Apesar de serem considerados essenciais (DE GIUSEPPE et al., 2014), apenas sete das 21 glicose-tolerantes apresentavam os dois resíduos juntos. Em algumas das glicose-tolerantes, a leucina (L174) é substituída por uma valina, um aspartato ou uma isoleucina. Em outras, o triptofano (W169) é substituído por uma asparagina ou uma leucina. E em outras, ambos são substituídos por uma asparagina e uma glutamina, respectivamente.

### **5.1.3 Mecanismo de glicose-tolerância**

A importância de resíduos tanto no meio desse canal quanto na entrada para glicose-tolerância já havia sido revelada através de mutações na  $\beta$ -glicosidases extraída do metagenoma de *China South Sea* (Figura 7b). Nesse estudo, a tolerância a glicose de uma  $\beta$ -glicosidase não tolerante (Bgl1B) foi aumentada através das mutações H228T e N301Q/V302F (YANG et al., 2015b). Esse resultado pode sugerir que sítios secundários a qual a celobiose se liga no caminho até o sítio ativo são de grande importância para o processo de liberação da glicose e a inevitável inibição. O resíduo correspondente a H228 na  $\beta$ -glicosidase de *T. Brockii* é T226. Analisando outras  $\beta$ -glicosidases glicose-tolerantes, não há ocorrência de uma histidina na mesma posição. Logo, o alinhamento múltiplo do bolsão catalítico das glicose-tolerantes confirma a importância desse resíduo, sugerindo que a mutação H228T é benéfica. Além de que a análise visual também demonstrou que a presença de uma histidina naquela posição interfere na alocação da celobiose no sítio ativo (Figura 29).

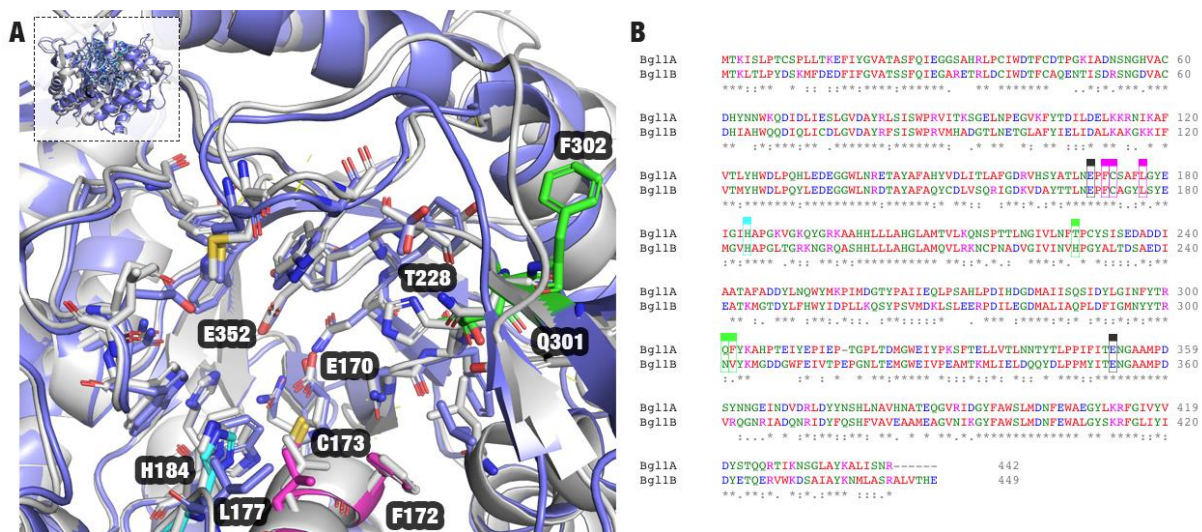
O estudo de YANG et al. (2015b) contrapõe a afirmação de que a largura e profundidade do canal sejam os únicos responsáveis pela glicose-tolerância conforme relatado por DE GIUSEPPE et al. (2014). Eles afirmam que Bgl1A (tolerante) e Bgl1B (não tolerante) possu-

em canais de largura e profundidade iguais, o que indica que existem outros mecanismos que levam à resistência a glicose. Eles propõem um mecanismo de atividade de  $\beta$ -glicosidases GH1: (i) glicose e o celobiose podem se ligar a diversos sítios ao longo do canal; (ii) a inibição ocorre quando a glicose prefere se ligar ao sítio ativo. Quando ela prefere se ligar a outros sítios ocorre a chamada tolerância; e (iii) ainda, a estimulação por glicose depende da geometria dos outros sítios, onde a ligação da glicose aumenta a atividade de clivagem de substrato por meio de transglicosilação ou de outro efeito alostérico. A transglicosilação pode estimular a clivagem do substrato se a glicose estiver ligada próximo ao substrato. Entretanto, eles afirmam que o mecanismo de transglicosilação ainda não está totalmente elucidado.

Assim, a glicose-tolerância pode resultar da ligação da glicose com outros sítios, provavelmente na entrada (apontam os resíduos 301 e 302) ou no meio (apontam os resíduos 184 e 228) do canal que leva ao sítio ativo. O resíduo na posição 184 já havia sido sugerido como importante para tolerância a glicose através da mutação H184F (LIU et al., 2011). O resíduo H184 (correspondente a H181 em *T. brockii*) aparece na maioria das  $\beta$ -glicosidases glicose-tolerantes reportadas. A mutação H184F mostrou um aumento na tolerância a inibição por glicose quando o substrato era pNPG. Fenilalanina é o segundo mais comum resíduo nesta posição (Figura 22b; Figura 21). Além disso, os resultados de *docking* demonstram que uma histidina realiza mais contatos com a celobiose do que uma fenilalanina na mesma posição. Entretanto, não há relato do impacto dessa mutação na tolerância a inibição por glicose quando o substrato era celobiose. Assim, mais experimentos com celobiose como substrato são necessários para fazer inferências. Ainda, FLORINDO et al. (2018) relatam que a mudança de um único aminoácido na posição correspondente a H181 nas  $\beta$ -glicosidases de *Trichoderma harzianum* (ThBgl1 e ThBgl2) de um aminoácido hidrofóbico (F180) para um aminoácido polar (N186) permitiu a formação de um túnel de água que leva ao sítio ativo. Eles ainda reforçam que o papel das moléculas de água na vizinhança do substrato é uma característica determinante para conduzir a reação para a hidrólise ou transglicosilação. Logo, a mutação de um aminoácido polar por um hidrofóbico favoreceria a transglicosilação (processo que pode unir uma celobiose a uma glicose formando uma nova molécula).

É importante ressaltar que Bgl1A e Bgl1B possuem identidade de 55% (resíduos idênticos 245/443; cobertura de 98%). Isso quer dizer que elas possuem 198 resíduos distintos (Figura 33). Logo, a proposta de mutações baseada em comparações de sequências deveria se atentar para 198 resíduos de aminoácidos que poderiam ser substituídos. Como se viu na Figura 28, os mutantes de Bgl1B apresentaram melhorias em sua resistência a inibição por glicose. Entretanto, Bgl1A ainda mantém atividade superior que todos os mutantes em altas con-

centrações de glicose. Logo, deve se questionar se seria possível melhorar ainda mais a atividade de Bgl1B em altas concentrações de glicose efetuando apenas algumas mutações em outros resíduos. Usando apenas alinhamento de sequência seria necessário testar todas essas mutações. Assim, torna-se claro a necessidade de métodos mais eficientes para reduzir o número de testes em bancada, como por exemplo as comparações por assinaturas estruturais propostas por SSV.



**Figura 33. Comparação entre estruturas de Bgl1A e Bgl1B.**

(A) Alinhamento estrutural entre Bgl1A (*cartoon* azul) e Bgl1B (*cartoon* branco). Estruturas apresentam um RMSD de 0,643. Resíduos do bolsão catalítico são mostrados como *sticks* (exceto F302). Resíduos destacados no trabalho de YANG et al. (2015b): T228, Q301 e F302 (verde). Resíduos destacados no trabalho LIU et al. (2011): H228 (azul ciano). Posições correspondentes aos resíduos destacados no trabalho de DE GIUSEPPE et al. (2014) e CAO et al. (2015): F172 (posição correspondente a W169), C173 e L177 (magenta). Resíduos catalítico E170 e E352 foram rotulados para servir como referência de posicionamento. Figura gerada com PyMOL (<http://pymol.org>). (B) Alinhamento entre sequências gerado com Clustal Omega v1.2.4 (<https://www.ebi.ac.uk/Tools/msa/clustalo>). Percentual de identidade: 55,43% (calculado por Clustal2.1). Fonte: próprio autor.

Infer-se então que não é apenas a acessibilidade da glicose ao sítio ativo que está relacionado à inibição. Afinidade, energia de ligação e preferência relativa da glicose a outros sítios é crucial para a  $\beta$ -glicosidases glicose-tolerantes. YANG et al. (2015b) ainda relatam que o mecanismo e modelo proposto foi suportado por validações bioquímicas e pode ser estendido a outras  $\beta$ -glicosidases. Entretanto, ainda seria possível aperfeiçoar a atividade de Bgl1B para que ela se aproxime de Bgl1A. No estudo de caso 1, SSV obteve um  $\Delta\Delta$ SSV de -231 para a mutação benéfica H228T, quando se esperava um valor de  $\Delta\Delta$ SSV  $< 0$  para mutações benéficas. No estudo de caso 2, em um teste cego, SSV foi capaz de prever três mutações que melhoram a atividade de Bgl1B (H228T, H228C e H228V). Logo, as outras mutações preditas para Bgl1B poderiam melhorar a atividade da enzima a níveis mais próximos de

Bgl1A, uma vez que SSV indica que tais mutações inserem padrões internos mais similares a Bgl1A. Entretanto, validações em bancada são necessárias para conclusões definitivas.

#### 5.1.4 Papel dos resíduos do bolsão catalítico na atividade

Em resumo, o estudo de estruturas das  $\beta$ -glicosidases glicose-tolerantes é importante para detecção de sítios para aplicação de mutações sítio-dirigidas que levem a construção de enzimas  $\beta$ -glicosidases mais eficazes para produção de biocombustíveis. Nesta seção, demonstrou-se que os resíduos H121, N166, E167, Y299, E355, W402, W122, N297, E409, W410 e F418 têm alta conservação nas estruturas de  $\beta$ -glicosidases glicose-tolerantes coletadas, e por essa razão presume-se que tenham papel vital para ligação do substrato. Também foi detectada uma aparente coocorrência do resíduo C170 quando os resíduos W169 e L174 aparecem na estrutura. Foi ainda proposta uma possível ligação disso com a tolerância à glicose. Também foi verificado que o resíduo T226 pode ser essencial para a tolerância à glicose. Além disso, foi detectado que o resíduo H181, descrito na literatura com importante para tolerância a inibição por glicose, realiza mais contatos com a celobiose do que quando é substituído por uma fenilalanina. Ainda, a mutação nesse resíduo pode estar relacionado com uma maior passagem de água até o sítio ativo, o que favoreceria a hidrólise. Por fim, foi proposto que o número de contatos de um resíduo em determinada posição pode ser uma boa métrica, além da conservação, para propor mutações para  $\beta$ -glicosidases não tolerantes. Entretanto, métodos que envolvam estrutura tridimensional podem ser de grande valor para a proposta de mutações mais eficientes. Os resultados das análises de sequências sugerem que existem certas conservações no bolsão catalítico das glicose-tolerantes e certos resíduos podem estar relacionados com as características dessas enzimas (Tabela 12). Pode-se citar por exemplo o resíduo na posição 224, que é majoritariamente uma asparagina nas glicose-tolerantes, mas pode apresentar diversos outros resíduos em outras proteínas da família GH1.

**Tabela 12. Importância dos resíduos do bolsão catalítico.**

<i>Resíduo</i> <sup>1</sup>	<i>GH1</i> <sup>2</sup>	<i>Glicose-tolerantes</i> <sup>3</sup>	<i>Importância</i> <sup>4</sup>
<i>H121</i>	H	H	Provavelmente está envolvido na ligação do substrato a enzima ou na estabilização do estado de transição, participando da ação catalítica junto a E167 e E355. Faz parte da região responsável pela ligação da celobiose.
<i>W122</i>	W	W	Faz parte da região responsável pela ligação da celobiose.
<i>N166</i>	N	N	Faz parte da região responsável pela ligação da celobiose.

<i>E167</i>	E	E	Sítio ativo (ácido/base catalítico)
<i>W169</i>	y G F W	N W L	Importantes para liberação da glicose. Entretanto, uma fenilalanina (F) aparece nesta posição em Bgl1A (também considerada glicose-tolerante).
<i>C170</i>	i a C V	C V T	Possivelmente requer um aminoácido pequeno nesta posição devido a presença de aminoácidos volumosos na vizinhança, como por exemplo W169 (ao lado).
<i>L174</i>	i E V L	L V D	Importante para a liberação da glicose.
<i>H181</i>	H	H F M	Pode estar relacionado a um túnel que permite a passagem de água até o sítio ativo, o que indicaria a preferência pela hidrólise ou pela transglicosilação.
<i>N224</i>	k i l t d g s a V N	A N	Possivelmente uma asparagina ou uma alanina nesta posição favorece a glicose-tolerância, uma vez que outras proteínas da família GH1 apresentam uma variada gama de aminoácidos nesta posição.
<i>L225</i>	f N a T M i V L	N F L G I	Apresenta uma grande variação de aminoácidos nesta posição. Realiza poucos ou nenhum contato com o ligante. Baixa importância para glicose-tolerância.
<i>T226</i>	i l r q v p d e k M G N A S T	T S A G D N	Apresenta uma grande variação de aminoácidos nesta posição. Entretanto, a substituição de uma histidina por treonina, citosina, aspartato, serina ou valina melhorou a resistência a inibição de Bgl1B. Logo, a presença de um aminoácido polar básico e mais volumoso pode não ser benéfico para a glicose-tolerância, sendo preferível um aminoácido capaz de agir como acceptor em ligações de hidrogênio. Resíduo responsável pelo processo de liberação da glicose (ver apêndices).
<i>A244</i>	c W p d M e k Q n g r t i S V A H L F Y	A M K R S	Apresenta uma grande variação de aminoácidos nesta posição. Realiza poucos ou nenhum contato com o ligante. Baixa importância para glicose-tolerância.
<i>N297</i>	N	N	Resíduo próximo ao sítio ativo. Aparenta ter alguma relação com a ligação da celobiose.
<i>Y298</i>	d N Y	Y W F H	Resíduo próximo ao sítio ativo. Nas glicose-tolerantes a posição foi ocupada por resíduos volumosos.
<i>Y299</i>	Y	Y	Faz parte da região responsável pela ligação da celobiose.
<i>T300</i>	S T	S T	Recebe uma serina ou uma treonina tanto nas GH1 quanto nas glicose-tolerantes. Seu papel precisa ser melhor estabelecido.
<i>S302</i>	c y f M H i p v l g N r T d Q S A k E	Q L N H A S	Apresenta uma grande variação de aminoácidos nesta posição. Baixa importância para glicose-tolerância.
<i>E355</i>	E	E	Sítio ativo (nucleófilo)
<i>W402</i>	W	W	Faz parte da região responsável pela ligação da celobiose.
<i>E409</i>	E	E	Faz parte da região responsável pela ligação da celobiose.

<i>W410</i>	W	W	Faz parte da região responsável pela ligação da celobiose.
<i>F418</i>	F	F	Faz parte da região responsável pela ligação da celobiose.

<sup>1</sup> Aminoácidos correspondentes ao bolsão catalítico de *T. brockii*.

<sup>2</sup> Aminoácidos encontrados nas posições correspondentes em outras proteínas da família GH1. Determinado pela ferramenta SIFT. Aminoácidos que aparecem poucas vezes para determinada posição foram descartados (*threshold* 0,05). Letras maiúsculas indicam que o aminoácido aparece no alinhamento de toda a família. Letras minúsculas indicam que é uma predição do SIFT (NG; HENIKOFF, 2003).

<sup>3</sup> Aminoácidos encontrados nas posições correspondentes em  $\beta$ -glicosidases glicose-tolerantes. Aminoácidos que aparecem em menos do que duas sequências foram descartados.

<sup>4</sup> Possível importância dos resíduos para a glicose-tolerância.

Assim, mesmo resíduos declarados como vitais para glicose-tolerância, como (W169 e L174), alguns não estão presentes em todas as glicose-tolerantes detectadas. Além disso, não se pode ter certeza se tais resíduos vitais agem em conjunto com outros resíduos próximos ou distantes. Logo, apesar de que a comparação com estruturas primárias de enzimas que apresentem características desejadas seja uma boa estratégia, como por exemplo a comparação entre Bgl1A e Bgl1B, o uso de algoritmos e métodos capazes de detectar detalhes às vezes imperceptíveis ao olhar humano pode ser importante. O uso de assinaturas estruturais pelo método SSV permite uma abstração do problema, uma vez que SSV implementa a transferência de características intrínsecas por meio da redução da diferença de variação de assinaturas. SVV permite a proposta de mutações eficientes mesmo sem uma explicação explícita dos motivos biológicos que levam aquela mutação a melhorar a atividade da enzima.

## 5.2 Comparação do SSV com outros métodos de proposta de mutações

Mutações podem afetar a estabilidade da proteína, influenciar na catálise, afetar sítios de ligação para substrato, causar regulações alostéricas ou modificações pós-traducionais (WANG; MOULT, 2001). Até mesmo o impacto de mutações pontuais, aquelas que causam a mudança de um único aminoácido, na função da proteína é difícil de ser previsto computacionalmente. Mudanças na estabilidade podem levar à disfunção de uma proteína, podendo causar até mesmo doenças. Entretanto, mudanças na estabilidade também são a base da engenharia de proteínas (LAIMER et al., 2015). Estima-se que mutações pontuais causem uma mudança na estabilidade da proteína entre um e três Kcal/mol, o que não seria suficiente para alterar o enovelamento da proteína, entretanto mutações podem afetar importantes funções das proteínas (WANG; MOULT, 2001). Por isso, propor mutações *in silico* em enzimas não é uma tarefa trivial e deve-se levar em consideração múltiplos fatores e estratégias.

Alinhamento de sequências é o mais tradicional método computacional para comparação de sequências e detecção de diferenças por meio das matrizes de substituição (CORPET, 1988). Alinhamentos de sequências podem ser realizados par-a-par ou com múltiplas sequências. Diversas ferramentas têm utilizado análises de sequências como estratégias para proposta de mutações. Por exemplo, a ferramenta SIFT utiliza alinhamentos múltiplos e homologia de sequências para detectar ocorrência de resíduos para determinada posição de uma proteína alvo e toda sua família (NG; HENIKOFF, 2001). Outro exemplo é a ferramenta ProSAR, que usa sequências para prever contribuições de mutações para funções de proteínas (BERLAND et al., 2014; FOX et al., 2007). Entretanto, tais métodos não levam em consideração a estrutura tridimensional da proteína.

Outra estratégia computacional para a proposta de mutações é a predição *in silico* da diferença da variação da energia livre de Gibbs ( $\Delta\Delta G$ ). O cálculo do  $\Delta\Delta G$  pode ser utilizado para prever mudanças na estabilidade que são desejadas nas abordagens de engenharia de proteínas. A teoria da variação de energia, proposta por Willard Gibbs, define que, durante reações químicas, a energia livre de Gibbs (G) pode ser calculada com base em três fatores: entalpia, entropia e temperatura. Em temperatura constante, a variação de energia livre ( $\Delta G$ ) é determinada pelo número de ligações químicas, formação e quebra de interações não covalentes (entalpia), como ligações de hidrogênio, e a variação da aleatoriedade do sistema (entropia). Em reações exergônica o valor de  $\Delta G$  é negativo, pois o sistema se transforma de modo a possuir menor energia (liberação de energia). Em reações endergônicas o valor de  $\Delta G$  é positivo, pois o sistema adquire energia livre. A diferença da energia livre de Gibbs ( $\Delta\Delta G$ ) corresponde a diferença de energia livre entre as estruturas selvagem e mutante. Valores mais baixos de  $\Delta\Delta G$  refletem uma menor dificuldade para assumir certa conformação e são termodinamicamente mais estáveis (NELSON et al., 2014). Ferramentas computacionais têm utilizado abordagens, como assinaturas estruturais e aprendizado de máquina, para predição da diferença da energia livre de Gibbs ( $\Delta\Delta G$ ). Como exemplo pode-se citar as ferramentas mCSM (PIRES; ASCHER; BLUNDELL, 2014) e Maestro (LAIMER et al., 2015). Diversos estudos têm mostrado que essas abordagens têm utilidade, mas também apresentam limitações (LAIMER et al., 2015). Por exemplo, levam em consideração apenas o impacto no ambiente local do resíduo mutado, além de ser difícil realizar previsões nas variadas condições em que a proteína está presente, como temperatura e pH, uma vez que tais ferramentas dependem de dados experimentais para serem treinadas. Além disso, não apresentam explicação semântica dos motivos biológicos que levaram a tal resultado.

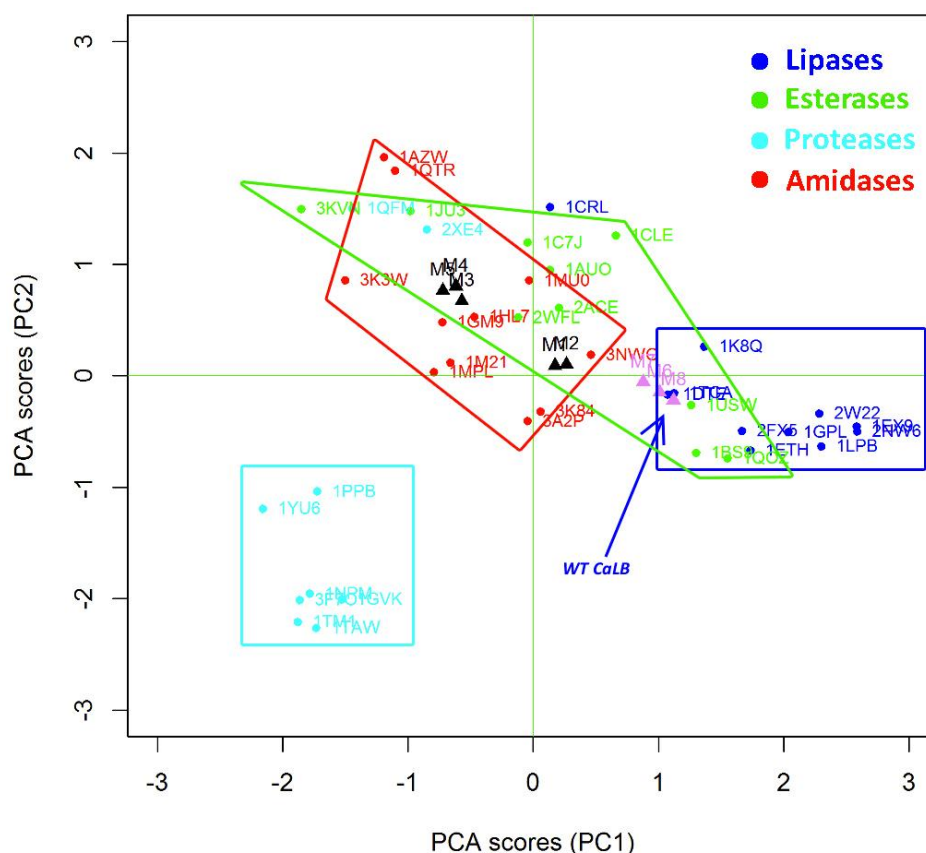


Métodos que levam em consideração a estrutura tridimensional podem ser uma alternativa robusta a métodos que trabalham apenas com sequência. Por exemplo, o método *active site constellations* usa uma abordagem baseada em estrutura que é independente de parâmetros de similaridade entre sequência e estrutura para sugerir promiscuidade catalítica, e assim, desenvolver e manipular novas enzimas (STEINKELLNER et al., 2014). BioGPS é um outro exemplo de metodologia *in silico* para engenharia de enzimas promíscuas usando propriedades físico-químicas e geométricas de estruturas tridimensionais (FERRARIO et al., 2014). BioGPS usa assinaturas estruturais (referidas como *fingerprints*) para comparar propriedades de sítios ativos levando em consideração mais do que a sequência. Portanto, considerou-se uma abordagem similar ao SSV.

Assim, realizou-se uma comparação de SSV com BioGPS usando os dados do estudo de caso com oito mutantes de CaLB (lipase B de *Candida antarctica*). Entretanto há algumas particularidades em comparação com o estudo de caso executado no trabalho de FERRARIO et al. (2014). Primeiramente, eles usaram uma base de dados heterogênea com 42 serina-hidrolases para treinamento e construção da assinatura. Entretanto, eles afirmaram que não foi feita uma curadoria nessa base de dados, baseando-se exclusivamente na anotação, o que poderia apresentar um viés no resultado, uma vez que não há uma garantia que as enzimas selecionadas apresentam a melhor performance para atividade requerida. Por isso decidiu-se utilizar o mutante M3 como *template* no experimento com SSV. M3 apresentava o maior fator de melhoria (atividade de amidase 11,2 vezes maior que a selvagem). Como experimento controle, usou-se o mutante M3 como teste e treinamento, e obteve-se o resultado esperado ( $\Delta\Delta\text{SSV} = -841$ , quando o esperado para uma mutação benéfica era um  $\Delta\Delta\text{SSV} < 0$ ). Das outras sete mutantes avaliadas, SSV obteve um resultado similar a BioGPS em cinco, errando apenas a predição dos mutantes M4 (G39A), que obteve um  $\Delta\Delta\text{SSV} = 150$ , mas por ter um FM de 2,8 esperava-se um  $\Delta\Delta\text{SSV} < 0$ ; e M6 (I189A), que obteve um  $\Delta\Delta\text{SSV} = -94$ , mas por ter um FM de 0,4 esperava-se um  $\Delta\Delta\text{SSV} > 0$ . É possível que com uma base de treinamento com mais *templates* seria possível obter-se uma maior acurácia. O mutante M8 (T103G) obteve um  $\Delta\Delta\text{SSV} = 0$ . Esse caso foi considerado um acerto devido ao mutante ter um FM de 1,1, o que indica uma baixa mudança na atividade de amidase da enzima, logo esperava-se  $\Delta\Delta\text{SSV} = 0$  ou  $\Delta\Delta\text{SSV} < 0$ .

No estudo de FERRARIO et al. (2014) é descrito que a base de treinamento é composta por quatro classes de serina-hidrolases: (i) lipases; (ii) esterases; (iii) proteases; e (iv) amidases. O objetivo deles era converter uma lipase (CaLB) em uma amidase, que tem maior utilidade na indústria. Assim, eles demonstraram a eficácia de BioGPS plotando os componentes

principais (PC1 e PC2) da base de treinamento e dos mutantes (Figura 34). A Figura 34 realmente demonstra que os mutantes benéficos estão dentro do grupo das amidases, enquanto a proteína selvagem está presente no grupo das lipases. Entretanto, algumas considerações precisam ser feitas. Os componentes principais dos mutantes M6 e M7 se aproximaram levemente ao grupo das amidases. Entretanto, os dados experimentais indicam que tais mutações reduzem a atividade de amidase. Ainda, o grupo das esterases parece ser bastante heterogêneo, se mesclando ao grupo das amidases e das lipases. E há diversos *outliers* presentes em grupos distintos, além de *outliers* não classificados em nenhum grupo.



**Figura 34. Projeção da análise de componentes principais (PC1 e PC2) dos mutantes CaLB no modelo BioGPS-UPCA (score global).**

Mutantes melhorados são destacados como triângulos pretos enquanto mutantes não benéficos estão em rosa. Proteína selvagem CaLB (1TCA) é indicado pela seta azul. As diferentes classes de serina hidrolases são relatadas em cores diferentes. Cada proteína é indicada pelo seu PDB ID. Fonte: adaptado de (FERRARIO et al., 2014).

Por fim, é necessário destacar que SSV apresenta algumas vantagens sobre BioGPS. Os autores de BioGPS submeteram as estruturas mutantes a 500 ns de dinâmica molecular usando GROMACS (versão 4) para prepará-las antes das análises. Dinâmica molecular tem um alto custo computacional, enquanto SSV usa assinaturas baseadas no método CSM, que tem um baixo custo computacional de processamento. Ferramentas baseadas nos métodos

CSM têm sido usadas com sucesso para muitas aplicações, como por exemplo predição de efeitos de mutações na estabilidade de proteínas (PIRES; ASCHER; BLUNDELL, 2014) e predição do impacto na afinidade entre anticorpos e antígenos (PIRES; ASCHER, 2016). Conclui-se que assinaturas estruturais fornecem um método computacional viável para identificar padrões em macromoléculas que podem ser importantes para definir estrutura e função.

## 6. Conclusões

Neste manuscrito propôs-se um método para sugestão de mutações em enzimas  $\beta$ -glicosidase utilizadas na produção de biocombustíveis de segunda geração. Realizou-se uma profunda busca na literatura através de uma revisão sistemática e criou-se uma base de dados com estruturas de  $\beta$ -glicosidases glicose-tolerantes, denominada BETAGDB. A característica da tolerância a inibição por glicose tem sido descrita como responsável por permitir que enzimas  $\beta$ -glicosidase tenham maior eficiência na conversão de celobiose em glicose. As análises dos resultados da revisão sistemática sugerem que os resíduos responsáveis pela característica da tolerância a glicose devem estar na concavidade principal de  $\beta$ -glicosidases: um canal que leva a passagem do substrato até o sítio ativo. Denominou-se bolsão catalítico a região e todos os resíduos localizados a uma distância de 6,5 Å da posição de ligação do substrato no sítio ativo. Propôs-se ainda, que a assinatura estrutural extraída do bolsão catalítico poderia caracterizar  $\beta$ -glicosidases glicose-tolerantes, além de permitir a comparação dessas enzimas com  $\beta$ -glicosidases não tolerantes. Para testar essa hipótese, criou-se um método para propor mutações baseado na diferença de variação de assinaturas estruturais, denominado SSV (*Structural Signature Variation*). A partir de três estudos de caso, avaliou-se a capacidade do método em prever se mutações seriam benéficas ou não, propôs-se mutações para uma  $\beta$ -glicosidase não tolerante comparando as mutações propostas com dados experimentais disponíveis, e por fim, comparou-se o método SSV com SVM e com a ferramenta BioGPS. SSV foi eficaz nos três estudos de caso.

Assinaturas estruturais podem ser uma alternativa complementar aos métodos baseados em sequência e às ferramentas de predição de diferença de variação de energia livre. Em geral, métodos baseados em sequência apresentam baixo custo computacional. SSV é um método baseado em comparações de estruturas tridimensionais com baixo custo computacional. O uso de estruturas tridimensionais tem sido adotado por outras ferramentas, modelos e algoritmos, como por exemplo os métodos *active site constellations* e *BioGPS Descriptors*. BioGPS reportou que seu método poderia ser usado com baixo custo computacional, entretanto nos estudos de caso apresentados foi necessário o uso de dinâmica molecular para preparar as estruturas, o que pode ter tornado o processo computacionalmente custoso.

SSV se mostra um eficiente método para proposta de mutações para  $\beta$ -glicosidases não tolerantes e pode ajudar na engenharia de enzimas com maior tolerância a inibição por

glicose para produção de biocombustíveis de segunda-geração. O método proposto aqui pode ser estendido para outras enzimas. Além disso, SSV pode ser utilizado em conjunto com outros métodos, ferramentas e algoritmos para sugerir mutações com maior confiabilidade para reduzir custos para experimentos *in vitro*. Uma descrição completa de SSV pode ser obtida em MARIANO et al. (2019). SSV está disponível em um website com um interface amigável no endereço: <<http://bioinfo.dcc.ufmg.br/ssv>>.

## 7. Perspectivas

A proposta de mutações que levem a uma enzima  $\beta$ -glicosidase que não sofra qualquer tipo de inibição durante o processo de sacarificação ainda é um problema em aberto e dificilmente será solucionado por completo, uma vez que a inibição ocorre por competição. Entretanto, este trabalho apresentou boas estratégias para melhorar enzimas pouco eficientes com base em assinaturas estruturais de  $\beta$ -glicosidases caracterizadas como eficientes (ver apêndice 9). Espera-se que as análises, estratégias e ferramentas aqui propostas possam levar a uma nova geração de enzimas mais eficientes para produção de biocombustíveis.

As análises de estruturas evidenciaram neste trabalho que  $\beta$ -glicosidases podem apresentar diversos mecanismos de glicose-tolerância. Análises da entrada e saída de substrato e ligante usando técnicas de dinâmica molecular podem ser úteis para melhor elucidar esses mecanismos (ver apêndice 8).

Poucas  $\beta$ -glicosidases com tolerância conhecida possuem dados disponíveis atualmente: 23 sequências ao todo, sendo 21 da família GH1 e duas da família GH3. Dessas apenas cinco possuem estrutura tridimensional revelada por experimentos de bancada. Espera-se que novas  $\beta$ -glicosidases glicose-tolerantes descobertas possam ajudar a melhorar o método proposto.

Além disso, a maior parte dos testes realizados com SSV utilizou de mutações pontuais. A combinação de diferentes mutações pode, eventualmente, produzir enzimas mais eficientes. Entretanto, deve-se levar em consideração que cada um dos resíduos do bolsão catalítico pode ser mutado por até 19 aminoácidos. Isso geraria um imenso número de possíveis combinações, o que dificultaria a análise. Entretanto, espera-se futuramente realizar testes *in silico* com mais mutações.

Validações em bancada estão sendo realizadas pela equipe do Prof. Dr. Luis Fernando Marins na FURG-RS, colaborador do projeto. Detalhes das mutações propostas foram omitidos desta tese por fazerem parte de patentes requisitadas.

## 8. Referências bibliográficas

- AKRAM, F. et al. Cloning with kinetic and thermodynamic insight of a novel hyperthermostable  $\beta$ -glucosidase from *Thermotoga naphthophila* RKU-10T with excellent glucose tolerance. **Journal of Molecular Catalysis B: Enzymatic**, v. 124, p. 92–104, fev. 2016.
- ALLEN, W. J. et al. DOCK 6: Impact of new features and current docking performance. **Journal of Computational Chemistry**, v. 36, n. 15, p. 1132–1156, 5 jun. 2015.
- AŠIĆ, A. et al. Purification and Characterization of  $\beta$ -Glucosidase from *Agaricus bisporus* (White Button Mushroom). **The Protein Journal**, v. 34, n. 6, p. 453–461, dez. 2015.
- BAI, A. et al. A novel thermophilic  $\beta$ -glucosidase from *Caldicellulosiruptor bescii*: Characterization and its synergistic catalysis with other cellulases. **Journal of Molecular Catalysis B: Enzymatic**, v. 85–86, p. 248–256, jan. 2013.
- BARRETT, T. et al. The crystal structure of a cyanogenic beta-glucosidase from white clover, a family 1 glycosyl hydrolase. **Structure (London, England: 1993)**, v. 3, n. 9, p. 951–960, 15 set. 1995.
- BÉGUIN, P.; AUBERT, J. P. The biological degradation of cellulose. **FEMS microbiology reviews**, v. 13, n. 1, p. 25–58, jan. 1994.
- BENOLIEL, B. et al. Expression of a Glucose-tolerant  $\beta$ -glucosidase from *Humicola grisea* var. *thermoidea* in *Saccharomyces cerevisiae*. **Applied Biochemistry and Biotechnology**, v. 160, n. 7, p. 2036–2044, abr. 2010.
- BERLAND, M. et al. A web-based tool for rational screening of mutants libraries using ProSAR. **Protein Engineering, Design and Selection**, v. 27, n. 10, p. 375–381, 1 out. 2014.
- BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 1 jan. 2000.
- BERRIN, J.-G. et al. Substrate (aglycone) specificity of human cytosolic beta-glucosidase. **Biochemical Journal**, v. 373, n. Pt 1, p. 41–48, 1 jul. 2003.
- BITAR, M.; FRANCO, G. R. A basic protein comparative three-dimensional modeling methodological workflow theory and practice. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)**, v. 11, n. 6, p. 1052–1065, 2014.
- BIVER, S. et al. Two promising alkaline  $\beta$ -glucosidases isolated by functional metagenomics from agricultural soil, including one showing high tolerance towards harsh detergents, oxidants and glucose. **Journal of Industrial Microbiology & Biotechnology**, v. 41, n. 3, p. 479–488, mar. 2014.
- BOUDABBOUS, M. et al. Trans-glycosylation capacity of a highly glycosylated multi-specific  $\beta$ -glucosidase from *Fusarium solani*. **Bioprocess and Biosystems Engineering**, v. 40, n. 4, p. 559–571, abr. 2017.

BP GLOBAL. **BP Statistical Review of World Energy. Inclui dados da F.O. Lichts; US Energy Information Administration.**, 6 jan. 2016. Disponível em: <<http://www.bp.com/content/dam/bp/pdf/energy-economics/statistical-review-2016/bp-statistical-review-of-world-energy-2016-full-report.pdf>>

BRAIUCA, P. et al. 3D-QSAR Applied to the Quantitative Prediction of Penicillin G Amidase Selectivity. **Advanced Synthesis & Catalysis**, v. 348, n. 6, p. 773–780, 2006.

BREVES, R. et al. Genes encoding two different beta-glucosidases of *Thermoanaerobacter brockii* are clustered in a common operon. **Applied and environmental microbiology**, v. 63, n. 10, p. 3902–3910, 1997.

BROWN, P. T.; CALDEIRA, K. Greater future global warming inferred from Earth's recent energy budget. **Nature**, v. 552, n. 7683, p. 45–50, 06 2017.

CAIRNS, J. R. K.; ESEN, A.  $\beta$ -Glucosidases. **Cellular and Molecular Life Sciences**, v. 67, n. 20, p. 3389–3405, out. 2010.

CANTAREL, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. **Nucleic Acids Research**, v. 37, n. Database issue, p. D233-238, jan. 2009.

CAO, L. et al. Engineering a novel glucose-tolerant  $\beta$ -glucosidase as supplementation to enhance the hydrolysis of sugarcane bagasse at high glucose concentration. **Biotechnology for Biofuels**, v. 8, n. 1, dez. 2015.

CHAMOLI, S. et al. Secretory expression, characterization and docking study of glucose-tolerant  $\beta$ -glucosidase from *B. subtilis*. **International Journal of Biological Macromolecules**, v. 85, p. 425–433, abr. 2016.

CHAN, C. S. et al. Characterization of a glucose-tolerant  $\beta$ -glucosidase from *Anoxybacillus* sp. DT3-1. **Biotechnology for Biofuels**, v. 9, n. 1, p. 174, 2016.

CHANDEL, A. K.; SILVA, S. S. DA. **Biochemistry, Genetics and Molecular Biology: "Sustainable Degradation of Lignocellulosic Biomass - Techniques, Applications and Commercialization"**. [s.l: s.n.].

CHAUVE, M. et al. Comparative kinetic analysis of two fungal beta-glucosidases. **Biotechnology for Biofuels**, v. 3, n. 1, p. 3, 11 fev. 2010.

CHOUDRI, B. S. et al. Bioenergy from Biofuel Residues and Wastes. **Water Environment Research: A Research Publication of the Water Environment Federation**, v. 89, n. 10, p. 1441–1460, 1 out. 2017.

CHUENCHOR, W. et al. Structural insights into rice BGluc1 beta-glucosidase oligosaccharide hydrolysis and transglycosylation. **Journal of Molecular Biology**, v. 377, n. 4, p. 1200–1215, 4 abr. 2008.

CORPET, F. Multiple sequence alignment with hierarchical clustering. **Nucleic Acids Research**, v. 16, n. 22, p. 10881–10890, 25 nov. 1988.



COTA, J. et al. Comparative analysis of three hyperthermophilic GH1 and GH3 family members with industrial potential. **New Biotechnology**, v. 32, n. 1, p. 13–20, 25 jan. 2015.

CRESPIM, E. et al. A novel cold-adapted and glucose-tolerant GH1  $\beta$ -glucosidase from *Exiguobacterium antarcticum* B7. **International Journal of Biological Macromolecules**, v. 82, p. 375–380, jan. 2016.

DE GIUSEPPE, P. O. et al. Structural basis for glucose tolerance in GH1  $\beta$ -glucosidases. **Acta Crystallographica Section D Biological Crystallography**, v. 70, n. 6, p. 1631–1639, 1 jun. 2014.

DE MELO-MINARDI, R. C. **Classificação Estrutural de Famílias de Proteínas com Base em Mapas de Contatos**. [s.l.] UFMG, 2008.

DECKER, C. H.; VISSER, J.; SCHREIER, P.  $\beta$ -Glucosidase multiplicity from *Aspergillus tubingensis* CBS 643.92: purification and characterization of four  $\beta$ -glucosidases and their differentiation with respect to substrate specificity, glucose inhibition and acid tolerance. **Applied Microbiology and Biotechnology**, v. 55, n. 2, p. 157–163, 1 mar. 2001.

DEL POZO, M. V. et al. Microbial  $\beta$ -glucosidases from cow rumen metagenome enhance the saccharification of lignocellulose in combination with commercial cellulase cocktail. **Biotechnology for Biofuels**, v. 5, p. 73, 21 set. 2012.

DESJARLAIS, R. L. et al. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. **Journal of Medicinal Chemistry**, v. 31, n. 4, p. 722–729, abr. 1988.

FANG, W. et al. Cloning and characterization of a beta-glucosidase from marine metagenome. **Sheng Wu Gong Cheng Xue Bao = Chinese Journal of Biotechnology**, v. 25, n. 12, p. 1914–1920, dez. 2009.

FANG, Z. et al. Cloning and Characterization of a  $\beta$ -Glucosidase from Marine Microbial Metagenome with Excellent Glucose Tolerance. **JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY**, v. 20, n. 9, p. 1351–1358, set. 2010.

FERRARIO, V. et al. BioGPS descriptors for rational engineering of enzyme promiscuity and structure based bioinformatic analysis. **PLoS One**, v. 9, n. 10, p. e109354, 2014a.

FERRARIO, V. et al. An integrated platform for automatic design and screening of virtual mutants based on 3D-QSAR analysis. **Journal of Molecular Catalysis B: Enzymatic**, v. 101, p. 7–15, 1 mar. 2014b.

FLORINDO, R. N. et al. Structural insights into  $\beta$ -glucosidase transglycosylation based on biochemical, structural and computational analysis of two GH1 enzymes from *Trichoderma harzianum*. **New Biotechnology**, v. 40, n. Pt B, p. 218–227, 25 jan. 2018.

FOX, R. J. et al. Improving catalytic function by ProSAR-driven enzyme evolution. **Nature Biotechnology**, v. 25, n. 3, p. 338–344, mar. 2007.

FRANK, E.; HALL, M. A.; WITTEN, I. H. The WEKA Workbench. In: **Data Mining: Practical Machine Learning Tools and Techniques**. 4. ed. [s.l.: s.n.].

FRUTUOSO, M. A. **Estudo das bases moleculares de reações de transglicosilação em Î<sup>2</sup>-glicosidases GH1 de Spodoptera frugiperda eTenebrio molitor**. text—[s.l.] Universidade de São Paulo, 18 fev. 2011.

GOPALAN, V. et al. Exolytic hydrolysis of toxic plant glucosides by guinea pig liver cytosolic  $\beta$ -glucosidase. **Journal of Biological Chemistry**, v. 267, n. 20, p. 14027–14032, 1992.

GOUJON, M. et al. A new bioinformatics analysis tools framework at EMBL–EBI. **Nucleic Acids Research**, v. 38, n. suppl 2, p. W695–W699, 7 jan. 2010.

GUEGUEN, Y. et al. Purification and characterization of an intracellular  $\beta$ -glucosidase from Botrytis cinerea. **Enzyme and Microbial Technology**, v. 17, n. 10, p. 900–906, out. 1995.

GUMEROV, V. M. et al. A Novel Highly Thermostable Multifunctional Beta-Glycosidase from Crenarchaeon Acidilobus saccharovorans, A Novel Highly Thermostable Multifunctional Beta-Glycosidase from Crenarchaeon Acidilobus saccharovorans. **Archaea, Archaea**, v. 2015, 2015, p. e978632, 11 out. 2015.

GUO, B.; AMANO, Y.; NOZAKI, K. Improvements in Glucose Sensitivity and Stability of Trichoderma reesei  $\beta$ -Glucosidase Using Site-Directed Mutagenesis. **PLOS ONE**, v. 11, n. 1, p. e0147301, 20 jan. 2016.

HARNPICHARNCHAI, P. et al. A thermotolerant beta-glucosidase isolated from an endophytic fungi, Periconia sp., with a possible use for biomass conversion to sugars. **Protein Expression and Purification**, v. 67, n. 2, p. 61–69, out. 2009.

HENRISSAT, B. A classification of glycosyl hydrolases based on amino acid sequence similarities. **Biochemical Journal**, v. 280, n. 2, p. 309–316, 1 dez. 1991.

HENRISSAT, B.; BAIROCH, A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. **Biochemical Journal**, v. 293, n. 3, p. 781–788, 1 ago. 1993.

HENRISSAT, B.; DAVIES, G. Structural and sequence-based classification of glycoside hydrolases. **Current Opinion in Structural Biology**, v. 7, n. 5, p. 637–644, out. 1997.

HO, D. P.; NGO, H. H.; GUO, W. A mini review on renewable sources for biofuel. **Biore-source Technology**, v. 169, p. 742–749, out. 2014.

HUANG, Y. et al. Identification of a  $\beta$ -glucosidase from the Mucor circinelloides genome by peptide pattern recognition. **Enzyme and Microbial Technology**, v. 67, p. 47–52, dez. 2014.

IRWIN, J. J. et al. ZINC: A Free Tool to Discover Chemistry for Biology. **Journal of Chemical Information and Modeling**, v. 52, n. 7, p. 1757–1768, 23 jul. 2012.

JABBOUR, D.; KLIPPEL, B.; ANTRANIKIAN, G. A novel thermostable and glucose-tolerant  $\beta$ -glucosidase from Fervidobacterium islandicum. **Applied Microbiology and Biotechnology**, v. 93, n. 5, p. 1947–1956, mar. 2012.

JAKALIAN, A.; JACK, D. B.; BAYLY, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. **Journal of Computational Chemistry**, v. 23, n. 16, p. 1623–1641, dez. 2002.

- JENG, W.-Y. et al. High-resolution structures of *Neotermes koshunensis*  $\beta$ -glucosidase mutants provide insights into the catalytic mechanism and the synthesis of glucoconjugates. **Acta Crystallographica. Section D, Biological Crystallography**, v. 68, n. Pt 7, p. 829–838, jul. 2012.
- JOHNSON, M. et al. NCBI BLAST: a better web interface. **Nucleic Acids Research**, v. 36, n. Web Server issue, p. W5-9, 1 jul. 2008.
- JUBB, H. C. et al. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. **Journal of Molecular Biology**, v. 429, n. 3, p. 365–371, 03 2017.
- KADAM, S. K.; DEMAIN, A. L. Addition of cloned beta-glucosidase enhances the degradation of crystalline cellulose by the *Clostridium thermocellum* cellulose complex. **Biochemical and Biophysical Research Communications**, v. 161, n. 2, p. 706–711, 15 jun. 1989.
- KHAIRUDIN, N. B. A.; MAZLAN, N. S. F. Molecular docking study of Beta-glucosidase with cellobiose, cellotetraose and cellotriose. **Bioinformation**, v. 9, n. 16, p. 813–817, 2013.
- KUMAR, R.; SINGH, S.; SINGH, O. V. Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. **Journal of Industrial Microbiology & Biotechnology**, v. 35, n. 5, p. 377–391, maio 2008.
- LAIMER, J. et al. MAESTRO--multi agent stability prediction upon point mutations. **BMC bioinformatics**, v. 16, p. 116, 16 abr. 2015.
- LEE, H.-L. et al. Mutations in the substrate entrance region of  $\beta$ -glucosidase from *Trichoderma reesei* improve enzyme activity and thermostability. **Protein Engineering Design and Selection**, v. 25, n. 11, p. 733–740, 1 nov. 2012.
- LI, G. et al. Molecular cloning and characterization of a novel  $\beta$ -glucosidase with high hydrolyzing ability for soybean isoflavone glycosides and glucose-tolerance from soil metagenomic library. **Bioresource Technology**, v. 123, p. 15–22, nov. 2012.
- LIEW, K. J. et al. Purification and characterization of a novel GH1 beta-glucosidase from *Jeotgalibacillus malaysiensis*. **International Journal of Biological Macromolecules**, v. 115, p. 1094–1102, ago. 2018.
- LIU, J. et al. The 184th residue of  $\beta$ -glucosidase Bgl1B plays an important role in glucose tolerance. **Journal of Bioscience and Bioengineering**, v. 112, n. 5, p. 447–450, nov. 2011.
- LOPEZ-CAMACHO, C. et al. Amino acid substitutions enhancing thermostability of *Bacillus polymyxa* beta-glucosidase A. **The Biochemical Journal**, v. 314 ( Pt 3), p. 833–838, 15 mar. 1996.
- LOVELL, S. C. et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. **Proteins**, v. 50, n. 3, p. 437–450, 15 fev. 2003.
- LU, J. et al. Expression and characterization of a novel highly glucose-tolerant  $\beta$ -glucosidase from a soil metagenome. **Acta Biochimica et Biophysica Sinica**, v. 45, n. 8, p. 664–673, 1 ago. 2013.

MAIER, J. A. et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. **Journal of Chemical Theory and Computation**, v. 11, n. 8, p. 3696–3713, 11 ago. 2015.

MALLEK-FAKHFAKH, H.; BELGHITH, H. Physicochemical properties of thermotolerant extracellular  $\beta$ -glucosidase from *Talaromyces thermophilus* and enzymatic synthesis of cello-oligosaccharides. **Carbohydrate Research**, v. 419, p. 41–50, jan. 2016.

MARANA, S. R. et al. Amino acid residues involved in substrate binding and catalysis in an insect digestive  $\beta$ -glycosidase. **Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology**, v. 1545, n. 1–2, p. 41–52, 9 fev. 2001.

MARCOLINO, L. S.; COUTO, B. R. G. M.; SANTOS, M. A. DOS. Genome Visualization in Space. In: ROCHA, M. P. et al. (Eds.). **Advances in Bioinformatics**. Advances in Intelligent and Soft Computing. [s.l.] Springer Berlin Heidelberg, 2010. p. 225–232.

MARIANO, D. et al. **A guide to performing systematic literature reviews in bioinformatics**. Belo Horizonte, MG, Brazil: Universidade Federal de Minas Gerais, 2017a. Disponível em:

<[https://www.researchgate.net/publication/318506189\\_A\\_guide\\_to\\_performing\\_systematic\\_literature\\_reviews\\_in\\_bioinformatics](https://www.researchgate.net/publication/318506189_A_guide_to_performing_systematic_literature_reviews_in_bioinformatics)>. Acesso em: 19 jul. 2017.

MARIANO, D. et al. A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). **International Journal of Molecular Sciences**, v. 20, n. 2, 15 jan. 2019.

MARIANO, D. C. B. et al. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002. **BMC Genomics**, v. 17, p. 315, 2016.

MARIANO, D. C. B. et al. Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review. **Genetics and Molecular Research**, v. 16, n. 3, 2017b.

MARTIN, M. A. First generation biofuels compete. **New Biotechnology**, v. 27, n. 5, p. 596–608, 30 nov. 2010.

MARTÍ-RENOM, M. A. et al. Comparative protein structure modeling of genes and genomes. **Annual Review of Biophysics and Biomolecular Structure**, v. 29, p. 291–325, 2000.

MATSUZAWA, T. et al. Crystal structure and identification of a key amino acid for glucose tolerance, substrate specificity, and transglycosylation activity of metagenomic  $\beta$ -glucosidase Td2F2. **The FEBS journal**, v. 283, n. 12, p. 2340–2353, jun. 2016.

MATSUZAWA, T.; YAOI, K. Screening, identification, and characterization of a novel saccharide-stimulated  $\beta$ -glycosidase from a soil metagenomic library. **Applied Microbiology and Biotechnology**, v. 101, n. 2, p. 633–646, jan. 2017.

MELEIRO, L. P. et al. A novel  $\beta$ -glucosidase from *Humicola insolens* with high potential for untreated waste paper conversion to sugars. **Applied Biochemistry and Biotechnology**, v. 173, n. 2, p. 391–408, maio 2014.

MELEIRO, L. P. et al. A *Neurospora crassa*  $\beta$ -glucosidase with potential for lignocellulose hydrolysis shows strong glucose tolerance and stimulation by glucose and xylose. **Journal of Molecular Catalysis B: Enzymatic**, v. 122, p. 131–140, dez. 2015.

MURPHY, L. et al. Product inhibition of five *Hypocrea jecorina* cellulases. **Enzyme and Microbial Technology**, v. 52, n. 3, p. 163–169, 5 mar. 2013.

NELSON, D. L. et al. **Princípios de bioquímica de Lehninger**. [s.l.] Artmed, 2014.

NG, P. C.; HENIKOFF, S. Predicting deleterious amino acid substitutions. **Genome Research**, v. 11, n. 5, p. 863–874, maio 2001.

NG, P. C.; HENIKOFF, S. SIFT: Predicting amino acid changes that affect protein function. **Nucleic Acids Research**, v. 31, n. 13, p. 3812–3814, 1 jul. 2003.

PEARSON, W. R. Finding Protein and Nucleotide Similarities with FASTA. **Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]**, v. 53, p. 3.9.1-3.9.25, 2016.

PEI, J. et al. *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase: a glucose-tolerant enzyme with high specific activity for cellobiose. **Biotechnol Biofuels**, v. 5, n. 31, p. 1–10, 2012.

PENTZOLD, S. et al. How insects overcome two-component plant chemical defence: plant  $\beta$ -glucosidases as the main target for herbivore adaptation. **Biological Reviews**, v. 89, n. 3, p. 531–551, 1 ago. 2014.

PERCIVAL ZHANG, Y.-H.; HIMMEL, M. E.; MIELENZ, J. R. Outlook for cellulase improvement: screening and selection strategies. **Biotechnology Advances**, v. 24, n. 5, p. 452–481, out. 2006.

PÉREZ-PONS, J. A.; REBORDOSA, X.; QUEROL, E. Properties of a novel glucose-enhanced beta-glucosidase purified from *Streptomyces* sp. (ATCC 11238). **Biochimica Et Biophysica Acta**, v. 1251, n. 2, p. 145–153, 6 set. 1995.

PETTERSEN, E. F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. **Journal of Computational Chemistry**, v. 25, n. 13, p. 1605–1612, out. 2004.

PIRES, D. E. et al. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. **BMC Genomics**, v. 12, n. Suppl 4, p. S12, 22 dez. 2011.

PIRES, D. E. V. **CSM: uma assinatura para grafos biológicos baseada em padrões de distâncias**. [s.l.] UFMG, 2012.

PIRES, D. E. V. et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics (Oxford, England)**, v. 29, n. 7, p. 855–861, 1 abr. 2013.

PIRES, D. E. V.; ASCHER, D. B. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. **Nucleic Acids Research**, v. 44, n. W1, p. W469-473, 8 jul. 2016.

- PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics (Oxford, England)**, v. 30, n. 3, p. 335–342, 1 fev. 2014.
- PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. **Scientific Reports**, v. 6, p. 29575, 7 jul. 2016.
- RAJASREE, K. P. et al. Highly glucose tolerant  $\beta$ -glucosidase from *Aspergillus unguis*: NII 08123 for enhanced hydrolysis of biomass. **Journal of Industrial Microbiology & Biotechnology**, v. 40, n. 9, p. 967–975, set. 2013.
- RAMAN, J. K.; GNANSOUNOU, E. Ethanol and lignin production from Brazilian empty fruit bunch biomass. **Bioresource Technology**, v. 172, p. 241–248, nov. 2014.
- RAMANI, G. et al. Molecular cloning and expression of thermostable glucose-tolerant  $\beta$ -glucosidase of *Penicillium funiculosum* NCL1 in *Pichia pastoris* and its characterization. **Journal of Industrial Microbiology & Biotechnology**, v. 42, n. 4, p. 553–565, abr. 2015a.
- RAMANI, G. et al. Transglycosylating glycoside hydrolase family 1  $\beta$ -glucosidase from *Penicillium funiculosum* NCL1: Heterologous expression in *Escherichia coli* and characterization. **Biochemical Engineering Journal**, Emerging Trends in Industrial Biotechnology (SI-ICETB-2014). v. 102, p. 6–13, 15 out. 2015b.
- REGO, N.; KOES, D. 3Dmol.js: molecular visualization with WebGL. **Bioinformatics**, v. 31, n. 8, p. 1322–1324, 15 abr. 2015.
- ROBAK, K.; BALCEREK, M. Review of Second Generation Bioethanol Production from Residual Biomass. **Food Technology and Biotechnology**, v. 56, n. 2, p. 174–187, jun. 2018.
- SAHA, B. C.; BOTHAST, R. J. Production, purification, and characterization of a highly glucose-tolerant novel beta-glucosidase from *Candida peltata*. **Applied and Environmental Microbiology**, v. 62, n. 9, p. 3165–3170, 1996.
- SALGADO, J. C. S. et al. Glucose tolerant and glucose stimulated  $\beta$ -glucosidases - A review. **Bioresource Technology**, v. 267, p. 704–713, nov. 2018.
- SALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. **Journal of Molecular Biology**, v. 234, n. 3, p. 779–815, 5 dez. 1993.
- SANSENAYA, S. et al. The crystal structure of rice (*Oryza sativa* L.) Os4BGlu12, an oligosaccharide and tuberonic acid glucoside-hydrolyzing  $\beta$ -glucosidase with significant thioglucohydrolase activity. **Archives of Biochemistry and Biophysics**, v. 510, n. 1, p. 62–72, 1 jun. 2011.
- SANSENAYA, S.; MANEESAN, J.; CAIRNS, J. R. K. Exchanging a single amino acid residue generates or weakens a +2 cellooligosaccharide binding subsite in rice  $\beta$ -glucosidases. **Carbohydrate Research**, v. 351, p. 130–133, 1 abr. 2012.
- SANZ-APARICIO, J. et al. Crystal structure of beta-glucosidase A from *Bacillus polymyxa*: insights into the catalytic activity in family 1 glycosyl hydrolases. **Journal of Molecular Biology**, v. 275, n. 3, p. 491–502, 23 jan. 1998.

SCHRÖDER, C. et al. Characterization of a heat-active archaeal  $\beta$ -glucosidase from a hydrothermal spring metagenome. **Enzyme and Microbial Technology**, v. 57, p. 48–54, 10 abr. 2014.

SHATSKY, M.; NUSSINOV, R.; WOLFSON, H. J. A method for simultaneous alignment of multiple protein structures. **Proteins**, v. 56, n. 1, p. 143–156, 1 jul. 2004.

SHEN, M.; SALI, A. Statistical potential for assessment and prediction of protein structures. **Protein Science: A Publication of the Protein Society**, v. 15, n. 11, p. 2507–2524, nov. 2006.

SHIPKOWSKI, S.; BRECHLEY, J. E. Characterization of an Unusual Cold-Active  $\beta$ -Glucosidase Belonging to Family 3 of the Glycoside Hydrolases from the Psychrophilic Isolate *Paenibacillus* sp. Strain C7. **Applied and Environmental Microbiology**, v. 71, n. 8, p. 4225–4232, ago. 2005.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Molecular Systems Biology**, v. 7, n. 1, p. 539, 1 jan. 2011.

SILVA, M. F. M. et al. Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes. **Entertainment Computing**, v. 29, p. 31–42, 1 mar. 2019.

SILVA, M. F. M.; LEIJOTO, L. F.; NOBRE, C. N. Algorithms Analysis in Adjusting the SVM Parameters: An Approach in the Prediction of Protein Function. **Applied Artificial Intelligence**, v. 0, n. 0, p. 1–16, 12 maio 2017.

SINGHANIA, R. R. et al. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. **Bioresource Technology**, v. 127, p. 500–507, jan. 2013.

SOLOMON, B. D. Biofuels and sustainability. **Annals of the New York Academy of Sciences**, v. 1185, p. 119–134, jan. 2010.

SØRENSEN, A. et al.  $\beta$ -Glucosidases from a new *Aspergillus* species can substitute commercial  $\beta$ -glucosidases for saccharification of lignocellulosic biomass. **Canadian Journal of Microbiology**, v. 57, n. 8, p. 638–650, 1 ago. 2011.

SOUZA, F. H. M. et al. Gene cloning, expression and biochemical characterization of a glucose- and xylose-stimulated  $\beta$ -glucosidase from *Humicola insolens* RP86. **Journal of Molecular Catalysis B: Enzymatic**, v. 106, p. 1–10, ago. 2014.

STEINKELLNER, G. et al. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. **Nature Communications**, v. 5, p. 4150, 23 jun. 2014.

SWANGKEAW, J. et al. Characterization of  $\beta$ -glucosidases from *Hanseniaspora* sp. and *Pichia anomala* with potentially aroma-enhancing capabilities in juice and wine. **World Journal of Microbiology and Biotechnology**, v. 27, n. 2, p. 423–430, 12 jun. 2010.

TEUGJAS, H.; VÄLJAMÄE, P. Selecting  $\beta$ -glucosidases to support cellulases in cellulose saccharification. **Biotechnology for biofuels**, v. 6, n. 1, p. 1, 2013.

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, 30 jan. 2010.

TSUKADA, T. et al. Role of subsite +1 residues in pH dependence and catalytic activity of the glycoside hydrolase family 1 beta-glucosidase BGL1A from the basidiomycete *Phanerochaete chrysosporium*. **Biotechnology and Bioengineering**, v. 99, n. 6, p. 1295–1302, 15 abr. 2008.

UCHIMA, C. A. et al. Heterologous expression and characterization of a glucose-stimulated  $\beta$ -glucosidase from the termite *Neotermes koshunensis* in *Aspergillus oryzae*. **Applied Microbiology and Biotechnology**, v. 89, n. 6, p. 1761–1771, mar. 2011.

UCHIMA, C. A. et al. Heterologous Expression in *Pichia pastoris* and Characterization of an Endogenous Thermostable and High-Glucose-Tolerant  $\beta$ -Glucosidase from the Termite *Nasutitermes takasagoensis*. **Applied and Environmental Microbiology**, v. 78, n. 12, p. 4288–4293, 15 jun. 2012.

UCHIYAMA, T.; MIYAZAKI, K.; YAOI, K. Characterization of a Novel  $\beta$ -Glucosidase from a Compost Microbial Metagenome with Strong Transglycosylation Activity. **Journal of Biological Chemistry**, v. 288, n. 25, p. 18325–18334, 21 jun. 2013.

UCHIYAMA, T.; YAOI, K.; MIYAZAKI, K. Glucose-tolerant  $\beta$ -glucosidase retrieved from a Kusaya gravy metagenome. **Frontiers in Microbiology**, v. 6, 16 jun. 2015.

VAZ, S. Sugarcane-Biorefinery. **Advances in Biochemical Engineering/Biotechnology**, v. 166, p. 125–136, 2019.

WANG, Z.; MOULT, J. SNPs, protein structure, and disease. **Human Mutation**, v. 17, n. 4, p. 263–270, abr. 2001.

WATANABE, T. et al. Purification and properties of *Aspergillus niger* beta-glucosidase. **European journal of biochemistry / FEBS**, v. 209, n. 2, p. 651–659, 15 out. 1992.

WEBB, B.; SALI, A. Comparative Protein Structure Modeling Using MODELLER. **Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]**, v. 47, p. 5.6.1-32, 2014.

WIERZBICKA-WOŚ, A. et al. Cloning and characterization of a novel cold-active glycoside hydrolase family 1 enzyme with  $\beta$ -glucosidase,  $\beta$ -fucosidase and  $\beta$ -galactosidase activities. **BMC biotechnology**, v. 13, p. 22, 15 mar. 2013.

WILSON, D. S.; KEEFE, A. D. Random mutagenesis by PCR. **Current Protocols in Molecular Biology**, v. Chapter 8, p. Unit8.3, maio 2001.

WITHERS, S. G. et al. Mechanistic consequences of mutation of the active site nucleophile Glu 358 in *Agrobacterium* beta-glucosidase. **Biochemistry**, v. 31, n. 41, p. 9979–9985, 20 out. 1992.

WOODWARD, J.; WISEMAN, A. Fungal and other  $\beta$ -d-glucosidases — Their properties and applications. **Enzyme and Microbial Technology**, v. 4, n. 2, p. 73–79, 1 mar. 1982.



XU, H. et al. Characterization of a Glucose-, Xylose-, Sucrose-, and d-Galactose-Stimulated  $\beta$ -Glucosidase from the Alkalophilic Bacterium *Bacillus halodurans* C-125. **Current Microbiology**, v. 62, n. 3, p. 833–839, 1 mar. 2011.

YANG, F. et al. Overexpression and characterization of a glucose-tolerant  $\beta$ -glucosidase from *T. aotearoense* with high specific activity for cellobiose. **Applied Microbiology and Biotechnology**, v. 99, n. 21, p. 8903–8915, nov. 2015a.

YANG, Y. et al. A mechanism of glucose tolerance and stimulation of GH1  $\beta$ -glucosidases. **Scientific Reports**, v. 5, p. 17296, 25 nov. 2015b.

YOON, J.-J.; KIM, K.-Y.; CHA, C.-J. Purification and characterization of thermostable beta-glucosidase from the brown-rot basidiomycete *Fomitopsis palustris* grown on microcrystalline cellulose. **Journal of Microbiology (Seoul, Korea)**, v. 46, n. 1, p. 51–55, fev. 2008.

ZHAO, L. et al. Enzymatic properties of *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase fused to *Clostridium cellulovorans* cellulose binding domain and its application in hydrolysis of microcrystalline cellulose. **BMC biotechnology**, v. 13, n. 1, p. 1, 2013.

ZOUHAR, J. et al. Insights into the functional architecture of the catalytic center of a maize beta-glucosidase Zm-p60.1. **Plant Physiology**, v. 127, n. 3, p. 973–985, nov. 2001.

## 9. Apêndices

### Apêndice 1. Prêmio: ISCB *Wikipedia Competition* (2018)

### Apêndice 2. Prêmio: *Best Poster Award X-Meeting 2016* na categoria “*Proteins and Proteomics*”

- Título: “*Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production*”

### Apêndice 3. Prêmio: *Best Poster Award X-Meeting 2016* na categoria “*Software Development and Databases*”

- Título: “*An approach for constructing a database of manually curated contacts in proteins*”

### Apêndice 4. Prêmio: *Best Paper Award - "Oral presentation", 2nd Brazilian Student Council Symposium*

- Título: “*A new method based on structural signatures to propose mutations for enzymes  $\beta$ -glucosidase used in biofuel production*”

### Apêndice 5. Currículo Lattes

### Apêndice 6. Artigo 1

- Título: “*Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review*”

### Apêndice 7. Artigo 2

- Título: “*A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV)*”

### Apêndice 8. Artigo 3

- Título: “*Molecular Dynamics Gives New Insights into the Glucose Tolerance and Inhibition Mechanisms on  $\beta$ -Glucosidases*”

### Apêndice 9. Estratégias para produção de $\beta$ -glicosidases glicose-tolerantes

### Apêndice 10. Número de contatos entre resíduos e celobiose em $\beta$ -glicosidases GH1 glicose-tolerantes

## Apêndice 1. Prêmio: *ISCB Wikipedia Competition* (2018)



**2018**



# Wikipedia Competition

**1<sup>st</sup> Place**

Diego César Batista Mariano, Renato Augusto Corrêa dos Santos,  
Cristiane Hayumi Taniguti, Fernando Henrique Correr, Jenifer Camila  
Godoy dos Santos and Ana Letycia Basso Garcia

*For Biostatistics*

A handwritten signature in blue ink that reads 'Thomas Lengauer'.

---

Thomas Lengauer, PhD  
Past-President, ISCB

A handwritten signature in blue ink that reads 'Diane E. Kovats'.

---

Diane E. Kovats, CMP, CAE  
Executive Director, ISCB

**Apêndice 2. Prêmio: *Best Poster Award X-Meeting 2016* na categoria “*Proteins and Proteomics*”**



Belo Horizonte | November 16th to 18th 12<sup>th</sup> International Conference of the AB3C

## Best Poster Award

This certifies that the Best Poster Award is given to the work intitled  
*"Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production"*, authored by **Diego Mariano, Thiago Silva Correia, José Renato Pereira de Moura Barroso and Raquel Melo Minardi**, at the poster session (Proteins and Proteomics) of the X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Belo Horizonte - Brazil between November 16 and 18 of 2016.

*Nicole Scherer*  
Nicole Scherer  
Poster Session Chair

*Mainá Bitar*  
Mainá Bitar  
Poster Session Co-Chair

*Glória Franco*  
Glória Franco  
AB3C President



**Apêndice 3. Prêmio: Best Poster Award X-Meeting 2016 na categoria “Software Development and Databases”**



**Apêndice 4. Prêmio: Best Paper Award - "Oral presentation",  
2nd Brazilian Student Council Symposium.**

International Society for Computational Biology

**2nd Brazilian Student Council Symposium**

4 October 2017, Hotel Colina Verde, São Pedro, São Paulo, Brazil




**Best Paper Award**

**Diego Mariano and Raquel Cardoso de Melo-Minardi**

for the paper entitled:

*A new method based on structural signatures to propose mutations for enzymes  $\beta$ -glucosidase used in biofuel production*

  
Nilson Da Rocha Coimbra  
President

  
Juliana Assis  
Secretary



## **Apêndice 5. Currículo *Lattes***



## Diego César Batista Mariano

Endereço para acessar este CV: <http://lattes.cnpq.br/2878991216624166>

ID Lattes: **2878991216624166**

Última atualização do currículo em 23/12/2019

Mestre e Doutor em Bioinformática pela Universidade Federal de Minas Gerais. Atualmente realiza estágio pós-doutoral no Departamento de Ciência da Computação da UFMG. Tem experiência nas áreas de desenvolvimento de sistemas Web, Data Science e visualização de dados, Bioinformática e machine learning. Tem conhecimentos nas linguagens: PHP, JavaScript, Python, R, Perl, HTML, CSS e SQL. **(Texto informado pelo autor)**

### Identificação

**Nome**

Diego César Batista Mariano

**Nome em citações bibliográficas**

MARIANO, D. C. B.;BATISTA MARIANO, D. C.;MARIANO, DIEGO CÉSAR BATISTA;MARIANO, DIEGO;MARIANO, DIEGO CB;MARIANO, DIEGO C. B.;MARIANO, D.C.B.;MARIANO, DIEGO C.B.

**Lattes ID**

<http://lattes.cnpq.br/2878991216624166>

**Orcid ID**

<https://orcid.org/0000-0002-5899-2052>

### Endereço

**Endereço Profissional**

Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.  
Laboratório de Bioinformática e Sistemas (DCC/UFMG - SALA 4340). Avenida Presidente Antônio Carlos, 6627 - Instituto de Ciências Exatas Pampulha  
31270010 - Belo Horizonte, MG - Brasil  
Telefone: (31) 34095810  
URL da Homepage: <http://www.lbs.dcc.ufmg.br/>

### Formação acadêmica/titulação

**2015 - 2019**

Doutorado em Bioinformática (Conceito CAPES 7).  
Universidade Federal de Minas Gerais, UFMG, Brasil.  
Título: Uso de assinaturas estruturais para proposta de mutações em enzimas  $\beta$ -glicosidase usadas na produção de biocombustíveis, Ano de obtenção: 2019.  
Orientador: Raquel Cardoso de Melo Minardi.  
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.

**2013 - 2015**

Palavras-chave: Bioinformática; Visualização de dados biológicos; biocombustíveis.  
Mestrado em Bioinformática (Conceito CAPES 7).  
Universidade Federal de Minas Gerais, UFMG, Brasil.  
Título: SIMBA: uma ferramenta Web para gerenciamento de montagens de genomas bacterianos, Ano de Obtenção: 2015.  
Orientador: Vasco Ariston de Carvalho Azevedo.  
Coorientador: Rommel Thiago Jucá Ramos.  
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.

**2009 - 2012**

Palavras-chave: Bioinformática; Genômica; Sistemas Web; Montagem.  
Graduação em Sistemas de Informação.  
Faculdade Anhanguera de Belo Horizonte, FABRAI, Brasil.  
Título: Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos.  
Orientador: Sandro Renato Dias.



<b>2010 - 2010</b>	Bolsista do(a): Programa Universidade para Todos, PROUNI, Brasil. Curso técnico/profissionalizante em Aprendizagem Industrial Gráfica.
<b>2008 - 2010</b>	SENAI - Departamento Regional de Minas Gerais, SENAI/DR/MG, Brasil. Curso técnico/profissionalizante em Redes computacionais. Serviço Nacional de Aprendizagem Comercial - SENAC Minas, SENAC/MG, Brasil. Bolsista do(a): Secretaria de Estado de Educação de Minas Gerais, SEEMG, Brasil.

## Pós-doutorado

<b>2019</b>	Pós-Doutorado. Universidade Federal de Minas Gerais, UFMG, Brasil. Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil. Grande área: Ciências Exatas e da Terra
-------------	---

## Formação Complementar

<b>2019 - 2019</b>	Data Science: Productivity Tools. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2019 - 2019</b>	Data Science: Wrangling. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2019 - 2019</b>	Data Science: Inference and Modeling. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2019 - 2019</b>	Data Science: Machine Learning. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2019 - 2019</b>	Data Science: Probability. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2019 - 2019</b>	Data Science: Linear Regression. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2018 - 2018</b>	Data Science: Visualization. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2018 - 2018</b>	Data Science: R Basics. (Carga horária: 16h). Harvard University (HarvardX), HARVARDX, Estados Unidos.
<b>2018 - 2018</b>	Intro to Python for Data Science. (Carga horária: 4h). DataCamp, DATACAMP, Estados Unidos.
<b>2018 - 2018</b>	Using Python for Research. (Carga horária: 4h). DataCamp, DATACAMP, Estados Unidos.
<b>2017 - 2017</b>	Extensão universitária em Laboratório de Criação de Materiais Didáticos para EaD. (Carga horária: 60h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2017 - 2017</b>	Módulo teórico do Curso de Verão de Engenharia de Máquinas Biológicas. (Carga horária: 12h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2017 - 2017</b>	Introdução a Modelagem Matemática. (Carga horária: 15h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2014 - 2017</b>	Extensão universitária em Língua Inglesa. (Carga horária: 420h). Centro Acadêmico de Ciências Sociais (FAFICH, UFMG), CACS, Brasil.
<b>2016 - 2016</b>	Extensão universitária em Língua Francesa. (Carga horária: 60h). Centro Acadêmico de Ciências Sociais (FAFICH, UFMG), CACS, Brasil.
<b>2015 - 2015</b>	Extensão universitária em Modelagem Molecular de Sistemas Biológicos. (Carga horária: 16h). Universidade Federal do Rio Grande, FURG, Brasil.
<b>2014 - 2014</b>	Extensão universitária em PATRIC: estudo de sistemas patogênicos. (Carga horária: 15h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2014 - 2014</b>	Extensão universitária em Anotação avançada de sequências com EGene2. (Carga horária: 15h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2013 - 2013</b>	Extensão universitária em Conversación y Redacción. (Carga horária: 15h). Universidad de Salamanca, USAL, Espanha.
<b>2013 - 2013</b>	Extensão universitária em Lengua Española. (Carga horária: 30h). Universidad de Salamanca, USAL, Espanha.
<b>2013 - 2013</b>	Bioinformática Estrutural e Análises de Proteoma. (Carga horária: 90h). Universidade Federal de Minas Gerais, UFMG, Brasil.
<b>2012 - 2012</b>	Uso do Google para visualizar dados geográficos. (Carga horária: 4h). Universidade Federal de Lavras, UFLA, Brasil.
<b>2012 - 2012</b>	Cloud computing. Dell Training Centre, DELL, Brasil.
<b>2012 - 2012</b>	

	Dell layered security solutions. Dell Training Centre, DELL, Brasil.
<b>2008 - 2008</b>	Extensão universitária em Infra-estrutura de TI (S2B Microsoft). (Carga horária: 84h). Pontifícia Universidade Católica de Minas Gerais, PUC Minas, Brasil.
<b>2005 - 2006</b>	Hardware - Montagem e Manutenção de computadores. (Carga horária: 72h). Microlins, MICROLINS, Brasil.

## Atuação Profissional

---

### Universidade Federal de Minas Gerais, UFMG, Brasil.

#### Vínculo institucional

**2015 - Atual**

Vínculo: Bolsista, Enquadramento Funcional: Estudante de doutorado, Carga horária: 40, Regime: Dedicção exclusiva.

#### Vínculo institucional

**2013 - 2015**

Vínculo: Bolsista, Enquadramento Funcional: Estudante de mestrado, Carga horária: 40, Regime: Dedicção exclusiva.

#### Atividades

**8/2018 - Atual**

Ensino, Ciências da Computação, Nível: Pós-Graduação  
Disciplinas ministradas

**03/2018 - 07/2018**

Ambientes para Computação (estágio à docência)

Ensino, Bioinformática, Nível: Pós-Graduação

Disciplinas ministradas

**8/2013 - 11/2013**

Introdução à Programação Python Aplicada à Bioinformática (estágio à docência)

Ensino, Bioinformática, Nível: Pós-Graduação

Disciplinas ministradas

Introdução à Linguagem de Programação Perl (estágio à docência)

### Evolua Comunicação, EVOLUA, Brasil.

#### Vínculo institucional

**2012 - 2013**

Vínculo: Colaborador, Enquadramento Funcional: Analista desenvolvedor, Carga horária: 40, Regime: Dedicção exclusiva.

### Serviço Nacional de Aprendizagem Comercial - SENAC Minas, SENAC/MG, Brasil.

#### Vínculo institucional

**2011 - 2012**

#### Outras informações

Vínculo: Colaborador, Enquadramento Funcional: Professor, Carga horária: 15

Professor do curso de Hardware - Manutenção de computadores

### Prefeitura Municipal de Lagoa Santa, PMLS, Brasil.

#### Vínculo institucional

**2011 - 2012**

Vínculo: , Enquadramento Funcional: Técnico em informática, Carga horária: 30

### Faculdade Anhanguera de Belo Horizonte, FABRAI, Brasil.

#### Vínculo institucional

**2011 - 2012**

Vínculo: Bolsista, Enquadramento Funcional: Iniciação científica, Carga horária: 6

### Universidad de Salamanca, USAL, Espanha.

#### Vínculo institucional

**2013 - 2013**

Vínculo: Estudante de Intercâmbio, Enquadramento Funcional: Estudante de Intercâmbio, Regime: Dedicção exclusiva.

## Projetos de pesquisa

---

**2016 - Atual**

ThAMES: Modelos, algoritmos e visualizações para o estudo de redes biológicas multivariadas dinâmicas

Projeto certificado pelo(a) coordenador(a) Raquel Cardoso de Melo Minardi em 26/02/2019.

Descrição: Bioinformática é uma área essencialmente interdisciplinar e agregadora de conhecimentos da Ciência da Computação, e de diversas áreas na solução de problemas biológicos. Sob a perspectiva de Computação, existem ainda grandes demandas de

novos modelos que possibilitem a geração de conhecimento útil para catalizar descobertas e invenções com potencial biotecnológico. Nossa atuação em Bioinformática consiste em trabalhar em diversos cenários biológicos que tratam de problemas em aberto na biologia ou com especial interesse biotecnológico. No presente projeto, temos um grande problema a ser estudado e uma interessante aplicação em um cenário de alta relevância biotecnológica: a engenharia de enzimas do tipo Beta-Glicosidases envolvidas na produção de biocombustíveis de segunda geração. Alguns problemas tem impedido a produção custo-eficiente desses biocombustíveis. Dentre eles, destaca-se a inibição da enzima Beta-Glicosidase por Glicose. Do ponto de vista de Bioinformática, esse projeto consiste então no estudo do impacto de mutações em proteínas para sua engenharia visando resolver o problema acima descrito. É importante destacar que a construção de enzimas mutantes com a eficiência necessária para os requisitos levantados anteriormente é um problema de grande complexidade, uma vez que uma proteína tem centenas de aminoácidos e para cada aminoácido há dezenove mutações possíveis. Dessa forma, fica evidente a necessidade de novos modelos, algoritmos e ferramentas que sejam capazes propor soluções para o problema descrito. A abordagem que aqui propomos consiste na modelagem dessas proteínas como grafos nos quais um nó pode ser um aminoácido (centenas) ou um átomo (milhares) e as arestas descrevem interações químicas entre eles. Essa seria uma rede multivariada e dinâmica, a qual pretendemos analisar através de modelos, algoritmos e visualizações que serão propostas nesse projeto e que podem ser amplamente utilizadas em outros cenários de aplicação.

Situação: Em andamento; Natureza: Pesquisa.

Alunos envolvidos: Graduação: (2) / Especialização: (0) / Mestrado acadêmico: (0) / Mestrado profissional: (7) / Doutorado: (6) .

Integrantes: Diego César Batista Mariano - Integrante / Raquel Cardoso de Melo Minardi - Coordenador.

**2014 - 2015**

Biologia Computacional-REDE DE COOPERAÇÃO ACADÊMICA PARA O ESTUDO E DESENVOLVIMENTO DE FERRAMENTAS PARA A GENÔMICA ESTRUTURAL E FUNCIONAL

Descrição: Descrição: Fortalecer e ampliar o intercâmbio acadêmico entre os programas inter-unidades de Pós-Graduação em Bioinformática da UFMG (CAPES 6) e da USP (5), o de Biotecnologia da UFPA (CAPES 5) e o de Bioinformática da UFPR (CAPES 3) com a criação de uma rede voltada a aumentar a formação de recursos humanos em Biologia Computacional, em resposta à presente chamada...

Situação: Concluído; Natureza: Pesquisa.

Alunos envolvidos: Mestrado acadêmico: (7) Doutorado: (6) .

Integrantes: Diego César Batista Mariano - Integrante / Edgar Lacerda de Aguiar - Integrante / SIOMAR C. SOARES - Integrante / VASCO A. C. AZEVEDO - Coordenador / Vinícius A. C. Abreu - Integrante / Carlos A. A. Diniz - Integrante / SS Hassan - Integrante / Alberto Fernandes de Oliveira Junior - Integrante / TIWARI, S. - Integrante / Edson Luiz Folador - Integrante / Jose Miguel Ortega - Integrante / Lucas Bleicher - Integrante / Sintia Silva de Almeida - Integrante / Rafaela Salgado Ferreira - Integrante / Liza Figueiredo Felicori Vilela - Integrante / Leticia de Oliveria Castro - Integrante / Lucas Amorim Gonçalves - Integrante / Raoni Almeida de Souza - Integrante / Benjamin Viart - Integrante / Camila Franco Batista de Oliveira - Integrante / Andrea Silveira Vilela - Integrante.

**2014 - Atual**

Bioinformática Estrutural de Proteínas: modelos, algoritmos e aplicações biotecnológicas Projeto certificado pelo(a) coordenador(a) Raquel Cardoso de Melo Minardi em 29/06/2016.

Descrição: Este projeto propõe-se enfrentar dois desafios biotecnológicos de grande relevância nacional. O primeiro envolve a ricina, uma potente fitotoxina encontrada na mamoneira. Co-produtos da produção do óleo de mamona podem conter quantidades letais de ricina. Além do risco de intoxicação animal e humana do bagaço, isso tem dificultado sua reutilização e reciclagem, em especial no semiárido nordestino. A ricina preocupa também pelo seu potencial uso como arma química. Portanto, há forte apelo para se encontrar meios efetivos de neutralizar, inibir e/ou detectar a ricina. O segundo envolve a modelagem de celulasas multifuncionais, capazes de degradar eficientemente a lignocelulose, composto base para a produção de biocombustíveis de segunda geração, os quais têm por princípio a utilização de subprodutos de plantas que não podem ser usados na alimentação humana. Ambos desafios exigem uma abordagem multidisciplinar sinérgica entre a modelagem teórica in silico e experimental in vitro.

Objetivos: 1) Prever in silico e validar in vitro ligantes de ricina 2) Simular in silico e caracterizar in vitro efeitos causados por intervenções estratégicas em beta-glicosidases modificadas. 3) Projetar, implementar e validar modelos, algoritmos e ferramentas de Quimioinformática e Bioinformática Estrutural de Proteínas que permitam dar suporte aos objetivos biotecnológicos previamente descritos. Espera-se que este projeto possa revelar novas plataformas químicas para o desenvolvimento de inibidores, kits de detecção, biossensores, neutralizadores (substratos suicidas) e outros produtos de

inovação biotecnológica baseados na ricina, com potencial para alavancar o agronegócio da mamona e a indústria nacional de fármacos. Espera-se também a modelagem de celulases multifuncionais, capazes especialmente de orientar a manipulação genética de algas produtoras de  $\beta$ -glicosidases, com características catalíticas eficientes e de tolerância à inibição por glicose e celobiose, permitindo seu uso industrial na produção de etanol de segunda geração. Espera-se, em fim, que esses desafios possam inspirar novas metodologias matemáticas, computacionais, químicas e bioquímicas integradas, de modo a gerar artefatos biotecnológicos inovadores, bem como contribuir para a formação de profissionais com perfil multidisciplinar altamente qualificados.

Situação: Em andamento; Natureza: Pesquisa.

Alunos envolvidos: Graduação: (4) / Mestrado acadêmico: (6) / Doutorado: (6) .

Integrantes: Diego César Batista Mariano - Integrante / Raquel Cardoso de Melo Minardi - Coordenador.

**2011 - 2012**

Comparando o desempenho de Bancos de Dados NoSQL e Relacionais manipulando dados biológicos

Projeto certificado pelo(a) coordenador(a) Sandro Renato Dias em 19/08/2013.

Descrição: Apresentar o uso de bancos de dados relacionais e não relacionais do tipo NoSQL para controlar e manipular grandes quantidades de dados, comparando o desempenho de um SGBD (Sistema de Gerenciamento de Bancos de Dados) não relacional, o MongoDB, com um SGBD relacional, o MySQL. Tem como objetivo verificar as vantagens e as desvantagens do uso de bancos de dados não relacionais em comparação aos bancos de dados relacionais, utilizando a base de dados biológicos PDB (Protein Data Bank) para testes de leitura de arquivos em disco e gravação de seu conteúdo nos SGBDs.

Situação: Concluído; Natureza: Pesquisa.

Alunos envolvidos: Graduação: (2) / Doutorado: (1) .

Integrantes: Diego César Batista Mariano - Coordenador / Sandro Renato Dias - Integrante / Edgar Lacerda de Aguiar - Integrante.

Número de produções C, T & A: 3

## Revisor de periódico

**2019 - Atual**

Periódico: Bioinformatics

## Áreas de atuação

1. Grande área: Ciências Biológicas / Área: Genética / Subárea: Bioinformática.
2. Grande área: Ciências Exatas e da Terra / Área: Ciência da Computação / Subárea: Programação para Web.

## Idiomas

<b>Espanhol</b>	Compreende Razoavelmente, Fala Razoavelmente, Lê Razoavelmente, Escreve Razoavelmente.
<b>Inglês</b>	Compreende Razoavelmente, Fala Razoavelmente, Lê Bem, Escreve Bem.
<b>Francês</b>	, Lê Pouco.
<b>Português</b>	Compreende Bem, Fala Bem, Lê Bem, Escreve Bem.

## Prêmios e títulos

<b>2018</b>	ISCB Wikipedia Competition Winner, ISCB.
<b>2017</b>	Best Paper Award - "Oral presentation", 2nd Brazilian Student Council Symposium.
<b>2016</b>	Best Poster Award - "Proteins and Proteomics", X-meeting 2016 - 12th International Conference of the AB3C.
<b>2016</b>	Best Poster Award - "Software Development and Databases", X-meeting 2016 - 12th International Conference of the AB3C.
<b>2014</b>	Best Poster Award - "People's choice", ISCB - Latin America X-Meeting on Bioinformatics with BSB and SoBio 2014.
<b>2013</b>	Prêmio Anhanguera de Mérito Científico e Acadêmico - Categoria: Programa de Iniciação Científica - Ciências Exatas, da Terra e Engenharias, Anhanguera Educacional.
<b>2006</b>	Menção honrosa - "Olimpíada Brasileira de Matemática", Instituto Nacional de Matemática Pura e Aplicada - IMPA.
<b>2004</b>	Medalha de bronze - "Olimpíada Brasileira de Astronomia", Agência Espacial Brasileira.

## Produção bibliográfica

### Artigos completos publicados em periódicos

Ordenar por

Ordem Cronológica ▼

1.  **MARIANO, DIEGO**; SANTOS, LUCIANNA ; MACHADO, KARINA ; WERHLI, ADRIANO ; DE LIMA, LEONARDO ; DE MELO-MINARDI, RAQUEL . A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES **JCR**, v. 20, p. 333, 2019.
2. **MARIANO, DIEGO**; MARTINS, PEDRO ; HELENE SANTOS, LUCIANNA ; DE MELO-MINARDI, RAQUEL CARDOSO . Introducing Programming Skills for Life Science Students. BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION **JCR**, v. 1, p. 21230, 2019.
3. COSTA, LEON SULFIERRY CORRÊA ; **MARIANO, DIEGO CÉSAR BATISTA** ; ROCHA, RAFAEL EDUARDO OLIVEIRA ; KRAML, JOHANNES ; SILVEIRA, CARLOS HENRIQUE DA ; LIEDL, KLAUS ROMAN ; DE MELO-MINARDI, RAQUEL CARDOSO ; LIMA, LEONARDO HENRIQUE FRANCA DE . Molecular Dynamics Gives New Insights into the Glucose Tolerance and Inhibition Mechanisms on  $\beta$ -Glucosidases. MOLECULES **JCR**, v. 24, p. 3215, 2019.
4. GOMIDE, ANNE CYBELLE PINTO ; IBRAIM, IZABELA COIMBRA ; ALVES, JORIANNE T.C. ; DE SÁ, PABLO GOMES ; DE OLIVEIRA SILVA, YURI RAFAEL ; SANTANA, MARIANA PASSOS ; SILVA, WANDERSON MARQUES ; FOLADOR, EDSON LUIZ ; **MARIANO, DIEGO C.B.** ; DE PAULA CASTRO, THIAGO LUIZ ; BARBOSA, SILVANIRA ; DORELLA, FERNANDA ALVES ; CARVALHO, ALEX F. ; PEREIRA, FELIPE L. ; LEAL, CARLOS A.G. ; FIGUEIREDO, HENRIQUE C.P. ; AZEVEDO, VASCO ; SILVA, ARTUR ; FOLADOR, ADRIANA RIBEIRO CARNEIRO . Transcriptome analysis of *Corynebacterium pseudotuberculosis* biovar Equi in two conditions of the environmental stress. GENE **JCR**, v. 8, p. 1, 2018.
5. SILVA, MARCOS F.M. ; MARTINS, PEDRO M. ; **MARIANO, DIEGO C.B.** ; SANTOSA, LUCIANNA HELENE ; PASTORINI, ISABELA ; PANTUZA, NAIARA ; NOBRE, CRISTIANE N. ; DE MELO-MINARDI, RAQUEL C. . Proteingo: motivation, user experience, and learning of molecular interactions in biological complexes. ENTERTAINMENT COMPUTING, v. 1, p. 001, 2018.
6. HURTADO, RAQUEL ENMA ; ABURJAILE, FLAVIA ; **MARIANO, DIEGO** ; CANÁRIO, MARCUS VINICIUS ; BENEVIDES, LEANDRO ; FERNANDEZ, DANIEL ANTONIO ; ALLASI, NATALY OLIVIA ; RIMAC, ROCIO ; JUSCAMAYTA, JULIO EDUARDO ; MAXIMILIANO, JORGE ENRIQUE ; ROSADIO, RAUL HECTOR ; AZEVEDO, VASCO ; MATURRANO, LENIN . Draft Genome Sequence of a Virulent Strain of *Pasteurella Multocida* Isolated From Alpaca. JOURNAL OF GENOMICS, v. 5, p. 68-70, 2017.
7. **MARIANO, D.C.B.**; LEITE, C. ; SANTOS, L.H.S. ; MARINS, L.F. ; MACHADO, K.S. ; WERHLI, A.V. ; LIMA, L.H.F. ; DE MELO-MINARDI, R.C. . Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review. GENETICS AND MOLECULAR RESEARCH **JCR**, v. 16, p. 1-19, 2017.  
**Citações:** **SCOPUS** 1
8. GUIMARÃES, LUIS C. ; VIANA, MARCUS V. C. ; BENEVIDES, LEANDRO J. ; **MARIANO, DIEGO C. B.** ; VERAS, ADOONEY A. O. ; SÁ, PABLO H. C. ; ROCHA, FLÁVIA S. ; VILAS BOAS, PRISCILLA C. B. ; SOARES, SIOMAR C. ; BARBOSA, MARIA S. ; GUIISO, NICOLE ; BADELL, EDGAR ; CARNEIRO, ADRIANA R. ; AZEVEDO, VASCO ; RAMOS, ROMMEL T. J. ; SILVA, ARTUR . Draft Genome Sequence of Toxigenic *Corynebacterium ulcerans* Strain 04-7514, Isolated from a Dog in France. Genome Announcements, v. 4, p. e00172-16, 2016.
9. GUIMARÃES, LUIS C. ; VIANA, MARCUS V. C. ; BENEVIDES, LEANDRO J. ; **MARIANO, DIEGO C. B.** ; VERAS, ADOONEY A. O. ; SÁ, PABLO H. C. ; ROCHA, FLÁVIA S. ; VILAS BOAS, PRISCILLA C. B. ; SOARES, SIOMAR C. ; BARBOSA, MARIA S. ; GUIISO, NICOLE ; BADELL, EDGAR ; CARNEIRO, ADRIANA R. ; AZEVEDO, VASCO ; RAMOS, ROMMEL T. J. ; SILVA, ARTUR . Draft Genome Sequence of *Corynebacterium ulcerans* Strain 04-3911, Isolated from Humans. Genome Announcements, v. 4, p. e00171-16, 2016.
10. ALMEIDA, SINTIA ; TIWARI, SANDEEP ; **MARIANO, DIEGO** ; SOUZA, FLÁVIA ; JAMAL, SYED BABAR ; COIMBRA, NILSON ; RAITTZ, ROBERTO TADEU ; DORELLA, FERNANDA ALVES ; CARVALHO, ALEX FIORINE DE ; PEREIRA, FELIPE LUIZ ; SOARES, SIOMAR DE CASTRO ; LEAL, CARLOS AUGUSTO GOMES ; BARH, DEBMALYA ; GHOSH, PREETAM ; FIGUEIREDO, HENRIQUE ; MOURA-COSTA, LÍLIA FERREIRA ; PORTELA, RICARDO WAGNER ; MEYER, ROBERTO ; SILVA, ARTUR ; AZEVEDO, VASCO . The genome anatomy of *Corynebacterium pseudotuberculosis* VD57 a highly virulent strain causing Caseous lymphadenitis. Standards in Genomic Sciences **JCR**, v. 11, p. 29, 2016.  
**Citações:** **WEB OF SCIENCE** 2 | **SCOPUS** 3
11.  **MARIANO, DIEGO CÉSAR BATISTA**; SOUSA, THIAGO DE JESUS ; PEREIRA, FELIPE LUIZ ; ABURJAILE, FLÁVIA ; BARH, DEBMALYA ; ROCHA, FLÁVIA ; PINTO, ANNE CYBELLE ; HASSAN, SYED SHAH ; SARAIVA, TESSÁLIA DINIZ LUERCE ; DORELLA, FERNANDA ALVES ; DE CARVALHO, ALEX FIORINI ; LEAL, CARLOS AUGUSTO GOMES ; FIGUEIREDO, HENRIQUE CÉSAR PEREIRA ; SILVA, ARTUR ; RAMOS, ROMMEL THIAGO JUCÁ ; AZEVEDO, VASCO ARISTON CARVALHO . Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of *Corynebacterium pseudotuberculosis* strain 1002. BMC Genomics **JCR**, v. 17, p. 315, 2016.  
**Citações:** **WEB OF SCIENCE** 2 | **SCOPUS** 2
- 12.

DE AGUIAR, EDGAR LACERDA ; **MARIANO, D. C. B.** ; VIANA, MARCUS VINÍCIUS CANÁRIO ; BENEVIDES, LEANDRO DE JESUS ; DE SOUZA ROCHA, FLÁVIA ; DE CASTRO OLIVEIRA, LETÍCIA ; PEREIRA, F. L. ; DORELLA, FERNANDA ALVES ; LEAL, CARLOS AUGUSTO GOMES ; DE CARVALHO, ALEX FIORINI ; SANTOS, GABRIELA SILVA ; MATTOS-GUARALDI, ANA LUIZA ; NAGAO, PRESCILLA EMY ; DE CASTRO SOARES, SIOMAR ; HASSAN, SYED SHAH ; PINTO, ANNE CYBELE ; FIGUEIREDO, HENRIQUE CÉSAR PEREIRA ; AZEVEDO, VASCO . Complete genome sequence of *Streptococcus agalactiae* strain GBS85147 serotype of type Ia isolated from human oropharynx. *Standards in Genomic Sciences JCR*, v. 11, p. 1, 2016.

13. GUIMARÃES, LUIS C. ; VIANA, MARCUS V. C. ; BENEVIDES, LEANDRO J. ; **MARIANO, DIEGO C. B.** ; VERAS, ADONNEY A. O. ; SÁ, PABLO H. C. ; ROCHA, FLÁVIA S. ; VILAS BOAS, PRISCILLA C. B. ; SOARES, SIOMAR C. ; BARBOSA, MARIA S. ; GUIISO, NICOLE ; BADELL, EDGAR ; AZEVEDO, VASCO ; RAMOS, ROMMEL T. J. ; SILVA, ARTUR . Draft Genome Sequence of Toxigenic Strain 03-8664 Isolated from a Human Throat. *Genome Announcements*, v. 4, p. e00719-16, 2016.
14. TIWARI, SANDEEP ; JAMAL, SYED BABAR ; OLIVEIRA, LETICIA CASTRO ; CLERMONT, DOMINIQUE ; BIZET, CHANTAL ; **MARIANO, DIEGO** ; DE CARVALHO, PAULO VINICIUS SANCHES DALTRO ; SOUZA, FLAVIA ; PEREIRA, FELIPE LUIZ ; DE CASTRO SOARES, SIOMAR ; GUIMARÃES, LUIS C. ; DORELLA, FERNANDA ; CARVALHO, ALEX ; LEAL, CARLOS ; BARH, DEBMALYA ; FIGUEIREDO, HENRIQUE ; HASSAN, SYED SHAH ; AZEVEDO, VASCO ; SILVA, ARTUR . Whole-Genome Sequence of Strain CIP 106629 Isolated from a Dog with Bilateral Otitis from the United Kingdom. *Genome Announcements*, v. 4, p. e00683-16, 2016.
15. ALMEIDA, SINTIA ; LOUREIRO, DAN ; PORTELA, RICARDO W. ; **MARIANO, DIEGO C. B.** ; SOUSA, THIAGO J. ; PEREIRA, FELIPE L. ; DORELLA, FERNANDA A. ; CARVALHO, ALEX F. ; MOURA-COSTA, LILIA F. ; LEAL, CARLOS A. G. ; FIGUEIREDO, HENRIQUE C. ; MEYER, ROBERTO ; AZEVEDO, VASCO . Complete Genome Sequence of the Attenuated Strain T1. *Genome Announcements*, v. 4, p. e00947-16, 2016.
16. ★ **MARIANO, DIEGO C. B.** ; PEREIRA, FELIPE L. ; AGUIAR, EDGAR L. ; OLIVEIRA, LETÍCIA C. ; BENEVIDES, LEANDRO ; GUIMARÃES, LUÍS C. ; FOLADOR, EDSON L. ; SOUSA, THIAGO J. ; GHOSH, PREETAM ; BARH, DEBMALYA ; FIGUEIREDO, HENRIQUE C. P. ; SILVA, ARTUR ; RAMOS, ROMMEL T. J. ; AZEVEDO, VASCO A. C. . SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. *BMC Bioinformatics JCR*, v. 17, p. 65-72, 2016.  
**Citações:** **WEB OF SCIENCE** 1 | **SCOPUS** 1
17. BENEVIDES, LEANDRO DE JESUS ; VIANA, MARCUS VINICIUS CANÁRIO ; **MARIANO, DIEGO CÉSAR BATISTA** ; ROCHA, FLÁVIA DE SOUZA ; BAGANO, PRISCILLA CAROLINNE ; FOLADOR, EDSON LUIZ ; PEREIRA, FELIPE LUIZ ; DORELLA, FERNANDA ALVES ; LEAL, CARLOS AUGUSTO GOMES ; CARVALHO, ALEX FIORINI ; SOARES, SIOMAR DE CASTRO ; CARNEIRO, ADRIANA ; RAMOS, ROMMEL ; BADELL-OCANDO, EDGAR ; GUIISO, NICOLE ; SILVA, ARTUR ; FIGUEIREDO, HENRIQUE ; AZEVEDO, VASCO ; GUIMARÃES, LUIS CARLOS . Genome Sequence of *Corynebacterium ulcerans* Strain FRC11. *Genome Announcements*, v. 3, p. e00112-15, 2015.  
**Citações:** **SCOPUS** 1
18. **MARIANO, DIEGO CB** ; PEREIRA, FELIPE L. ; GHOSH, PREETAM ; BARH, DEBMALYA ; FIGUEIREDO, HENRIQUE CP ; SILVA, ARTUR ; RAMOS, ROMMEL TJ ; AZEVEDO, VASCO AC . MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. *BIOINFORMATION (ONLINE) (CHENNAI)*, v. 11, p. 276-279, 2015.  
**Citações:** **SCOPUS** 7
19. SOUSA, THIAGO JESUS ; **MARIANO, DIEGO** ; PARISE, DOGLAS ; PARISE, MARIANA ; VIANA, MARCUS VINICIUS CANÁRIO ; GUIMARÃES, LUIS CARLOS ; BENEVIDES, LEANDRO JESUS ; ROCHA, FLÁVIA ; BAGANO, PRISCILLA ; RAMOS, ROMMEL ; SILVA, ARTUR ; FIGUEIREDO, HENRIQUE ; ALMEIDA, SINTIA ; AZEVEDO, VASCO . Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 12C. *Genome Announcements*, v. 3, p. e00759-15, 2015.
20. ABREU, VINICIUS A. C. ; ALMEIDA, SINTIA ; TIWARI, SANDEEP ; HASSAN, SYED SHAH ; **MARIANO, DIEGO** ; SILVA, ARTUR ; BAUMBACH, JAN ; AZEVEDO, VASCO ; RÖTTGER, RICHARD . CMRegNet-An interspecies reference database for corynebacterial and mycobacterial regulatory networks. *BMC Genomics JCR*, v. 16, p. 452, 2015.  
**Citações:** **WEB OF SCIENCE** 1 | **SCOPUS** 2
21. OLIVEIRA, L. C. SARAIVA, T. D. L. SOARES, S. C. RAMOS, R. T. J. SA, P. H. C. G. CARNEIRO, A. R. MIRANDA, F. FREIRE, M. RENAN, W. JUNIOR, A. F. O. SANTOS, A. R. PINTO, A. C. SOUZA, B. M. CASTRO, C. P. DINIZ, C. A. A. ROCHA, C. S. **MARIANO, D. C. B.** DE AGUIAR, E. L. FOLADOR, E. L. BARBOSA, E. G. V. ABURJAILE, F. F. GONCALVES, L. A. GUIMARAES, L. C. AZEVEDO, M. AGRETI, P. C. M. , *et al.* ; Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain. *Genome Announcements*, v. 2, p. e00980-14-e00980-14, 2014.  
**Citações:** **SCOPUS** 1
22. VIANA, M. V. C. ; DE JESUS BENEVIDES, L. ; **BATISTA MARIANO, D. C.** ; DE SOUZA ROCHA, F. ; BAGANO VILAS BOAS, P. C. ; FOLADOR, E. L. ; PEREIRA, F. L. ; ALVES DORELLA, F. ; GOMES LEAL, C. A. ; FIORINI DE CARVALHO, A. ; SILVA, A. ; DE CASTRO SOARES, S. ; PEREIRA FIGUEIREDO, H. C. ; AZEVEDO, V. ; GUIMARAES, L. C. . Genome Sequence of *Corynebacterium ulcerans* Strain 210932. *Genome Announcements*, v. 2, p. e01233-14-e01233-14, 2014.
23. SILVA, A. C. B. ; VALGAS, B. O. ; **MARIANO, D. C. B.** ; DIAS, S. R. . Just-in-Biblio: website para formatação de referências bibliográficas usando PHP e MySQL. *Anais do Seminário de Produção Acadêmica da Anhanguera*, v. 3, p. 1, 2012.
24. **MARIANO, D. C. B.** ; AGUIAR, E. L. ; DIAS, S. R. . Comparando o desempenho de Bancos de Dados NoSQL e Relacionais manipulando dados biológicos. *Anais do Seminário de Produção Acadêmica da Anhanguera*, v. 3, p. 1, 2012.

## Livros publicados/organizados ou edições

1. **MARIANO, DIEGO**; de MELO-MINARDI, R. C. . Introdução à Programação Web para Bioinformática: HTML, CSS, PHP & JavaScript. 1. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2017. v. 3. 403p .
2. **MARIANO, DIEGO CÉSAR BATISTA**; de MELO-MINARDI, R. C. . Introdução à Programação para Bioinformática com Perl. 1. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2016. v. 2. 200p .
3. **MARIANO, DIEGO CB**; RAMOS, ROMMEL TJ ; AZEVEDO, V. A. C. . Montagem e finalização de genomas procariotos com mapeamento óptico. 1. ed. Saarbrücken, Alemanha: Novas Edições Acadêmicas, 2016. v. 1. 76p .
4. **MARIANO, D. C. B.**; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . Introdução à Programação para Bioinformática com Biopython. 3. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2015. v. 1. 230p .

## Capítulos de livros publicados

1. OLIVEIRA JUNIOR, A. F. ; BENEVIDES, LEANDRO DE JESUS ; **MARIANO, D. C. B.** ; AGUIAR, E. L. ; SOUSA, T. J. ; SILVA, A. ; AZEVEDO, V. A. C. . Bioinformatics. In: Zahoorullah S MD. (Org.). A Textbook of Biotechnology. 1ed.: SMGroup, 2015, v. , p. 15-32.

## Trabalhos completos publicados em anais de congressos

1. **MARIANO, D. C. B.**; SILVA, A. C. B. ; VALGAS, B. O. ; DIAS, S. R. . Sistema para formatação e gerenciamento de referências bibliográficas. In: XII Congresso de Iniciação Científica da FACECA, 2012, Varginha. XII CONGRESSO DE INICIAÇÃO CIENTÍFICA, 2012. v. 1.
2. **MARIANO, D. C. B.**; DIAS, S. R. ; AGUIAR, E. L. . Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos. In: 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012, São Paulo. 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012.

## Resumos publicados em anais de congressos

1. **MARIANO, DIEGO CÉSAR BATISTA**; de MELO-MINARDI, R. C. . A new method based on structural signatures to propose mutations for enzymes beta-glucosidase used in biofuel production. In: 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), 2017, São Pedro. Anais X-Meeting 2017, 2017.
2. **MARIANO, D. C. B.**; CORREIA, T. S. ; BARROSO, J. R. P. M. ; de MELO-MINARDI, R. C. . Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production. In: X-Meeting 2016 - 12th International Conference of the AB3C, 2016, Belo Horizonte. Anais X-Meeting 2016, 2016.
3. SOUSA, THIAGO JESUS ; PARISE, DOGLAS ; PEREIRA, FELIPE L ; PEREIRA FIGUEIREDO, H. C. ; COSTA, D. A. ; AZEVEDO, V. ; SILVA, A. ; RAMOS, R. T. J. ; **MARIANO, D. C. B.** . Detection and correction mis-assemblies in genome of Corynebacterium pseudotuberculosis. In: X-Meeting 2016 - 12th International Conference of the AB3C, 2016, Belo Horizonte. Anais X-Meeting 2016 - 12th International Conference of the AB3C, 2016.
4. VIANA, MARCUS V. C. ; PARISE, DOGLAS ; SOUSA, THIAGO J. ; BENEVIDES, LEANDRO J. ; **MARIANO, D. C. B.** ; ROCHA, FLÁVIA ; BAGANO, PRISCILLA ; GUIMARÃES, LUIS C. ; FIGUEIREDO, H. ; AZEVEDO, V. . Genome sequence of Corynebacterium pseudotuberculosis 32. In: Brazilian-International Congress of Genetics, 2016, Caxambu-MG. Anais Brazilian-International Congress of Genetics, 2016.
5. SILVA, M. F. M. ; MARTINS, P. M. ; **MARIANO, D. C. B.** ; PASTORINI, I. ; PANTUZA, N. ; de MELO-MINARDI, R. C. . An approach for constructing a database of manually curated contacts in proteins. In: X-meeting 2016 - 12th International Conference of the AB3C, 2016, Belo Horizonte. Anais X-meeting 2016 - 12th International Conference of the AB3C, 2016.
6. **MARIANO, D. C. B.**; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . Data visualization for sequence comparison. In: X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015, São Paulo. Anais do X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015.
7. BARROSO, J. R. P. M. ; **MARIANO, D. C. B.** ; CORREIA, T. S. ; RODRIGUES, L. M. ; FASSIO, A. V. ; MARTINS, P. M. ; LEITE, C. ; SOUSA, T. J. ; POVOA, F. ; FERREIRA, R. S. ; BLEICHER, L. ; de MELO-MINARDI, R. C. . POTTER: a web tool for protein point mutation modelling and analysis. In: X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015, São Paulo. Anais do X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015.
8. CORREIA, T. S. ; BARROSO, J. R. P. M. ; **MARIANO, D. C. B.** ; de MELO-MINARDI, R. C. . Detecting  $\beta$ -glucosidases with high catalytic efficiency for cellulose degradation using singular value decomposition. In: X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015, São Paulo. Anais do X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015.
9. **MARIANO, D. C. B.**; PEREIRA, F. L. ; AGUIAR, E. L. ; OLIVEIRA, L. C. ; BENEVIDES, L. ; GUIMARÃES, LUIS CARLOS ; FOLADOR, EDSON LUIZ ; SOUSA, T. J. ; GHOSH, PREETAM ; BARH, DEBMALYA ; FIGUEIREDO, H. C. P. ; SILVA, A. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: a web tool for managing bacterial genome assembly. In: X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015, São Paulo. Anais do X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015.
10. SOUSA, T. J. ; **MARIANO, D. C. B.** ; ABURJAILE, F. F. ; ROCHA, F. ; PEREIRA, F. L. ; FIGUEIREDO, H. ; SILVA, A. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . Optical mapping to detect misassemblies in genome of Corynebacterium pseudotuberculosis strain 1002. In: X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015, São Paulo. Anais do X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics, 2015.
11. CORREIA, T. S. ; **MARIANO, D. C. B.** ; BARROSO, J. R. P. M. ; de MELO-MINARDI, R. C. . Classificação de famílias de proteínas usando decomposição em valores singulares: enzima  $\beta$ -glucosidase como estudo de caso. In: XXIV Semana de

iniciação científica / PRPQ / UFMG, 2015, Belo Horizonte. Anais da XXIV Semana de iniciação científica / PRPQ / UFMG, 2015.

12. **MARIANO, D. C. B.**; OLIVEIRA, L. C. ; FOLADOR, E. L. ; AGUIAR, E. L. ; BENEVIDES, L. ; PEREIRA, F. L. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: A Web Tool for Complete Assembly of Bacterial Genomes. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio, 2014, Belo Horizonte. ISCB - LA / X-Meeting / BSB / SoBio, 2014.
13. **MARIANO, D. C. B.**; OLIVEIRA, L. C. ; FOLADOR, E. L. ; AGUIAR, E. L. ; BENEVIDES, L. ; PEREIRA, F. L. ; VIANA, M. V. C. ; SOUSA, R. T. J. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: A Simple Way to Make Complete Assemblies of Bacterial Genomes. In: First ISCB Latin American Student Council Symposium, 2014, Belo Horizonte. Program Booklet - ISCB LA Student Council Symposium, 2014. p. 24-25.
14. AGUIAR, E. L. ; **MARIANO, D. C. B.** ; OLIVEIRA, L. C. ; AMORIM, L. G. ; OLIVEIRA JUNIOR, A. F. ; ROCHA, F. ; PEREIRA, F. L. ; SOARES, S. C. ; DORELLA, F. ; LEAL, C. ; FIGUEIREDO, H. ; AZEVEDO, V. A. C. . The complete genome sequence of Streptococcus agalactiae strain GBS85147. In: First ISCB Latin American Student Council Symposium, 2014, Belo Horizonte. Program Booklet - ISCB LA Student Council Symposium, 2014. p. 24-25.
15. BENEVIDES, L. ; VIANA, M. V. C. ; **MARIANO, D. C. B.** ; ROCHA, F. ; FOLADOR, E. L. ; PEREIRA, F. L. ; DORELLA, F. A. ; CARVALHO, A. ; LEAL, C. ; SILVA, A. ; SOARES, S. C. ; FIGUEIREDO, H. ; AZEVEDO, V. A. C. ; GUIMARAES, L. C. . Complete Genome Sequence of Corynebacterium ulcerans strain 210931. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio, 2014, Belo Horizonte. ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio, 2014.
16. AGUIAR, E. L. ; **MARIANO, D. C. B.** ; OLIVEIRA, L. C. ; AMORIM, L. G. ; OLIVEIRA JUNIOR, A. F. ; ROCHA, F. ; PEREIRA, F. L. ; SOARES, S. C. ; DORELLA, F. A. ; LEAL, C. ; FIGUEIREDO, H. C. P. ; AZEVEDO, V. A. C. . The complete genome sequence of Streptococcus agalactiae strain GBS85147. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio, 2014, Belo Horizonte. ISCB - LA / X-Meeting / BSB / SoBio, 2014.
17. **MARIANO, D. C. B.**; AGUIAR, E. L. ; SOUSA, T. J. ; RODRIGUES, L. M. ; MELO-MINARDI, R. C. ; AZEVEDO, V. A. C. . Uso da visualização de dados como elemento transformador na etapa de scaffolding dos processos de montagem de genomas procariotos através do software KION. In: InWeb - CETIC Workshop sobre Visualização de Dados e Informações, 2014, Belo Horizonte. InWeb - CETIC Workshop sobre Visualização de Dados e Informações, 2014.
18. AGUIAR, E. L. ; FASSIO, A. V. ; **MARIANO, D. C. B.** ; MARTINS, P. M. ; MELO-MINARDI, R. C. ; AZEVEDO, V. A. C. . EC Number Viewer: um Website para visualização de evoluções na classificação de EC Numbers. In: InWeb - CETIC Workshop sobre Visualização de Dados e Informações, 2014, Belo Horizonte. InWeb - CETIC Workshop sobre Visualização de Dados e Informações, 2014.
19. VIANA, M. V. C. ; BENEVIDES, L. ; **MARIANO, D. C. B.** ; ROCHA, F. ; FOLADOR, E. L. ; PEREIRA, F. L. ; DORELLA, F. ; LEAL, C. ; CARVALHO, A. ; SILVA, A. ; SOARES, S. C. ; FIGUEIREDO, H. ; AZEVEDO, V. A. C. ; GUIMARAES, L. C. . Complete Genome Sequence of Corynebacterium ulcerans strain 210932. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio, 2014, Belo Horizonte. ISCB - LA / X-Meeting / BSB / SoBio, 2014.
20. OLIVEIRA, L. C. ; AMORIM, L. G. ; **MARIANO, D. C. B.** ; SANTOS, M. A. ; SOARES, S. C. ; MIYOSHI, A. ; AZEVEDO, V. A. C. . Phylogenetic Inference Of Bacterial Evolutionary Relationship From The Analysis Of Genomic Signature Using Singular Value Decomposition (SVD). In: 59º Congresso Brasileiro de Genética, 2013, Águas de Lindóia. Anais do 59º Congresso Brasileiro de Genética, 2013.
21. AMORIM, L. G. ; **MARIANO, D. C. B.** ; OLIVEIRA, L. C. ; SANTOS, M. A. ; SOARES, S. C. ; MIYOSHI, A. ; AZEVEDO, V. A. C. . Análise de bactérias baseada na decomposição por valores singulares de frequências de padrões de assinatura genômica. In: 27º Congresso Brasileiro de Microbiologia, 2013, Natal. Anais do 27º Congresso Brasileiro de Microbiologia, 2013.
22. **MARIANO, D. C. B.**; OLIVEIRA, L. C. ; AMORIM, L. G. ; SANTOS, M. A. ; SOARES, S. C. ; MIYOSHI, A. ; AZEVEDO, V. A. C. . A New Approach For Classification And Phylogenetic Analysis Of Bacteria Using Singular Value Decomposition (SVD). In: X-meeting BSB 2013, 2013, Recife. Anais do X-meeting BSB 2013, 2013.
23. ABREU, V. A. C. ; ALMEIDA, S. S. ; **MARIANO, D. C. B.** ; OLIVEIRA, L. C. ; AMORIM, L. G. ; DINIZ, C. A. A. ; AGUIAR, E. L. ; SOARES, S. C. ; HASSAN, S. ; BAUMBACH, J. ; AZEVEDO, V. A. C. . CMRegNet: A database of transcriptional regulatory network. In: X-meeting BSB 2013, 2013, Recife. Anais X-meeting BSB 2013, 2013.
24. **MARIANO, D. C. B.**; DIAS, S. R. ; AGUIAR, E. L. . Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos. In: XXI Congresso de Pós-graduação da UFLA, 2012, Lavras. Anais - XXI Congresso, 2012.
25. **MARIANO, D. C. B.**; DIAS, S. R. ; AGUIAR, E. L. . Comparing the performance of databases relational and NoSQL for manipulating biological data. In: X-meeting 2012 - 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas. AbstractBook - X-meeting 2012, 2012.

## Apresentações de Trabalho

1. **BATISTA MARIANO, D. C.**; MARTINS, P. M. ; FASSIO, A. V. ; de MELO-MINARDI, R. C. . A new method based on structural signatures to propose mutations for enzymes  $\beta$ -glucosidase used in biofuel production. 2017. (Apresentação de Trabalho/Conferência ou palestra).
2. **MARIANO, DIEGO CÉSAR BATISTA**. Programação web com PHP. 2016. (Apresentação de Trabalho/Conferência ou palestra).
3. **MARIANO, D. C. B.**; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . Data visualization for sequence comparison. 2015. (Apresentação de Trabalho/Congresso).
4. **MARIANO, D. C. B.**; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: A Web Tool for Complete Assembly of Bacterial Genomes. 2014. (Apresentação de Trabalho/Seminário).



## Programas de computador sem registro

1. MARTINS, P. M. ; **MARIANO, DIEGO C. B.** ; PASTORINI, I. ; PANTUZA, N. ; SILVA, M. F. M. ; de MELO-MINARDI, R. C. . Proteingo. 2016.
2. **MARIANO, D. C. B.**; SOARES, SIOMAR DE CASTRO ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . Dinnotator. 2014.
3. **MARIANO, D. C. B.**; AGUIAR, E. L. ; SOUSA, T. J. ; RODRIGUES, L. M. ; MELO-MINARDI, R. C. ; AZEVEDO, V. A. C. . Kion. 2014.
4. **MARIANO, D. C. B.**; PEREIRA, F. L. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA. 2014.

## Trabalhos técnicos

1. **MARIANO, DIEGO**; LEITE, C. ; SANTOS, L. H. S. ; ROCHA, R. E. O. ; de MELO-MINARDI, R. C. . A guide to performing systematic literature reviews in bioinformatics. 2017.

## Entrevistas, mesas redondas, programas e comentários na mídia

1. **MARIANO, DIEGO CÉSAR BATISTA**. Equipe brasileira vence competição mundial de Biologia Computacional. 2018. (Programa de rádio ou TV/Entrevista). 📻
2. **MARIANO, DIEGO CÉSAR BATISTA**. Doutorando da Bioinformática é coautor de artigo vencedor de competição da Wikipedia. 2018. (Programa de rádio ou TV/Entrevista). 📻

## Redes sociais, websites e blogs

1. **MARIANO, DIEGO CÉSAR BATISTA**; MARTINS, P. M. . LBS. 2017. (Site).

### Demais tipos de produção técnica

1. **MARIANO, DIEGO**. Introdução à Linguagem HTML. 2018. (Curso de curta duração ministrado/Outra).
2. **MARIANO, DIEGO C. B.**. Introdução ao Framework Bootstrap. 2018. (Curso de curta duração ministrado/Outra).
3. **MARIANO, DIEGO**. Introdução ao jQuery. 2018. (Curso de curta duração ministrado/Outra).
4. **MARIANO, DIEGO**. Introdução à linguagem JavaScript. 2018. (Curso de curta duração ministrado/Outra).
5. **MARIANO, DIEGO C. B.**. Introdução à linguagem CSS. 2018. (Curso de curta duração ministrado/Outra).
6. **MARIANO, D.C.B.**. Introdução ao PHP orientado a objetos. 2018. (Curso de curta duração ministrado/Outra).
7. **MARIANO, DIEGO**. Curso de programação com Perl. 2018. (Curso de curta duração ministrado/Outra).
8. **MARIANO, D.C.B.**. Introdução à Linguagem Python. 2018. (Curso de curta duração ministrado/Outra).
9. **MARIANO, DIEGO**. Boas práticas em PHP. 2018. (Curso de curta duração ministrado/Outra).
10. **MARIANO, D. C. B.**. Introdução a banco de dados com MySQL & PHPMyAdmin. 2018. (Curso de curta duração ministrado/Outra).
11. **MARIANO, DIEGO CÉSAR BATISTA**. Introdução à programação de computadores. 2018. (Curso de curta duração ministrado/Outra).
12. **MARIANO, DIEGO**. Introduction to HTML Language. 2018. (Curso de curta duração ministrado/Outra).
13. **MARIANO, DIEGO**. BLAST: Ferramenta de Alinhamentos Locais de Sequências. 2018. (Curso de curta duração ministrado/Outra).
14. **BATISTA MARIANO, D. C.**. Modelagem de proteínas por homologia. 2018. (Curso de curta duração ministrado/Outra).
15. **MARIANO, DIEGO CB**. Data Science: Visualização de Dados com Python. 2018. (Curso de curta duração ministrado/Outra).
16. **MARIANO, DIEGO CB**. Introdução ao Sistema Operacional Linux. 2018. (Curso de curta duração ministrado/Outra).
17. **MARIANO, DIEGO CB**. Terminal Linux. 2018. (Curso de curta duração ministrado/Outra).
18. **BATISTA MARIANO, D. C.**. Introdução à Criação de Sites Dinâmicos com PHP. 2018. (Curso de curta duração ministrado/Outra).
19. OLIVEIRA JUNIOR, A. F. ; **MARIANO, D. C. B.** ; AMORIM, L. G. ; OLIVEIRA, L. C. . Interação entre Microorganismos - Planta & Métodos Moleculares Aplicados (monitoria). 2013. (Curso de curta duração ministrado/Outra).
20. **MARIANO, D. C. B.**. Hardware - Montagem e manutenção de computadores. 2011. .

## Eventos

---

### Participação em eventos, congressos, exposições e feiras

1. 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C),. A new method based on structural signatures to propose mutations for enzymes beta-glucosidase used in biofuel production. 2017. (Congresso).
2. 2nd Brazilian Student Council Symposium.A new method based on structural signatures to propose mutations for enzymes beta-glucosidase used in biofuel production. 2017. (Simpósio).
3. III Curso de Verão de Engenharia de Máquinas Biológicas. 2017. (Seminário).
4. I Semana de Engenharia, Tecnologia e Computação.Programação web com PHP. 2016. (Seminário).
5. RSG Brazil - 1st Student Council Symposium. 2016. (Simpósio).
6. X-Meeting 2016 - 12th International Conference of the AB3C. Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production. 2016. (Congresso).

7. X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics. Data visualization for sequence comparison. 2015. (Congresso).
8. InWeb - CETIC Workshop sobre Visualização de Dados e Informações. Uso da visualização de dados como elemento transformador na etapa de scaffolding dos processos de montagem de genomas procariotos através do software KION.. 2014. (Oficina).
9. ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoBio. SIMBA: A Web Tool for Complete Assembly of Bacterial Genomes. 2014. (Congresso).
10. Latin American Council Symposium. SIMBA: a simple way to make complete assemblies of bacterial genomes. 2014. (Simpósio).
11. Terceira Escola de Verão em Computação do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais,. 2014. (Oficina).
12. 59º Congresso Brasileiro de Genética. Phylogenetic inference of bacterial evolutionary relationship from the analysis of genomic signature using singular value decomposition (SVD). 2013. (Congresso).
13. Microbiota: the bacteria that govern us. 2013. (Simpósio).
14. Programa TOP Espanha 2013 Santander Universidades. 2013. (Outra).
15. X-meeting BSB 2013. 2013. (Simpósio).
16. Feira do empreendedor SEBRAE. 2012. (Outra).
17. XII Congresso de Iniciação Científica da FACECA. Sistema para formatação e gerenciamento de referências bibliográficas. 2012. (Congresso).
18. XXI Congresso de Pós-graduação da UFLA. Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos. 2012. (Congresso).
19. Infratec - Feira de tecnologia do Senac Minas. Redes Wireless. 2010. (Outra).
20. Microsoft TechNet. 2010. (Encontro).
21. INFORUSO. 2009. (Outra).
22. OBMEP - Olimpíada brasileira de matemática. 2006. (Olimpíada).
23. OBA - Olimpíada Brasileira de Astronomia. 2004. (Olimpíada).

#### Organização de eventos, congressos, exposições e feiras

1. **MARIANO, D. C. B.**. Infratec - Feira de Tecnologia do Senac Minas. 2010. (Festival).

## Orientações

---

#### Orientações e supervisões em andamento

#### Dissertação de mestrado

1. Naiara Pantuza. Modelos e algoritmos para proposta de mutações em proteínas:  $\beta$ -glicosidases como estudo de caso. Início: 2017. Dissertação (Mestrado profissional em Bioinformática) - Universidade Federal de Minas Gerais. (Coorientador).

#### Orientações e supervisões concluídas

#### Iniciação científica

1. Thiago Da Silva Correia. Classificação de famílias de proteínas usando decomposição em valores singulares: enzima beta-glicosidase como estudo de caso. 2015. Iniciação Científica. (Graduando em Matemática Computacional) - Universidade Federal de Minas Gerais. Orientador: Diego César Batista Mariano.

## Inovação

---

#### Programa de computador sem registro

1. **MARTINS, P. M.** ; **MARIANO, DIEGO C. B.** ; PASTORINI, I. ; PANTUZA, N. ; SILVA, M. F. M. ; de MELO-MINARDI, R. C. . Proteingo. 2016.

## Educação e Popularização de C & T

---

## Livros e capítulos

1. **MARIANO, D. C. B.**; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . Introdução à Programação para Bioinformática com Biopython. 3. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2015. v. 1. 230p .
2. **MARIANO, DIEGO CÉSAR BATISTA**; de MELO-MINARDI, R. C. . Introdução à Programação para Bioinformática com Perl. 1. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2016. v. 2. 200p .
3. **MARIANO, DIEGO**; de MELO-MINARDI, R. C. . Introdução à Programação Web para Bioinformática: HTML, CSS, PHP & JavaScript. 1. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2017. v. 3. 403p .

## Outras informações relevantes

---

Aprovado em 1º lugar no processo seletivo de Doutorado em Bioinformática da UFMG (2015). Realizou intercâmbio na Universidad de Salamanca (Espanha) com enfoque em "Lengua Española, Conversación y Redacción" (2013). Aprovado em 1º lugar no processo seletivo SENAI - Aprendizagem Industrial Gráfica (2010).

## **Apêndice 6. Artigo 1**

**Título:** “*Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review*”

**Journal:** *Genetics and Molecular Research*

**Fator de impacto (2016):** 0,765

**Ano de publicação:** 2017

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/319197000>

# Characterization of glucose-tolerant $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: A systematic review

Article in *Genetics and molecular research: GMR* · August 2017

DOI: 10.4238/gmr16039740

CITATIONS

5

READS

523

8 authors, including:



**Diego Mariano**

Federal University of Minas Gerais

26 PUBLICATIONS 88 CITATIONS

[SEE PROFILE](#)



**Lucianna Helene Santos**

Federal University of Minas Gerais

36 PUBLICATIONS 27 CITATIONS

[SEE PROFILE](#)



**Karina S. Machado**

Universidade Federal do Rio Grande (FURG)

69 PUBLICATIONS 186 CITATIONS

[SEE PROFILE](#)



**Adriano Velasque Werhli**

Universidade Federal do Rio Grande (FURG)

69 PUBLICATIONS 621 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Computational study of industrially promissor mutations in beta glucosidase enzymes aiming protein engineering [View project](#)



Genome assembly [View project](#)

# Characterization of glucose-tolerant $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: a systematic review

D.C.B. Mariano<sup>1</sup>, C. Leite<sup>1</sup>, L.H.S. Santos<sup>1</sup>, L.F. Marins<sup>2</sup>, K.S. Machado<sup>3</sup>, A.V. Werhli<sup>3</sup>, L.H.F. Lima<sup>4</sup> and R.C. de Melo-Minardi<sup>1</sup>

<sup>1</sup>Laboratório de Bioinformática e Sistemas,  
Departamento de Ciência da Computação,  
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

<sup>2</sup>Laboratório de Biologia Molecular, Instituto de Ciências Biológicas,  
Universidade Federal de Rio Grande, Rio Grande, RS, Brasil

<sup>3</sup>Grupo de Biologia Computacional, Centro de Ciências Computacionais-C3,  
Universidade Federal de Rio Grande, Rio Grande, RS, Brasil

<sup>4</sup>Universidade Federal de São João Del-Rei, Campus Sete Lagoas,  
Sete Lagoas, MG, Brasil

Corresponding authors: D.C.B. Mariano / R.C. de Melo-Minardi

E-mail: diegomariano@ufmg.br / raquelcm@dcc.ufmg.br

Genet. Mol. Res. 16 (3): gmr16039740

Received May 26, 2017

Accepted July 28, 2017

Published August 17, 2017

DOI <http://dx.doi.org/10.4238/gmr16039740>

Copyright © 2017 The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution ShareAlike (CC BY-SA) 4.0 License.

**ABSTRACT.**  $\beta$ -glucosidases are enzymes that catalyze the hydrolysis of oligosaccharides and disaccharides, such as cellobiose. These enzymes play a key role in cellulose degrading, such as alleviating product inhibition of cellulases. Consequently, they have been considered essential for the biofuel industry. However, the majority of the characterized  $\beta$ -glucosidases is inhibited by glucose. Hence, glucose-tolerant  $\beta$ -glucosidases have been targeted to improve the production of second-generation biofuels. In this

paper, we proceeded a systematic literature review (SLR), collected protein structures and constructed a database of glucose-tolerant  $\beta$ -glucosidases, called betagdb. SLR was performed at PubMed, ScienceDirect and Scopus Library databases and conducted according to PRISMA framework. It was conducted in five steps: i) analysis of duplications, ii) title reading, iii) abstract reading, iv) diagonal reading, and v) full-text reading. The second, third, fourth, and fifth steps were performed independently by two researchers. Besides, we performed bioinformatics analysis on the collected data, such as structural and multiple alignments to detect the most conserved residues in the catalytic pocket, and molecular docking to characterize essential residues for substrate recognizing, glucose tolerance, and the  $\beta$ -glucosidase activity. We selected 27 papers, 23 sequences, and 5 PDB files of glucose-tolerant  $\beta$ -glucosidases. We characterized 11 highly conserved residues: H121, W122, N166, E167, N297, Y299, E355, W402, E409, W410, and F418. The presence of these residues may be essential for  $\beta$ -glucosidases. We also discussed the importance of residues W169, C170, L174, H181, and T226. Furthermore, we proposed that the number of contacts for each residue in the catalytic pocket might be a metric that could be used to suggest mutations. We believe that the herein propositions, together with the sequence and structural data collection, might be helpful for effective engineering of  $\beta$ -glucosidases for biofuel production and may help to shed some light on the degradation of cellulosic biomass.

**Key words:** Biofuel; Glucose-tolerant  $\beta$ -glucosidases; Bioinformatics; Systematic literature review

## INTRODUCTION

$\beta$ -glucosidase (E.C. 3.2.1.21) is a class of heterologous enzymes that hydrolyze glycosidic bonds of disaccharides, oligosaccharides, alkyl- and aryl- $\beta$ -glycosides (Cairns and Esen, 2010). They have been found in metagenomes (Uchiyama et al., 2015) and several organisms, such as animals (Uchima et al., 2011), fungi (Saha and Bothast, 1996), plants (Pentzold et al., 2014), and bacteria (Crespim et al., 2016). In animals, these enzymes help in the metabolism of glycolipids and digestive functions (Cairns and Esen, 2010). In plants, they play several roles, such as defense, the release of flavor compounds, cell wall catabolism, and lignification (Pentzold et al., 2014). In bacteria and fungi, they are essential components of cellulose hydrolysis (Ramani et al., 2015a).  $\beta$ -glucosidases have applications in several areas of biotechnological industries, such as aroma improvement of juices and wine (Swangkeaw et al., 2010), hydrolysis of soybean isoflavone glycosides (Singhania et al., 2013), toxicity reduction of animal feed (Cota et al., 2015), and cellulose degradation for biomass conversion in biofuel production (Singhania et al., 2013).

Based on substrate specificity, these enzymes were classified into three main groups: i) aryl- $\beta$ -glucosidases, ii) cellobiases, and iii) broad-specificity  $\beta$ -glucosidases (Singhania et al.,

2013; Ramani et al., 2015a; Yang et al., 2015a). However, the limitations of this classification required a new classification method based on sequence similarities (Henrissat, 1991). Since then,  $\beta$ -glucosidases have been classified into glycoside hydrolase (GH) families 1, 3, 5, 9, and 30 (Cairns and Esen, 2010). However, the majority of the  $\beta$ -glucosidases belongs to families 1 and 3 (Singhania et al., 2013; Crespim et al., 2016).  $\beta$ -glucosidases from the GH3 family present an aspartate as a catalytic nucleophile and a glutamate as a catalytic acid/base. On the other hand, enzymes of the GH1 family belong to the clan GH-A, that group proteins with  $(\alpha/\beta)_8$  TIM barrel fold and conserved catalytic amino acids on  $\beta$ -strands 4 and 7 of the barrel (Cairns and Esen, 2010; Jabbour et al., 2012; Crespim et al., 2016). GH1  $\beta$ -glucosidases hydrolyze through the mechanism of retention of the anomeric carbon, using two glutamates as a catalytic nucleophile and catalytic acid/base. GH1  $\beta$ -glucosidases have attracted attention due to the high resistance to product inhibition, which has many applications for cellulose degradation (Yang et al., 2015b).

Cellulose is the major source of biomass on the Earth, accounting for around 40 to 50% of the plant biomass weight (Uchima et al., 2012; Uchiyama et al., 2015). Cellulose is constituted of glucose monomers connected by  $\beta$ -1,4 glucosidic bonds (Ramani et al., 2015b). The glucose obtained by cellulose degradation may be fermented to produce bioethanol, a promising green alternative and renewable source for the production of fuels (Teugjas and Våljamäe, 2013).  $\beta$ -glucosidases act synergistically with endoglucanases (E.C. 3.2.1.4) and exoglucanases or cellobiohydrolases (E.C. 3.2.1.91) to compose the enzymatic system for cellulose bioconversion (Kumar et al., 2008). While endoglucanases act in the cellulose chain producing oligosaccharides of variable length, exoglucanases act in the oligosaccharides producing mainly cellobiose.  $\beta$ -glucosidases cleave the link  $\beta$ -1,4 glucosidic bonds with the help of a water molecule, producing two glucose molecules (Béguin and Aubert, 1994). It is well established that  $\beta$ -glucosidases have a pivotal role in this enzymatic system removing cellobiose, which is a strong inhibitor of both endoglucanases and exoglucanases (Murphy et al., 2013; Zhao et al., 2013; Chamoli et al., 2016). However, most of the  $\beta$ -glucosidases reported are inhibited by the increase of glucose concentration (Teugjas and Våljamäe, 2013; Yang et al., 2015b). Hence, there is the growing interest in searching for thermostable and glucose-tolerant  $\beta$ -glucosidases. Their production may help to shed some light on the degradation of cellulosic biomass and may improve the saccharification process for biofuel production (Pei et al., 2012).

In this paper, we conducted a systematic literature review (SLR) to collect, analyze, and summarize the state-of-the-art research about glucose-tolerant  $\beta$ -glucosidases. We aimed to identify research trends about  $\beta$ -glucosidase enzymes used to improve the biofuel production. Although the mechanisms and molecular basis for glucose tolerance have not been completely enlightened, several studies have presented insights about the role of glucose tolerance, as well as structural aspects of interaction with the substrates and products, and also some discoveries on more efficient  $\beta$ -glucosidases in biomass degradation (Singhania et al., 2013). Also, we collected protein structures and constructed a glucose-tolerant  $\beta$ -glucosidase database. The construction of glucose-tolerant  $\beta$ -glucosidase structure database may be useful for bioinformatics analysis, such as the characterization of residues responsible for glucose tolerance in  $\beta$ -glucosidases. We also performed molecular docking and characterized important amino acids near to the cellobiose-binding region for the occurrence of hydrolysis. The provided information may be useful for engineering efficient enzymes for second-generation biofuel production.



## MATERIAL AND METHODS

The SLR protocol was created based on the guide for performing SLR in bioinformatics (Mariano et al., 2017), the guideline for systematic reviews of Kitchenham (2004), and the PRISMA statement (Liberati et al., 2009).

### Search strategy

Data collection for the SLR was performed from November 2015 until February 2016. Search terms were defined based on interviews with researchers in the area, the study of Pei et al. (2012), and were iteratively improved based on the results obtained from the first queries. PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>), ScienceDirect (<http://www.sciencedirect.com/>), and Scopus (<http://www.scopus.com/>) databases were used to collect the studies. The query used was (“beta-glucosidases” or “beta-glucosidase”) and (“glucose-tolerant” or “glucose tolerance” or “glucose insensitive” or “product tolerant” or “product insensitive”). The period of publication for the three databases was “all the time”. For ScienceDirect, we applied a filter to return only content type declared as “journal”. For Scopus, we applied filters to return manuscripts in English, to search in any part of the document, and with document type declared as “article”. No filter was applied for PubMed. The filters were applied to remove undesirable formats, such as book chapters or articles in idioms different from English.

### Eligibility criteria and selection

The SLR was performed in five steps: i) exclusion of duplications, ii) analysis of titles, iii) analysis of abstracts, iv) diagonal reading (when only the introduction, titles of the figures and conclusion are read), and v) full-text reading. Second, third, fourth, and fifth steps were performed independently by two reviewers to minimize the risk of bias or mistakes. In steps two, three and four, a paper was kept for the next step if at least one of the evaluators had approved. In these steps, the reviewers evaluate if the paper presented a  $\beta$ -glucosidase applied for biofuel production and if glucose tolerance was cited as crucial for improving glucose production. We prioritized articles containing publicly available protein structural and kinetic data, such as inhibition constant for glucose and affinity for cellobiose. Other criteria, such as culture mediums and reagents used were not evaluated. In the last step, we performed a full-text reading. For each paper, we answered five questions: i) does the study present a  $\beta$ -glucosidase with the enzyme kinetic patterns, such as inhibition constant ( $K_i$ ) for glucose, Michaelis constant ( $K_m$ ) for cellobiose, catalytic constant ( $K_{cat}$ ), characterized? ii) Is the  $\beta$ -glucosidase presented as a possible enzyme for biofuel production and it hydrolyzes cellobiose as a substrate? iii) Does the study show a genetic engineering strategy, such as mutagenesis or fusion, to solve the problem of glucose inhibition? iv) Does the paper seek to explain the mechanisms of inhibition by glucose in  $\beta$ -glucosidases? v) Does the paper present  $\beta$ -glucosidase sequence or three-dimensional structure data publicly available for bioinformatics analysis?

For each question, both evaluators, in consensus, gave scores: (0) if it does not attend requirements of the question; (1) if it attends the requirements of the question partially; and (2) if it attends requirements of the question. A paper was included in the review if it has a score equal or higher than six (60%). We also collected the following information from the papers: i) author names, ii) title, iii) sequence IDs, iv) PDB IDs, and v) enzyme kinetic information (if available).

## Biological data collection

The sequences were collected in the databases UniProt (<http://www.uniprot.org/>) and GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>). Structural files were obtained in the Protein Data Bank (PDB; Berman et al., 2000). Protein three-dimensional structures are crucial to infer about the glucose tolerance mechanism in  $\beta$ -glucosidases. However, few experimental structures of glucose-tolerant were found.

## Bioinformatics analysis

Homology modeling has been used to obtain three-dimensional structures for sequences of  $\beta$ -glucosidases (Yang et al., 2015b; Chamoli et al., 2016; Crespim et al., 2016). To make available as much information as possible to enable studies of glucose-tolerant  $\beta$ -glucosidases, we constructed 18 three-dimensional models to represent the sequences without a structure available. The sequences were modeled by homology using a consolidated pipeline of the literature (Bitar and Franco, 2014). For each protein, 100 models were built using MODELLER (Webb and Sali, 2014). NCBI BLAST web interface (Johnson et al., 2008) was used for template selection. The best model was selected based on the lowest value of MODELLER objective function and by the highest number of residues in favored region calculated by the RAMPAGE software (Lovell et al., 2003). Sequences were compared with a global alignment using the ggsearch36 software (Pearson, 2016) that implements the Needleman-Wunsch algorithm. Results with e-value lower than 0.001 were eliminated. An identity matrix was constructed using in-house scripts. We also built a database and website to make available visualizations of the three-dimensional structures and characteristics of the PDB files collected and modeled using 3Dmol (Rego and Koes, 2015).

We used the  $\beta$ -glucosidase from the termite *Neotermes koshunensis* in complex with cellobiose (Jeng et al., 2012) to characterized the catalytic pocket of GH1  $\beta$ -glucosidases. We considered the catalytic pocket as the residues present in the channel that guides the substrate to the two catalytic amino acids. We selected 24 residues whose distances from ligand were less than or equal to 6.5 Å from the ligand. The distance was chosen based on previous metrics to determine residues of pockets (Pires et al., 2013). Then, we performed structural alignments among the 21 collected glucose-tolerant GH1  $\beta$ -glucosidases collected and the *N. koshunensis*  $\beta$ -glucosidase to obtain the corresponding residues of the catalytic pocket using MultiProt (Shatsky et al., 2004). We also performed sequence alignments using Clustal Omega (Sievers et al., 2011) to detect a consensus subsequence and constructed a residue conservation visualization using D3 JavaScript library (<https://d3js.org>).

Molecular docking has been used to understand cellobiose binding in  $\beta$ -glucosidases (Khairudin and Mazlan, 2013). We performed molecular docking to verify which residues of the 21 glucose-tolerant GH1  $\beta$ -glucosidases were directly related to the recognizing of the substrate and to determine the number of contacts carried out for each residue with the cellobiose. DOCK 6 (version 6.7) was used to perform molecular docking (Allen et al., 2015). DOCK 6 uses an incremental construction method as a sampling algorithm for flexible ligand docking, the so-called anchor-and-grow algorithm (Ewing and Kuntz, 1997). In this approach, the ligand is separated in layers, and the major rigid substructure of the ligand is primarily recognized as the anchor, the anchor is then rigidly oriented in the binding site. Subsequently, each layer of flexible bonds is grown from each cluster, minimized, ranked, and clustered

again. The latter procedure is repeated until the molecule is fully constructed.

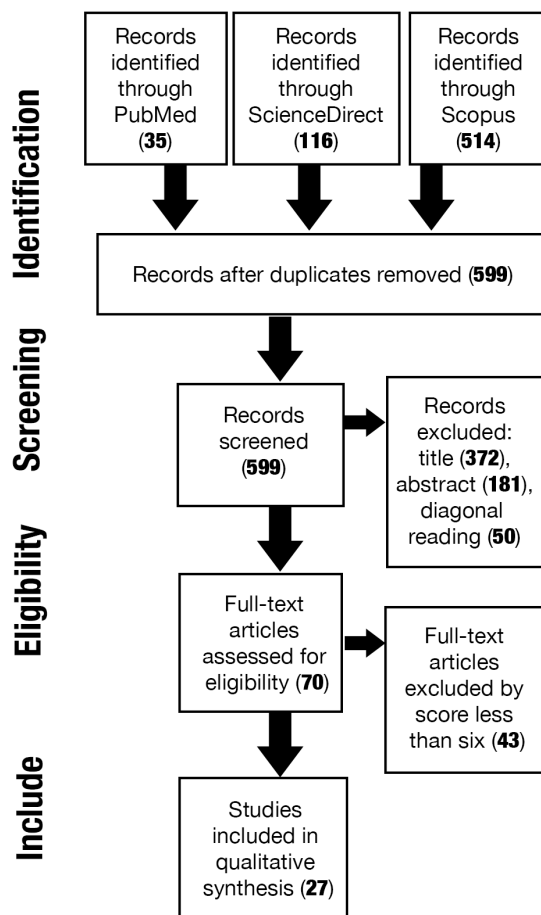
Before molecular docking with DOCK 6, all systems (receptor and ligand molecules) were evenly prepared. The crystal structure, PDB code: 3VIK (Jeng et al., 2012), bound to the cellobiose (CBI) ligand, was chosen as a reference, and all structures were superimposed on it, using the matchmaker tool in Chimera program (Pettersen et al., 2004). The superimposing of structures before molecular docking facilitates comparison between the crystallographic CBI and pose outcomes (**Figure S1**). For the structures extracted from the PDB database, ions, and other co-factors were removed. Hydrogens were added at physiological pH, and standard receptor residues were assigned AMBER ff14sb atomic partial charges (Maier et al., 2015), while CBI was assigned AM1-BCC charges (Jakalian et al., 2002). All receptor-CBI systems were subjected to energy minimization, 500 steepest descent minimization steps, where all residues were unrestrained using Chimera's Minimize Structure module (Pettersen et al., 2004). This minimization step allows the newly added hydrogen atoms of receptor and ligand to adjust in physically moderate positions, while also transporting the protein to a potential energy point inferior to the one from before.

Receptor and CBI were separated so that a specific DOCK preparation could be performed. DOCK's sphgen program (DesJarlais et al., 1988) was used to generate spheres within the binding site. The spheres that were within 8.0 Å of the crystallographic CBI position (PDB code: 3VIK) were kept for docking. A box around the spheres plus a 5.0 Å margin in all directions was used to restrict the receptor space for energy grid calculations. Lastly, energy interactions between a dummy probe atom and all receptor atoms on a 0.3 Å resolution grid within the box were calculated with DOCK's GRID program (Meng et al., 1992). In the calculated grid, van der Waals interactions were modeled through the Lennard-Jones potential with 6-12 attractive and repulsive exponents, respectively, whereas, a distance-dependent dielectric coefficient was used to represent the electrostatic interactions. The grid files are essential for rapid score evaluation using DOCK's native energy-based score GridScore. CBI was treated as flexible in all docking outcomes, according to the "standard flexible docking (FLX) protocol" described in Allen et al. (2015). A maximum of 300 ligand conformations was retained and clustered (RMSD cut-off of 2.0Å) for all receptor-CBI systems to compose our docking analysis. We used all the structures to calculate the number of contacts between residues and substrate. A contact is defined as all kinds of direct interactions: polar, nonpolar, favorable, and unfavorable (including clashes).

## RESULTS

The results of SLR were described according to PRISMA workflow and checklist (Figure 1 and **Table S1**). We collected 665 papers from three databases: PubMed, ScienceDirect, and Scopus. We performed five steps to evaluate the quality of the publications according to predefined objectives of the present SLR. After the analysis of two evaluators, 27 papers were included in the SLR (**Table S2**). We grouped the 27 papers into three categories: 1) papers that report a new glucose-tolerant  $\beta$ -glucosidases with proper application for biofuel production and with sequences or three-dimensional structures publicly available; 2) papers that report an engineering genetic technique being used to improve the  $\beta$ -glucosidase activity; 3) papers that report a comprehensive explanation about glucose tolerance mechanism and try to determine the related structural aspects (**Table S2**). We also collected five protein structural files, 23  $\beta$ -glucosidase sequences, and their kinetic parameters (**Table S3**). Besides, 18 three-

dimensional models of the sequences without a PDB file available in the PDB database were constructed by homology (**Figure S2**). We decided to perform the next steps only with the 21 GH1  $\beta$ -glucosidases because glucose-tolerant  $\beta$ -glucosidases (from now called GTBGL) have been described only in the GH1 family.

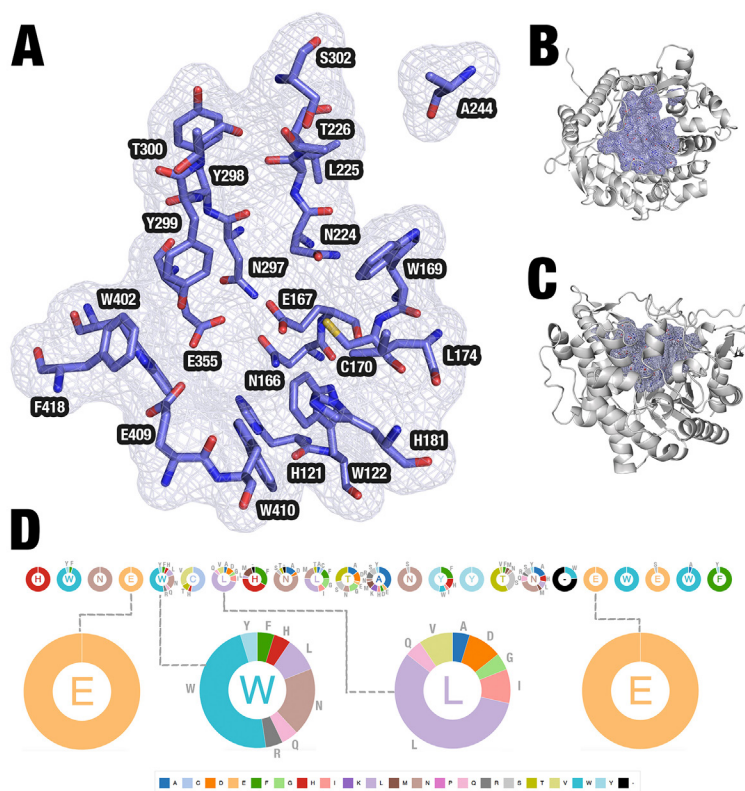


**Figure 1.** PRISMA workflow describing the number of papers collected during the systematic literature review based on the PRISMA statement.

### Conserved residues in the catalytic pocket of GTBGL

Based on the information collected in the SLR, we hypothesized that the key amino acids responsible for regulating the glucose tolerance process should be in the channel that guides to the active site. We called this region as the catalytic pocket (Figure 2A-C). To verify this, we performed alignments among the residues in the pocket and detected a GTBGL consensus subsequence composed of 22 most conserved amino acids near to the likely region of cellobiose binding: “HWNEWCLHNLNTANYTNEWWEWF” (Figure 2D). These amino acids corresponded in the *Thermoanaerobacter brockii*  $\beta$ -glucosidase to the residues: H121,

W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410, and F418 (Figure 2). The catalytic pocket of the *T. brockii*  $\beta$ -glucosidase presented the highest quantity of similarities with the GTBGL consensus subsequence. For this reason, from now on it was used as a reference (the corresponding residues for all GH1 glucose-tolerant  $\beta$ -glucosidases are available in **Table S4**).



**Figure 2.** Conserved residues in the catalytic pocket of glucose-tolerant GH1  $\beta$ -glucosidases. **A.** Residues to 6.5 ångström of the ligand in the GH1  $\beta$ -glucosidase from *Thermoanaerobacter brockii*. **B.** Enzyme top view. **C.** Enzyme side view. **D.** Conserved residues in the 21 GH1 glucose-tolerant  $\beta$ -glucosidases. The conservation of the amino acid is shown in the pie chart. We highlighted the two catalytic residues, E167 and E355, and the residues W169 and L173 described in the literature as essential for glucose tolerance. The amino acids “HWNEWCLHNLNTANYTNEWWF” correspond in the  $\beta$ -glucosidase from *T. brockii* to the residues: H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302 (correspond to an asparagine), E355, W402, E409, W410, and F418. Images generated by PyMOL (<http://pymol.org>) and D3.js (<http://d3js.org>).

Six residues were conserved in all glucose-tolerant  $\beta$ -glucosidases: H121, N166, E167, Y299, E355, and W402 (Table 1). E167 and E355 were acid/base catalytic and nucleophile catalytic residues, respectively. The other residues were located near to both catalytic residues. The conserved residues were essential to recognize the substrate. The residues W122, N297, E409, W410, and F418 were conserved in more than 90% of the sequences. Also, W169,

C170, L174, N224, A244, Y299, and T300 were conserved in the majority of the sequences. The low conservation of amino acids L225, T226, and S302 (an asparagine in the consensus subsequence) suggested minor importance to these residues.

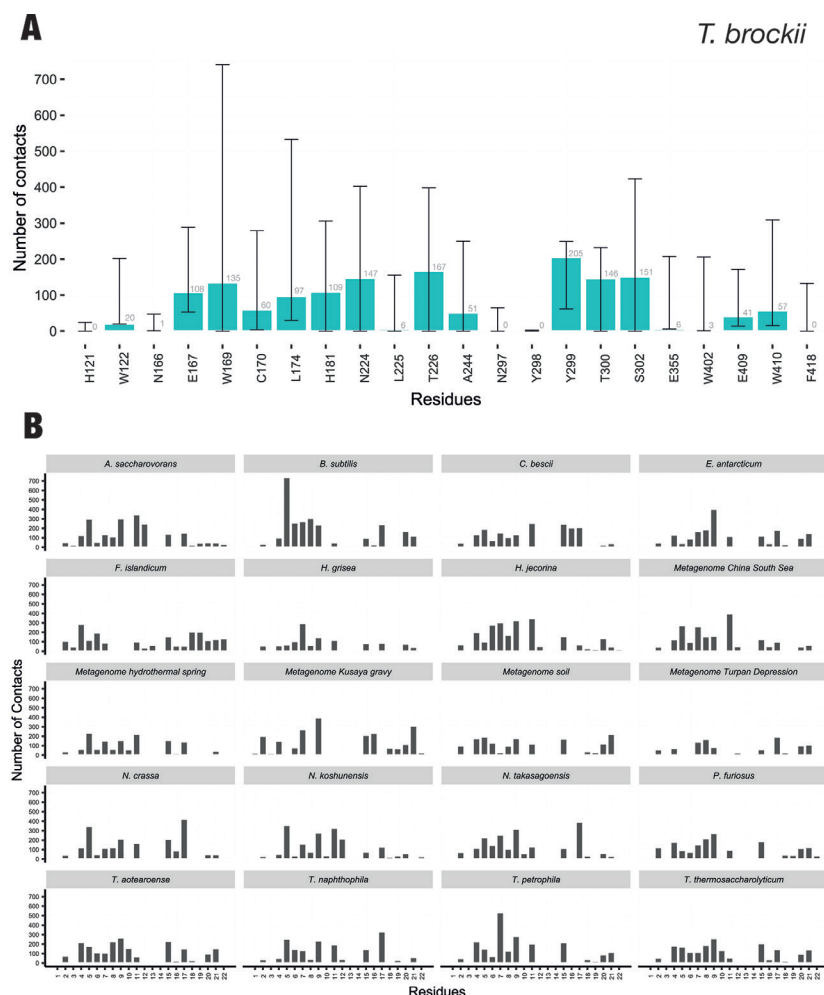
**Table 1.** Residue conservation and contacts with cellobiose.

#	Residue ( <i>T. brockii</i> )	Residue conservation		Contacts with cellobiose	
		>90% conserved	100% conserved	Corresponding residue makes contact in all GTBGL	No contacts when substituted (reference: residue of <i>T. brockii</i> )
1	H121	x	x		
2	W122	x		x	
3	N166	x	x	x	
4	E167	x	x	x	
5	W169				x
6	C170			x	
7	L174			x	
8	H181				x
9	N224				x
10	L225				
11	T226				x
12	A244				
13	N297	x			
14	Y298				
15	Y299	x	x	x	
16	T300				
17	S302				x
18	E355	x	x	x	
19	W402	x	x	x	
20	E409	x		x	
21	W410	x		x	
22	F418	x			

The table corresponds to the 22 corresponding residues in the catalytic pocket of the 21 glucose-tolerant  $\beta$ -glucosidases (GTBGL).

## Cellobiose docking

From the retained and clustered docking conformations, we calculated the number of contacts between all receptor residues and cellobiose (**Figure S3**). The corresponding residues to W122, N166, E167, C170, L174, Y299, E355, W402, E409, and W410 formed contacts with the cellobiose in every GTBGL (Table 1 and Figure 3). Moreover, the corresponding residues to H181 and N224 established contacts in almost every GTBGL. However, they are not present in the *Bacillus subtilis*  $\beta$ -glucosidase. Also, some residues established several contacts, except when they were substituted by other amino acids, such as W169 when substituted by a glutamine or T226 when substituted by a glycine. Furthermore, the residue found in the 17th position of the catalytic pocket consensus subsequence, which in the *T. brockii*  $\beta$ -glucosidase was S302, presented many contacts when it was a serine but mostly none when the residue was an alanine. When this residue was an asparagine, the most conserved amino acid in this position, in some  $\beta$ -glucosidases, it performed several contacts while in another no contact at all, such as in *Neurospora crassa* and *Trichoderma reesei*  $\beta$ -glucosidases, respectively. The corresponding residues to L225, A244, N297, Y298, T300, and F418 established few or no contacts (Table 1 and Figure 3). However, the residues in the corresponding position of N297 and F418 established several contacts when they were substituted by serine and tyrosine, respectively.



**Figure 3.** Contacts between residues and the cellobiose docked in glucose-tolerant GH1  $\beta$ -glucosidases. **A.** Contacts between residues and the cellobiose docked in the GH1  $\beta$ -glucosidase from *T. brockii*. A black line indicates the minor and maximum value detected in the corresponding position of other glucose-tolerant  $\beta$ -glucosidases. **B.** Contacts with the ligand that occurs in the other glucose-tolerant  $\beta$ -glucosidases. A complete table of the corresponding values can be obtained in **Table S4**.

## DISCUSSION

Glucose inhibits the  $\beta$ -glucosidase activity competing with the substrate (cellobiose) to bind in the active site. Hence, glucose-tolerant  $\beta$ -glucosidases may increase the biofuel production capacity and reduce costs (Yang et al., 2015b). An efficient hydrolysis of biomass requires active  $\beta$ -glucosidases at higher glucose concentration (Singhania et al., 2013). For this reason, glucose-tolerant  $\beta$ -glucosidases are important. However, it has been hard to identify a  $\beta$ -glucosidase completely tolerant to product inhibition, due to the similarities between product and substrate. It is observed that when the affinity to the product is reduced,

the substrate affinity is reduced as well. However,  $\beta$ -glucosidases with low sensitivity to inhibition have been found. Moreover, the enzymatic hydrolysis of cellulose that produces glucose concentrations may reach 650-1000 mM (Meleiro et al., 2015). Hence, an ideal  $\beta$ -glucosidase for biofuel production should have elevated levels of glucose tolerance and also a high catalytic efficiency.

In addition to glucose tolerance, a stimulatory effect of glucose for  $\beta$ -glucosidases was reported in several studies (Pei et al., 2012; Yang et al., 2015b; Crespim et al., 2016; Guo et al., 2016). This effect has been reported exclusively in some GH1  $\beta$ -glucosidases (de Giuseppe et al., 2014; Yang et al., 2015b). However, not every glucose-tolerant  $\beta$ -glucosidases present this stimulatory effect (Meleiro et al., 2015). Glucose stimulation consists in an improvement in the  $\beta$ -glucosidase activity in a particular range of glucose concentrations. For instance, it has been reported that in the presence of 50 mM glucose the activity of a  $\beta$ -glucosidase from *Humicola insolens* RP86 was stimulated about 1.8-fold (Souza et al., 2014). Glucose stimulation occurs due to an allosteric effect by glucose binding in a secondary site or by transglycosylation (Cao et al., 2015). It has been suggested that when glucose concentration increases during the saccharification process, substrate inhibition is gradually prevented (Guo et al., 2016), supporting the idea that glucose stimulation occurs not due to the presence of certain glucose quantity, but due to the reduction of substrate concentration in the reaction environment. Besides glucose inhibition, some  $\beta$ -glucosidases are inhibited by cellobiose.

### Conserved patterns in the catalytic pocket

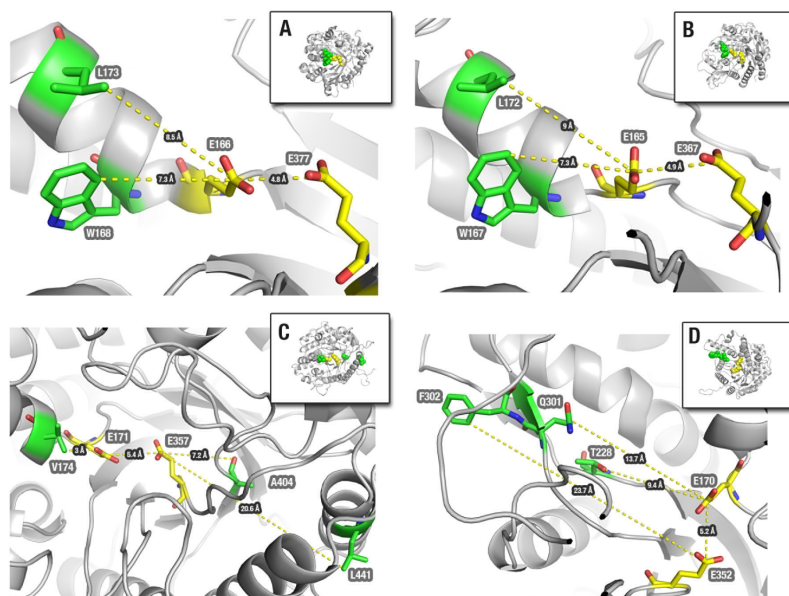
$\beta$ -glucosidase structures collected during the SLR were used to characterize conserved residues present in the catalytic pocket. Residue conservation is a primary indication of conserved patterns that could be used to propose mutations for non-tolerant  $\beta$ -glucosidases. The residues H121, N166, E167, Y299, E355, W402, W122, N297, E409, W410, and F418 are conserved in more than 90% of the sequences (Table 1). Hence, any mutation of these residues could not be indicated for engineering more efficient  $\beta$ -glucosidases once conservation is an indicative of the residue importance. On the other hand, some residues appear in the majority of the sequences. However, they are not highly conserved, such as W169, C170, L174, N224, A244, Y299, and T300. Other residues are poorly conserved, such as L225, T226, and S302, what can indicate that these positions have little importance.

Also, the docking results (Figure 3) showed that some residues are important for the substrate recognizing. The corresponding residues to W122, N166, E167, C170, L174, Y299, W402, E409, and W410 performed several contacts with the ligand. The residue E167 is one of the catalytic glutamates and performs several contacts with the ligand. Moreover, E355, the other catalytic glutamate, produced few contacts with the ligand. The H121 is another residue that is highly conserved but performs few contacts with the ligand. The H121 probably acts together with E167 and E355 in the catalytic activity. It has been highlighted that H121 is somehow involved in substrate binding or transition state stabilization (Barrett et al., 1995; Sanz-Aparicio et al., 1998).

We also detected a possible co-occurrence of a cysteine (C170) when appearing both a tryptophan (W169) and leucine (L174). Glucose-tolerant GH1  $\beta$ -glucosidases present a deep and narrow channel that limits the glucose access to the active site. It has been suggested that the channel shape guides the substrate to the active site and it is responsible for reducing the glucose access and, consequently, glucose tolerance in GH1  $\beta$ -glucosidases (de Giuseppe et



al., 2014). Equally, some amino acids located near to the active site were reported in some studies as essential for the release of glucose (Figure 4; de Giuseppe et al., 2014; Guo et al., 2016). In the  $\beta$ -glucosidase from *Humicola grisea* var. *thermoidea* (PDB: 4MDO), the amino acids W168 and L173 (W169 and L174 of *T. Brockii*) were reported as contributors for relieving enzyme inhibition. They apply constraints at the +2 subsite, and consequently, cause to glucose access limitation at the -1 subsite (Figure 4A). In another study, the site-directed mutagenesis L167W and P172L (corresponding residues of W169 and L174) were used to improve the glucose tolerance, pH and thermostability of  $\beta$ -glucosidase from *T. reesei* (Figure 4B; Guo et al., 2016). The W169 and L174 were conserved in the majority of glucose-tolerant  $\beta$ -glucosidases. This conservation highlights the importance of these amino acids for the regulation of entrance and exit of products and substrate, already shown in other articles (de Giuseppe et al., 2014; Guo et al., 2016). Likewise, the residue C170 appears in the majority of the catalytic pockets, which indicates that mutating an amino acid in this corresponding position for a cysteine may be beneficial, as reported in the literature (Figure 4C; Cao et al., 2015). In a recent study, the authors used random mutagenesis and detected three beneficial mutations: V174C, A404V, and L441F (where V174 corresponds to C170 in the *T. Brockii*  $\beta$ -glucosidase). These mutations allowed the construction of a new glucose-tolerant and thermostable mutant, which enhanced sugarcane bagasse conversion by 14-35%.



**Figure 4.** Structural data of  $\beta$ -glucosidases, amino acids related to glucose tolerance, and the active site. **A.** *Humicola grisea*  $\beta$ -glucosidase presents a narrow and deep channel that guide to the active site. The acid/base catalytic and nucleophile amino acids, E166 and E377, are  $\sim 5$  Å away. The residues W168 and L173 are responsible for contributing to relieving enzyme inhibition. **B.** *Trichoderma reesei*  $\beta$ -glucosidase. The active site residues are E165 and E367. Mutations L167W and P172L improved the activity. **C.**  $\beta$ -glucosidase from metagenome Turpan Depression. The active site residues are E357 and E171. Mutations V174C, A404V, and L441F improved the catalytic activity, although glucose-tolerance reduced of 3.5 M to 3 M. **D.**  $\beta$ -glucosidase from metagenome China South Sea. The active site residues are E170 and E352. Mutations H228T, N301Q, and V302F were performed in the entrance and middle of the channel that guides to the active site; Images generated by PyMOL (<http://pymol.org>).

Our results suggest that the appearing of a cysteine in this position seems to be correlational with the appearing of tryptophan and leucine in a nearby region, such as W169 and L174. Furthermore, according to the molecular docking results, when the three residues tryptophan, cysteine, and leucine appear together, they perform fewer contacts than when other residues substitute them; this suggests that co-occurrence may be related to the glucose tolerance. For instance, the residues tryptophan, cysteine, and leucine of the *Exiguobacterium antarcticum*  $\beta$ -glucosidase performed 347, 51, and 117 contacts, respectively (**Table S5**). In the same positions, the *Nasutitermes takasagoensis*  $\beta$ -glucosidase presents an arginine, an aspartate, and a threonine, that performed 740, 261, and 275 contacts, respectively (**Table S5**). The *E. antarcticum*  $\beta$ -glucosidase showed IC<sub>50</sub> for glucose of 1000 mM (Crespim et al., 2016). On the other hand, the *N. takasagoensis*  $\beta$ -glucosidase showed a glucose tolerance of 600 mM (Uchima et al., 2012). These values are not full evidence that these amino acids are related to a higher glucose tolerance. However, other glucose-tolerant  $\beta$ -glucosidase, such as the extracted of metagenome from China South Sea (Yang et al., 2015b), Turpan Depression (Cao et al., 2015), and Kusaya gravy (Uchiyama et al., 2015), present fewer contacts in these positions and IC<sub>50</sub> of 1000, 3500, and 750 mM, respectively (**Table S5**). This information might suggest residues that perform fewer contacts with the cellobiose in the middle of the catalytic pocket channel and may be beneficial for glucose tolerance.

Furthermore, mutagenesis of residues at the entrance and in the middle of the same channel of a GH1  $\beta$ -glucosidases isolated from a metagenome of the China South Sea has been used to characterize other sites that glucose has more preference than the active site (Figure 4D; Yang et al., 2015b). The glucose tolerance on a non-tolerant  $\beta$ -glucosidase (bgl1B) was increased through the mutations H228T and N301Q/V302F (Figure 4D; Yang et al., 2015b). This outcome may suggest that secondary sites, which cellobiose binds the path of the active site, are of great importance in the release of glucose process and the binding inhibition. The corresponding residue to H228 in the *T. Brockii*  $\beta$ -glucosidase is the T226. Hence, the mutation H228T appears to be beneficial. Although T226 was not highly conserved (Figure 2), its substitution for another amino acid could change the number of contacts (Table 1). Indeed, a histidine does not appear in this position in any glucose-tolerant  $\beta$ -glucosidase.

Moreover, the importance of the amino acid 184 of a  $\beta$ -glucosidase obtained from a marine microbial metagenomic library had been revealed earlier through the H184F mutation (Liu et al., 2011). The residue H184 (corresponding to H181 in *T. Brockii*  $\beta$ -glucosidase) appears in the majority of the tolerant  $\beta$ -glucosidases reported. The mutation H184F showed an increase in the glucose tolerance when the substrate was pNPG (4-nitrophenyl- $\beta$ -D-glucopyranoside; Liu et al., 2011). Indeed, phenylalanine is the second more common residue in this position based on multiple alignments (**Figure S4**). We detected that a histidine performs more contacts with the cellobiose than a phenylalanine in the same position. However, there is no report of the impact of this mutation on the glucose tolerance when the substrate was cellobiose. Hence, more experiments are necessary to make inferences.

## SLR evaluation

Although SLR was used originally to medical research, they have been used to perform unbiased reviews in various areas (Kitchenham, 2004). Recently, several reviews have covered: i) general aspects of  $\beta$ -glucosidases (Cairns and Esen, 2010); ii) efficiency of  $\beta$ -glucosidases from fungus (Tiwari et al., 2013); iii) the role of  $\beta$ -glucosidases in the

hydrolysis of cellulose (Singhania et al., 2013); iv) reduction of product inhibition during enzymatic lignocellulose hydrolysis (Andrić et al., 2010); and v) advances in enzymes for lignocellulosic biomass conversion (Saha and Bothast, 1997). However, to our knowledge, this is the first report of an SLR that analyzes the role of glucose tolerance in  $\beta$ -glucosidases for biofuel production focusing on bioinformatics analysis.

Most of the glucose-tolerant  $\beta$ -glucosidases reported in this SLR belong to the GH1 family (21 of 23). Indeed,  $\beta$ -glucosidases from GH1 family have been reported to be 10- to 1000-fold more glucose tolerant than the ones from the GH3 family (de Giuseppe et al., 2014). The potential for industrial use of  $\beta$ -glucosidase enzymes belonging to the GH1 family has been highlighted due to its broad substrate specificity and weak inhibition by glucose (Cota et al., 2015). Although we have detected two glucose-tolerant GH3  $\beta$ -glucosidases that showed activity in glucose concentrations of 140 mM (Huang et al., 2014) and 400 mM (Ramani et al., 2015b), this SLR confirms that GH1  $\beta$ -glucosidases are more promising for the second-generation biofuel production. However, more works with GH3  $\beta$ -glucosidases in the future may bring new conclusions.

In the SLR, we detected that glucose inhibition constant ( $K_i$ ) was the main parameter used to measure product inhibition. For example, the  $\beta$ -glucosidase from *Bacillus subtilis* showed the highest glucose  $K_i$  value (1.9 M) among the sequences collected (**Table S2**). When the  $K_i$  for glucose was not available, the papers reported another tolerance parameter, such as  $IC_{50}$ . For instance, the *Mucor circinelloides*  $\beta$ -glucosidase retained 84% activity at glucose concentrations up to 140 mM (Huang et al., 2014), a high value for a  $\beta$ -glucosidase from GH3 family. It has been suggested that an ideal  $\beta$ -glucosidase could be obtained by improving the tolerance of GH3 or improving the specificity constant ( $k_{cat} / K_m$ ) of the GH1 (Cao et al., 2015). Moreover, it has been reported two  $\beta$ -glucosidases from the GH1 and GH3 families present in the same operon of *Thermoanaerobacter brockii* (Breves et al., 1997). Hence, the diversity of  $\beta$ -glucosidases from different families may be beneficial for the organism to perform the biomass degradation.

### Bias risk

The different methods used to measure the glucose inhibition may result in a risk of bias for the SLR. The inhibition constant ( $K_i$ ) is an effective metric to measure the inhibition. However, it is harder to be determined experimentally. For this reason, the majority of the works has preferred to use  $IC_{50}$  to determine the glucose inhibition in  $\beta$ -glucosidases. In this SLR, we noticed that the  $IC_{50}$  had been determined by different methods, which can be a problem for comparisons among the results. Comparisons among  $IC_{50}$  results would depend on enzymatic assays, different from  $K_i$ . Also,  $IC_{50}$  may vary at different enzyme concentrations, even if the parameters considered to measure the enzyme activity were the same. In the various studies accessed in this SLR, enzymatic concentrations used in  $IC_{50}$  assays were not the same and were only slightly related.

Another risk of bias concerns substrates used in glucose inhibition tests. Although several works have presented  $\beta$ -glucosidases with resistance at high glucose concentrations and affinity tests with cellobiose, most studies have used pNPG (4-nitrophenyl- $\beta$ -D-glucopyranoside) as a substrate for glucose tolerance tests. The pNPG substrate concentration can be measured without interference from glucose added, which is more convenient for bench experiments. However, for biofuel production, the  $\beta$ -glucosidase natural substrate is

cellobiose (Bohlin et al., 2010). The inhibition constant for glucose and substrate specificity constant for cellobiose are the most important metrics to select  $\beta$ -glucosidases (Teugjas and Väljamäe, 2013).

## CONCLUSIONS

In summary, we performed a broad search in the literature guided by the restrictions of a systematic literature review. We selected 27 papers that report the state-of-art of glucose-tolerant  $\beta$ -glucosidases used in the production of second-generation biofuel and collected structural data. We constructed a database with five three-dimensional structure files, 23 glucose-tolerant  $\beta$ -glucosidases sequences, and their kinetic information, and also modeled by homology 18 glucose-tolerant  $\beta$ -glucosidase sequences. We identified the most conserved residues in the catalytic pocket and the residues that perform more and fewer contacts with the cellobiose. This information may be substantial for understanding the  $\beta$ -glucosidases mechanisms, identifying sites for mutations and engineering novel and more efficient  $\beta$ -glucosidases for biofuel production. Based on the SLR results, we conclude that site-directed mutagenesis seems to be a great strategy to produce more efficient  $\beta$ -glucosidases for biofuel production. For this reason, it is important to determine the residues related to the glucose tolerance. We detected that the residues H121, N166, E167, Y299, E355, W402, W122, N297, E409, W410, and F418 are highly conserved, and for this reason, they probably are critical for substrate recognizing. We also detected an apparent co-occurrence of the residue C170 when both residues W169 and L174 appear, which may be related to the glucose tolerance. We also verify that the residue T226 may be essential for the glucose tolerance. In addition, we detected that the residue H181, which has been described as important for glucose tolerance, perform more contacts with the cellobiose than when is substituted by phenylalanine. However, the use of experimental methods is necessary to infer about the importance of this residue for GTBGL. We also proposed that the number of contacts of a residue in a determined position could be used as a metric to suggest mutation in non-tolerant  $\beta$ -glucosidases.

To the best of our knowledge, this is the first report of an SLR about glucose tolerance in  $\beta$ -glucosidases and a database of glucose-tolerant  $\beta$ -glucosidases. The results of this work may be useful for *in silico*, *in vitro* and *in vivo* experiments and may help shed light on the production of second-generation biofuel. The sequences and structural data were organized in a database, called betagdb, which is available at: <<http://bioinfo.dcc.ufmg.br/betagdb>>.

## ACKNOWLEDGMENTS

The authors thank the funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The authors thank Larissa Leijôto for the help in the conception of Figure 4.

## REFERENCES

- Akram F, Haq I ul, Khan MA, Hussain Z, Mukhtar H, et al. (2016). Cloning with kinetic and thermodynamic insight of a novel hyperthermostable  $\beta$ -glucosidase from *Thermotoga naphthophila* RKU-10T with excellent glucose tolerance. *J. Mol. Catal. B Enzym.* 124: 92-104.

- Allen WJ, Balias TE, Mukherjee S, Brozell SR, et al. (2015). DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* 36: 1132-1156 <https://doi.org/10.1002/jcc.23905>.
- Andrić P, Meyer AS, Jensen PA and Dam-Johansen K (2010). Reactor design for minimizing product inhibition during enzymatic lignocellulose hydrolysis: I. Significance and mechanism of cellobiose and glucose inhibition on cellulolytic enzymes. *Biotechnol. Adv.* 28: 308-324 <https://doi.org/10.1016/j.biotechadv.2010.01.003>.
- Bai A, Zhao X, Jin Y, Yang G, et al. (2013). A novel thermophilic  $\beta$ -glucosidase from *Caldicellulosiruptor bescii*: Characterization and its synergistic catalysis with other cellulases. *J. Mol. Catal. B Enzym.* 85-86: 248-256.
- Badal CS and Bothast RJ (1997). Enzymes in Lignocellulosic Biomass Conversion. In *Fuels and Chemicals from Biomass* (Vol. 666, pp. 46-56). American Chemical Society.
- Barrett T, Suresh CG, Tolley SP, Dodson EJ, et al. (1995). The crystal structure of a cyanogenic beta-glucosidase from white clover, a family 1 glycosyl hydrolase. *Struct. Lond. Engl.* 3: 951-960.
- Béguin P and Aubert JP (1994). The biological degradation of cellulose. *FEMS Microbiol. Rev.* 13: 25-58. <https://doi.org/10.1111/j.1574-6976.1994.tb00033.x>
- Benoliel B, Poças-Fonseca MJ, Torres FAG and de Moraes LMP (2010). Expression of a Glucose-tolerant  $\beta$ -glucosidase from *Humicola grisea* var. *thermoidea* in *Saccharomyces cerevisiae*. *Appl. Biochem. Biotechnol.* 160: 2036-2044.
- Berman HM, Westbrook J, Feng Z, Gilliland G, et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28: 235-242. <https://doi.org/10.1093/nar/28.1.235>
- Bitar M and Franco GR (2014). A basic protein comparative three-dimensional modeling methodological workflow theory and practice. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 11: 1052-1065. <https://doi.org/10.1109/TCBB.2014.2325018>
- Bohlin C, Olsen SN, Morant MD, Patkar S, et al. (2010). A comparative study of activity and apparent inhibition of fungal  $\beta$ -glucosidases. *Biotechnol. Bioeng.* 107: 943-952 <https://doi.org/10.1002/bit.22885>.
- Breves R, Bronnenmeier K, Wild N, Lottspeich F, et al. (1997). Genes encoding two different beta-glucosidases of *Thermoanaerobacter brockii* are clustered in a common operon. *Appl. Environ. Microbiol.* 63: 3902-3910.
- Cao LC, Wang ZJ, Ren GH, Kong W, et al. (2015). Engineering a novel glucose-tolerant  $\beta$ -glucosidase as supplementation to enhance the hydrolysis of sugarcane bagasse at high glucose concentration. *Biotechnol. Biofuels* 8: 202 <https://doi.org/10.1186/s13068-015-0383-z>.
- Chamoli S, Kumar P, Navani NK and Verma AK (2016). Secretory expression, characterization and docking study of glucose-tolerant  $\beta$ -glucosidase from *B. subtilis*. *Int. J. Biol. Macromol.* 85: 425-433 <https://doi.org/10.1016/j.jbiomac.2016.01.001>.
- Cota J, Corrêa TLR, Damásio ARL, Diogo JA, et al. (2015). Comparative analysis of three hyperthermophilic GH1 and GH3 family members with industrial potential. *N. Biotechnol.* 32: 13-20 <https://doi.org/10.1016/j.nbt.2014.07.009>.
- Crespim E, Zanphorlin LM, de Souza FHM, Diogo JA, et al. (2016). A novel cold-adapted and glucose-tolerant GH1  $\beta$ -glucosidase from *Exiguobacterium antarcticum* B7. *Int. J. Biol. Macromol.* 82: 375-380 <https://doi.org/10.1016/j.jbiomac.2015.09.018>.
- de Giuseppe PO, Souza T de ACB, Souza FHM, Zanphorlin LM, et al. (2014). Structural basis for glucose tolerance in GH1  $\beta$ -glucosidases. *Acta Crystallogr. D Biol. Crystallogr.* 70: 1631-1639 <https://doi.org/10.1107/S1399004714006920>.
- DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, et al. (1988). Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* 31: 722-729. <https://doi.org/10.1021/jm00399a006>
- Ewing TJA and Kuntz ID (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* 18: 1175-1189 [https://doi.org/10.1002/\(SICI\)1096-987X\(19970715\)18:9<1175::AID-JCC6>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1096-987X(19970715)18:9<1175::AID-JCC6>3.0.CO;2-O).
- Fang Z, Fang W, Liu J, Hong Y, et al. (2010). Cloning and characterization of a  $\beta$ -glucosidase from marine microbial metagenome with excellent glucose tolerance. *J. Microbiol. Biotechnol.* 20: 1351-1358.
- Gumerov VM, Rakitin AL, Mardanov AV, Ravin NV, et al. (2015). A novel highly thermostable multifunctional beta-glucosidase from crenarchaeon *Acidilobus saccharovorans*, A novel highly thermostable multifunctional beta-glucosidase from crenarchaeon *Acidilobus saccharovorans*. *Archaea* 2015: e978632.
- Guo B, Amano Y and Nozaki K (2016). Improvements in glucose sensitivity and stability of *Trichoderma reesei*  $\beta$ -glucosidase using site-directed mutagenesis. *PLoS One* 11: e0147301 <https://doi.org/10.1371/journal.pone.0147301>.
- Henrissat B (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* 280: 309-316 <https://doi.org/10.1042/bj2800309>.
- Huang Y, Busk PK, Grell MN, Zhao H, et al. (2014). Identification of a  $\beta$ -glucosidase from the *Mucor circinelloides* genome by peptide pattern recognition. *Enzyme Microb. Technol.* 67: 47-52 <https://doi.org/10.1016/j.enzmictec.2014.09.002>.
- Jabbour D, Klippel B and Antranikian G (2012). A novel thermostable and glucose-tolerant  $\beta$ -glucosidase from

- Fervidobacterium islandicum*. *Appl. Microbiol. Biotechnol.* 93: 1947-1956 <https://doi.org/10.1007/s00253-011-3406-0>.
- Jakalian A, Jack DB and Bayly CI (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 23: 1623-1641 <https://doi.org/10.1002/jcc.10128>.
- Jeng W-Y, Wang N-C, Lin C-T, Chang W-J, et al. (2012). High-resolution structures of *Neotermes koshunensis*  $\beta$ -glucosidase mutants provide insights into the catalytic mechanism and the synthesis of glucoconjugates. *Acta Crystallogr. D Biol. Crystallogr.* 68: 829-838 <https://doi.org/10.1107/S0907444912013224>.
- Johnson M, Zaretskaya I, Raytselis Y, Merezukh Y, et al. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.*, 36(Web Server issue), W5-9.
- Cairns JR and Esen A (2010).  $\beta$ -Glucosidases. *Cell. Mol. Life Sci.* 67: 3389-3405 <https://doi.org/10.1007/s00018-010-0399-2>.
- Khairudin NBA and Mazlan NSF (2013). Molecular docking study of Beta-glucosidase with cellobiose, cellotetraose and cellotriose. *Bioinformation* 9: 813-817 <https://doi.org/10.6026/97320630009813>.
- Kitchenham B (2004). Procedures for Performing Systematic Reviews. Keele University, Keele 33: 1-26.
- Kumar R, Singh S and Singh OV (2008). Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. *J. Ind. Microbiol. Biotechnol.* 35: 377-391 <https://doi.org/10.1007/s10295-008-0327-8>.
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* 6: e1000100 <https://doi.org/10.1371/journal.pmed.1000100>.
- Liu J, Zhang X, Fang Z, Fang W, et al. (2011). The 184th residue of  $\beta$ -glucosidase Bgl1B plays an important role in glucose tolerance. *J. Biosci. Bioeng.* 112: 447-450 <https://doi.org/10.1016/j.jbiosc.2011.07.017>.
- Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PIW, et al. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* 50: 437-450 <https://doi.org/10.1002/prot.10286>.
- Lu J, Du L, Wei Y, Hu Y, et al. (2013). Expression and characterization of a novel highly glucose-tolerant  $\beta$ -glucosidase from a soil metagenome. *Acta Biochim. Biophys. Sin.* 45: 664-673.
- Maier JA, Martinez C, Kasavajhala K, Wickstrom L, et al. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* 11: 3696-3713 <https://doi.org/10.1021/acs.jctc.5b00255>.
- Mariano D, Leite C, Santos LHS, Rocha REO, et al. (2017). A guide to performing systematic literature reviews in bioinformatics (Technical Report No. RT.DCC.002/2017) (p. 63). Belo Horizonte, MG, Brazil: Universidade Federal de Minas Gerais.
- Meleiro LP, Salgado JCS, Maldonado RF, Alponi JS, et al. (2015). A *Neurospora crassa*  $\beta$ -glucosidase with potential for lignocellulose hydrolysis shows strong glucose tolerance and stimulation by glucose and xylose. *J. Mol. Catal., B Enzym.* 122: 131-140 <https://doi.org/10.1016/j.molcatb.2015.09.003>.
- Meng EC, Shoichet BK and Kuntz ID (1992). Automated docking with grid-based energy evaluation. *J. Comput. Chem.* 13: 505-524 <https://doi.org/10.1002/jcc.540130412>.
- Murphy L, Bohlin C, Baumann MJ, Olsen SN, et al. (2013). Product inhibition of five *Hypocrea jecorina* cellulases. *Enzyme Microb. Technol.* 52: 163-169 <https://doi.org/10.1016/j.enzmictec.2013.01.002>.
- Pearson WR (2016). Finding Protein and Nucleotide Similarities with FASTA. *Curr. Protoc. Bioinforma.* Ed. Board Andreas Baxevanis AI, 53, 3.9.1-3.9.25.
- Pei J, Pang Q, Zhao L, Fan S, et al. (2012). *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase: a glucose-tolerant enzyme with high specific activity for cellobiose. *Biotechnol. Biofuels* 5: 31. <https://doi.org/10.1186/1754-6834-5-31>
- Pentzold S, Zagrobelyny M, Rook F and Bak S (2014). How insects overcome two-component plant chemical defence: plant  $\beta$ -glucosidases as the main target for herbivore adaptation. *Biol. Rev. Camb. Philos. Soc.* 89: 531-551 <https://doi.org/10.1111/brv.12066>.
- Pettersen EF, Goddard TD, Huang CC, Couch GS, et al. (2004). UCSF Chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25: 1605-1612 <https://doi.org/10.1002/jcc.20084>.
- Pires DEV, de Melo-Minardi RC, da Silveira CH, Campos FF, et al. (2013). aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics* 29: 855-861 <https://doi.org/10.1093/bioinformatics/btt058>.
- Ramani G, Meera B, Rajendhran J and Gunasekaran P (2015a). Transglycosylating glycoside hydrolase family 1  $\beta$ -glucosidase from *Penicillium funiculosum* NCL1: Heterologous expression in *Escherichia coli* and characterization. *Biochem. Eng. J.* 102: 6-13 <https://doi.org/10.1016/j.bej.2015.03.018>.
- Ramani G, Meera B, Vanitha C, Rajendhran J, et al. (2015b). Molecular cloning and expression of thermostable glucose-tolerant  $\beta$ -glucosidase of *Penicillium funiculosum* NCL1 in *Pichia pastoris* and its characterization. *J. Ind. Microbiol. Biotechnol.* 42: 553-565 <https://doi.org/10.1007/s10295-014-1549-6>.

- Rego N and Koes D (2015). 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* 31: 1322-1324 <https://doi.org/10.1093/bioinformatics/btu829>.
- Saha BC and Bothast RJ (1996). Production, purification, and characterization of a highly glucose-tolerant novel beta-glucosidase from *Candida peltata*. *Appl. Environ. Microbiol.* 62: 3165-3170.
- Saha BC and Bothast RJ (1997). Enzymes in Lignocellulosic Biomass Conversion. *Fuels and Chemicals from Biomass* 666: 46-56.
- Sanz-Aparicio J, Hermoso JA, Martínez-Ripoll M, Lequerica JL, et al. (1998). Crystal structure of beta-glucosidase a from *Bacillus polymyxa*: insights into the catalytic activity in family 1 glycosyl hydrolases. *J. Mol. Biol.* 275: 491-502 <https://doi.org/10.1006/jmbi.1997.1467>.
- Shatsky M, Nussinov R and Wolfson HJ (2004). A method for simultaneous alignment of multiple protein structures. *Proteins* 56: 143-156 <https://doi.org/10.1002/prot.10628>.
- Sievers F, Wilm A, Dineen D, Gibson TJ, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7: 539 <https://doi.org/10.1038/msb.2011.75>.
- Singhania RR, Patel AK, Sukumaran RK, Larroche C, et al. (2013). Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. *Bioresour. Technol.* 127: 500-507 <https://doi.org/10.1016/j.biortech.2012.09.012>.
- Schröder C, Elleuche S, Blank S and Antranikian G (2014). Characterization of a heat-active archaeal  $\beta$ -glucosidase from a hydrothermal spring metagenome. *Enzyme Microb. Technol.* 57: 48-54.
- Souza FHM, Meleiro LP, Machado CB, Zimbaridi ALRL, et al. (2014). Gene cloning, expression and biochemical characterization of a glucose- and xylose-stimulated  $\beta$ -glucosidase from *Humicola insolens* RP86. *J. Mol. Catal., B Enzym.* 106: 1-10 <https://doi.org/10.1016/j.molcatb.2014.04.007>.
- Swangkeaw J, Vichitphan S, Butzke CE and Vichitphan K (2010). Characterization of  $\beta$ -glucosidases from *Hanseniaspora* sp. and *Pichia anomala* with potentially aroma-enhancing capabilities in juice and wine. *World J. Microbiol. Biotechnol.* 27: 423-430 <https://doi.org/10.1007/s11274-010-0474-8>.
- Teugjas H and Väljamäe P (2013). Selecting  $\beta$ -glucosidases to support cellulases in cellulose saccharification. *Biotechnol. Biofuels* 6: 105. <https://doi.org/10.1186/1754-6834-6-105>
- Tiwari P, Misra BN and Sangwan NS (2013).  $\beta$ -Glucosidases from the fungus trichoderma: an efficient cellulase machinery in biotechnological applications. *BioMed Res. Int.* 2013: 203735 <https://doi.org/10.1155/2013/203735>.
- Uchima CA, Tokuda G, Watanabe H, Kitamoto K, et al. (2011). Heterologous expression and characterization of a glucose-stimulated  $\beta$ -glucosidase from the termite *Neotermes koshunensis* in *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* 89: 1761-1771 <https://doi.org/10.1007/s00253-010-2963-y>.
- Uchima CA, Tokuda G, Watanabe H, Kitamoto K, et al. (2012). Heterologous expression in *Pichia pastoris* and characterization of an endogenous thermostable and high-glucose-tolerant  $\beta$ -glucosidase from the termite *Nasutitermes takasagoensis*. *Appl. Environ. Microbiol.* 78: 4288-4293 <https://doi.org/10.1128/AEM.07718-11>.
- Uchiyama T, Yaoi K and Miyazaki K (2015). Glucose-tolerant  $\beta$ -glucosidase retrieved from a Kusaya gravity metagenome. *Front. Microbiol.* 6: 548 <https://doi.org/10.3389/fmicb.2015.00548>.
- Webb B and Sali A (2014). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* Ed. Board Andreas Baxevasis AI, 47, 5.6.1-32.
- Yang F, Yang X, Li Z, Du C, et al. (2015a). Overexpression and characterization of a glucose-tolerant  $\beta$ -glucosidase from *T. aotearoense* with high specific activity for cellobiose. *Appl. Microbiol. Biotechnol.* 99: 8903-8915 <https://doi.org/10.1007/s00253-015-6619-9>.
- Yang Y, Zhang X, Yin Q, Fang W, et al. (2015b). A mechanism of glucose tolerance and stimulation of GH1  $\beta$ -glucosidases. *Sci. Rep.* 5: 17296 <https://doi.org/10.1038/srep17296>.
- Zhao L, Pang Q, Xie J, Pei J, et al. (2013). Enzymatic properties of *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase fused to *Clostridium cellulovorans* cellulose binding domain and its application in hydrolysis of microcrystalline cellulose. *BMC Biotechnol.* 13: 101. <https://doi.org/10.1186/1472-6750-13-101>

## Supplementary material

[Figure S1](#). Redocking of cellobiose using the flexible ligand docking protocol.

[Figure S2](#). Crystallographic structures and models of  $\beta$ -glucosidases.

[Figure S3](#). Docking of cellobiose using the flexible ligand docking protocol in all 21 structures.

[Figure S4](#). Alignment among the catalytic pocket of glucose-tolerant  $\beta$ -glucosidases.

[Table S1](#). PRISMA checklist.

[Table S2](#). List of references, with main topics, groups, and scores.

[Table S3](#). List of glucose-tolerant  $\beta$ -glucosidases collected by organism.

[Table S4](#). Number of contacts between residues and the cellobiose docked in glucose-tolerant GH1  $\beta$ -glucosidases.

[Table S5](#). Matrix of the identity percentage among sequences using global alignment (algorithm Needleman-Wunsch).



## **Apêndice 7. Artigo 2**

**Título:** *“A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV)”*

**Journal:** *International Journal of Molecular Sciences*

**Fator de impacto (2017):** 4,183

**Ano de publicação:** 2019



Article

# A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV)

Diego César Batista Mariano <sup>1,\*</sup>, Lucianna Helene Santos <sup>1</sup>, Karina dos Santos Machado <sup>2</sup>, Adriano Velasque Werhli <sup>2</sup>, Leonardo Henrique França de Lima <sup>3</sup> and Raquel Cardoso de Melo-Minardi <sup>1</sup>

<sup>1</sup> Laboratório de Bioinformática e Sistemas (LBS), Department of Computer Science, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, Brazil; luciannahss@gmail.com (L.H.S.); raquelcm@dcc.ufmg.br (R.C.d.M.-M.)

<sup>2</sup> Laboratório de Biologia Computacional (COMBI-L). Centro de Ciências Computacionais-C3, Universidade Federal do Rio Grande, 96203-900 Rio Grande, Brazil; karinaecomp@gmail.com (K.d.S.M.); werhli@gmail.com (A.V.W.)

<sup>3</sup> Laboratório de Modelagem Molecular e Bioinformática (LAMMB), Departamento de Ciências Exatas e Biológicas (DECEB). Universidade Federal de São João Del-Rei, Campus Sete Lagoas, 35701-970 Sete Lagoas, Brazil; leofrancelima@ufsj.edu.br

\* Correspondence: diegomariano@ufmg.br; Tel.: +55-31-3409-5896

Received: 20 November 2018; Accepted: 6 January 2019; Published: 15 January 2019



**Abstract:** With the use of genetic engineering, modified and sometimes more efficient enzymes can be created for different purposes, including industrial applications. However, building modified enzymes depends on several in vitro experiments, which may result in the process being expensive and time-consuming. Therefore, computational approaches could reduce costs and accelerate the discovery of new technological products. In this study, we present a method, called structural signature variation (SSV), to propose mutations for improving enzymes' activity. SSV uses the structural signature variation between target enzymes and template enzymes (obtained from the literature) to determine if randomly suggested mutations may provide some benefit for an enzyme, such as improvement of catalytic activity, half-life, and thermostability, or resistance to inhibition. To evaluate SSV, we carried out a case study that suggested mutations in  $\beta$ -glucosidases: Essential enzymes used in biofuel production that suffer inhibition by their product. We collected 27 mutations described in the literature, and manually classified them as beneficial or not. SSV was able to classify the mutations with values of 0.89 and 0.92 for precision and specificity, respectively. Then, we used SSV to propose mutations for Bgl1B, a low-performance  $\beta$ -glucosidase. We detected 15 mutations that could be beneficial. Three of these mutations (H228C, H228T, and H228V) have been related in the literature to the mechanism of glucose tolerance and stimulation in GH1  $\beta$ -glucosidase. Hence, SSV was capable of detecting promising mutations, already validated by in vitro experiments, that improved the inhibition resistance of a  $\beta$ -glucosidase and, consequently, its catalytic activity. SSV might be useful for the engineering of enzymes used in biofuel production or other industrial applications.

**Keywords:** enzymes; prediction of mutations; second-generation biofuel

## 1. Introduction

Enzymes, in most cases, are proteins that accelerate biochemical reactions. They have applications in several fields of the industry, such as the production of drugs, food, beverage, biofuel, and so

on [1,2]. Moreover, genetic engineering has been used to construct more efficient enzymes for industrial applications through mutations [3].

Techniques, such as error-prone PCR (epPCR), have been used to evaluate mutations systematically in several works. In this technique, a modified DNA polymerase inserts random mutations in the gene that codifies an enzyme during the replication process [4]. For instance, an epPCR library was used to identify three efficient mutations for an enzyme used in biofuel production. The combination of these mutations allowed the construction of a mutant enzyme that increased sugarcane bagasse conversion to fermentable sugars by 14–35% [3]. However, the proposal of modified enzymes depends on several *in vitro* and *in vivo* experiments, which may result in the process being expensive and time-consuming due to the vast number of possible mutations. For example, a protein with approximately 400 residues may present a total of  $20^{400}$  residue combinations, which corresponds to  $2.58 \times 10^{520}$  possible mutations. From all possible mutations, experimental techniques can evaluate only hundreds of them. Therefore, a previous selection with a computational method may reduce costs and allow a higher number of tests, with promising mutated enzymes.

When comparing proteins, sequence alignment is the most traditional computational method. It identifies similar regions between proteins using substitution matrices [5]. For instance, an approach based on protein sequence activity relationships (ProSAR) uses sequences to predict the contributions of mutations on protein functions [6,7]. However, it does not consider the impact of the three-dimensional structure or the physicochemical proprieties of the mutated residues, which may be a limitation when suggesting mutations. Another approach to propose mutations is the evaluation of the variation of free energy of Gibbs difference ( $\Delta\Delta G$ ) to analyze the thermostability of molecules. However, these computations are not feasible for all cases [8]. Hence, free energy calculations are not able to estimate with accuracy the impact of a mutation in an enzyme, the interaction with substrates and products, and the protein motion for more than a few examples. Hence, computational methods to propose and to evaluate mutations in enzymes at a large scale are still necessary.

Structural signatures, also called fingerprints, may be an alternative to analyze the impact of mutations as they provide a computationally feasible method to identify patterns of macromolecular structural features that may be important for structure and function. They have been successfully used in classification and automatic annotation of proteins [9,10], prediction of mutation effects on protein stability [11], prediction of the impact of mutations on the affinity between protein and ligands [12], and prediction of the mutation impact on the affinity between an antibody and antigen [13]. The aCSM (atomic Cutoff Scanning Matrix) method, based on structural signatures, calculates a structural signature, which is based on atomic pairwise distances, also considering their physicochemical properties [14]. It was also successfully used for the prediction of protein-ligand interactions. Hence, it may be used to characterize important regions that interact with the ligand.

In this paper, we propose a method based on structural signatures variation (SSV) to suggest mutations for improving the activity of enzymes. Our method can be applied to several types of enzymes. Despite the genericity of our method, we present a case study to demonstrate it and suggest mutations in  $\beta$ -glucosidase enzymes used in second-generation biofuel production. In addition, we carried out a comparative case study to analyze SSV performance to a similar structure-based approach called BioGPS [15].

## 2. Results

### 2.1. SSV Definition

The structural signature variation (SSV) method is based on computing Euclidean distances between signatures of: (i) A wild enzyme and an enzyme model with the most similar signature to the wild type (called wild template); and (ii) a mutant enzyme and an enzyme model with the most similar signature to the mutant (called mutant template). The difference between the two distances (herein called the  $\Delta\Delta SSV$  score) may be used to predict the impact of the mutation. The SSV method requires

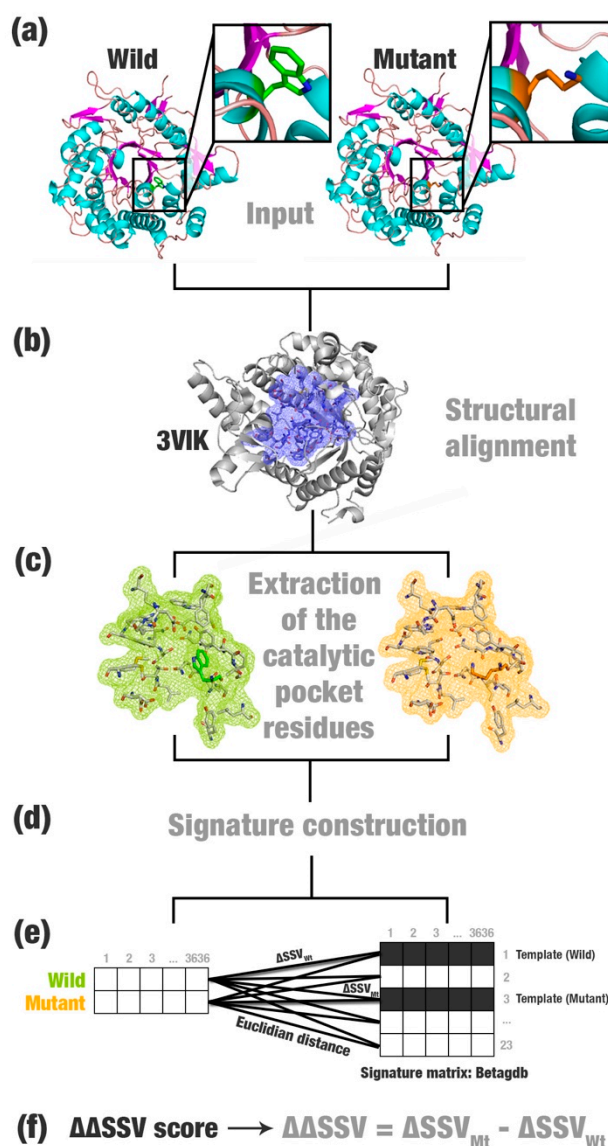
as input three-dimensional structures of a wild enzyme, a mutant enzyme (that can be modeled in silico), and enzyme models (herein called templates, i.e., proteins with positive characteristics that you want to transfer to other enzymes). SSV is computed using the following steps:

1. Most relevant residues' extraction: For wild, mutated, and templates' structures the most relevant residues are extracted and saved in a new Protein Data Bank (PDB) file (Figure 1a–c). This selection depends on the application and can be modified according to users' needs. This step is optional.
2. Structural signature construction: For every PDB file, we compute a vector with the cumulative distribution of the pairwise distances among all pairs of atoms and their physicochemical properties (aCSM algorithm) (Figure 1d).
3. Template definition: A template definition depends on a high-curated database of enzymes with beneficial characteristics. This database should be manually and previously defined. We selected as a template, proteins with the closest signature to wild and mutant proteins analyzed (Figure 1e).
4. Comparison between signatures: A distance matrix among all signatures is constructed (a similar matrix is used to define the template). The Euclidean distance between two signatures is called signature variation ( $\Delta$ SSV). The Euclidean distance between signatures of a wild enzyme and its template is called ( $\Delta$ SSV<sub>Wt</sub>). The Euclidean distance between signatures of a mutant enzyme and its template is called ( $\Delta$ SSV<sub>Mt</sub>). The difference between both values is the  $\Delta$ SSV score. If the  $\Delta$ SSV score is lower than zero, the mutant's signature is more alike to the template signature than to the wild's signature, suggesting that the mutation is beneficial. If the  $\Delta$ SSV score is higher than zero, the mutant's signature is more distant from the template signature than from wild's signature, suggesting that the mutation is not beneficial (Figure 1f).

## 2.2. Case Study 1: Evaluating Mutations for $\beta$ -Glucosidase Collected from the Literature

$\beta$ -glucosidases (E.C. 3.2.1.21) are enzymes that perform the hydrolysis of glucosidic bonds, mainly in disaccharides [16,17]. They act in synergy with exoglucanases (E.C. 3.2.1.91) and endoglucanases (E.C. 3.2.1.4) in the second-generation biofuel production process. In the biomass degradation, endoglucanases attack the cellulose chain, releasing oligosaccharides of various lengths. Then, exoglucanases act, producing mainly cellobiose.  $\beta$ -glucosidases hydrolyze the cellobiose in two glucose molecules, which will be used in the fermentation process for ethanol production [18]. They have a key role in this process by removing the cellobiose, which is a potent inhibitor of exoglucanases and endoglucanases [19–23]. However, the majority of the known  $\beta$ -glucosidases has been described as being inhibited by high concentrations of glucose [24–27]. Hence, the production of  $\beta$ -glucosidases with a high tolerance for glucose inhibition may improve biofuel production [28].

To evaluate our method, we present a first case study for proposing mutations to improve the activity of  $\beta$ -glucosidase enzymes even in high glucose concentrations. We compared wild and mutant  $\beta$ -glucosidases with templates obtained in a manually curated database of glucose-tolerant  $\beta$ -glucosidases [29]. The database holds a group of  $\beta$ -glucosidases with high resistance to glucose inhibition and high industrial applications. However, few glucose-tolerant  $\beta$ -glucosidases have been described in the literature [30]. We hypothesized that glucose-tolerant and non-tolerant  $\beta$ -glucosidases have discriminant signatures. Hence, the signature of glucose-tolerant  $\beta$ -glucosidases previously characterized can be used to define if mutations in non-tolerant  $\beta$ -glucosidases make their signature similar to a tolerant  $\beta$ -glucosidase or not.



**Figure 1.** The structural signature variation (SSV) schema. (a) SSV receives as input PDB files containing mutant and wild protein (in this example, a  $\beta$ -glucosidase enzyme). (b) The structures are structurally superposed to a template (*Neotermes koshunensis*  $\beta$ -glucosidase; PDB ID: 3VIK). (c) The corresponding residues of the catalytic pocket in the wild, mutant, and their templates' structures are extracted. (d) Structural signatures for all files are computed using the aCSM (atomic Cutoff Scanning Matrix) algorithm. (e) The Euclidean distance is computed for every line of the templates' signature matrix. The lowest values define the templates for wild and mutant. (f) The difference between the two distances ( $\Delta\Delta\text{SSV}$ ) is calculated.

### 2.2.1. Data Collection and Manual Classification of Mutation Effects

We collected 27 mutations in  $\beta$ -glucosidases from the literature and the UniProt database (<https://uniprot.org>) (Table 1). Every mutation was manually classified as beneficial or not according to the impact description in the  $\beta$ -glucosidase activity. We classified as “beneficial” mutations that tend to improve the saccharification process, such as mutations reported as responsible for improving the glucose tolerance, increasing optimal temperature, increasing the catalytic efficiency, reducing the affinity for the product, or improving the affinity for the substrate. On the other hand, we classified as “not beneficial” mutations that tend to reduce the saccharification process, such as mutations reported as responsible for decreasing the affinity for the substrate, increasing the affinity for the product, or

reducing the catalytic activity. For example, the mutation, H228T, in the  $\beta$ -glucosidase, Bgl1B, has been described as responsible for improving the glucose tolerance [27]. Hence, we classified it as beneficial. On the other hand, the mutation, V168Y, in the human cytosolic  $\beta$ -glucosidase has been described as responsible for reducing the specific activity [31]. Hence, we classified it as not beneficial.

**Table 1.** Mutations collected from the literature and UniProt.

ID	Mutation	Effect	Classification	Source
1	H228T	Improves glucose tolerance.	Beneficial	[27]
2	V174C/A404V/L441F	Increases the optimal temperature of 50 °C to 60 °C, reduces the optimal pH of 6 to 5.5.	Beneficial	[3]
3	H184F	Increases the inhibition constant for glucose.	Beneficial	[32]
4	P172L	Increases catalytic efficiency.	Beneficial	[33]
5	P172L/F250A	Increases catalytic efficiency.	Beneficial	[33]
6	L167W	Increases the optimal temperature and glucose tolerance.	Beneficial	[33]
7	L167W/P172L	Increases the activity (2 $\times$ ).	Beneficial	[34]
8	L167W/P172L/P338F	Increases the activity (1,3 $\times$ ).	Beneficial	[34]
9	V168Y	Reduction in the specific activity.	Not beneficial	[31]
10	F225S	Reduction in the specific activity.	Not beneficial	[31]
11	Y308F	Reduction in the specific activity.	Not beneficial	[31]
12	Y308A	Reduction in the specific activity.	Not beneficial	[31]
13	I207V	Increases the specificity constant ( $K_{cat}/K_m$ ).	Beneficial	[35]
14	N218H	Decreases the $K_m$ about 2-fold.	Beneficial	[36]
15	N273V	Increases the $K_m$ about 5-fold.	Not beneficial	[36]
16	F252I	Reduces substrate affinity.	Not beneficial	[37]
17	F252W	Reduces substrate affinity.	Not beneficial	[37]
18	F252Y	Reduces substrate affinity.	Not beneficial	[37]
19	M284N	Reduction of $K_{cat}/K_m$ 7 to 30-fold depending on the substrate.	Not beneficial	[35]
20	H276M	Reduction of $K_{cat}/K_m$ 2 to 6-fold depending on the substrate.	Not beneficial	[38]
21	V173C	Decreases affinity for cellobiose.	Not beneficial	[39]
22	M177L	Decreases affinity for cellobiose (small reduction).	Not beneficial	[39]
23	D229N	Decreases affinity for cellobiose (high reduction).	Not beneficial	[39]
24	H231D	Decreases affinity for cellobiose.	Not beneficial	[39]
25	E96K	Improves the thermostability.	Beneficial	[40]
26	N223G	Reduction of transglycosylation, glucose tolerance, and activity.	Not beneficial	[41]
27	N223Q	Reduction of transglycosylation, glucose tolerance, and activity.	Not beneficial	[41]

### 2.2.2. Predicting the Impact of Mutations

We performed the SSV method (Figure 1), evaluated the  $\Delta\Delta$ SSV score for the 27 mutations in  $\beta$ -glucosidases, and compared them to the expected results. For mutations classified as beneficial, we expected a negative  $\Delta\Delta$ SSV score; and for mutations classified as not beneficial, in turn, a positive  $\Delta\Delta$ SSV score.

SSV predicted correctly eight in a total of nine beneficial mutations (Table 2). For the non-beneficial mutations, where the expected  $\Delta\Delta$ SSV was higher than zero, SSV predicted correctly 12 out of 18.

**Table 2.**  $\Delta\Delta$ SSV score expected and the value predicted by SSV.

ID	Mutation	$\Delta\Delta$ SSV Expected	$\Delta\Delta$ SSV Score	Hit
1	H228T	$\Delta\Delta$ SSV < 0	−186.18	✓
2	V174C/A404V/L441F	$\Delta\Delta$ SSV < 0	−246.22	✓
3	H184F	$\Delta\Delta$ SSV < 0	100.37	
4	P172L	$\Delta\Delta$ SSV < 0	−6.29	✓
5	P172L/F250A	$\Delta\Delta$ SSV < 0	−6.29	✓
6	L167W	$\Delta\Delta$ SSV < 0	−602.80	✓
7	L167W/P172L	$\Delta\Delta$ SSV < 0	−615.46	✓
8	L167W/P172L/P338F	$\Delta\Delta$ SSV < 0	−615.46	✓
9	V168Y	$\Delta\Delta$ SSV > 0	330.56	✓
10	F225S	$\Delta\Delta$ SSV > 0	−365.07	
11	Y308F	$\Delta\Delta$ SSV > 0	34.19	✓
12	Y308A	$\Delta\Delta$ SSV > 0	−108.62	
13	I207V	$\Delta\Delta$ SSV < 0	−71.56	✓
14	N218H	$\Delta\Delta$ SSV < 0	−230.61	✓
15	N273V	$\Delta\Delta$ SSV > 0	−55.26	
16	F252I	$\Delta\Delta$ SSV > 0	86.70	✓
17	F252W	$\Delta\Delta$ SSV > 0	129.97	✓
18	F252Y	$\Delta\Delta$ SSV > 0	37.86	✓
19	M284N	$\Delta\Delta$ SSV > 0	−127.35	
20	H276M	$\Delta\Delta$ SSV > 0	−501.32	
21	V173C	$\Delta\Delta$ SSV > 0	13.59	✓
22	M177L	$\Delta\Delta$ SSV > 0	20.86	✓
23	D229N	$\Delta\Delta$ SSV > 0	18.11	✓
24	H231D	$\Delta\Delta$ SSV > 0	−54.22	
25	E96K	$\Delta\Delta$ SSV < 0	−31.08	✓
26	N223G	$\Delta\Delta$ SSV > 0	39.37	✓
27	N223Q	$\Delta\Delta$ SSV > 0	264.34	✓

### 2.2.3. Comparison with Other Methods

We compared our method to the support vector machine (SVM) implemented on the Weka (Waikato Environment for Knowledge Analysis) tool [42]. SVM is a learning algorithm for classification. We performed four experiments: (i) SSV; (ii) SVM using as input only wild signatures; (iii) SVM using as input only mutant signatures; and (iv) SVM using as input the difference of the wild vector and mutant vector. For these experiments, we evaluated the following metrics: Precision, accuracy, specificity, sensibility, and the F-measure [43].

We observed that the precision and specificity of SSV were superior to the other method. SSV obtained a precision of 0.89 and a specificity of 0.92 (Table 3). It also performed better in the prediction of beneficial mutations than the SVM.

**Table 3.** Metrics used to evaluate SSV.

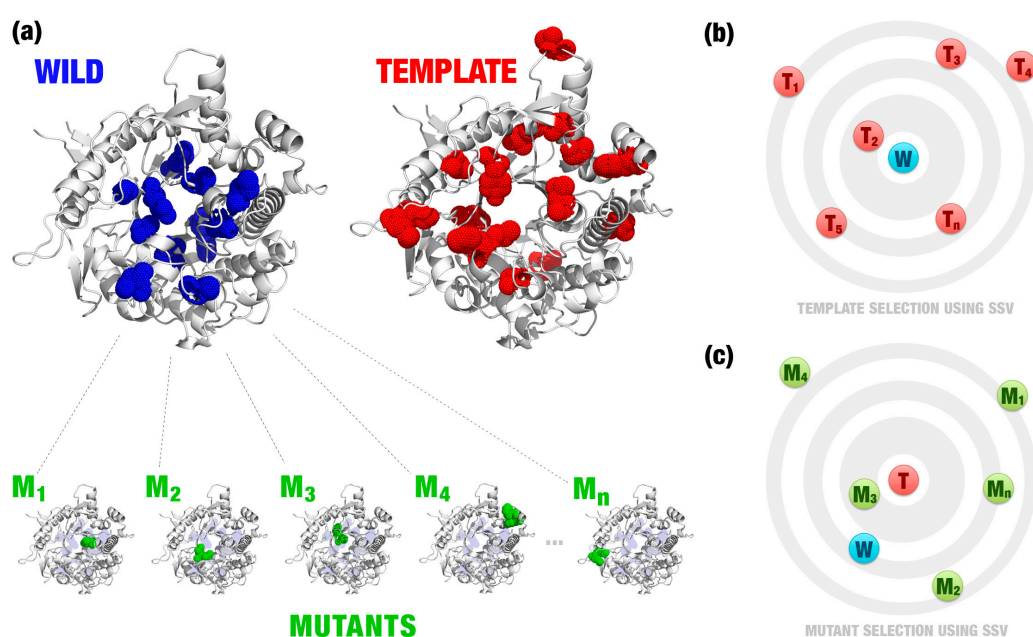
Metric	SSV	SVM (Wild)	SVM (Mutant)	SVM (Wild-Mutant)
Precision	0.89	0.64	0.36	0.36
Accuracy	0.74	0.81	0.74	0.74
Specificity	0.92	0.79	0.70	0.70
Sensitivity	0.57	0.88	1.00	1.00
F-measure	0.70	0.74	0.53	0.53

### 2.3. Case Study 2: Proposing Mutations for a Non-Tolerant $\beta$ -Glucosidase

In the second case study, we described a real application for the method SSV. We chose a non-tolerant  $\beta$ -glucosidase, Bgl1B (UniProt accession number: D0VEC8), to suggest mutations using SSV. Bgl1B was extracted from a marine metagenome and presented the half maximal inhibitory concentration (IC<sub>50</sub>) of 50 mM for glucose [44]. For comparison, Bgl1A, a glucose-tolerant  $\beta$ -glucosidase

also extracted from a marine metagenome, presented  $IC_{50}$  of 1000 mM [45]. In a recent study, several mutations for improving the activity in higher glucose concentrations were proposed for Bgl1B [27]. This study will be used to compare the results of the mutations proposed by the SSV method.

We modeled point mutations by homology for all residues of the catalytic pocket (composed by 22 residues around the active site). For each residue, 19 mutations were proposed, in a total of 418 mutants (Figure 2a). Then, we defined the template with the most similar signature (Figure 2b). Also, we used this template to evaluate the mutant that inserts more similar characteristics to the template (Figure 2c). Note that, in this example, the wild and template have a similar folding, but different sequences (Figure 2a). Wild (Bgl1B) and template (Bgl1A) have an identity of 55% (243 similar residues in a total of 443). Thus, it is necessary to evaluate hundreds of mutations to detect beneficial mutations using simply sequence alignment. SSV takes into consideration the changes in the protein environment, for example, changes in the residues volume, atoms distances, and their pharmacophoric properties.



**Figure 2.** (a) Wild and template have a similar folding, but differences in the sequence (illustrated by blue dots in the wild enzyme and by red dots in the template enzyme). Several point mutations were proposed for the wild enzyme (green dots). The template enzyme is defined based on a curated database of enzymes with desired characteristics (in this case study, Betagdb). For instance, in (b) the template, T2 was defined as the template (T) for the wild enzyme (W). SSV is illustrated by a two-dimensional visualization in (b) and (c). Euclidean distances between signatures of the wild/mutants and the template (signature variation) are used to define the best template (b) and mutant (c). In this example, the mutant, M3, was defined as the mutation that best inserts characteristics similar to the template (c). Images generated using PyMOL software (<http://pymol.org>).

After running SSV, we detected 86 mutations with negative  $\Delta\Delta SSV$  (available in the Supplementary File). In a real application, this could still be a high value of mutations for a bench test. Hence, we proposed additional steps to limit the number of promising mutations (a detailed description is available in the Section 4). We removed nine mutations that occurred in the residues, H125, N169, E170, Y298, E353, and W399, because they were conserved in 100% of glucose-tolerant  $\beta$ -glucosidases. We also removed 58 mutations indicated as being not allowed in the GH1 family by the SIFT (Sorting Intolerant From Tolerant) software [46]. SIFT uses the physical properties of amino acids and sequence homology to predict the effect of an amino acid substitution on the protein function. Then, we analyzed the mutation impact in the structure using mCSM (mutation Cutoff



Scanning Matrix) [11]. The mCSM software uses graph-based signatures to predict the effect of mutations in proteins. For 19 remaining mutations, mCSM considered four as highly destabilizing. In the end, 15 mutations were proposed for Bgl1B (Table 4). These mutations affect five residues: F172 (three mutations), G246 (two mutations), H228 (eight mutations), T299 (one mutation), and V227 (one mutation).

**Table 4.** Mutations proposed by the SSV method for the non-tolerant  $\beta$ -glucosidase, Bgl1B. Mutations underlined and in bold were found in the literature.

F172	G246	H228	T299	V227
F172I	G246S	H228A	T299S	V227M
F172K	G246T	<b><u>H228C</u></b>		
F172V		H228M		
		H228N		
		H228P		
		H228Q		
		<b><u>H228T</u></b>		
		<b><u>H228V</u></b>		

Experimental data is available in the literature for three proposed mutations: H228C, H228T, and H228V [27]. These single-point mutants keep the relative activity even in higher glucose concentrations than wild Bgl1B. This suggests that the SSV method can be promising to propose beneficial mutations for  $\beta$ -glucosidases.

#### 2.4. Case Study 3: Comparing to BioGPS Descriptors

In this case study, we compared SSV to the analysis performed in the BioGPS study [15]. BioGPS is a bioinformatics methodology for rational engineering of enzyme promiscuity that uses chemical, geometrical, and physical-chemical features of three-dimensional structures. BioGPS compares active sites' properties, taking into consideration more than the sequence structure. Therefore, we considered a similar approach to SSV.

In the BioGPS study, eight mutants experimentally evaluated (Table 5) for a lipase B from *Candida antarctica* (CaLB) were used to validate the method [47]. CaLB is a stable lipase that belongs to the serine-hydrolases super-family. The insertion of amidase activity in CaLB has many applications for the industry [15,47,48]. BioGPS classified the mutations based on the improvement factor (IF) referred to CaLB wild-type activity. The IF is equal to the amidase activity of the mutant, divided by the amidase activity of the CaLB wild [15]. We considered  $IF > 1$  as beneficial mutations, and  $IF < 1$  as not beneficial mutations (Table 5). Also, SSV considered the mutant, M8, as a possible neutral mutation for presenting an IF slightly over 1.

**Table 5.** CaLB's mutants evaluated by SSV for comparison to BioGPS. These values were obtained from references [15,47,48].

Mutant	Mutation	IF	Classification	$\Delta\Delta SSV$	Hit
M1	G39A/W104F/L278A	6.3	Beneficial	-841	✓
M2	G39A/T103G/L278A	3.8	Beneficial	-121	✓
M3	G39A/T103G/W104F/L278A	11.2	Beneficial	-841	✓
M4	G39A	2.8	Beneficial	150	
M5	G39A/L278A	3.3	Beneficial	-121	✓
M6	I189A	0.4	Not beneficial	-94	
M7	T40A	0.4	Not beneficial	40	✓
M8	T103G	1.1	Neutral/Beneficial	0	✓

We collected the residues presented in the region near the catalytic triad and ran SSV using M3 as the template (see the Section 4 for details). For the seven mutations validated by BioGPS, SSV

correctly predicted five (M3 was tested as a control experiment and should not be considered in the calculation of accuracy). However, this case study could present some biases that will be discussed in the next section.

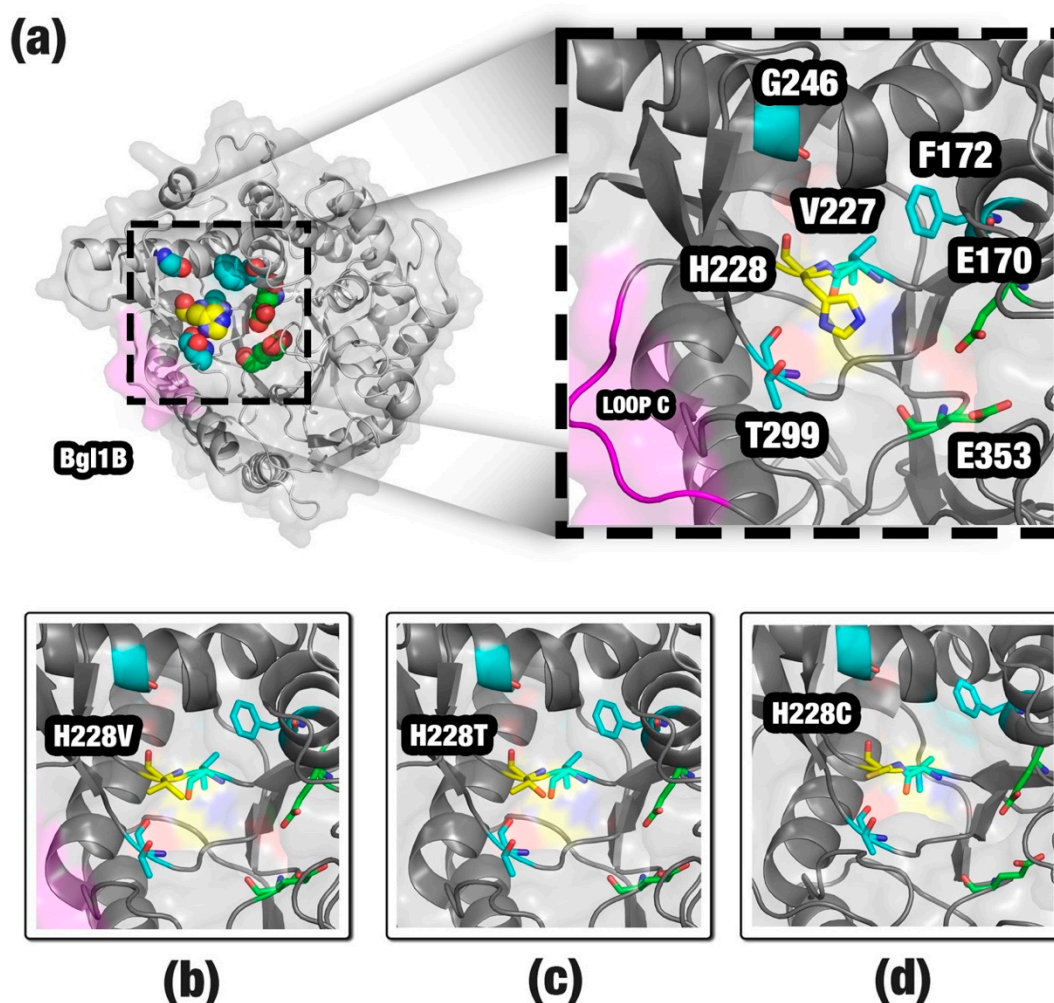
### 3. Discussion

We hypothesized that the more similar the signature of a  $\beta$ -glucosidase is to another  $\beta$ -glucosidase, classified as tolerant, the more they will preserve common characteristics. Hence, if a mutation turns the signature of a  $\beta$ -glucosidase more similar to the signature of a glucose-tolerant  $\beta$ -glucosidase, it might show comparable characteristics for biofuel production. The same could be inferred if the method was applied to another enzyme.

To validate our method, we collected 27 mutations from the literature, manually classified as beneficial or not, submitted it to three other methods, and compared it with the expected  $\Delta\Delta$ SSV score. We highlighted that our method does not have a direct competitor or another method that does exactly the same thing. Thus, three alternative methods based on SVM are proposed, which is the state of the art in machine learning for comparison. We attained 0.89 and 0.92 for the precision and specificity, respectively (Table 3). Precision is an appropriate metric to evaluate this case study as it emphasizes hits in beneficial mutations. This value of precision indicated that out of the nine beneficial mutations for  $\beta$ -glucosidases, SSV predicted eight correctly. The results showed that the Euclidean distance, implemented by SSV, achieved better results in the beneficial impact of mutation prediction than SVM (specificity and precision). However, SSV is not directly comparable to SVM. SSV is a simple strategy to model and compare the impact of mutations based on efficient proteins for a pre-established activity detected in nature. It uses the Euclidean distance to construct a score that will be used to compare structural signatures. SVM is a learning algorithm for supervised classification. In the case study, SVM received as input the structural signature matrix calculated by a step of the SSV method. We comprehend that this is not a straightforward comparison, but our intention is to demonstrate that our method is capable of classifying beneficial mutations correctly and achieves better results than using a model based on an SVM classifier.

#### 3.1. Improving the Activity of a Non-Tolerant $\beta$ -Glucosidase

A total of 15 mutations was proposed for improvement in the activity of Bgl1B (Figure 3a). The principal mutation site appeared to be the H228 residue. Our method proposed eight mutations for this site. Also, we found experimental data for three of these mutations: H228C, H228T, and H228V (Figure 3b–d). These mutations showed an activity improvement of Bgl1B even in higher glucose concentrations. Histidine is an amino acid classified as positively charged and bulky. The substitution of a histidine by an amino acid of a shorter side chain, such as cysteine, threonine, or valine, would provide a space that could allow a better allocation for glucose, agreeing with the study of Yang et al. [27]. Most of the other mutations proposed for H228 by SSV also provide a reduction in the side chain. Hence, we suggest that they could provide the same effect. The F172, G246, T299, and V227 residues are in the neighborhood of H228 (Figure 3a). We suppose that mutations in these sites could affect the exit pathway of the glucose from the active site. Also, these sites are near the loop C, a region in the entrance of the channel that guides to the active site. The geometrical differences around the loop C were described by Fang et al. [45] as being probably responsible for the characteristic of the glucose tolerance in  $\beta$ -glucosidase enzymes (Figure 3a). Taken together, the SSV results might indicate that our method was able to find some of the same beneficial mutations obtained by in vitro experiments and propose new ones to be tested.



**Figure 3.** Structure of Bgl1B (a), sites pointed by SSV as a target for mutations: H228 residue (yellow); F172, V227, G246, and T299 (cyan). The catalytic residues (E170 and E353) are shown in green. Additionally, loop C is in magenta. For comparison, we highlighted the mutations, H228V (b), H228T (c) and H228C (d), considered by the literature as being responsible for glucose tolerance. Images generated using PyMOL software (<http://pymol.org>).

### 3.2. Evaluating Mutations in CaLB

Using structural bioinformatics strategies for proposing mutations appears to complement sequence strategies. In general, methods based on sequences present a lower computational cost, such as the one implemented in ProSAR [6]. However, SSV is a method based on structural comparisons, with low computational costs. Other tools, models, and algorithms have been reported to use three dimensional structures with similar approaches to SSV to propose mutations, such as the active site constellations method [49], where distances between functional groups of the protein active site and the substrate are calculated and used as the template in a search for matches in structural databases, and BioGPS descriptors [15].

We analyzed eight mutations assessed in the BioGPS study using SSV. To construct our case study, we performed some modifications in the methodology. Ferrari et al. [15] used a database composed of 42 serine-hydrolases to construct the BioGPS fingerprint. However, the selection was performed according to their annotated E.C. number, which is a target of debate among the enzymologist community due to the lack of quality control. Despite the dubious quality, the authors considered the database as consistent to their research. However, SSV requires high-accuracy databases of templates.

Hence, we used the M3 mutant as a unique template. M3 was the mutant that inserted the highest value in the improvement factor in CaLB.

In addition, we evaluated the M3 mutant using the same file as the template for producing a control experiment. Indeed, the negative  $\Delta\Delta$ SSV value for the M3 mutation demonstrates that SSV correctly predicted the structural similarities between the mutant and template (Table 5).

From the SSV results, we could infer that W104F appear to be the most important mutation for improving the activity of CaLB. Although G39A presents some improvements in CaLB activity according to BioGPS, SSV was not able to detect the improvement. We can hypothesize that the substitution of a glycine by an alanine, the change of a hydrogen by a CH<sub>3</sub> group, is not sufficient to perform large modifications in the cumulative distribution of pairwise atoms calculated by aCSM. However, the substitution of a glycine could affect the mobility of secondary structures in the region, which would be detected using high-cost computation strategies, such as molecular dynamics. Indeed, the authors of BioGPS used 500 ns of molecular dynamics using the software, GROMACS [50], to construct and evaluate the mutants' fingerprints. Molecular dynamics have a high computational cost, and their use could make the assessment of mutations on a large scale not feasible.

Interestingly, the T103G mutation (found in M8) occurs in a region distant to the active site. For this reason, SSV predicted a neutral impact in the activity. Indeed, T103G proposed a slight improvement in the mutant activity (IF: 1.1), hence we consider this prediction correct.

The SSV mistake for mutant M6 could be related to the small number of elements in the template database. SSV depends on enzymes with efficient catalytic activities previously reported to be used as templates. For the  $\beta$ -glucosidase case study, we previously performed a systematic literature review, collected several mutations that were beneficial and not beneficial, and constructed a highly accurate database (Betagdb). However, a systematic literature review demands great effort, and the necessity to perform this kind of the previous study to construct a template database may be a negative point of the SSV approach.

Lastly, SSV presents a user-friendly interface, which could be easily run by users. Therefore, it could be used together with other strategies, such as BioGPS, ProSAR, or active site constellations, to aid in the proposition of more efficient mutations before performing in vitro experiments.

### 3.3. Important Issues before Using SSV

The use of SSV may present some drawbacks. First, the method depends on three-dimensional structure models to determine the structural signature. Models are obtained by computational heuristics and, for this reason, they can present differences to structures obtained by experimental methods, such as X-ray crystallography. However, achieving structures by experimental methods may be time consuming and expensive. Also, to propose mutations, SSV depends on templates with favorable characteristics, for example, mutations described in the literature, which are responsible for improvements in thermostability or catalytic activity, which may be hard to find.

SSV uses structural signature variations to detect patterns in enzymes with appropriate industrial applications and transfer them, testing random point mutations, to other enzymes that do not present similar behavior.

A final difficulty is a need for a curated database with positive and negative examples. In this work, we presented a case study, where we used a database obtained by a systematic literature review. Reviews like that take a long time to prepare and they are expensive. The SSV method may be reproduced using the three basic inputs: (i) A wild enzyme; (ii) a mutant of this enzyme; (iii) a template enzyme with positive characteristics for some industrial application that you desire to transmit for the mutant. Furthermore, we believe that in real scenarios, researchers involved in protein engineer processes should know interesting positive and negative examples to use as templates.

## 4. Materials and Methods

### 4.1. Method Description

#### 4.1.1. Extraction of the Catalytic Pocket

The residues of the catalytic pocket were collected from every  $\beta$ -glucosidase structure (Figure 1a). The catalytic pocket consists in the channel region that guides to the active site. This channel has been described as being responsible for the characteristic of glucose tolerance for  $\beta$ -glucosidases [51].

We extracted the residues up to 6.5 Å of the ligand using in-house scripts. This distance was selected based on a cutoff to characterize pockets for the structural signature [14]. Pires et al. [14] performed tests with 35,000 pockets to define how far from the ligand are the most important residues to construct a representative signature. They observed that all signature methods of aCSM present high  $p$ -values cutoff between 6.0 Å and 7.0 Å. Thus, they concluded that 6.0 Å was the best atomic cutoff for the pocket definition for their classification system. We extended the distance to 6.5 Å to include the corresponding residues to TRP169, an important amino acid for the glucose tolerance of  $\beta$ -glucosidases described in some studies [51].

We used, as a reference, the  $\beta$ -glucosidase in complex with cellobiose extracted from the termite, *Neotermes koshunensis* (PDB ID: 3VIK; [52]; Figure 1b). The residues of the 3VIK catalytic pocket are Q45, H148, W149, N192, S193, L195, T196, D199, M207, N253, I254, N255, Y273, N335, F336, Y337, T338, L340, W374, E402, W444, E451, W452, and F460. Then, we performed structural alignments between 3VIK and the  $\beta$ -glucosidases evaluated using the MultiProt tool [53] and selected the corresponding residues (Figure 1c). Optionally, the entire protein could be used in this step. However, we believe that calculating the signature of a specific region could improve the results.

#### 4.1.2. Structural Signature Construction

Structural signatures were constructed using aCSM [14]. The aCSM tool (UFMG, Belo Horizonte, Brazil) creates graph-based signatures to describe proteins. We used the version, aCSM-ALL, that also includes the pharmacophore classes: Hydrophobic, positively charged, negatively charged, hydrogen acceptor, hydrogen donor, aromatic, sulfur, and neutral. For each protein, aCSM-ALL calculates the pairwise distances among all pairs of atoms and constructs a distance matrix with the cumulative distribution. We used the cutoff range of 0 to 10 Å, and the cutoff step of 0.1 Å. For each protein, aCSM-ALL returns a vector with 3636 columns. The vector represents a unique structural signature, which may be used to identify the protein or compare it with other similar proteins.

In the aCSM-ALL matrix, the lines represent the protein, and the columns represent the cumulative distribution of pairwise atoms. Hence, for a cutoff of 0–10 Å and a step of 0.1 Å, aCSM-ALL calculates the number of atom pairs at cutoff distances of 0 to 0.1 Å, 0.1 to 0.2 Å, 0.2 to 0.3 Å, ( . . . ), 9.8 to 9.9 Å, and 9.9 to 10 Å. For example, a protein could present 100, 200, 50, 300, and 20 pairs of hydrophobic residues at cutoff distances of 2.0 to 2.1 Å, 3.0 to 3.1 Å, 5.3 to 5.4 Å, 7.4 to 7.5 Å, and 9.7 to 9.8 Å, respectively. All these numbers and other cutoffs were included in the matrix. Also, aCSM-ALL verified some combinations of residues, for instance, how many atom pairs of positively charged and negatively charged there were for all cutoffs' values. For this reason, each line of the aCSM-ALL matrix presented 3636 columns.

#### 4.1.3. Template Definition

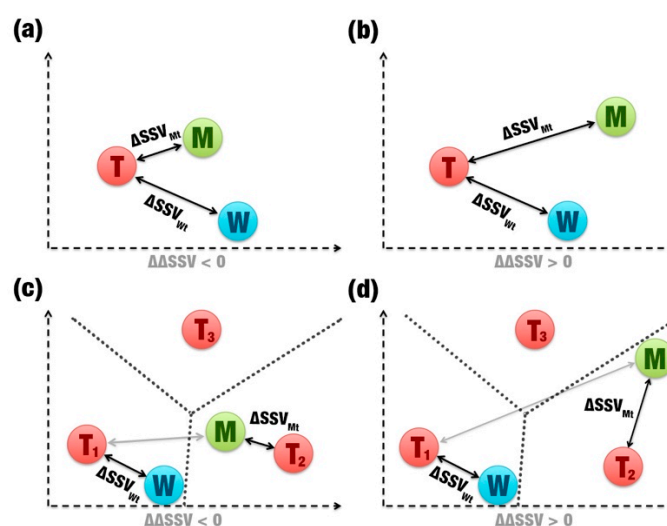
Templates are a three-dimensional structure of glucose-tolerant  $\beta$ -glucosidases that are used as models by SSV to define if mutations are beneficial or not. SSV depends on good templates to perform comparisons between signatures. Templates should be empirically selected based on the literature information.

We collected 23 PDB files of glucose-tolerant  $\beta$ -glucosidases from Betagdb (a list is available in the Supplementary File). Betagdb (<http://bioinfo.dcc.ufmg.br/betagdb>) is a database that contains

structures of  $\beta$ -glucosidases with high efficiency for biofuel production collected from a systematic literature review [29]. We previously calculated the structural signature of every glucose-tolerant  $\beta$ -glucosidase using the same parameters for wild and mutant signatures and stored it in the Betagdb signature matrix. We used the Euclidean distance to calculate the signature variation for each wild ( $\Delta\text{SSV}_{Wt}$ ) and mutant ( $\Delta\text{SSV}_{Mt}$ ) protein. The lowest value for the distance defines the template (Figure 1e). Wild and mutant  $\beta$ -glucosidases may have the same template or different templates.

#### 4.1.4. Comparison between Signatures

The  $\Delta\Delta\text{SSV}$  score is calculated from the comparison between signature variations (Figure 1f). This score is binary: If it is positive, the mutation is not beneficial (Figure 4b,d); if it is negative, the mutation is beneficial (Figure 4a,c). When wild and mutants have the same template (Figure 4a,b), SSV performs a simple distance comparison between the Euclidean distances of wild's and mutant's signatures to the template's signature. However, if a mutation causes a large change in the  $\beta$ -glucosidase signature, the mutant can show greater similarity in its signature to a second template (Figure 4c,d). The  $\Delta\Delta\text{SSV}$  is calculated using the difference of the distance variation for the mutant and the second template by the distance variation for the wild and the first template. In this case, the change in the signature is significant, which should indicate that the mutation is not beneficial. However, a significant signature change also can indicate that the mutant's signature is closer to another template. Therefore, high impacting mutations also may be beneficial (Figure 4c).



**Figure 4.** Two-dimensional representation of comparisons between signatures. (a) and (b) show simple comparisons (same template). (c) and (d) show comparisons with two different templates ( $T_1$ ,  $T_2$ , and  $T_3$ ). The gray arrows highlight that a second template was used. The dotted lines are used to show whether a mutation becomes more similar to a second template than the original. (a) and (c) represent beneficial mutations. (b) and (d) represent non-beneficial mutations.

#### 4.2. Case Study 1

We collected 27 mutations for  $\beta$ -glucosidases in the literature (Table S1), applied the calculations of signature variations, and evaluated the method's precision, accuracy, specificity, sensibility, and F-measure. Sequences were collected in the databases, GenBank (<http://www.ncbi.nlm.nih.gov/genbank>) and UniProt (<http://www.uniprot.org>). Three-dimensional structures were collected in the Protein Data Bank (PDB) [54]. The sequences without available three-dimensional structures were modeled by homology [55]. We selected the templates for modeling using the NCBI BLAST web interface [56] and built 100 models for each protein using MODELLER [57–59]. The best models were selected using the DOPE score. Mutations were modeled using the script for point mutations from MODELLER. For each of the 27 mutations, we extracted the catalytic pockets using in-house

scripts and constructed the structural signature. Then, we determined the templates and calculated the  $\Delta\Delta\text{SSV}$  score.

#### 4.3. Case Study 2

The sequence of Bgl1B was obtained in UniProt (accession number: D0VEC8). We constructed 100 models using MODELLER. We used as a model the GH1  $\beta$ -glucosidase from *Exiguobacterium antarcticum* B7 (PDB ID: 5DT5; coverage: 96%; and identity: 44%). We selected the best model using the DOPE score [57–59]. Point mutations were performed in the residues of Bgl1B's catalytic pocket. Each one of the 22 residues was mutated according to 19 possibilities using MODELLER's mutation script, resulting in 418 mutant proteins. We aligned the PDB files with the  $\beta$ -glucosidase in the complex with cellobiose (3VIK) and extracted the residues of the catalytic pocket based on residues established previously. We generated the structural signatures for all files and calculated the  $\Delta\Delta\text{SSV}$  score (Table S2).

In addition, we proposed additional steps to limit the number of mutations proposed. We removed mutations proposed based on three evaluations: (i) Mutations in conserved residues; (ii) residues that are not found in a specific position in the family; and (iii) mutations that potentially cause high destabilization in the protein structure.

Residue conservation is an important metric used to evaluate mutations. Highly conserved residues tend to present essential functions for the protein activity. We performed sequence alignment of catalytic pocket residues among Bgl1B and the  $\beta$ -glucosidases of Betagdb using Clustal Omega [60,61]. We detected six conserved residues: H125, N169, E170, Y298, E353, and W399. We removed mutations in these residues indicated by SSV.

Then, we used the SIFT Sequence [46] to analyze the substitution allowed in the GH1 family for every residue of the catalytic pocket (Table S3). We removed mutations not detected in that position for the GH1 family.

Mutations can affect the protein structure, causing a destabilization that may compromise the protein activity. We evaluated the impact of mutations in the protein structure using mCSM (FIOCRUZ MINAS, Belo Horizonte, Brazil), which predicts the variation of free energy ( $\Delta\Delta G$ ) [11]. Indeed, most of the mutations cause destabilization; however, some can cause high destabilization, which may change the protein folding state. We removed the mutations indicated by mCSM as highly destabilizing (Table S4). The remaining mutants were the final mutations proposed by our workflow for tests in vitro. Lastly, we compared the results with the mutations tested experimentally in the literature.

#### 4.4. Case Study 3

The three-dimensional structure of CaLB was obtained from the PDB (PDB ID: 1TCA). The mutants from M1 to M8 were constructed using the mutagenesis tool of the software, PyMOL (<http://pymol.org>). Water molecules were removed. To detect the residues of the pocket near the active site, we performed molecular docking in the wild-type and mutants using the software, AutoDock Vina (The Scripps Research Institute, La Jolla, CA, USA) [62]. We used *N*-benzyl-2-chloroacetamide, the same ligand used to determine amidase activities in CaLB [47]. The ligand was collected from the Zinc database [63]. We used parameter exhaustiveness = 50, a box of 15 Å × 15 Å × 15 Å, and the box center was defined based on the position of the last atom of the catalytic serine (residue S105; atom OG). We used the first conformation obtained by docking and collected all residues at the distance of 6.5 Å from any atom of the ligand. Then, we removed the ligand and saved the structures as PDB files. We performed tests in the SSV web tool using the wild-type, the eight mutants, and the template database (for this step, we compressed the mutant, M3, in a zip file). The links for the projects created in the SSV tool are available in the Supplementary Material (Table S5; Figure S1).

## 5. Conclusions

In this paper, we proposed structural signature variation (SSV), which is a novel method to compute and compare structural and physicochemical signatures of proteins, with the purpose of proposing beneficial mutations to support protein engineering processes. SSV can be used together with other methods, tools, and algorithms to suggest mutations with greater reliability for reducing costs of in vitro experiments.

We evaluated the quality of the predictions through two case studies with realistic examples for the protein engineering of  $\beta$ -glucosidases, enzymes involved in biofuel production. SSV presented a high precision for 27 mutations collected from the literature and was capable of detecting beneficial mutations already proposed in the literature for Bgl1B, starting from random point mutations. SSV was shown to be an efficient method to propose mutations for non-tolerant  $\beta$ -glucosidases and may help yield enzymes with more glucose tolerance for second-generation biofuel production.

In addition, we constructed a website, with a user-friendly interface, that implements the SSV method. It is available at (<http://bioinfo.dcc.ufmg.br/ssv>).

**Supplementary Materials:** Supplementary materials can be found at <http://www.mdpi.com/1422-0067/20/2/333/s1>.

**Author Contributions:** Conceptualization, D.C.B.M. and R.C.d.M.-M.; methodology and writing—original draft preparation, D.C.B.M.; writing—review and editing, L.H.S., K.d.S.M., A.V.W., L.H.F.d.L., R.C.d.M.-M.; supervision, project administration, and funding acquisition, R.C.d.M.-M.

**Funding:** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. Process number 51/2013 - 23038.004007/2014-82.

**Acknowledgments:** The authors thank the funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The authors thank Marcos Felipe and Rafael Rocha.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

CaLB	<i>Candida antarctica</i> lipase B
epPCR	Error-prone PCR
GH1	Glycoside hydrolase family 1
IF	Improvement Factor
SSV	Structural Signature Variation
SVM	Support Vector Machine

## References

1. Chaudhary, N.; Gupta, A.; Gupta, S.; Sharma, V.K. BioFuelDB: A database and prediction server of enzymes involved in biofuels production. *PeerJ* **2017**, *5*, e3497. [[CrossRef](#)] [[PubMed](#)]
2. Egerton, S.; Culloty, S.; Whooley, J.; Stanton, C.; Ross, R.P. Characterization of protein hydrolysates from blue whiting (*Micromesistius poutassou*) and their application in beverage fortification. *Food Chem.* **2018**, *245*, 698–706. [[CrossRef](#)] [[PubMed](#)]
3. Cao, L.; Wang, Z.; Ren, G.; Kong, W.; Li, L.; Xie, W.; Liu, Y. Engineering a novel glucose-tolerant  $\beta$ -glucosidase as supplementation to enhance the hydrolysis of sugarcane bagasse at high glucose concentration. *Biotechnol. Biofuels* **2015**, *8*, 202. [[CrossRef](#)] [[PubMed](#)]
4. Cirino, P.C.; Mayer, K.M.; Umeno, D. Generating mutant libraries using error-prone PCR. *Methods Mol. Biol.* **2003**, *231*, 3–9. [[PubMed](#)]
5. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **1988**, *16*, 10881–10890. [[CrossRef](#)] [[PubMed](#)]
6. Fox, R.J.; Davis, S.C.; Mundorff, E.C.; Newman, L.M.; Gavrilovic, V.; Ma, S.K.; Chung, L.M.; Ching, C.; Tam, S.; Muley, S.; et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25*, 338–344. [[CrossRef](#)] [[PubMed](#)]



7. Berland, M.; Offmann, B.; André, I.; Remaud-Siméon, M.; Charton, P. A web-based tool for rational screening of mutants libraries using ProSAR. *Protein Eng. Des. Sel.* **2014**, *27*, 375–381. [[CrossRef](#)]
8. Christ, C.D.; Fox, T. Accuracy assessment and automation of free energy calculations for drug design. *J. Chem. Inf. Model.* **2014**, *54*, 108–120. [[CrossRef](#)]
9. Pires, D.E.; de Melo-Minardi, R.C.; dos Santos, M.A.; da Silveira, C.H.; Santoro, M.M.; Meira, W. Cutoff Scanning Matrix (CSM): Structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics* **2011**, *12*, S12. [[CrossRef](#)]
10. Pires, D.E.V. CSM: Uma assinatura para grafos biológicos baseada em padrões de distâncias. Ph.D. Thesis, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte-MG, Brazil, 2012.
11. Pires, D.E.V.; Ascher, D.B.; Blundell, T.L. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinform. Oxf. Engl.* **2014**, *30*, 335–342. [[CrossRef](#)]
12. Pires, D.E.V.; Blundell, T.L.; Ascher, D.B. mCSM-lig: Quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* **2016**, *6*, 29575. [[CrossRef](#)] [[PubMed](#)]
13. Pires, D.E.V.; Ascher, D.B. mCSM-AB: A web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.* **2016**, *44*, W469–W473. [[CrossRef](#)] [[PubMed](#)]
14. Pires, D.E.V.; de Melo-Minardi, R.C.; da Silveira, C.H.; Campos, F.F.; Meira, W. aCSM: Noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinform. Oxf. Engl.* **2013**, *29*, 855–861. [[CrossRef](#)] [[PubMed](#)]
15. Ferrario, V.; Siragusa, L.; Ebert, C.; Baroni, M.; Foscatto, M.; Cruciani, G.; Gardossi, L. BioGPS descriptors for rational engineering of enzyme promiscuity and structure based bioinformatic analysis. *PLoS ONE* **2014**, *9*, e109354. [[CrossRef](#)] [[PubMed](#)]
16. Cairns, J.R.K.; Esen, A.  $\beta$ -Glucosidases. *Cell. Mol. Life Sci.* **2010**, *67*, 3389–3405. [[CrossRef](#)] [[PubMed](#)]
17. Lu, J.; Du, L.; Wei, Y.; Hu, Y.; Huang, R. Expression and characterization of a novel highly glucose-tolerant  $\beta$ -glucosidase from a soil metagenome. *Acta Biochim. Biophys. Sin.* **2013**, *45*, 664–673. [[CrossRef](#)] [[PubMed](#)]
18. Béguin, P.; Aubert, J.P. The biological degradation of cellulose. *FEMS Microbiol. Rev.* **1994**, *13*, 25–58. [[CrossRef](#)]
19. Murphy, L.; Bohlin, C.; Baumann, M.J.; Olsen, S.N.; Sørensen, T.H.; Anderson, L.; Borch, K.; Westh, P. Product inhibition of five *Hypocrea jecorina* cellulases. *Enzyme Microb. Technol.* **2013**, *52*, 163–169. [[CrossRef](#)]
20. Chamoli, S.; Kumar, P.; Navani, N.K.; Verma, A.K. Secretory expression, characterization and docking study of glucose-tolerant  $\beta$ -glucosidase from *B. subtilis*. *Int. J. Biol. Macromol.* **2016**, *85*, 425–433. [[CrossRef](#)]
21. Kadam, S.K.; Demain, A.L. Addition of cloned beta-glucosidase enhances the degradation of crystalline cellulose by the *Clostridium thermocellum* cellulose complex. *Biochem. Biophys. Res. Commun.* **1989**, *161*, 706–711. [[CrossRef](#)]
22. Watanabe, T.; Sato, T.; Yoshioka, S.; Koshijima, T.; Kuwahara, M. Purification and properties of *Aspergillus niger* beta-glucosidase. *Eur. J. Biochem. FEBS* **1992**, *209*, 651–659. [[CrossRef](#)]
23. Zhao, L.; Pang, Q.; Xie, J.; Pei, J.; Wang, F.; Fan, S. Enzymatic properties of *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase fused to *Clostridium cellulovorans* cellulose binding domain and its application in hydrolysis of microcrystalline cellulose. *BMC Biotechnol.* **2013**, *13*, 101. [[CrossRef](#)] [[PubMed](#)]
24. Gueguen, Y.; Chemardin, P.; Arnaud, A.; Galzy, P. Purification and characterization of an intracellular  $\beta$ -glucosidase from *Botrytis cinerea*. *Enzyme Microb. Technol.* **1995**, *17*, 900–906. [[CrossRef](#)]
25. Teugjas, H.; Våljamäe, P. Selecting  $\beta$ -glucosidases to support cellulases in cellulose saccharification. *Biotechnol. Biofuels* **2013**, *6*, 105. [[CrossRef](#)] [[PubMed](#)]
26. Rajasree, K.P.; Mathew, G.M.; Pandey, A.; Sukumaran, R.K. Highly glucose tolerant  $\beta$ -glucosidase from *Aspergillus unguis*: NII 08123 for enhanced hydrolysis of biomass. *J. Ind. Microbiol. Biotechnol.* **2013**, *40*, 967–975. [[CrossRef](#)] [[PubMed](#)]
27. Yang, Y.; Zhang, X.; Yin, Q.; Fang, W.; Fang, Z.; Wang, X.; Zhang, X.; Xiao, Y. A mechanism of glucose tolerance and stimulation of GH1  $\beta$ -glucosidases. *Sci. Rep.* **2015**, *5*, 17296. [[CrossRef](#)] [[PubMed](#)]
28. Pei, J.; Pang, Q.; Zhao, L.; Fan, S.; Shi, H. *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase: A glucose-tolerant enzyme with high specific activity for cellobiose. *Biotechnol. Biofuels* **2012**, *5*, 1–10. [[CrossRef](#)]

29. Mariano, D.C.B.; Leite, C.; Santos, L.H.S.; Marins, L.F.; Machado, K.S.; Werhli, A.V.; Lima, L.H.F.; de Melo-Minardi, R.C. Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: A systematic review. *Genet. Mol. Res.* **2017**, *16*. [[CrossRef](#)]
30. Salgado, J.C.S.; Meleiro, L.P.; Carli, S.; Ward, R.J. Glucose tolerant and glucose stimulated  $\beta$ -glucosidases—A review. *Bioresour. Technol.* **2018**, *267*, 704–713. [[CrossRef](#)]
31. Berrin, J.-G.; Czjzek, M.; Kroon, P.A.; McLauchlan, W.R.; Puigserver, A.; Williamson, G.; Juge, N. Substrate (aglycone) specificity of human cytosolic beta-glucosidase. *Biochem. J.* **2003**, *373*, 41–48. [[CrossRef](#)]
32. Liu, J.; Zhang, X.; Fang, Z.; Fang, W.; Peng, H.; Xiao, Y. The 184th residue of  $\beta$ -glucosidase Bgl1B plays an important role in glucose tolerance. *J. Biosci. Bioeng.* **2011**, *112*, 447–450. [[CrossRef](#)] [[PubMed](#)]
33. Lee, H.-L.; Chang, C.-K.; Jeng, W.-Y.; Wang, A.H.-J.; Liang, P.-H. Mutations in the substrate entrance region of  $\beta$ -glucosidase from *Trichoderma reesei* improve enzyme activity and thermostability. *Protein Eng. Des. Sel.* **2012**, *25*, 733–740. [[CrossRef](#)] [[PubMed](#)]
34. Guo, B.; Amano, Y.; Nozaki, K. Improvements in Glucose Sensitivity and Stability of *Trichoderma reesei*  $\beta$ -Glucosidase Using Site-Directed Mutagenesis. *PLoS ONE* **2016**, *11*, e0147301. [[CrossRef](#)] [[PubMed](#)]
35. Sansenya, S.; Maneesan, J.; Cairns, J.R.K. Exchanging a single amino acid residue generates or weakens a +2 cellobiosaccharide binding subsite in rice  $\beta$ -glucosidases. *Carbohydr. Res.* **2012**, *351*, 130–133. [[CrossRef](#)] [[PubMed](#)]
36. Chuenchor, W.; Pengthaisong, S.; Robinson, R.C.; Yuvaniyama, J.; Oonanant, W.; Bevan, D.R.; Esen, A.; Chen, C.-J.; Opassiri, R.; Svasti, J.; et al. Structural insights into rice BGLu1 beta-glucosidase oligosaccharide hydrolysis and transglycosylation. *J. Mol. Biol.* **2008**, *377*, 1200–1215. [[CrossRef](#)] [[PubMed](#)]
37. Zouhar, J.; Vévodová, J.; Marek, J.; Damborský, J.; Su, X.D.; Brzobohatý, B. Insights into the functional architecture of the catalytic center of a maize beta-glucosidase Zm-p60.1. *Plant Physiol.* **2001**, *127*, 973–985. [[CrossRef](#)] [[PubMed](#)]
38. Sansenya, S.; Opassiri, R.; Kuaprasert, B.; Chen, C.-J.; Cairns, J.R.K. The crystal structure of rice (*Oryza sativa* L.) Os4BGLu12, an oligosaccharide and tuberonic acid glucoside-hydrolyzing  $\beta$ -glucosidase with significant thioglucohydrolase activity. *Arch. Biochem. Biophys.* **2011**, *510*, 62–72. [[CrossRef](#)]
39. Tsukada, T.; Igarashi, K.; Fushinobu, S.; Samejima, M. Role of subsite +1 residues in pH dependence and catalytic activity of the glycoside hydrolase family 1 beta-glucosidase BGL1A from the basidiomycete *Phanerochaete chrysosporium*. *Biotechnol. Bioeng.* **2008**, *99*, 1295–1302. [[CrossRef](#)]
40. Sanz-Aparicio, J.; Hermoso, J.A.; Martínez-Ripoll, M.; Lequerica, J.L.; Polaina, J. Crystal structure of beta-glucosidase A from *Bacillus polymyxa*: Insights into the catalytic activity in family 1 glycosyl hydrolases. *J. Mol. Biol.* **1998**, *275*, 491–502. [[CrossRef](#)]
41. Matsuzawa, T.; Jo, T.; Uchiyama, T.; Manninen, J.A.; Arakawa, T.; Miyazaki, K.; Fushinobu, S.; Yaoi, K. Crystal structure and identification of a key amino acid for glucose tolerance, substrate specificity, and transglycosylation activity of metagenomic  $\beta$ -glucosidase Td2F2. *FEBS J.* **2016**, *283*, 2340–2353. [[CrossRef](#)]
42. Frank, E.; Hall, M.A.; Witten, I.H. The WEKA Workbench. In *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2016.
43. Silva, M.F.M.; Leijoto, L.F.; Nobre, C.N. Algorithms Analysis in Adjusting the SVM Parameters: An Approach in the Prediction of Protein Function. *Appl. Artif. Intell.* **2017**, *31*, 1–16. [[CrossRef](#)]
44. Fang, W.; Fang, Z.; Liu, J.; Hong, Y.; Peng, H.; Zhang, X.; Sun, B.; Xiao, Y. Cloning and characterization of a beta-glucosidase from marine metagenome. *Sheng Wu Gong Cheng Xue Bao* **2009**, *25*, 1914–1920. [[PubMed](#)]
45. Fang, Z.; Fang, W.; Liu, J.; Hong, Y.; Peng, H.; Zhang, X.; Sun, B.; Xiao, Y. Cloning and Characterization of a  $\beta$ -Glucosidase from Marine Microbial Metagenome with Excellent Glucose Tolerance. *J. Microbiol. Biotechnol.* **2010**, *20*, 1351–1358. [[CrossRef](#)] [[PubMed](#)]
46. Ng, P.C.; Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **2001**, *11*, 863–874. [[CrossRef](#)] [[PubMed](#)]
47. Ferrario, V.; Ebert, C.; Svendsen, A.; Besenmatter, W.; Gardossi, L. An integrated platform for automatic design and screening of virtual mutants based on 3D-QSAR analysis. *J. Mol. Catal. B Enzym.* **2014**, *101*, 7–15. [[CrossRef](#)]
48. Braiuca, P.; Boscarol, L.; Ebert, C.; Linda, P.; Gardossi, L. 3D-QSAR Applied to the Quantitative Prediction of Penicillin G Amidase Selectivity. *Adv. Synth. Catal.* **2006**, *348*, 773–780. [[CrossRef](#)]

49. Steinkellner, G.; Gruber, C.C.; Pavkov-Keller, T.; Binter, A.; Steiner, K.; Winkler, C.; Łyskowski, A.; Schwamberger, O.; Oberer, M.; Schwab, H.; et al. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. *Nat. Commun.* **2014**, *5*, 4150. [[CrossRef](#)]
50. Berendsen, H.J.C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56. [[CrossRef](#)]
51. de Giuseppe, P.O.; Souza, T.d.A.C.B.; Souza, F.H.M.; Zaphorlin, L.M.; Machado, C.B.; Ward, R.J.; Jorge, J.A.; Furriel, R.d.P.M.; Murakami, M.T. Structural basis for glucose tolerance in GH1  $\beta$ -glucosidases. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70*, 1631–1639. [[CrossRef](#)]
52. Jeng, W.-Y.; Wang, N.-C.; Lin, C.-T.; Chang, W.-J.; Liu, C.-I.; Wang, A.H.-J. High-resolution structures of *Neotermes koshunensis*  $\beta$ -glucosidase mutants provide insights into the catalytic mechanism and the synthesis of glucoconjugates. *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68*, 829–838. [[CrossRef](#)]
53. Shatsky, M.; Nussinov, R.; Wolfson, H.J. A method for simultaneous alignment of multiple protein structures. *Proteins* **2004**, *56*, 143–156. [[CrossRef](#)] [[PubMed](#)]
54. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
55. Bitar, M.; Franco, G.R. A basic protein comparative three-dimensional modeling methodological workflow theory and practice. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 1052–1065. [[CrossRef](#)] [[PubMed](#)]
56. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezuk, Y.; McGinnis, S.; Madden, T.L. NCBI BLAST: A better web interface. *Nucleic Acids Res.* **2008**, *36*, W5–W9. [[CrossRef](#)] [[PubMed](#)]
57. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinform.* **2014**, *47*. [[CrossRef](#)] [[PubMed](#)]
58. Martí-Renom, M.A.; Stuart, A.C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325. [[CrossRef](#)]
59. Sali, A.; Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815. [[CrossRef](#)]
60. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [[CrossRef](#)] [[PubMed](#)]
61. Goujon, M.; McWilliam, H.; Li, W.; Valentin, F.; Squizzato, S.; Paern, J.; Lopez, R. A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res.* **2010**, *38*, W695–W699. [[CrossRef](#)]
62. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [[CrossRef](#)]
63. Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768. [[CrossRef](#)] [[PubMed](#)]



## **Apêndice 8. Artigo 3**

**Título:** *“Molecular Dynamics Gives New Insights into the Glucose Tolerance and Inhibition Mechanisms on  $\beta$ -Glucosidases”*

**Journal:** *Molecules*

**Fator de impacto (2017):** 3,060

**Ano de publicação:** 2019

Article

# Molecular Dynamics Gives New Insights into the Glucose Tolerance and Inhibition Mechanisms on $\beta$ -Glucosidases

Leon Sulfierry Corrêa Costa <sup>1,†</sup>, Diego César Batista Mariano <sup>2,†</sup>,  
Rafael Eduardo Oliveira Rocha <sup>2,3</sup>, Johannes Kraml <sup>4</sup>, Carlos Henrique da Silveira <sup>5</sup>,  
Klaus Roman Liedl <sup>4</sup> , Raquel Cardoso de Melo-Minardi <sup>2</sup> and  
Leonardo Henrique Franca de Lima <sup>1,\*</sup>

<sup>1</sup> Laboratory of Molecular and Bioinformatics Modeling, Department of Exact and Biological Sciences (DECEB), Universidade Federal de São João Del-Rei, Campus Sete Lagoas, Sete Lagoas 35701-970, Brazil

<sup>2</sup> Laboratory of Bioinformatics and Systems (LBS), Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil

<sup>3</sup> Laboratory of Molecular Modeling and Drug Design, Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil

<sup>4</sup> Institute of General, Inorganic and Theoretical Chemistry (IGITC), Center for Molecular Biosciences Innsbruck (CMBI), Leopold-Franzens-Universität-Innsbruck, Innrain 82, 6020 Innsbruck, Austria

<sup>5</sup> Institute of Technological Sciences, Universidade Federal de Itajubá, Campus Itabira, Itabira 35903-087, Brazil

\* Correspondence: leofrancalima@ufsj.edu.br; Tel.: +55-31-3775-5500

† These authors contributed equally to this work.

Received: 29 July 2019; Accepted: 23 August 2019; Published: 4 September 2019



**Abstract:**  $\beta$ -Glucosidases are enzymes with high importance for many industrial processes, catalyzing the last and limiting step of the conversion of lignocellulosic material into fermentable sugars for biofuel production. However,  $\beta$ -glucosidases are inhibited by high concentrations of the product (glucose), which limits the biofuel production on an industrial scale. For this reason, the structural mechanisms of tolerance to product inhibition have been the target of several studies. In this study, we performed *in silico* experiments, such as molecular dynamics (MD) simulations, free energy landscape (FEL) estimate, Poisson–Boltzmann surface area (PBSA), and grid inhomogeneous solvation theory (GIST) seeking a better understanding of the glucose tolerance and inhibition mechanisms of a representative GH1  $\beta$ -glucosidase and a GH3 one. Our results suggest that the hydrophobic residues Y180, W350, and F349, as well the polar one D238 act in a mechanism for glucose releasing, herein called “slingshot mechanism”, dependent also on an allosteric channel (AC). In addition, water activity modulation and the protein loop motions suggest that GH1  $\beta$ -Glucosidases present an active site more adapted to glucose withdrawal than GH3, in consonance with the GH1s lower product inhibition. The results presented here provide directions on the understanding of the molecular mechanisms governing inhibition and tolerance to the product in  $\beta$ -glucosidases and can be useful for the rational design of optimized enzymes for industrial interests.

**Keywords:**  $\beta$ -Glucosidases; GH1; GH3; glucose tolerance; slingshot mechanism; allosteric channel; molecular dynamics simulation; free energy landscape; Poisson–Boltzmann surface area; grid inhomogeneous solvation theory

## 1. Introduction

$\beta$ -Glucosidases are enzymes with great importance for many industrial processes, such as the production of wine [1], animal feed [2], and biofuel [3,4]. In the second-generation biofuel production,

they act in synergy with endoglucanases and exoglucanases for the conversion of lignocellulose into fermentable sugars [5,6].  $\beta$ -Glucosidases catalyze the last and limiting step, converting cellobiose into two molecules of glucose [7]. However, most  $\beta$ -glucosidases are inhibited in high concentrations of glucose, which provides an accumulation of cellobiose [8,9]. Moreover, accumulated cellobiose may cause inhibition of other enzymes that catalyze the first reactions of cellulose hydrolysis [10,11]. Hence,  $\beta$ -glucosidases have been considered key enzymes in the second-generation biofuel process, and they have been the target of many studies that aimed to elucidate the inhibition mechanisms and also propose modifications that could improve the catalytic activity, thermostability, and tolerance to glucose inhibition [3,12–22].

Most  $\beta$ -glucosidases have been classified into the glycoside hydrolase families 1 (GH1) and 3 (GH3) [20].  $\beta$ -glucosidases from GH1 family have been described as more efficient for biofuel production due to a higher resistance to glucose inhibition and a more conserved structure, which allowed more studies of beneficial mutations [6,13,20,23]. Giuseppe et al. [12] argued that GH1  $\beta$ -glucosidases are tenfold to 1000-fold more glucose-tolerant than GH3 due to their active site be located in a deep and narrow cavity, while the GH3  $\beta$ -glucosidases present a shallow pocket. For instance, they compare the three-dimensional structures of *Humicola insolens* GH1  $\beta$ -glucosidase (here called HiBG), hence more resistant to glucose inhibition, and the *Aspergillus aculeatus* GH3  $\beta$ -glucosidase (here called AaBG), thus, more susceptible to glucose inhibition. They supposed that the difference in the inhibitory behavior of these two enzymes could be a consequence of the difference in the accessibility of glucose to the catalytic site. HiBG presents an active site at 17 Å of depth, while AaBG at 11.3 Å. However, an explanation of why the accessibility mechanism would be more relevant for the product than to the substrate entrance is still missing.

Yang et al. [13] suggested that relative binding affinity of glucose to some sites at the entrance and middle of the substrate channel modulates the glucose dependence and regulates the effects of product inhibition in the GH1 enzymes. Using site-directed mutations, they proposed that three residues 228, 301, and 302 of a marine metagenome GH1  $\beta$ -glucosidase (Bgl1A) could be crucial to glucose tolerance. They detected that mutations in analogous residues (H228T and N301Q/V302F) of a homologous non-tolerant  $\beta$ -glucosidase (Bgl1B) led to glucose tolerance. By inspection virtual docking analysis, they have advocated that this effect would be due to the favoring of the glucose binding to these sites considerably distant from the catalytic cleft. The same study has pointed the position 228 as presenting higher significance on the tolerance acquiring effect. The secondary effect of the other two would be due to their proximity of the first. The authors also have suggested that this region is an allosteric subsite (AS).

In a more recent study, the comparison between crystallographic structures of tolerant and non-tolerant GH1  $\beta$ -glucosidases has shown that the tolerant seems to have the region close to the AS prolonged on a kind of allosteric channel (AC), providing an alternative thermodynamically favorable binding site that is not the same as catalytic cleft [24]. Moreover, a computational approach detected mutations at the residue 228 and the AS/AC neighborhood from the metagenomic Bgl1B lead to structural signatures more similar to glucose tolerant  $\beta$ -glucosidases [25]. In another study, Mariano et al. [3] analyzed 21 glucose-tolerant  $\beta$ -glucosidases and detected a consensus sequence of 22 amino acids in the substrate channel. Using molecular docking of cellobiose at a modeled structure of a *Thermoanaerobacter brockii*  $\beta$ -glucosidase [26], they detected that the residues W122, N166, E167, C170, L174, Y299, E355, W402, E409, and W410 perform contacts with the ligand. These residues are located in the substrate channel, with a considerable part of them on the AC/AS neighborhood, which highlights the importance of this region for glucose tolerance.

In the mentioned studies, a dynamic depiction of the possible mechanisms in which the AS/AC region would guide the glucose to exit or thermodynamically compete with its binding at the catalytic cleft was never described. In this way, a considerable set of studies seems to point to glucose tolerance/inhibition paradox. The confront between how favorably glucose binds at the catalytic cleft against how favorably it binds at alternative regions appears to be the answer. However, these

studies are based on the analysis of static structures or the functional interpretation of site-directed mutagenesis. Hence, to better understand the expulsion mechanisms, the dynamics aspects of protein and ligand, as well the dynamics behavior of the water energetic along with the active site, must be considered [27]. In this sense, molecular dynamic simulations (MD) are a very contumacious way to get dynamic and energetic information of biological systems at the atomic scale at the same time and consider the solvent effects explicitly.

In this study, we performed a set of MD simulations for the GH1  $\beta$ -glucosidase of *Humicola insolens* (HiBG) and the GH3  $\beta$ -glucosidase of *Aspergillus aculeatus* (AaBG). We analyzed the cellobiose (substrate) and glucose (product) mobility in each system, protein–ligand interactions among different regions of the active site, and we calculated the ligand-free energy landscape (FEL) for the protein active site. We also analyzed the role of the water favorability at different active subsites by the confront of grid inhomogeneous theory (GRID) and Poisson–Boltzmann surface analysis (PBSA) for representative protein conformations. The results presented here provide the first dynamic depiction of the glucose withdrawal mechanism in GH1 and an explanation about the lower adequacy of this mechanism in GH3 enzymes. They supply guidelines to the understanding of the molecular mechanisms governing inhibition and tolerance to the product in  $\beta$ -glucosidases of industrial interest. We hope they provide means for obtaining new biotechnologically optimized enzymes, with an imminent impact on the improvement in second-generation bioethanol production.

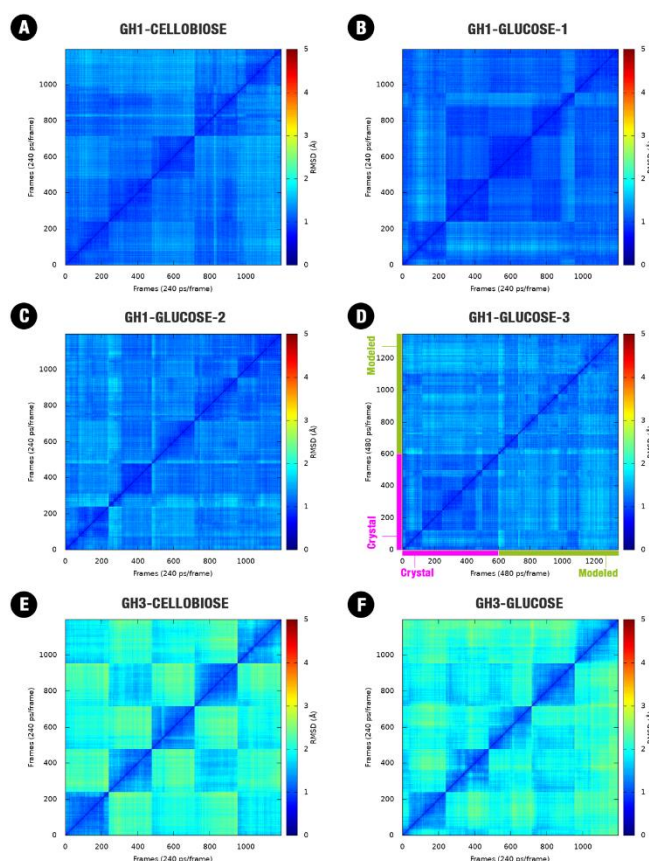
## 2. Results

### 2.1. Equilibration and Conformational Sampling from the MD Sets

To probe the equilibration and sampling convergence for different MD simulations and systems, we constructed two-dimensional root-mean-square deviation plots (2D-RMSD; Figure 1). We observed convergence in the protein dynamics for GH1s in the presence of the substrate and the product (Figure 1A–D). Additionally, the protein conformational sampling at the GH1-glucose system does not change significantly independent of the glucose starting position (Figure 1B–D).

For AaBG, protein conformational convergence is less evident in the presence of cellobiose or glucose (Figure 1E,F). In some cases, we can notice a conformational superposition between different simulations (with an RMSD  $< 2 \text{ \AA}$ ) still at the relaxing phase (not shown). While the GH1 enzymes are globular and relatively compact, presenting the mobile loops with size considerably smaller than the GH3 enzymes, these are considerably higher in terms of mass, present an elongated shape and bigger loops (Figure 2E). All these aspects allow a higher movement amplitude for the most mobile parts of the AaBG when compared to HiBG. This higher amplitude for the GH3 movements led to conformational convergence at the same scale as GH1.

The ligand movements present a higher convergence and reduced mobility at the GH3 enzyme than at GH1, even though the mobility of GH3 is higher than GH1 (Figure 2). At this lower mobility condition, the glucose on the AaBG active site and along our MD simulations has established hydrogen bonds mainly with the polar residues at the  $-1$  subsite, overall residues D73, K170, H171, and D261, beyond hydrophobic interactions with M226 and W262 (Figure S4, Figure S9, and Tables S4–S6).

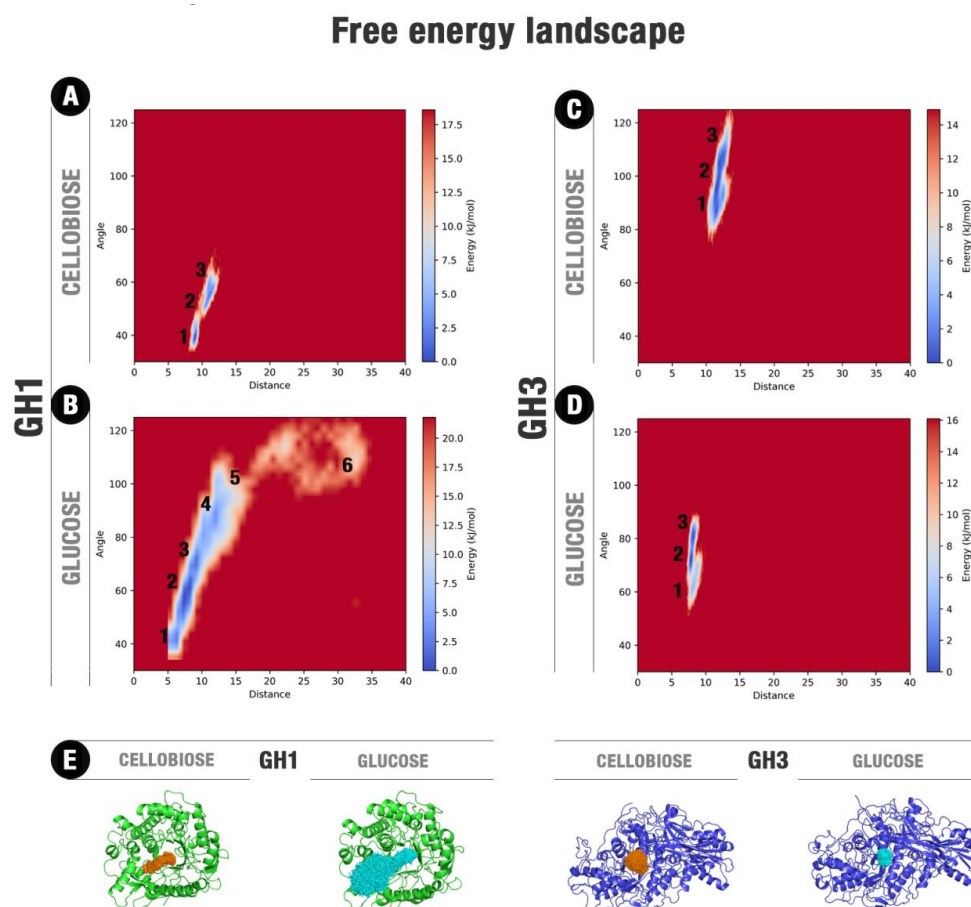


**Figure 1.** Two-dimensional root-mean-square deviation (RMSD) plot for the combined five trajectories (60 ns each one) for each one of the starting systems: (A) GH1 (HiBG)–Cellobiose complex, (B) GH1–glucose complex with glucose starting from the crystallographic pose, (C) GH1–glucose complex with glucose starting from the modeled pose, (D) GH1–glucose crystallographic complex (magenta) and the complex starting from the modeled pose (green). (E) GH3 (AaBG)–Cellobiose complex, (F) GH3–glucose complex. RMSD post the alignment of all the frames and considering just the backbone atoms (C $\alpha$ , C, O, N).

The comparative principal component analysis (PCA) between the movements of the protein and the ligand shows that the ligand freedom inside the GH1 enzyme is higher than in GH3. Moreover, we observed a more upper coupling between the ligand and protein movements in the GH1 when compared to the GH3 (Figure S10). This was observed for GH1 simulations with cellobiose and glucose. For the last ligand, such effect accentuates so its mobility than culminates with its escaping from the active site in two simulations with different starting conditions.

We also performed the free energy landscape (FEL) analysis considering the ligand space inside the active site to verify the comparative energetic freedom of cellobiose and glucose on each system (Figure 2). It is noteworthy, the considerably high freedom of glucose inside the GH1 enzyme compared to the other systems. For HiBG complexed with glucose (Figure 2A) and AaBG complexed with cellobiose and glucose (Figure 2C,D), the ligands cannot access positions significantly far from the initial point. On the other hand, for HiBG complexed with glucose, the ligand can access several regions of the energy landscape, consonant with the higher glucose mobility in this system, as observed in Figure 2E and Figures S2–S5. The gain in translational mobility in this system culminates with the access of the energetic region six on Figure 2B, depicting the glucose withdrawal.



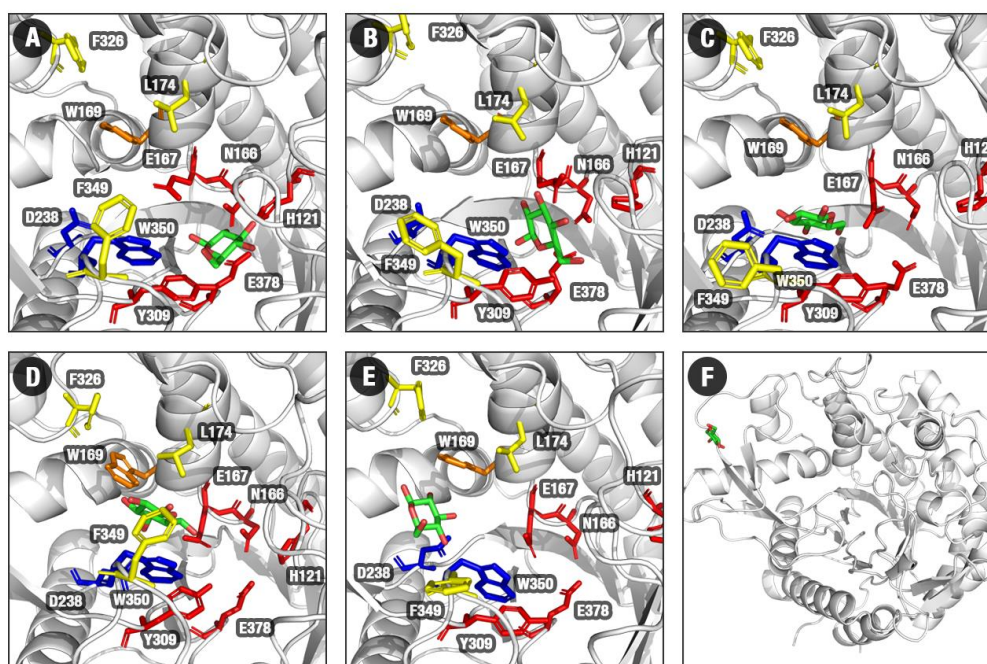


**Figure 2.** Free energy landscapes (FEL) for cellobiose and glucose at their respective and individual complexes with GH1 and GH3. (A) HiBG complexed with cellobiose, (B) HiBG complexed with glucose. This FEL was depicted considering together the respective GH1–glucose-1 and GH1–glucose-2 sets of simulations. The respective starting points at the simulations GH1–glucose-1 and GH1–glucose-2 are at the regions 3 and 1, (C) AaBG complexed with cellobiose, (D) AaBG complexed with glucose. The distance is relative to the geometric center of each ligand and the catalytic E378 (in GH1) or H121 (in GH3). The angle is formed by these two geometric centers and the geometric center of each protein. (E) Superposition of the ligand poses around all the respective trajectory ensembles for each system. Protein average frames are shown in green (HiBG) and blue (AaBG) cartoons. All frames of ligand positions for cellobiose (orange) and glucose (cyan) were overlapped.

## 2.2. MD Shows Glucose Releasing for HiBG (GH1)

For the MD of the GH1–glucose complex, we observed the complete ligand releasing to the bulk in two respective individual trajectories. The first one starting from the modeled catalytic pose (with total duration of 90 ns up the complete glucose exit) and the other starting from the crystallographic one (with glucose initially at the intermediary site, packed against residues W169, L174, Y180, F349, and W350; the last with the default duration of 60 ns). For the remaining replicates, glucose presented itself significant freedom to occupy different subsites (Figure S2), but it was positioned with considerably higher frequency at the middle portion of the substrate channel between subsites –1 and +1 (region 2 at the FEL from Figure 2B, and pose in Figure 3B). This is an intermediary region, and its relatively higher occupancy shows the high facility of glucose to pass through that from one subsite to the other.

### Glucose binding modes in HiBG



**Figure 3.** Binding modes for HiBG in complex with glucose of the GH1–glucose FEL (Figure 2B) for the regions 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), and 6 (F). Protein structures are shown in gray, subsite –1 (red sticks), subsite +1/+2 (yellow sticks), D238/W350 (blue sticks), W169 (orange sticks). The poses (selected from the minima at the FEL from Figure 2B) show a probable exit path for glucose from the HiBG active site.

By the profiles of Figure 2A,B and the superposition of cellobiose and glucose poses in Figure 2E and Figures S2 and S3, we observed that ligand translation (and consequent glucose exit) along the GH1 active site is biased in terms of angle relative to the catalytic cleft and the protein geometric center. There is a preferential path from where the ligand moves and, eventually, escapes (Figure 3).

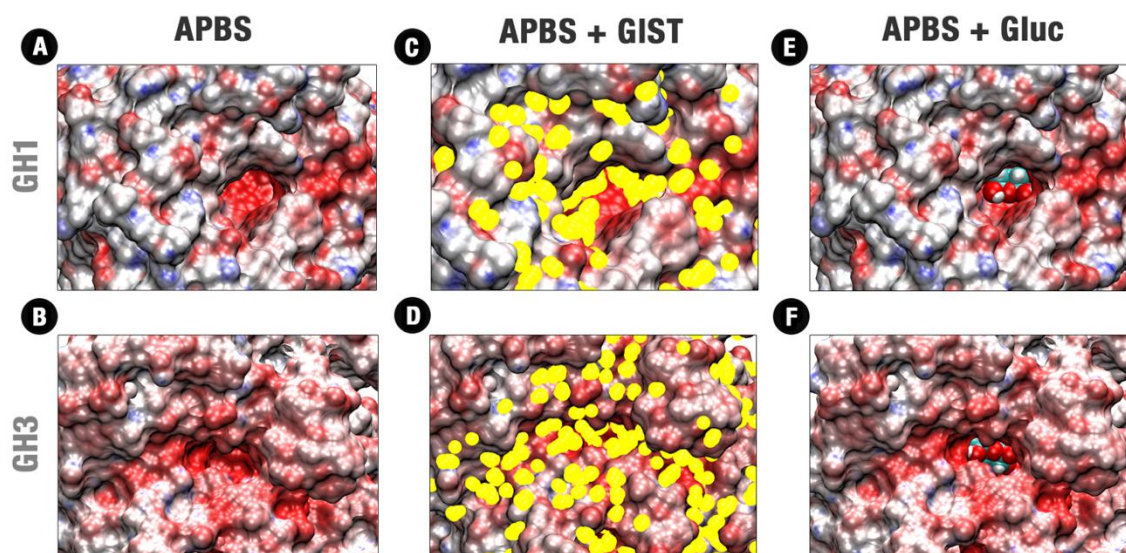
In pose one (the closest to the active site), glucose interacts with the polar –1 subsite (N166, E167, H121, and E378, Figure 3A, Figure S6, Tables S1–S3). Then, it leaves the catalytic cleft and goes to the middle of the substrate channel at the minimum two (the most statistically populated position according to Figure 2A,B FEL, Figure 3B). It is worth noting that, while on the less energetically favorable catalytic pose, the water-mediated hydrogen bonds prevalence, at the more favorable pose two, the participation of these water-mediated bonds decreases and the sites for direct hydrogen bonds increase (Figure S7, Tables S1 and S2). Concerning the behavior of glucose at the GH3 catalytic cleft, the participation of the direct hydrogen bonds is higher than the water-mediated hydrogen bonds (Figure S9, Tables S4 and S5). This indicates that direct interactions with the active site wall and water exclusion are important to determine the favorability of the glucose retaining at different sites at the  $\beta$ -glucosidases here studied.

Subsequently, glucose is trapped by hydrophobic interactions with residues between the middle and the entrance of the channel, where this ligand interacts mainly with W350, and in a minor scale with L174, W169, W436, F349, and Y181 (Figure 3C,D, Figure S7, Tables S1–S3). This state is corresponding to region three in Figure 2B FEL. This set of hydrophobic residues is denoted in the literature as gatekeeper residues, acting as bottlenecks for the substrate channel, limiting the entrance and exit of molecules [2]. Their roles have been correlated with glucose tolerance [2,3]. Finally, glucose is released from the hydrophobic region, mainly due to interactions with D238 (Figure 3E,F). This residue represents the analog to position 228, described as the allosteric site (AS) in [13,25,26]. Our studies suggest this amino

acid has significant importance for glucose releasing in glucose-tolerant  $\beta$ -glucosidases, in a process herein described as “slingshot mechanism”, which will be better presented in the next sections.

### 2.3. APBS vs. GIST vs. Ligand Fitting: Differences on the Electrostatic Distribution, Water Activity and Stereochemical Adjustments Around the Active Sites of Hibg and Aabg and Ligand Dynamics

The APBS and GIST analyses of the decoys recovered from the MD FELs (Figure 2) show a high negative electrostatic potential as well a significant trend to favorable water retention at the catalytic cleft of both enzymes (Figure 4A,B,D). This is consonant with the presence of the two catalytic acids, the polar residues of the  $-1$  subsite and the necessary adequation of this position to the polar substrate. However, a distinction in topology and distribution of polarity is found at the catalytic cleft and neighborhood between the two enzymes. The catalytic cleft at HiBG is located at the bottom of a 17–18 Å deep tunnel, with decreasing polarity up to the nonpolar loops at the surface. On the other hand, AaBG presents a constrict catalytic cleft surrounded by a flat and shallower intermediary region, around 6 Å distance from the catalytic residues, equally charged and hydrophilic.



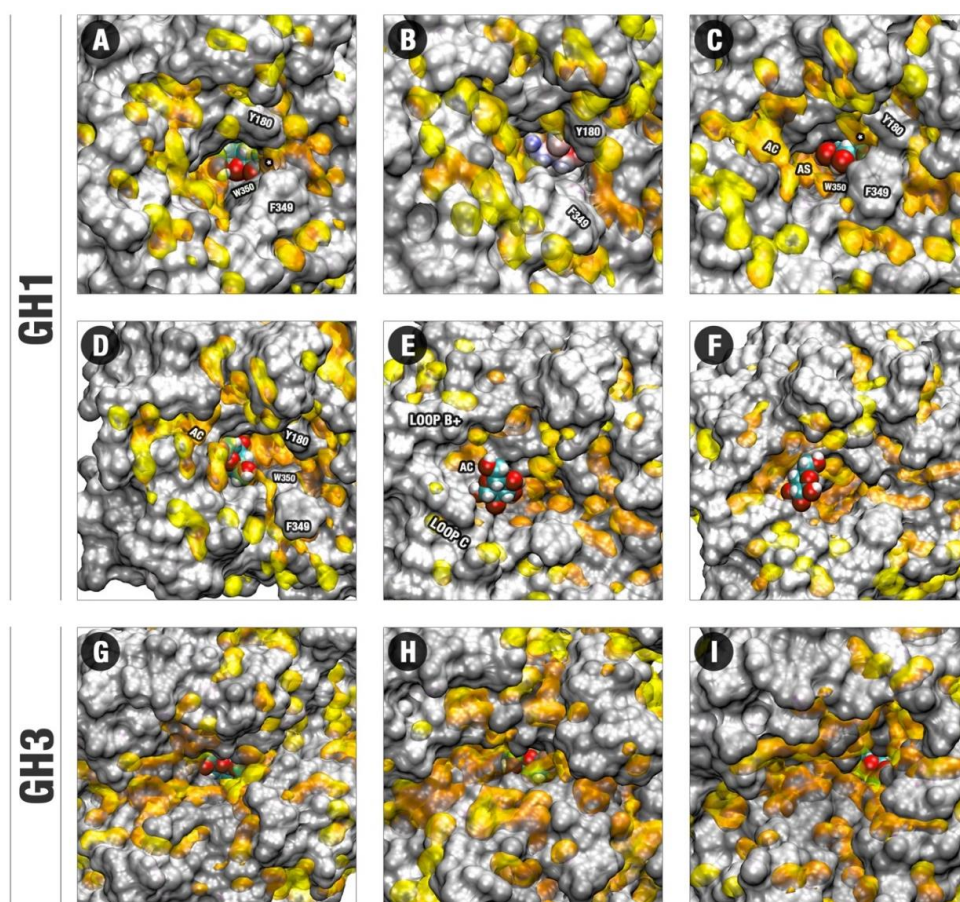
**Figure 4.** Confront between APBS, and GIST data and the glucose fitting in GH1 and GH3 representative poses at the catalytic cleft. (A,B) APBS respectively for HiBG at the pose representative of the minimum 1 in Figure 2B and AaBG at the pose representative of the minimum 2 in Figure 2C; (C,D) APBS and GIST results respectively for the previous pose of HiBG and AaBG (E,F) APBS and glucose fitting respectively for the previous pose of HiBG and AaBG. APBS scale in red, white, and blue corresponding to  $\psi$  values of  $-20.00:0.00: +20.00$ , respectively. GIST results are shown in yellow dots for water interacting centers with  $g^v(O) \geq 10.00$  units of the bulk density and  $\Delta G_{\text{solv}}^v \geq +3.0$  kcal/mol. Glucose is shown in cyan spheres with carbon atoms in cyan, oxygens in red, and hydrogens in white.

The GIST analysis of the same decoys shows that both catalytic pockets are also centered with high potential for water exclusion (Figure 4C,D). Considering just sites with occupancy  $g^v(O) \geq 10$  units of the bulk density, only unfavored water sites ( $\Delta G_{\text{solv}}^v \geq +3$  kcal/mol) are recovered in our analyses for both enzymes. This indicates that the significant losses in terms of entropy (for the strongly retained water molecules) as well as in water–water interactions are, in general, not enthalpically compensated enough by the new water–solute interactions at this same site. This is also in agreement with the known importance of the solvent exclusion on the binding mechanisms in the active sites [28].

Despite the similarities mentioned above, the differences in topology and polarity distribution also induce distinctions on the GIST recovered profiles on the active sites of both enzymes. The differentially distributed polarity of the HiBG active site and the width of its catalytic cleft tend to distribute the unfavorable water retention sites majorly on the sides, forming hydrophobic paths that

go from the bottom to the surface (Figure 4C,D). On the other hand, the shallower catalytic cleft of the AaBG tends to be uniformly filled with unfavorable hydration sites for all its extension, the same considering the flat and equally confined region around it and the apolar loops at the surface.

Figure 4E,F compares the respective allocations of glucose at the HiBG and AaBG post catalytic poses. Glucose is relatively loose inside the deeper catalytic cleft of the GH1 enzyme, contacting the unfavored water regions of Figure 4C majorly by the ligand extremities (Figure 5A). On the other hand, glucose is more stereochemically fitted by the polar fence at the shallower catalytic cleft of AaBG. This high stereochemical adequacy agrees with a higher enclosure of glucose by the polar residues of the −1 site in the GH3 (Figure S4) that in the GH1 enzyme (Figure 3A,B and Figure S2). The allocation of glucose at the GH3 catalytic cleft is also more superposed to the unfavored hydration sites at this cleft in Figure 4D in all their extension (Figure 5G–I).



**Figure 5.** Conformations recovered by the FEL profiles of the glucose positioning in GH1 and GH3, colored by the GIST data. (A–F) GIST results and glucose positions occupancy at the respective FEL regions 1–4 and two different samples of region 5 in GH1 in Figure 2B. (G–I) The analog profiles for the FEL respective regions 1–3 in GH3 in Figure 2C. GIST positions are shown in yellow transparent surfaces for water-interacting centers with  $g^v(O) \geq 10.00$  units of the bulk density and  $\Delta G_{\text{solv}}^v \geq +3.0$  kcal/mol. Glucose, shown in spheres in all the figures, is colored with carbons in cyan, oxygens in red, and hydrogens in white (except for B, which was colored by APBS). Residues involved in the establishment of the hydrophobic cage and regions important for the glucose escaping in GH1 are highlighted. AS: Allosteric site. AC: Allosteric channel. A white asterisk (\*) is used to point the site for water exclusion confined by the hydrophobic cage between Y180, F349, and W350. The same figure colored by APBS and GIST is available in the Supplementary Materials (Figure S13).

Figure 5 suggests that the distinctions of topology, distribution of charge and unfavorable hydration sites, as well of stereochemical adequation are straightly related to the dynamic differences of glucose on the GH1 and GH3 enzymes. The movement of glucose in a direction outside the active site in the HiBG occurs according to this ligand, loosely bound at the catalytic cleft, diffuses along the hydrophobic paths and between the different hydrophobic subsites connecting the bottom to the surface (Figure 5A–F). In Figure 5A,B, the respective decoys corresponding to the minima 1 and 2 in Figure 2B and the poses in Figure 3A,B are depicted. It can be noted that from pose A to B, glucose diffuses from the catalytic cleft through a hydrophobic path that brings this ligand in proximity to the apolar residues W350, F349, and Y180 (hydrophobic bottleneck). Then, subtle reorientations of the side chains (majorly F349) promote a kind of hydrophobic cage, in which glucose tends to be retained, being also impaired to return by the same path to the catalytic cleft (Figure 5C). The decoy here analyzed corresponds to the minimum of region three from Figure 2B and the pose in Figure 3D. At this position, glucose is orientated in a way that facilitates the establishment of hydrogen bonds with the D238 residue (AS, Figure S7, Tables S1 and S2). At the next decoy (corresponding to the local minimum four in Figure 2B and the pose in Figure 3E), glucose is completely displaced to this AS (Figure 5D). It can be noticed at this pose that glucose is positioned on a site with significant potential for water exclusion, indicating that the binding to the AS is favorable due to the hydrogen bonds with D238 and hydrophobic effects. Continuous to this allosteric site, there is an allosteric channel (AC) delimited by side chains from the loop C and an insertion of the loop B not usually present at the normal non-tolerant GH1s, here called loop B+. This allosteric channel is filled by unfavorable hydration sites, forming a hydrophobic path that connects the AS to the external environment. In Figure 5E, depicting a next decoy from the region 5 of Figure 2B FEL, it can be noted that subtle movements of the residues at the loops C and B+ promote at the same time a higher opening of the AC and a significant reduction at the potential for unfavorable water retention at the region of the AS. In this way, the hydrophobic effect at this region diminishes, glucose becomes free to unbind the AS and follow by the enlarged AC. However, the AC never loses its hydrophobicity completely. In Figure 5F, depicting a subsequent decoy on the region five from Figure 2B, it can be noted that a set of hydrophobic patches remains, so that glucose maintains itself at the proximity, establishing occasional CH/ $\pi$  hydrophobic interactions. This effect, in turn, draws back the return of this ligand inside the active site. On the subsequent events at the escaping MDs, glucose gradually unbinds this channel, or it is guided through the same channel to outside, in both cases resulting in the exit depicted in Figure 3F.

For GH3, the unfavorable water sites also form narrow hydrophobic channels, some of them being, in principle, compatible with possible guidance of glucose outside the active site in a similar way as on GH1 (Figure 5G–I). However, in all representative decoys, this ligand presents higher stereochemical adequacy to the polar catalytic cleft compared to the GH1 system. Also, the region where the ligand binds in this cleft has higher superposition with unfavorable hydration sites. The catalytic cleft and its immediate surroundings can be considered as hot spots of unfavorable water sites on the GH3 pocket. In this way, the permanency of glucose at the catalytic cleft on the non-tolerant enzyme seems to be favored both by the direct protein–ligand interactions and by the hydrophobic water exclusion effects (in consonance with the reduced number of water-mediated hydrogen bonds depicted for this system at Table S5), so that glucose remains tightly bound at the GH3-glucose system along all the MDs set.

The same analysis concerning the protein–cellobiose MD simulations shows similar behavior to the respective GH1–glucose and GH3–glucose complexes concerning the water retaining sites, but with crucial distinctions relative to the presence of the two glucose monomers establishing interactions with different subsites at the same time (Figure S8). For GH1, considering together the two respective glucosides (the reducing and the non-reducing), a comparable trend can be seen where the ligand occupies the same major subsites of the hydrophobic paths as in the glucose simulations. However, a critical difference is precisely the fact that the interactions of the non-reducing and the reducing glucosides with the active site occur at the same time, reinforcing each other. This impairs the complete withdrawal of the substrate from the active site, despite the relative mobility of the same. This same

mobility, in the case of cellobiose inside the GH1 active site seems to allow a better synergic exploring of the active pocket environment by the two glucosides. For instance, in Figure S11C, it can be noted that the movement of cellobiose allows a synergic binding of the ligand reducing glucoside at the AS and of the non-reducing close to the hydrophobic cage. Both regions are hot spots for hydrophobic interactions according to the GIST results, as already described. The synergic interaction of both glucosides at the two hot spot regions impairs the ligand to follow the AC or unbind the active site on a similar way that glucose when interacting with the AS.

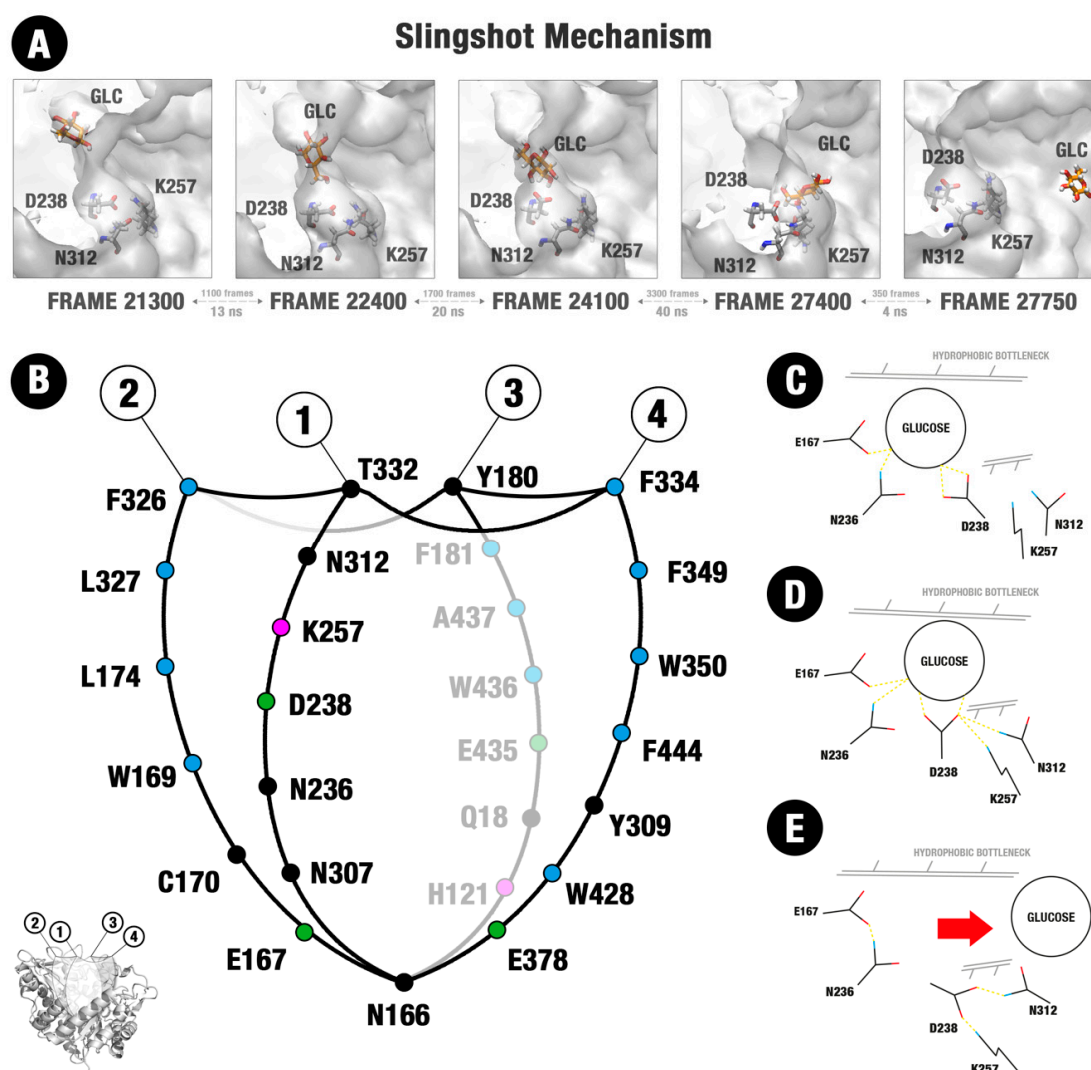
For the GH3, the additional reducing glucoside is considerably less fitted to the active site wall and less superposed to the water exclusion sites than the non-reducing or the individual glucose molecule (Figure S11D–F). This agrees with the high rotational mobility of the reducing glucoside inside GH3 compared to the non-reducing one, as depicted in Figure 2E (third image) and Figure S5. It is also in consonance with the lower number of contacts depicted for the same reducing glucoside in Figure S8 and Tables S4–S6. Beside this, the presence of this extremity and its contacts with the external loops (even though sparse) reduce the average enclosure of the catalytic fence and of the –1 subsite residues around the non-reducing extremity, resulting in a higher mobility also for this extremity (as can be noted in Figure S5) compared to the individual glucose in Figure S4.

In this way, the analysis of the data related to APBS, GIST and stereochemical fitting here presented suggests an active pocket more suited to the residence of the substrate than the product on the tolerant GH1 here studied, while there is apparent qualitatively opposite behavior for the non-tolerant GH3.

#### *2.4. Evidence of the Participation of the D238 Residue and Neighborhood on the Glucose Catapulting Outside the GH1 Enzyme*

An interesting aspect is noted concerning the dynamics of D238 and its neighbor residues at the two simulations in which glucose escapes completely from the active site. When glucose is far from D238, this residue tends to make polar interactions majorly with the dyad K257 and N312 (Figure 6A, first snapshot). As glucose is guided to the residue D238 by the hydrophobic paths and the protein movements, as well other auxiliary polar interactions, the residue D238 interrupts its interactions with the polar dyad and establishes hydrogen bonds with the glucose (Figure 6A, second and third snapshots). In the next snapshots, the diffusion of glucose in the direction of the AC and posterior withdrawal of the active site is accompanied by a movement of D238 back to the K257-N312 dyad (Figure 6A, fourth and fifth snapshots). The sequence of movements resembles a slingshot catapult, as if the D238 caught the glucose and then, being attracted again to its polar partners, it threw the ligand in the direction of the AC region. Although such movement has been observed at the two simulations in which glucose escaped from the active site, presenting a relatively small sampling, its apparent relationship with the withdrawal mechanism has not gone unnoticed. Beyond that, it involves a region known for its importance in glucose tolerance [13,25,26].

In this way, the possible implications of such movement on the intrinsic mechanism of glucose escaping from the tolerant GH1 enzyme will be included in the discussion section.



**Figure 6.** Structure of the substrate channel of HiGB and the mechanism of glucose release. **(A)** Five snapshots of the MD simulation that illustrate the slingshot mechanism and residues involved. **(B)** Structure of the substrate channel organized into four sectors (1–4). A set of hydrophobic residues is located in the entrance of this channel, mainly in the sectors 2, 3, and 4. Circles represent apolar (blue), polar neutral (black), polar positively charged (magenta), and polar negatively charged (green) amino acids. **(C–E)** Amino acid residues involved in the slingshot mechanism: E167, N236, D238, K257, and N312.

### 3. Discussion

#### 3.1. Tolerance and Inhibition are Dependent on a Set of Topological and Physical-Chemical Distinctions Between the Respective GH1 and GH3 Active Sites

Many aspects of glucose tolerance in GH1s are not fully understood up to now. Previously published studies attribute the tolerance of HiBG (and other glucose-tolerant GH1  $\beta$ -glucosidases) to a higher difficulty of glucose to have access to the catalytic cleft compared to the non-tolerant enzymes, as AaBG (GH3) [12]. The issues described as limiting for the glucose access are the higher depth of the substrate channel and the presence of gatekeeper hydrophobic residues at the tunnel entrance. However, it is not clear why these aspects would not impair the access of the substrate to the catalytic region equally.

Another hypothesis points to the presence of an alternative allosteric binding site (AS) specific to glucose that would impair the access of this ligand to the catalytic cleft, facilitating its withdrawal from the active pocket. It is not evident why this alternative site would be more acceptable to glucose than to cellobiose binding and what would be the dynamic/energetic mechanisms that could facilitate the driving of glucose from the catalytic position to the same site and from the AS to outside the pocket. Without such final liberation of glucose from this site, the retention of this ligand at the same would favor more a transglycosylation process than tolerance [13].

Our computational results suggest a mechanism of glucose tolerance based on a combination of the aforementioned attributes. Data here presented shed light on how these characteristics combined allow a ligand launching mechanism intrinsic to the GH1 and more adapted to glucose than to the substrate exit. The mechanism involves a set of particular characteristics of the tolerant GH1 active site distinct from the non-tolerant GH3, each one of them in concurrence with the previous issues pointed in the literature: (i) A different substrate/product stereochemical fitting relationship, favoring more the substrate than the product fitting on the GH1 pocket and, apparently, the opposite in GH3. This differential fitting results on significantly higher mobility of glucose and greater responsivity of this ligand to the protein dynamics inside the GH1 active site. In this way, glucose is easily guided by the protein dynamics to the sites that favor its withdrawal, (ii) a different distribution of charge and hydrophobicity at the GH1 and GH3 active pocket which tends to generate water constraining paths in GH1, along which glucose tend to be hydrophobically driven to the exit zones, while in GH3 there is a favoring to the residence of this ligand on the catalytic cleft. Considering cellobiose, the water-constraining paths in GH1 seem to create zones that facilitate the hydrophobic adjustment of this larger ligand along different regions of the active site in simultaneous. In this way, the same mechanism seems to contribute more efficiently to the substrate permanency and the product escaping, and (iii) the action of an allosteric site (AS) continuous to an allosteric channel (AC) typical of tolerant GH1 enzymes. This allosteric site, together with the physical-chemical attributes, promotes a kind of slingshot mechanism. In this mechanism, glucose is first guided to the AS. The small polar residue at this site catches glucose and, boosted by the opening of the AC and by a set of local interactions, catapults this ligand outside the active site.

### 3.2. *The High Mobility of Glucose Inside the GH1 Active Site, Corroboration with the Sparse Electronic Density of HiBG*

In our MD sets, glucose inside the HiBG's active site has demonstrated significantly energetic accessibility to different regions (Figure 2B) corroborated by its considerable mobility (Figure 3) and culminating with withdrawal in two independent MDs starting by two different poses (Figure 3F). This high mobility agrees with the observed significantly lower stereochemical adjustment of glucose to the catalytic cleft of the GH1 enzyme compared to GH3 (Figure 4E,F). This set of clues agrees with the experimental evidence of the sparse electronic density, as well the atypical position of glucose at the crystallographic structure of the HiBG-glucose complex compared to other non-tolerant GH1s (Figure S1). The typical GH1-glucose crystallographic complexes present a well solved electronic density (describing clearly all the heavy atoms with  $\sigma = 1.5$ ) at the  $-1$  site (Figure S1A,B). However, glucose in HiBG is located close to the exit (at the hydrophobic cage visualized at our MD studies) and with a scarcer electronic density, hardly describing just the regions around the carbons 5 and 6, the oxygen at the position 1 and the carbon-oxygen pair at which would be the position 2 (Figure S1C,D). This evidence, per se, indicates higher mobility of glucose inside the HiBG compared to usual non-tolerant GH1s, confirming our computational data. However, we must stress the possibility that the electronic density described as from glycerol at the usual glucose position in HiBG should be from a still less solved (and so, most mobile) glucose at this site (Figure S1D). Considering the sparse electronic density that has allowed the attribution of the glucose identity at the hydrophobic cage at this structure, it cannot be attributed to the uniqueness of the definition of the electronic density at the  $-1$  site as if in fact from glycerol or a second (and most mobile) glucose molecule. This absence of uniqueness persists



even if we carry out an analysis with less strict  $\sigma$  criteria ( $\sigma \leq 1$ , not shown). The presence of less solved glucose at this site, together with a subtly better solved at the hydrophobic cage, would be in conformity with the relative FEL values at the two sites in Figure 2 and with the lower stereochemical adjustment at the  $-1$  subsite depicted in Figure 4E.

### 3.3. The Importance of the Hydrophobic Subsites at the Exit and the Water Activity Modulation on The Glucose Withdrawal

An unclear point in the importance of the hydrophobic gatekeeper regions described in the study of Giuseppe et al. [12] is: if the hydrophobic residues restrict the entrance of glucose in the substrate channel, why would they not restrict the entrance of cellobiose? Or even, would they be responsible for retention of glucose in the middle of the substrate channel, which could cause the enzyme inhibition?

Our computational data show that the zones propense to ligand binding due to water exclusion effects in GH1 do not restrict just to the hydrophobic residues at the entrance of the channel. These zones form true paths that connect from the bottom of the substrate channel (the catalytic cleft at the  $-1$  subsite) to the exterior region. In the case of the less stereochemically fitted product (glucose), the movement of the protein and the modulation of the hydrophobicity along these paths is easily able to conduct this ligand along them up to the AS/AC region and finally to the exit. In the case of the two glucoside rings, the same movement of the protein leads this ligand to orientations that allow the simultaneous interactions of both of those regions highly propense to water exclusion at the same time that reinforces the interactions at each site in a cooperative way. The same physical-chemical characteristics of the water-constrained paths mediate the paradoxical effect of facilitating the glucose diffusion to outside and cellobiose permanency at the active site.

It is interesting to highlight that a significantly important region at the hydrophobic mediation of the glucose withdrawal and cellobiose permanency, the hydrophobic cage composed by the residues W350, F349, W169, L174, and Y180 (Figure 3C,D, Figure 5D, Figure S2, Figure S3, Figure S11B,C), has a substantial corroboration from the literature. This is the region where the electronic density of glucose (even than sparse) is solved in HiBG. The residues W169 and L174 are described in the literature as promoters of glucose tolerance [3]. The W350 is a conserved residue in GH1 enzymes. The establishment of the hydrophobic cage is strongly determined by movements and packing of the residues F349 and Y180. The F349 is located on a small insertion at the loop C, found on the HiBG but not at the usual non-tolerant GH1s. Considering the positions topologically analogous to these two hydrophobic residues and based on the five GH1 PDBs used as a comparison in Figure S1, we usually find hydrophobic residues, but relatively smaller, such as alanine or leucine at the analogous position to 180 and valine or methionine on the analogous to 349. While the characteristic remains, the volume and coverage at these hydrophobic regions can be a target in the improvement of glucose tolerance in usual GH1 enzymes.

For the GH3 enzyme, the catalytic cleft seems to represent a hotspot for water exclusion that, together with higher stereochemical adjustment to glucose, seems to be a determinant factor in the product permanency on the catalytic site in our simulations. Future studies of rational design of more glucose-tolerant GH3s can take the enhancing of the hydrophobicity at the flat region immediately outside the catalytic cleft and promote a higher water exclusion effect compared to the cleft itself. This could be of interest due to the higher usual catalytic activity of this class of enzymes [29,30]. Another possibility could be the shortening of the external loops, once their approximation contributes to bringing the  $-1$  subsite residues on the imprisonment position around glucose (Figure S4). This position seems to contribute to the high stereochemical fitting of the product at the catalytic cleft and the high constraining of unfavorable water molecules at the apo state (Figure 4D,F, Figure 5G–I), both contributing to the product retention.

Due to the idiosyncratic polarity characteristics of carbohydrates (i.e., being polar ligands, but with CH/ $\pi$  interactions accounting to their interactions with protein-binding sites), it is not surprising that the ability to interact with hydrophobic paths has an influence on their residence or escaping in  $\beta$ -glucosidases [31].

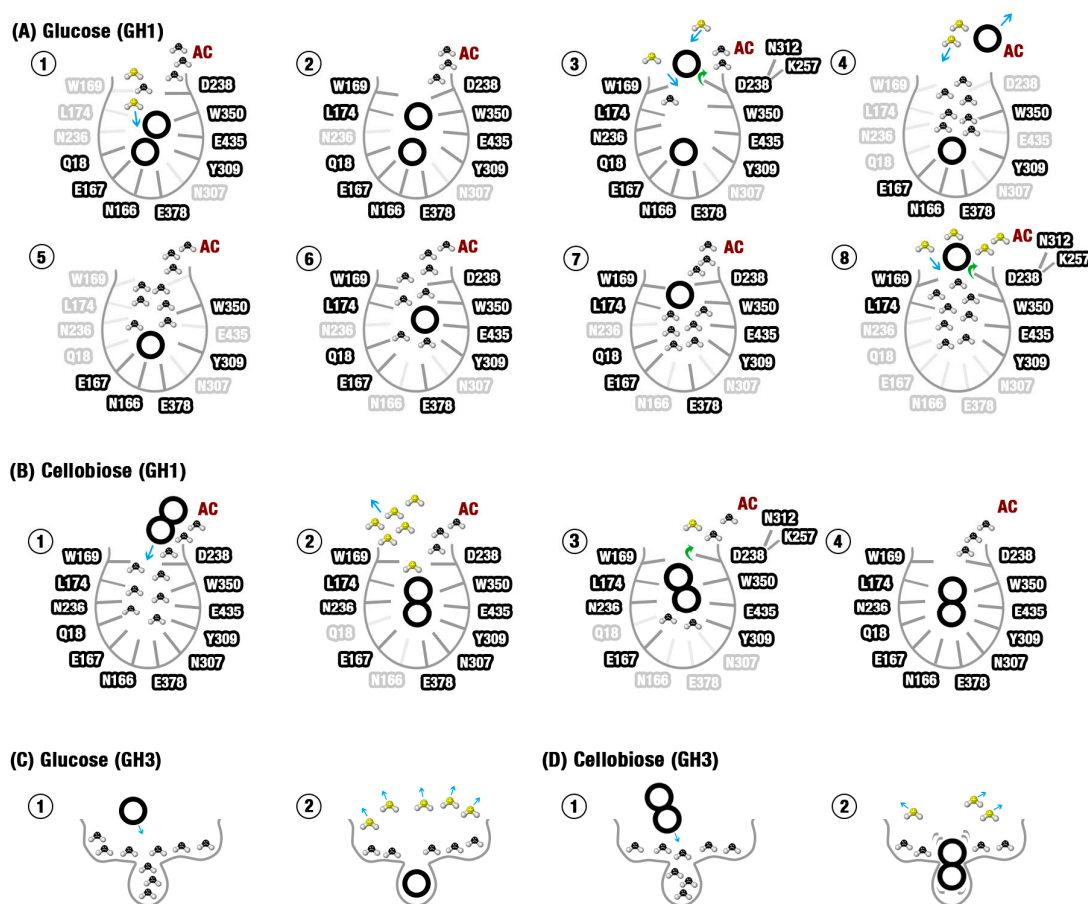
#### 3.4. The Importance of an Allosteric Site/Channel and a Slingshot Mechanism for the Glucose Withdrawal

Our MD data confirm the role of D238, the analogous position of the residue 228 previously described in [13,25], as an allosteric sub-site responsible for impairing the occupancy of glucose at the catalytic cleft. More than this, the data here presented depicts the first dynamic/energetic description of the mechanism by which this site promotes such effect.

The fundamental role of this position at the glucose immobilization and driving to the exit seems to be dependent of previously described hydrophobic paths, of its continuity to an allosteric channel (AC) formed between the loops C and B+ (Figure 5D–F), as well as local interactions at its neighborhood that promote a kind of slingshot, catapulting glucose outside (Figure 6). The effect of mutations that enhance the access to this site on the increase of glucose tolerance, as described in [13,25], can be better understood in the face of the present data. Additionally, it is important to highlight the participation of loop C in the formation of the AC from where glucose finally escapes in our simulations. The loop C topology has been already described as a determinant for glucose tolerance in GH1 enzymes [3]. Moreover, in loop C are located the W350 and F349 residues, both fundamental to the establishment of the hydrophobic cage.

Beyond the role of the hydrophobic paths, other polar interactions seem to be relevant for the guidance of glucose up to the AS position at the D238 residue (Figure 6B–E). If we divide the catalytic pocket of the GH1 enzyme into a set of sections (based on the statistically relevant sites depicted in Figures S6–S9 and Tables S1–S6 and the active site topology), we obtain the Sections 2–4 majorly hydrophobic and the Section 1 (where the D238 residue is located), majorly polar (Figure 6B). Sections 2–4 participate more actively on the establishment of the water constraining zones that create a hydrophobic way to conduct glucose from the catalytic cleft to the hydrophobic cage. However, along the glucose path to the hydrophobic cage and from there to the AS, this ligand is also guided by hydrophilic interactions with residues from the Section 1 and neighbors, overall, the catalytic E167 and N236 (Figure 6C–E). Between them, N236 has been already described as presenting significant importance in the glucose tolerance mechanism [32].

The results here presented shed light on the molecular mechanisms that lead to the glucose tolerance/inhibition paradox between GH1 and GH3 enzymes. They allow the proposal of a detailed description of the mechanism of selective glucose escaping in glucose-tolerant  $\beta$ -Glucosidases (here called slingshot mechanism) and why the same mechanism is not facilitated in an intolerant GH3. We expect that new industrially promising modifications on GH1 and GH3 enzymes can be formulated in the future. Figure 7 presents a model of the tolerance/inhibition mechanisms here proposed, including all the physical-chemical elements.



**Figure 7.** The sequence of molecular events that lead to glucose tolerance at the GH1 enzyme and inhibition at the GH3. **(A)** Slingshot mechanism influences the glucose tolerance of the HiBG GH1 enzyme. Figure 7A1–8 shows the most probable sequence of events (based on the computational issues here recovered), from the hydrolysis to the final elimination of the non-reducing glucose. **(B)** Cellobiose cannot so easily escape from the HiBG substrate channel. The sequence of events depicted in B1–4 is similar to the events described for glucose. However, the higher set of synergic interactions involving the protein and the two glucose rings in cellobiose draws back the substrate. **(C)** At the shallower and more charged pocket from the AaBG GH3 enzyme, the catalytic cleft is more stereochemically fitted to glucose. Besides this, the same constricted cleft tends to retain a higher set of unfavorable water molecules that are more efficiently liberated with the glucose binding. **(D)** The relatively small cleft of AaBG is less stereochemically suited to cellobiose than to glucose, resulting in higher ligand mobility and less unfavorable water elimination. AC: Allosteric channel. Unfavorable waters (black) are energetically constrained water molecules, favorable waters (yellow) depict water molecules liberated or energetically relaxed, bold countered residue names depict residues involved in interactions at the represented state, transparent residue names are residues not involved in interactions at the represented state.

## 4. Materials and Methods

### 4.1. Protein and Ligand Structures

The crystallographic structures of  $\beta$ -glucosidase A from *Humicola insolens* in complex with glucose (PDB: 4MDP, with a 2.05 Å resolution) [12], as well  $\beta$ -glucosidase 1 from *Aspergillus aculeatus* in complex with glucose (PDB: 4IIG, 2.30 Å resolution) and cellobiose (PDB: 4IIH, 2.00 Å resolution) [33] were both collected from the Protein Data Bank [34].

For 4MDP structure, glucose is not found on the catalytic cleft, where it use to be found on other GH1-glucose crystallographic complexes (as in PDB: 3VIJ, PDB: 4PTX, PDB: 3WH6, PDB: 2JIE, PDB: 2O9T, PDB: 2E40, the last, complexed with gluconolactone), but packed between the residues W169, L174, F349, and W350, on an intermediary position between the catalytic cleft and the active site exit, henceforth cited as the intermediary site (IS, Figure S1A). On the other hand, close to the usual position of the catalytic cleft (depicted by the E167 and E377 catalytic acids), there is crystallographic glycerol in HiBG (Figure S1B). While the electronic density of the PDB:4MDP glucose at the IS is significantly sparse compared to the glucose on the catalytic cleft on the other structures (Figure S1C), suggesting that a glucose transition site, the electronic density of HiBG crystallographic glycerol is approximately as high as that from the glucose usually found at this position (Figure S1C,D). Considering the relative chemical similarities between glycerol (a poly-alcohol) and carbohydrates, as well as the catalytic cleft, the usual position of glucose in GH1 enzymes, we have used two different starting structures for our GH1 MD simulations: The crystallographic one (with the original glucose starting at the experimental IS position) and a prepared one, with glucose at the catalytic position. This last structure was built by structural alignment with the *Neotermes koshunensis*  $\beta$ -glucosidase A complexed with glucose (PDB: 3VIJ, resolution of 1.03 Å), conserving the glucose from this last complex and the protein from the PDB: 4MDP one. The alignment was carried out considering just the protein backbone (RMSD of 0.6 Å) using the PyMOL software. For the GH1-cellobiose complex, once there is no experimental structure of the HiBG complexed with cellobiose, we have carried out just the structural backbone alignment with the PDB: 3VIK (*N. koshunensis*  $\beta$  glucosidase in complex with cellobiose, with a resolution of 1.10 Å and alignment RMSD of 0.611 Å), conserving the cellobiose from the PDB: 3VIK structure. For the simulations of the GH3 systems, both the protein structures and the respective glucose and cellobiose at the catalytic clefts of each original PDB were properly used.

For all the respective systems, the protonation states of the histidines were estimated by the H++ online server using the default salinity and dielectric parameters (respectively, 0.15 M, the internal dielectric of 10 and 80) and pH 7.00 [35].

#### 4.2. Molecular Dynamics Simulations

All the systems were solvated with a TIP3P cubic box with a 12 Å padding (measured from the outermost point of the protein) using the plugin tleap of the AmberTools package [36]. Na<sup>+</sup> and Cl<sup>-</sup> ions were added until neutrality and ionic strength of 0.15M. In each system, the protein, water, and ions were described by the ff99SB AMBER force field, while the respective carbohydrates were described by GLYCAM06 [37,38]. All the simulations were carried out at the NPT ensemble, keeping the temperature at 300 K by a Langevin thermostat and the pressure at 1 bar with an isotropic implementation of the Berendsen barostat. Periodic boundary conditions were carried out, using a cutoff of 10 Å for the nonbonded interactions, and solving the long-range electrostatic interactions by particle mesh Ewald (PME). The respective 1–4 interactions for protein and carbohydrate were scaled according to the specific default politics of each force field. All the simulations were executed using as a numeric integrator the AMBER16 software [36], a numeric time step of 2 fs and treating the hydrogen coordinates by the SETTLE constraint algorithm.

After the above-mentioned system prepare, a minimization/relaxing protocol was carried out consisting of: (i) 1000 steps of energy minimization by conjugate gradient, (ii) 300 ps of NPT pre-relaxation with harmonic restraint for protein and carbohydrate, (iii) 200 ps NPT relaxation removing the carbohydrate restrains, (iv) 200 ps of NPT relaxation without restrictions to the side chains of the residues 5 Å around the ligand, (v) 200 ps of NPT relaxation with no restrictions for the whole residues around the ligand, (vi) 10 ns of unprecedented NPT pre-production. This protocol was carried out in five replicates for the respective starting structures (after their respective protonation, solvation, and ionization) from each one of the five systems: HiBG-cellobiose, HiBG-glucose (crystallographic position), HiBG-glucose (catalytic position), AaBG-cellobiose, and AaBG-glucose. Henceforth, these specific starting systems will be called as GH1-cellobiose, GH1-glucose-1, GH1-glucose-2, GH3-cellobiose,

and GH3-glucose, unless previously mentioned in the opposite. After that, from each one of these respective five replicates per system, an independent productive MD protocol was carried out with the reinitializing of the velocities for a 300 K compatible distribution, the carrying out of a 60 ns NPT simulation (300 K and 1 bar) and saving the coordinates at every 4 ps. Altogether, five completely independent sampling simulations were carried out (from the minimization/relaxation to the productive step) for each respective GH1–cellobiose, GH3–cellobiose, GH3–glucose system and ten completely independent for the GH1–glucose system (considering the two different starting points for glucose at this system). For the GH1–glucose-2 system, the specific simulation in which the glucose releasing has turned evident was extended up to 90 ns (30 ns more) in order to allow the complete exit of this ligand. In this way, and taking together all the independent simulations, a sampling corresponding to 300 ns of productive MD (encompassing 75,000 frames) was collected for each respective GH1–cellobiose, GH3–glucose and GH3–cellobiose system and a 630 ns one (corresponding to 157,700 frames) for the GH1–glucose.

We also carried out GIST (grid inhomogeneous solvation theory) designated MD simulations just for the protein chain of the representative conformers selected from the protein–ligand FEL profiles. Firstly, we extracted the ligand and solvated the protein on an octahedral box using the AmberTools package with a minimum distance between the box edge and the protein of 12 Å [39]. We used the same TIP3P water model and described the protein–water system with the same ff99SB force field and used the same H++ estimated protonation states previously applied at the sampling MD. In sequence, we equilibrated and simulated the systems at respective 100 ns of productive MDs, but with 50 kcal mol<sup>-1</sup> coordinate restraints for the entire protein, as required by GIST analysis [40] at AMBER18 [39]. Moreover, we did describe the ions explicitly for the GIST designated MDs, having neutralized the system by a uniform plasma algorithm, as implemented in AMBER18. This last approach is recommended, in order to reduce the solvent complexity and allows the use of the respective references for bulk density and water–water interaction energy as the one of pure TIP3P simulations (see Section 4.6 below). All the remaining NPT and general MD conditions were maintained according to the sampling MD methodology above described. The coordinates were saved at every 100 ps for the GIST analysis itself.

### 4.3. MD Trajectories Analysis

2D-RMSD plots (all the frames against all the frames) related to the protein backbone were carried out using the cpptraj plugin from AmberTools [36] to check convergence along the different simulations for each different system. For the GH1–glucose system, in special, this analysis was carried out considering in separate simulations the two respective starting orientations of the glucose as these two systems together.

Principal component analysis (PCA), considering the protein backbone as the ligand heavy atoms were carried out using WORDOM [41]. In both cases, the covariance matrix calculation, diagonalization, and estimation of the eigenvector projections were made aligning the protein backbone of all the MD frames to the average structure. Principal components (PC) of proteins represent internal (conformational) movements, while the ligand PC is dominated by the translational and rotational motions of the ligand inside the active site.

To analyze the pattern of the different contact types between the ligand and protein (direct hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic contacts) along the different active site sub-regions, we used the LIGPLOT software [42]. For hydrogen-bond calculations, we considered 2.7 Å as the maximum hydrogen–acceptor distance and 3.35 Å as the maximum acceptor–donor one. Non-bonded contact parameters were performed using 2.90 Å as the minimum-contact distance and 3.90 Å as the maximum-contact one.

#### 4.4. Ligand–Protein Free Energy Landscape Estimations

To get a glimpse of the product/substrate energetic freedom inside and outside (for the system in that the ligand is released) the active site of each protein, we estimated the Gibbs free energy landscape (FEL) from the pose density using the plugin GMX from Gromacs 2.1 [43]. This algorithm uses a histogram method to estimate the relative  $\Delta G$  at different subregions of a 2D sampling space of two collective variables (respectively  $r_1$  and  $r_2$ ), descriptors of the phenomena of interest (in this case the ligand position along the trajectories) [44]. The  $\Delta G$  value at each subregion of the 2D plot is estimated by the Boltzmann distribution probabilities along this space according to the (1):

$$\Delta G(r_1, r_2) = -k_b T \ln P(r_1, r_2) \quad (1)$$

For the purpose of this study,  $r_1$  was set as the cartesian distance between the respective geometric centers from the ligand and a reference residue at the deeper part of the catalytic site, where  $r_2$  was the angle involving these two respective geometric centers and the geometric center of the protein. For the GH1 protein studied here, the reference residue was the catalytic E378, while for GH3 it was the H121 at the  $-1$  site.

Finally, from the FEL minima, the representative poses of the ligand pathway along the active site were selected using a homemade algorithm that located the more densely populated grids along different subregions of the 2D histogram and the respective trajectory frames that more approximated from these values (in  $r_1$  and  $r_2$ ).

#### 4.5. Poisson–Boltzmann Surface Area

To get insights into the mechanisms of energetic complementarity of the protein active site to the substrate and product at the representative poses accessed along the MD simulations, both the electrostatic potential distribution along the protein surface (with special attention to the active site) and the water energetics along the same active site were estimated and compared.

In order to have a glimpse of the electrostatic potential distribution, we estimated the Poisson–Boltzmann surface area, with the integrated use of the online versions of the PDB2PQR and the adaptive Poisson–Boltzmann solver (APBS) tools, respectively [45–48]. Basically, the PDB2PQR software was used to generate the input files for the PBSA calculations with APBS, i.e., the pqr file, containing the PDB atomic coordinates as well the atomic charges and radii at the respective beta and gamma columns, and an input file containing the grid dimensions, external and internal dielectrics, solvent probe radius, as well the remaining parameters for the Poisson–Boltzmann equation approximation. In sequence, APBS was used to estimate the surface electrostatic potential along the protein surface according to the linearized numerical approximation of the differential Poisson–Boltzmann equation, the original differential equation depicted in (2):

$$\nabla^2 \psi = -\frac{c_0 \beta}{\epsilon_{solv} \cdot \epsilon_{sol}} \left[ e^{\frac{-\beta \psi(x,y,z)}{k_B T}} - e^{\frac{\beta \psi(x,y,z)}{k_B T}} \right] \quad (2)$$

where  $\psi$  is the electric three-dimensional potential along the solute surface,  $c_0$  is the solvent ionic concentration (here considered at the default of 0.15 M),  $\epsilon_{solv}$  and  $\epsilon_{sol}$  are the respective dielectric values of the solvent (here considered at the default of 78.54) and of the solute (i.e., the protein, here considered at the default of 2),  $\beta$  is the absolute value of the charge of an electron ( $1.602 \times 10^{-19}$  coulombs).

In order to be self-consistent, the atomic charges were attributed according to the same AMBER force fields previously used on the MD simulations (ff99SB for the protein and GLYCAM06 at the single calculation made for glucose) and the same protonation states were equally preserved. The remaining PBSA parameters were taken to APBS calculations by the software default.

The PBSA maps generated by the grid files were analyzed as images at the VMD software [49]. To analyze the electrostatic distribution in a comparative approach, the electrostatic maps were

superposed at the respective structures using a color scale in the range of  $-20.00: 0.00: 20.00$  for the dimensionless surface electrostatic potential  $\psi$  in all the systems.

#### 4.6. Grid Inhomogeneous Solvation Theory Analyses

The structural water occupancy and energetics around the active site was analyzed by the grid inhomogeneous solvation theory (GIST) of the representative conformers of the minima obtained from the FEL analysis. The concept of GIST is briefly outlined in this section. For a more detailed discussion of the theoretical background, we recommend the studies described in [40,50,51]. For a better glimpse of the state of the art concerning the recent usability of this technique in different and complex biochemical systems, accurate examples can be found in [50,52,53].

Basically, GIST is a method to calculate the position-dependent free energy of the water molecules in a system on a grid-based approach. The method is made in combination with molecular dynamics simulation and takes into consideration only the phase space of the water molecules, restraining the solute to a single conformation. As this is a rough simplification of a mobile solute, multiple GIST calculations can be used to estimate the free energy of solvation to different conformers  $q$  of this same solute  $\Delta G_{\text{sol}v}(q)$ . The final results according to the differential probability of each conformer  $p(q)$  by the summation (3):

$$\Delta G_{\text{sol}v} \approx \sum \Delta G_{\text{sol}v}(q)p(q) \quad (3)$$

The objective of the GIST calculations was to compare the differences in the energetic behavior of the water along the extension of the protein surface at the active pocket. This aimed to estimate the contribution of this same water energetic differences on the accessibility of the ligand to migrate from one subsite of the active pocket to the other as well from each subsite to the bulk or vice-versa (looking for possible paths of ligand escaping or retention based on the water exclusion principle). In this way we did not integrate the grid estimated  $\Delta G_{\text{sol}}$  values, analyzing the results on a per voxel approach, i.e., we estimated the free energy of water molecules transfer from the bulk to different subsites of the studied proteins at different decoys from their respective FELs. These subsites, in turn, are computationally described as different voxels  $v$ . This per voxel  $\Delta G_{\text{sol}}$  value (here called  $\Delta G^v_{\text{sol}}$ ), in turn, can be naturally split in two respective enthalpic and two respective entropic per voxel terms as shown in (4):

$$\Delta G^v_{\text{sol}v} = \Delta E^v_{\text{sw}} + \Delta E^v_{\text{ww}} - T\Delta S^v_{\text{trans}} - T\Delta S^v_{\text{orient}} \quad (4)$$

where the enthalpic terms  $\Delta E_{\text{sw}}$  and  $\Delta E_{\text{ww}}$  are the respective solute–water and water–water interaction energies per voxel, as calculated by the molecular mechanics' force field. The solvation entropy terms, in turn, encompass the respective contributions of the translational  $\Delta S_{\text{trans}}$  and orientational (i.e., angular)  $\Delta S_{\text{orient}}$  of the water molecules around the different voxels, defined as Shannon entropies and algorithmically calculated via the nearest neighbor estimats and orientational (i.e., angular)  $\Delta S_{\text{orient}}$  of the water molecules around the different voxels, defined as Shannon entropies and algorithmically calculated via the nearest neighbor estimation [42,52–54]. All terms depicted in (4) are state functions, and therefore, need a reference state. In our case, the simulation of pure water model TIP3P was done without any solute. The reference value for the water–water interaction is, therefore,  $-9.533$  kcal/mol, whereas the reference values for the remaining terms  $\Delta E_{\text{sw}}$ ,  $\Delta S_{\text{trans}}$ ,  $\Delta S_{\text{orient}}$  are zero. At the specific case of the  $\Delta E^v_{\text{ww}}$ , with a non-zero reference value, and considering our per voxel approach, it is also necessary to consider how much each voxel is populated on average by water molecules along the simulation frames compared to the pure water. In this way, the reference water–water interaction in each voxel  $E^v_{\text{ww}}(\text{Ref})$  was calculated multiplying the respective numerical values of the reference energy in units of kcal/mol on a simple water–water interaction ( $-9.533$ ) by the reference water density in units of water molecules/ $\text{\AA}^3$  on a pure TIP3P simulation (0.0329) and the average per frame number

density of water oxygen centers found in each voxel compared to the bulk density ( $g^v(O)$  in units of  $\text{\AA}^3/0.0329$  molecules), resulting in the equality 5:

$$E^v_{ww}(Ref) = -0.3136 g^v(O) \text{ kcal/mol} \quad (5)$$

We have carried out the GIST analyses by the algorithm implemented at the cpptraj tool of the most recent version of the AmberTools package [36,55]. To proceed with the calculations of the respective total values of  $\Delta E^v_{sw}$ ,  $\Delta E^v_{ww}$ ,  $\Delta S^v_{trans}$ , and  $\Delta S^v_{orient}$  from the density-weighted values firstly returned by the GIST tool from AMBER, as well to combine these values in order to obtain the  $\Delta G^v_{solv}$  values, we used the GISTPP tool [55]. The protocol used was that described in the same reference, as well as in the related tutorial provided by the authors' website. The only addendum was the procedure to subtract the per voxel water–water interaction reference  $E^v_{ww}(Ref)$  estimated according to (5) in order to obtain the correctly bulk referenced  $\Delta G^v_{solv}$  values.

Looking for a higher simplification of the analysis, in all the calculations, only the protein was considered in the GIST preparative MDs and the GIST calculations. In this way, the effect of the carbohydrate ligand was considered as responsive to the water energetic effects occasioned by the protein macro-environment. Although this approach carries the error of depreciate possible modulations of the local water energetic by the ligand (overall dealing with hygroscopic ligands as carbohydrates), to introduce a rigid carbohydrate (being a naturally high flexible ligand) on the GIST designed MDs, as well on the GIST calculations themselves, could introduce more local artifacts than accuracy. Besides this, it is reasonable to suppose, overall on the semi-quantitative per site analysis here carried out, that the protein macro-environment will provide a more deterministic influence on the water energetics and that the small glucose or cellobiose molecule will respond to that.

The reference solvent density (as mentioned above) was considered as  $0.0329$  water molecules/ $\text{\AA}^3$  (adequate to the TIP3P model), and the default value of  $0.5 \text{ \AA}$  was used for the grid size (volume of  $0.125 \text{\AA}^3$  for each voxel). For all the GH1 conformers, analyses were carried out at the same box containing grid dimensions equal to 70, 90 and 80 at the respective X Y Z axes. Analogously, the GIST analyses were carried out at the same 90, 90, 90 grid box for all the GH3 conformers. In both cases, the box was placed in such a way to encompass the catalytic residues, the +1 and -1 sites, as well the +2 one in the case of the GH1 enzyme, beyond all the extension of the exit channel and loops around.

Finally, for the analysis of the significantly favorable or unfavorable (i.e., constrained) hydration regions, only the sites containing water molecules considerably retained (respective  $g^v(O)$  values  $\geq 10$  units of the bulk density) and with  $\Delta G^v_{solv}$  values respectively  $\leq -3 \text{ kcal}\cdot\text{mol}^{-1}$  or  $\geq 3 \text{ kcal}\cdot\text{mol}^{-1}$  were considered.

## 5. Conclusions

We used molecular dynamics simulations to understand the behavior of glucose and cellobiose in a GH1 (HiBG) and a GH3 (AaBG)  $\beta$ -glucosidase. We hypothesize that the explanation for glucose tolerance is related to a sophisticated mechanism of ligand release by the substrate channel, herein called "slingshot mechanism", that affects glucose more than cellobiose in GH1 enzymes. The proposed mechanism is in agreement with previous studies [3,12,13,25], confirming that the shape of the substrate channel of glucose-tolerant enzymes, the pharmacophoric properties of amino acids, the affinity of glucose by allosteric sites, and the protein motion act together to allow the glucose tolerance. It suggests that ligands in glucose-tolerant  $\beta$ -glucosidases are guided to a hydrophobic bottleneck region in the entrance of the substrate channel and released by interactions with a hydrogen bond acceptor amino acid residue present in this region (D238 in HiBG, Videos S1 and S2). Furthermore, the dynamics of favorable and unfavorable waters, the electrostatic distribution, as well as the product/substrate fitting relationship contribute to explain the glucose tolerance of GH1  $\beta$ -glucosidases and the non-tolerance of the GH3s. The results presented in this paper open new pathways for elucidating the mechanisms of tolerance and inhibition by the product. Beyond that, they give new insights for the rational design



of new enzymes resistant to inhibition and, consequently, the viability of the large-scale production of biofuels.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/1420-3049/24/18/3215/s1>. Supplementary PDF file (Figures S1–S13, Tables S1–S6). Video S1: MD of the glucose exit in a glucose-tolerant GH1  $\beta$ -glucosidase (also available at <<https://youtu.be/fzWynXUbdcl>>). Video S2: Interactions among glucose, D228, K257, and N312 (also available at <<https://youtu.be/7VzR0CVag6I>>).

**Author Contributions:** Conceptualization, L.H.F.L.; methodology, L.S.C.C.; data curation, L.S.C.C., D.C.B.M., R.E.O.R. and L.H.F.L.; data visualization, D.C.B.M.; writing—original draft preparation, L.H.F.L.; writing—review and editing, L.S.C.C., D.C.B.M., J.K., C.H.S., K.R.L. and R.C.M.-M.; project administration, L.H.F.L., K.R.L., C.H.S. and R.C.M.-M.; funding acquisition, R.C.M.-M.

**Funding:** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Project number: 51/2013 - 23038.004007/2014-82.

**Acknowledgments:** The authors thank the Brazilian funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Swangkeaw, J.; Vichitphan, S.; Butzke, C.E.; Vichitphan, K. Characterization of  $\beta$ -glucosidases from *Hanseniaspora* sp. and *Pichia anomala* with potentially aroma-enhancing capabilities in juice and wine. *World J. Microbiol. Biotechnol.* **2010**, *27*, 423–430. [[CrossRef](#)]
- Cota, J.; Corrêa, T.L.R.; Damásio, A.R.L.; Diogo, J.A.; Hoffmam, Z.B.; Garcia, W.; Oliveira, L.C.; Prade, R.A.; Squina, F.M. Comparative analysis of three hyperthermophilic GH1 and GH3 family members with industrial potential. *New Biotechnol.* **2015**, *32*, 13–20. [[CrossRef](#)] [[PubMed](#)]
- Mariano, D.C.B.; Leite, C.; Santos, L.H.S.; Marins, L.F.; Machado, K.S.; Werhli, A.V.; Lima, L.H.F.; de Melo-Minardi, R.C. Characterization of glucose-tolerant  $\beta$ -glucosidases used in biofuel production under the bioinformatics perspective: A systematic review. *Genet. Mol. Res.* **2017**, *16*, 1–19. [[CrossRef](#)] [[PubMed](#)]
- Singhania, R.R.; Patel, A.K.; Sukumaran, R.K.; Larroche, C.; Pandey, A. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. *Bioresour. Technol.* **2013**, *127*, 500–507. [[CrossRef](#)] [[PubMed](#)]
- Kumar, R.; Singh, S.; Singh, O.V. Bioconversion of lignocellulosic biomass: Biochemical and molecular perspectives. *J. Ind. Microbiol. Biotechnol.* **2008**, *35*, 377–391. [[CrossRef](#)] [[PubMed](#)]
- Cairns, J.R.K.; Esen, A.  $\beta$ -Glucosidases. *Cell. Mol. Life Sci.* **2010**, *67*, 3389–3405. [[CrossRef](#)] [[PubMed](#)]
- Béguin, P.; Aubert, J.P. The biological degradation of cellulose. *FEMS Microbiol. Rev.* **1994**, *13*, 25–58. [[CrossRef](#)]
- Teugjas, H.; Väljamäe, P. Selecting  $\beta$ -glucosidases to support cellulases in cellulose saccharification. *Biotechnol. Biofuels* **2013**, *6*, 105. [[CrossRef](#)]
- Yang, F.; Yang, X.; Li, Z.; Du, C.; Wang, J.; Li, S. Overexpression and characterization of a glucose-tolerant  $\beta$ -glucosidase from *T. aotearoense* with high specific activity for cellobiose. *Appl. Microbiol. Biotechnol.* **2015**, *99*, 8903–8915. [[CrossRef](#)]
- Zhao, L.; Pang, Q.; Xie, J.; Pei, J.; Wang, F.; Fan, S. Enzymatic properties of *Thermoanaerobacterium thermosaccharolyticum*  $\beta$ -glucosidase fused to *Clostridium cellulovorans* cellulose binding domain and its application in hydrolysis of microcrystalline cellulose. *BMC Biotechnol.* **2013**, *13*, 101. [[CrossRef](#)]
- Chamoli, S.; Kumar, P.; Navani, N.K.; Verma, A.K. Secretory expression, characterization and docking study of glucose-tolerant  $\beta$ -glucosidase from *B. subtilis*. *Int. J. Biol. Macromol.* **2016**, *85*, 425–433. [[CrossRef](#)] [[PubMed](#)]
- de Giuseppe, P.O.; Souza, T.; de, A.C.B.; Souza, F.H.M.; Zanphorlin, L.M.; Machado, C.B.; Ward, R.J.; Jorge, J.A.; dos Furriel, R.P.M.; Murakami, M.T. Structural basis for glucose tolerance in GH1  $\beta$ -glucosidases. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70*, 1631–1639. [[CrossRef](#)]
- Yang, Y.; Zhang, X.; Yin, Q.; Fang, W.; Fang, Z.; Wang, X.; Zhang, X.; Xiao, Y. A mechanism of glucose tolerance and stimulation of GH1  $\beta$ -glucosidases. *Sci. Rep.* **2015**, *5*, 17296. [[CrossRef](#)] [[PubMed](#)]

14. Pei, J.; Pang, Q.; Zhao, L.; Fan, S.; Shi, H. Thermoanaerobacterium thermosaccharolyticum  $\beta$ -glucosidase: A glucose-tolerant enzyme with high specific activity for cellobiose. *Biotechnol Biofuels* **2012**, *5*, 1–10. [[CrossRef](#)]
15. Guo, B.; Amano, Y.; Nozaki, K. Improvements in Glucose Sensitivity and Stability of *Trichoderma reesei*  $\beta$ -Glucosidase Using Site-Directed Mutagenesis. *PLoS ONE* **2016**, *11*, e0147301. [[CrossRef](#)] [[PubMed](#)]
16. Cao, L.; Wang, Z.; Ren, G.; Kong, W.; Li, L.; Xie, W.; Liu, Y. Engineering a novel glucose-tolerant  $\beta$ -glucosidase as supplementation to enhance the hydrolysis of sugarcane bagasse at high glucose concentration. *Biotechnol. Biofuels* **2015**, *8*. [[CrossRef](#)] [[PubMed](#)]
17. Uchima, C.A.; Tokuda, G.; Watanabe, H.; Kitamoto, K.; Arioka, M. Heterologous expression and characterization of a glucose-stimulated  $\beta$ -glucosidase from the termite *Neotermes koshunensis* in *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* **2011**, *89*, 1761–1771. [[CrossRef](#)]
18. Souza, F.H.M.; Nascimento, C.V.; Rosa, J.C.; Masui, D.C.; Leone, F.A.; Jorge, J.A.; Furriel, R.P.M. Purification and biochemical characterization of a mycelial glucose- and xylose-stimulated  $\beta$ -glucosidase from the thermophilic fungus *Humicola insolens*. *Process Biochem.* **2010**, *45*, 272–278. [[CrossRef](#)]
19. Souza, F.H.M.; Inocentes, R.F.; Ward, R.J.; Jorge, J.A.; Furriel, R.P.M. Glucose and xylose stimulation of a  $\beta$ -glucosidase from the thermophilic fungus *Humicola insolens*: A kinetic and biophysical study. *J. Mol. Catal. B Enzym.* **2013**, *94*, 119–128. [[CrossRef](#)]
20. Crespim, E.; Zanphorlin, L.M.; de Souza, F.H.M.; Diogo, J.A.; Gazolla, A.C.; Machado, C.B.; Figueiredo, F.; Sousa, A.S.; Nóbrega, F.; Pellizari, V.H.; et al. A novel cold-adapted and glucose-tolerant GH1  $\beta$ -glucosidase from *Exiguobacterium antarcticum* B7. *Int. J. Biol. Macromol.* **2016**, *82*, 375–380. [[CrossRef](#)]
21. Meleiro, L.P.; Salgado, J.C.S.; Maldonado, R.F.; Alpointi, J.S.; Zimbardi, A.L.R.L.; Jorge, J.A.; Ward, R.J.; Furriel, R.P.M. A *Neurospora crassa*  $\beta$ -glucosidase with potential for lignocellulose hydrolysis shows strong glucose tolerance and stimulation by glucose and xylose. *J. Mol. Catal. B Enzym.* **2015**, *122*, 131–140. [[CrossRef](#)]
22. Salgado, J.C.S.; Meleiro, L.P.; Carli, S.; Ward, R.J. Glucose tolerant and glucose stimulated  $\beta$ -glucosidases-A review. *Bioresour. Technol.* **2018**, *267*, 704–713. [[CrossRef](#)] [[PubMed](#)]
23. Jabbour, D.; Klippel, B.; Antranikian, G. A novel thermostable and glucose-tolerant  $\beta$ -glucosidase from *Fervidobacterium islandicum*. *Appl. Microbiol. Biotechnol.* **2012**, *93*, 1947–1956. [[CrossRef](#)] [[PubMed](#)]
24. Pang, P.; Cao, L.; Liu, Y.; Xie, W.; Wang, Z. Structures of a glucose-tolerant  $\beta$ -glucosidase provide insights into its mechanism. *J. Struct. Biol.* **2017**, *198*, 154–162. [[CrossRef](#)] [[PubMed](#)]
25. Mariano, D.C.B.; Santos, L.H.; Machado, K.D.S.; Werhli, A.V.; de Lima, L.H.F.; de Melo-Minardi, R.C. A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). *Int. J. Mol. Sci.* **2019**, *20*, 333. [[CrossRef](#)] [[PubMed](#)]
26. Breves, R.; Bronnenmeier, K.; Wild, N.; Lottspeich, F.; Staudenbauer, W.L.; Hofemeister, J. Genes encoding two different beta-glucosidases of *Thermoanaerobacter brockii* are clustered in a common operon. *Appl. Environ. Microbiol.* **1997**, *63*, 3902–3910.
27. Chen, W.; Enck, S.; Price, J.L.; Powers, D.L.; Powers, E.T.; Wong, C.-H.; Dyson, H.J.; Kelly, J.W. The Structural and Energetic Basis of Carbohydrate Aromatic Packing Interactions in Proteins. *J. Am. Chem. Soc.* **2013**, *135*, 9877–9884. [[CrossRef](#)] [[PubMed](#)]
28. Duff, M.R.; Howell, E.E. Thermodynamics and solvent linkage of macromolecule-ligand interactions. *Methods San Diego Calif* **2015**, *76*, 51–60. [[CrossRef](#)]
29. Geronimo, I.; Ntarima, P.; Piens, K.; Gudmundsson, M.; Hansson, H.; Sandgren, M.; Payne, C.M. Kinetic and molecular dynamics study of inhibition and transglycosylation in *Hypocrea jecorina* family 3  $\beta$ -glucosidases. *J. Biol. Chem.* **2019**, *294*, 3169–3180. [[CrossRef](#)]
30. Bergmann, J.C.; Costa, O.Y.A.; Gladden, J.M.; Singer, S.; Heins, R.; D'haeseleer, P.; Simmons, B.A.; Quirino, B.F. Discovery of two novel  $\beta$ -glucosidases from an Amazon soil metagenomic library. *FEMS Microbiol. Lett.* **2014**, *351*, 147–155. [[CrossRef](#)]
31. Spiwok, V. CH/ $\pi$  Interactions in Carbohydrate Recognition. *Molecules* **2017**, *22*, 1038. [[CrossRef](#)] [[PubMed](#)]
32. Matsuzawa, T.; Jo, T.; Uchiyama, T.; Manninen, J.A.; Arakawa, T.; Miyazaki, K.; Fushinobu, S.; Yaoi, K. Crystal structure and identification of a key amino acid for glucose tolerance, substrate specificity, and transglycosylation activity of metagenomic  $\beta$ -glucosidase Td2F2. *FEBS J.* **2016**, *283*, 2340–2353. [[CrossRef](#)] [[PubMed](#)]

33. Suzuki, K.; Sumitani, J.-I.; Nam, Y.-W.; Nishimaki, T.; Tani, S.; Wakagi, T.; Kawaguchi, T.; Fushinobu, S. Crystal structures of glycoside hydrolase family 3  $\beta$ -glucosidase 1 from *Aspergillus aculeatus*. *Biochem. J.* **2013**, *452*, 211–221. [[CrossRef](#)] [[PubMed](#)]
34. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
35. Anandakrishnan, R.; Aguilar, B.; Onufriev, A.V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, 537–541. [[CrossRef](#)] [[PubMed](#)]
36. Case, D.A.; Betz, R.M.; Cerutti, D.S. *Proceedings of the AMBER 16*; University of California: San Francisco, CA, USA, 2016.
37. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958. [[CrossRef](#)] [[PubMed](#)]
38. Kirschner, K.N.; Yongye, A.B.; Tschampel, S.M.; González-Outeiriño, J.; Daniels, C.R.; Foley, B.L.; Woods, R.J. GLYCAM06: A generalizable biomolecular force field. *Carbohydrates. J. Comput. Chem.* **2008**, *29*, 622–655. [[CrossRef](#)] [[PubMed](#)]
39. Case, D.A.; Ben-Shalom, I.Y.; Brozell, S.R. *Proceedings of the AMBER 18*; University of California: San Francisco, CA, USA, 2018.
40. Nguyen, C.N.; Kurtzman Young, T.; Gilson, M.K. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril. *J. Chem. Phys.* **2012**, *137*, 044101. [[CrossRef](#)]
41. Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. Wordom: A program for efficient analysis of molecular dynamics simulations. *Bioinforma. Oxf. Engl.* **2007**, *23*, 2625–2627. [[CrossRef](#)]
42. Wallace, A.C.; Laskowski, R.A.; Thornton, J.M. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **1995**, *8*, 127–134. [[CrossRef](#)]
43. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.E.; Berendsen, H.J.C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718. [[CrossRef](#)] [[PubMed](#)]
44. Stock, G.; Jain, A.; Riccardi, L.; Nguyen, P.H. Exploring the Energy Landscape of Small Peptides and Proteins by Molecular Dynamics Simulations. In *Protein and Peptide Folding, Misfolding, and Non-Folding*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2012; pp. 55–77. ISBN 978-1-118-18337-3.
45. Dolinsky, T.J.; Nielsen, J.E.; McCammon, J.A.; Baker, N.A. PDB2PQR: An automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, 665–667. [[CrossRef](#)] [[PubMed](#)]
46. Baker, N.A.; Sept, D.; Joseph, S.; Holst, M.J.; McCammon, J.A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Nat. Acad. Sci. USA* **2001**, *98*, 10037–10041. [[CrossRef](#)] [[PubMed](#)]
47. Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L.E.; Brookes, D.H.; Wilson, L.; Chen, J.; Liles, K.; et al. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **2018**, *27*, 112–128. [[CrossRef](#)] [[PubMed](#)]
48. Unni, S.; Huang, Y.; Hanson, R.M.; Tobias, M.; Krishnan, S.; Li, W.W.; Nielsen, J.E.; Baker, N.A. Web servers and services for electrostatics calculations with APBS and PDB2PQR. *J. Comput. Chem.* **2011**, *32*, 1488–1491. [[CrossRef](#)] [[PubMed](#)]
49. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [[CrossRef](#)]
50. Nguyen, C.N.; Cruz, A.; Gilson, M.K.; Kurtzman, T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput.* **2014**, *10*, 2769–2780. [[CrossRef](#)]
51. Schauerl, M.; Podewitz, M.; Waldner, B.J.; Liedl, K.R. Enthalpic and Entropic Contributions to Hydrophobicity. *J. Chem. Theory Comput.* **2016**, *12*, 4600–4610. [[CrossRef](#)]
52. Schauerl, M.; Podewitz, M.; Ortner, T.S.; Waibl, F.; Thoeny, A.; Loerting, T.; Liedl, K.R. Balance between hydration enthalpy and entropy is important for ice binding surfaces in Antifreeze Proteins. *Sci. Rep.* **2017**, *7*, 11901. [[CrossRef](#)]
53. Schauerl, M.; Czodrowski, P.; Fuchs, J.E.; Huber, R.G.; Waldner, B.J.; Podewitz, M.; Kramer, C.; Liedl, K.R. Binding Pose Flip Explained via Enthalpic and Entropic Contributions. *J. Chem. Inf. Model.* **2017**, *57*, 345–354. [[CrossRef](#)]

54. Pathria, R.K.; Beale, P.D. *Statistical Mechanics*; Academic Press: Cambridge, MA, USA, 2011; ISBN 978-0-12-382189-8.
55. Ramsey, S.; Nguyen, C.; Salomon-Ferrer, R.; Walker, R.C.; Gilson, M.K.; Kurtzman, T. Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. *J. Comput. Chem.* **2016**, *37*, 2029–2037. [[CrossRef](#)] [[PubMed](#)]

**Sample Availability:** Not available.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

## Apêndice 9. Estratégias para produção de $\beta$ -glicosidases glicose-tolerantes

Partindo de uma estrutura de  $\beta$ -glicosidase experimentalmente caracterizada como pouco eficiente para produção de biocombustíveis, sugere-se melhorar sua atividade seguindo os seguintes passos:

1. Determine experimentalmente a estrutural tridimensional da enzima;
  - a. Caso não seja possível, pode-se determinar a estrutura primária por sequenciamento;
  - b. Modelar por comparação;
2. Determine os resíduos do bolsão catalítico;
  - a. Os resíduos presentes no canal que leva ao sítio ativo têm sido reportados como de grande importância para glicose-tolerância;
  - b. Para determiná-los, pode-se realizar alinhamento estrutural de sua  $\beta$ -glicosidase com a  $\beta$ -glicosidase de *N. kosshunensis* (3VIK) e extrair os resíduos da sua enzima a uma distância de 6,5 Å da celobiose de 3VIK;
  - c. É importante ressaltar que foram reportadas algumas mutações que afetam a termoestabilidade de  $\beta$ -glicosidases em sítios distantes do bolsão catalítico, como por exemplo a mutação M416I na  $\beta$ -glicosidase de *Bacillus polymyxa* (LOPEZ-CAMACHO et al., 1996);
3. Verifique a subsequência consenso “HWNEWCLHNL TANYYTNEW EWF”;
  - a. Ela é composta pelos resíduos que mais aparecem nos bolsões catalíticos das glicose-tolerantes;
  - b. Para obter-se os resíduos correspondentes pode-se realizar:
    - i. Alinhamento de sequências (programas: Clustal ou BLAST);
    - ii. Alinhamento estrutural (programas: MultiProt ou PyMOL);
  - c. A subsequência corresponde na  $\beta$ -glicosidase de *T. brockii* aos resíduos: H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410 e F418;
  - d. Atenção ao aminoácido na posição 224. Essa posição varia muito em enzimas da família GH1, enquanto nas glicose-tolerantes é ocupada por uma asparagina ou uma alanina;

- e. A substituição de um aminoácido polar, básico ou volumoso na posição 228 por resíduos pouco volumosos capazes de agir como acceptor em ligações de hidrogênio, como serina, treonina ou cisteína pode aumentar a resistência a inibição;
  - f. A troca de um aminoácido hidrofóbico por um aminoácido polar na posição 181 também pode ser benéfica para a hidrólise, uma vez que permite a formação de um túnel de água até o sítio ativo;
  - g. A presença de um triptofano e uma leucina nas posições 169 e 174 pode favorecer a liberação da glicose e reduzir a inibição;
  - h. Mutações nos resíduos do sítio ativo E167 e E355 pode resultar na completa inativação da enzima (WITHERS et al., 1992), portanto não são recomendadas (o mesmo é válido para H121);
  - i. Além dos resíduos H121, E167 e E355, os resíduos W122, N166, N297, Y299, W402, E409, W410 e F418 aparecem em mais de 90% das glicose-tolerantes, logo podem ter algum papel importante na ligação do substrato. Entretanto, deve-se levar em consideração que uma mutação nesses resíduos pode gerar características não encontradas em outras  $\beta$ -glicosidases glicose-tolerantes (ou pelo menos ainda não reportadas na literatura), que podem ser benéficas ou não;
  - j. Padrões que caracterizam a glicose-tolerância em sequências são complexos de ser analisados pelo olhar humano. Logo a combinação de possíveis mutações benéficas pode ser transferida por meio de técnicas que usam informações de estruturas tridimensionais. Esta tese sugere o uso do SSV;
4. Construa mutantes *in silico*;
- a. A validação de mutantes *in silico* pode ajudar a determinar melhores mutantes para serem testados em bancada, o que permitirá uma redução de custos;
  - b. Podem ser implementadas com ferramentas como PyMOL ou MODELLER;
  - c. Aplicação de minimização e dinâmica molecular pode gerar melhores resultados, entretanto aumenta os custos computacionais (tempo de processamento);
5. Determine proteínas *template*;
- a. Para  $\beta$ -glicosidases recomenda-se o uso da base BETAGDB como referência para *templates*;
  - b. Entretanto, outras enzimas podem ser utilizadas como referência. SSV irá detectar padrões na assinatura da base de referência e avaliará quais mutações

aproximam a proteína selvagem da proteína *template*. Por esse motivo, SSV pode ser generalizado para outros tipos de enzimas;

6. Execute SSV;
  - a. SSV pode ser executado online pelo site: <<http://bioinfo.dcc.ufmg.br/ssv>>. Entretanto, o site permite que apenas uma mutação seja testada por vez;
  - b. Para execução em larga escala pode-se fazer o *download* do código-fonte do SSV e dependências e executar localmente (requer Python e Perl instalados). Neste caso, recomenda-se realizar todas as mutações possíveis nos resíduos do bolsão catalítico e, eventualmente, testar algumas combinações de mutações;
7. Filtre os resultados;
  - a. Dependendo da quantidade de mutações testadas, SSV pode sugerir muitas mutações;
  - b. Assim, pode-se combinar SSV com outras estratégias para propor mutações, como por exemplo, estratégias baseadas sequências (alinhamentos com Clustal Omega, análises com o ProSAR), conservação na família (SIFT), impacto na diferença de variação de energia livre (mCSM) ou outras análises baseadas em estrutura tridimensional (*active site constellations* e BioGPS);
  - c. SSV pode ser utilizado em conjunto com outras ferramentas para obter sugestões de mutações mais recomendadas;
8. Teste em bancada;
  - a. Após definidas possíveis sugestões de mutações é importante que sejam testadas em bancada;
  - b. Este trabalho usou dados obtidos da literatura nos estudos de caso que validaram o modelo proposto. Entretanto, reforça-se que é de vital importância que dados sejam validados em bancada.

Por fim, algumas considerações devem ser descritas. A glicose-tolerância é a característica descrita na literatura como de maior importância para a melhoria de enzimas utilizadas na produção de biocombustíveis de segunda-geração. Porém, para a indústria é de vital importância que as mutações gerem um real aumento na conversão da matéria-prima em açúcares fermentáveis, como o aumento da conversão do bagaço da cana entre 14 e 35% detectado no trabalho de CAO et al. (2015). Logo, para a seleção de  $\beta$ -glicosidases para uma produção mais eficiente deve-se considerar diversos fatores, como o tipo de processo que será adotado (SHF ou SFF), o ambiente em que a  $\beta$ -glicosidase será utilizada (temperatura e pH), a maté-

ria-prima usada, o organismo que irá produzir as  $\beta$ -glicosidases e se ele estará no meio ou a produção será separada e haverá a suplementação da enzima. Ou seja, deve-se levar em consideração muitos fatores, que podem ser utilizados na etapa de definição dos *templates*.

Além de tudo isso, deve-se apontar que é possível sintetizar proteínas já reportadas como glicose-tolerantes com base na sequência disponível em bancos de dados públicos. Entretanto, muitos grupos de pesquisa preferem trabalhar com proteínas cujo protocolo de manipulação em bancada já esteja estabelecido em seu grupo, devido a dificuldades técnicas em etapas como expressão e purificação. Para esse público-alvo, as mutações pontuais avaliadas por SSV podem ser de grande utilidade.



## Apêndice 10. Número de contatos entre resíduos e celobiose em $\beta$ -glicosidasas GH1 glicose-tolerantes.

<i>Organismo/Metagenoma</i>	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
<i>Thermoanaerobacter brockii</i>	Resíduos	H	W	N	E	W	C	L	H	N	L	T	A	N	Y	Y	T	S	E	W	E	W	F
	Contatos	0	20	1	108	135	60	97	109	147	6	167	51	0	0	205	146	151	6	3	41	57	0
<i>Nasutitermes takasagoensis</i>	Resíduos	H	W	N	E	R	T	D	M	D	N	A	F	N	F	Y	T	N	E	W	E	W	F
	Contatos	7	38	7	103	740	261	275	306	238	0	51	8	0	0	100	28	242	17	13	172	121	6
<i>Metagenome China South Sea</i>	Resíduos	H	W	N	E	F	C	L	H	N	F	T	A	N	F	Y	T	Q	E	W	E	W	F
	Contatos	3	49	8	131	44	91	171	189	402	0	118	0	0	0	121	42	183	30	9	100	147	0
<i>Thermoanaerobacterium thermosaccharolyticum</i>	Resíduos	H	W	N	E	W	C	L	H	N	L	N	A	N	Y	Y	T	A	E	W	E	W	F
	Contatos	5	46	5	124	271	97	262	154	162	0	398	51	0	0	126	52	100	12	6	47	66	6
<i>Metagenome Turpan Depression</i>	Resíduos	H	W	N	E	W	V	G	L	N	I	E	R	N	W	Y	T	A	A	W	E	W	F
	Contatos	8	100	14	176	194	129	30	97	177	0	121	0	0	0	172	3	0	41	30	122	222	16
<i>Trichoderma reesei</i>	Resíduos	H	W	N	E	W	C	L	F	N	G	D	R	N	H	Y	T	N	E	W	E	W	F
	Contatos	11	124	16	179	93	76	153	219	273	0	98	3	0	0	188	0	0	45	41	116	125	38
<i>Caldicellulosiruptor bescii</i>	Resíduos	H	W	N	E	Y	C	L	H	N	L	T	S	N	Y	Y	T	A	E	W	E	W	F
	Contatos	3	51	11	226	151	73	533	128	284	0	204	0	10	0	218	6	0	40	21	90	115	9
<i>Thermoanaerobacterium aotea-roense</i>	Resíduos	H	W	N	E	W	C	L	H	N	L	T	A	N	F	Y	S	S	E	W	E	W	F
	Contatos	0	47	3	136	192	73	154	106	136	1	255	3	0	0	249	206	211	18	6	24	42	4
<i>Metagenome soil</i>	Resíduos	H	W	N	E	L	C	I	M	N	L	S	A	N	Y	Y	F	A	E	W	E	W	F
	Contatos	24	202	22	150	3	80	273	12	397	2	6	0	1	0	213	232	0	77	71	116	309	28
<i>Metagenome Kusaya gravy</i>	Resíduos	H	W	N	E	Q	C	L	H	T	C	G	E	N	I	Y	N	Y	E	W	E	W	F
	Contatos	5	60	8	73	0	4	141	168	83	2	0	26	0	0	62	3	193	28	14	102	112	9
<i>Bacillus subtilis</i>	Resíduos	H	Y	N	E	N	H	A	-	-	A	T	N	S	Y	Y	M	M	E	W	S	A	Y
	Contatos	14	107	47	288	118	194	91	0	0	4	101	38	65	3	155	57	56	208	206	116	127	133
<i>Neotermes koshunensis</i>	Resíduos	H	W	N	E	L	T	D	M	N	I	N	Y	N	F	Y	F	L	E	W	E	W	F
	Contatos	10	73	10	200	101	279	303	171	326	5	346	53	0	0	157	0	72	30	22	134	47	21
<i>Exiguobacterium antarcticum</i>	Resíduos	H	W	N	E	W	C	L	H	N	L	A	A	N	F	Y	S	N	E	W	E	W	F

<i>B7</i>	<b>Contatos</b>	4	43	4	123	347	51	117	124	213	5	170	4	0	0	211	91	423	19	8	51	50	13
<i>Fervidobacterium islandicum</i>	<b>Resíduos</b>	<b>H</b>	<b>F</b>	<b>N</b>	<b>E</b>	<b>H</b>	<b>V</b>	<b>L</b>	<b>F</b>	<b>S</b>	<b>F</b>	<b>T</b>	<b>A</b>	<b>N</b>	<b>W</b>	<b>Y</b>	<b>V</b>	<b>R</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	19	39	6	53	257	147	137	43	236	15	197	42	0	0	146	14	333	10	30	14	62	12
<i>Pyrococcus furiosus</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>N</b>	<b>V</b>	<b>Q</b>	<b>F</b>	<b>A</b>	<b>F</b>	<b>A</b>	<b>I</b>	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>S</b>	<b>L</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	6	58	2	59	70	105	296	64	148	17	118	0	1	0	85	0	88	6	11	79	44	6
<i>Neurospora crassa</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>C</b>	<b>L</b>	<b>F</b>	<b>N</b>	<b>G</b>	<b>D</b>	<b>K</b>	<b>N</b>	<b>H</b>	<b>Y</b>	<b>T</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	11	53	23	128	301	55	137	114	303	18	346	250	0	0	141	0	153	24	47	52	51	35
<i>Humicola grisea var. thermoidea</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>C</b>	<b>L</b>	<b>F</b>	<b>N</b>	<b>G</b>	<b>D</b>	<b>K</b>	<b>N</b>	<b>H</b>	<b>Y</b>	<b>T</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	10	29	17	53	359	37	162	77	277	39	327	215	0	0	77	16	129	24	38	60	15	28
<i>Acidilobus saccharovorans</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>N</b>	<b>V</b>	<b>L</b>	<b>F</b>	<b>A</b>	<b>N</b>	<b>S</b>	<b>A</b>	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>T</b>	<b>Q</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	3	74	6	117	228	149	255	106	318	60	130	0	0	0	116	0	392	35	9	62	31	7
<i>Metagenome hydrothermal spring</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>N</b>	<b>V</b>	<b>I</b>	<b>F</b>	<b>A</b>	<b>T</b>	<b>S</b>	<b>S</b>	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>S</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	0	40	2	65	234	65	152	65	156	61	223	0	0	0	156	21	142	15	1	17	45	0
<i>Thermotoga naphthophila</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>V</b>	<b>V</b>	<b>H</b>	<b>N</b>	<b>N</b>	<b>G</b>	<b>M</b>	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>S</b>	<b>H</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	2	54	4	184	171	116	116	189	261	136	56	2	0	0	208	39	145	22	7	96	141	3
<i>Thermotoga petrophila</i>	<b>Resíduos</b>	<b>H</b>	<b>W</b>	<b>N</b>	<b>E</b>	<b>W</b>	<b>C</b>	<b>C</b>	<b>H</b>	<b>N</b>	<b>N</b>	<b>G</b>	<b>M</b>	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>S</b>	<b>H</b>	<b>E</b>	<b>W</b>	<b>E</b>	<b>W</b>	<b>F</b>
	<b>Contatos</b>	9	76	12	219	181	110	106	227	269	156	68	0	0	0	230	24	152	28	12	99	153	9

Estes aminoácidos correspondem à  $\beta$ -glicosidase da *Thermoanaerobacter brockii* aos resíduos: H121, W122, N166, E167, W169, C170, L174, H181, N224, L225, T226, A244, N297, Y298, Y299, T300, S302, E355, W402, E409, W410 e F418. Na 18ª posição um triptofano aparece nas  $\beta$ -glicosidasas de *Nasutitermes takasagoensis* (825 contatos), *Trichoderma reesei* (969 contatos), *Neotermes koshunensis* (655 contatos), *Neurospora crassa* (751 contatos) e *Humicola grisea var. Thermoidea* (732 contatos).